



Machine Learning-Based Demand Forecasting in Supply Chains: A Systematic Review of Models, Applications, and Research Challenges

Aimad Saber^{1*}, Taibi Saoudi², Aziz Soulhi³

¹ QSM Laboratory, Department of Industrial Engineering, École Mohammadia d'Ingénieurs, Mohammed V University in Rabat, Rabat 10050, Morocco

² QSM Laboratory, Department of Mechanical Engineering, École Mohammadia d'Ingénieurs, Mohammed V University in Rabat, Rabat 10050, Morocco

³ Department of Industrial Engineering, École Nationale Supérieure des Mines de Rabat, Rabat 10000, Morocco

Corresponding Author Email: aimad.saber@research.emi.ac.ma

Copyright: ©2026 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/isi.310910>

ABSTRACT

Received: 9 July 2026

Revised: 25 August 2026

Accepted: 1 September 2026

Available online: 30 September 2026

Keywords:

machine learning, demand forecasting, supply chain analytics, deep learning, ensemble learning, forecasting models

Machine learning (ML) has become increasingly important in demand forecasting due to its ability to capture complex patterns from large-scale supply chain data. However, the effectiveness of ML models varies considerably across forecasting scenarios, datasets, and evaluation settings. This study presents a systematic review of empirical research on ML-based demand forecasting in supply chains, following the PRISMA 2020 framework. A total of 40 peer-reviewed studies published between 2015 and 2025 were analyzed to identify prevailing modelling approaches, application domains, evaluation practices, and implementation challenges. The reviewed studies were categorized into supervised learning, deep learning, and hybrid/ensemble approaches. Hybrid and ensemble models represented the largest group, followed by supervised learning and deep learning methods. Most applications focused on retail and e-commerce environments, while manufacturing, spare parts, and logistics applications remained comparatively limited. The analysis indicates that no single ML approach consistently outperforms others across all forecasting contexts. Model performance is strongly influenced by data characteristics, demand patterns, forecasting horizons, baseline selection, and evaluation metrics. Furthermore, practical adoption depends not only on predictive accuracy but also on data quality, interpretability, computational requirements, and integration with operational decision processes. This review provides a structured overview of current evidence and identifies future research directions for developing reliable and context-aware ML-based forecasting systems in supply chain environments.

1. INTRODUCTION

1.1 Importance of demand forecasting in supply chain management

Demand forecasting supports inventory, production, sourcing and distribution decisions. Forecast errors can translate into excess stock, lost sales, and unstable replenishment, so the value of a method lies not only in statistical accuracy but also in its consequences for planning and service. Forecast errors can propagate across supply-chain stages and amplify order variability. This makes consistency across product, store, and temporal hierarchies, together with explicit treatment of promotions and external drivers, important for operational use. Moving averages, exponential smoothing, and Autoregressive Integrated Moving Average (ARIMA) remain useful because they are transparent and economical. Their assumptions can, however, be restrictive for high-dimensional, nonlinear, intermittent, or promotion-driven demand, especially when many related series and

exogenous variables are available [1, 2].

1.2 Emergence of machine learning in demand forecasting

Recent empirical work has expanded from high-dimensional retail regression [3] to calendar-aware machine learning (ML) [4], hybrid Long Short-Term Memory (LSTM)-random forest models [5], and transformer architectures [6]. The evidence does not support universal superiority: model rankings vary with demand pattern, sample size, horizon, covariates, baseline, and evaluation metric [7-10]. Comparable model-comparison studies published in this journal report the same context-dependence for Seasonal Autoregressive Integrated Moving Average (SARIMA), random forest, and LSTM forecasters [11, 12].

1.3 Positioning relative to existing reviews

Existing reviews map predictive big-data analytics, ML demand-forecasting applications, AI across supply-chain

functions, and retail forecasting practice [1, 2, 13, 14]. The present review narrows that scope to empirical demand, sales, and inventory forecasting and contributes a study-level comparison of data, horizons, baselines, metrics, findings, and limitations. The reconstructed extraction matrix also makes the descriptive counts and discussion traceable to individual studies.

1.4 Objectives and questions

This review groups the predominant modelling approach of each included study into supervised learning, deep learning, or hybrid/ensemble methods. It compares these families on evidence, data requirements, and deployment constraints, without treating unlike error metrics or datasets as directly commensurable.

RQ1. Which factors determine the selection of a supervised, deep learning, or hybrid/ensemble approach for a given demand forecasting scenario in supply chain management?

RQ2. Under what data, forecasting-horizon, and evaluation conditions do ML models demonstrate a measurable advantage over traditional statistical approaches, and where does this advantage diminish?

RQ3. What data availability and quality requirements shape the practical applicability of ML-based demand forecasting

models?

RQ4. What computational, organizational, and deployment trade-offs limit the industrial adoption of ML models for demand forecasting across different supply chain contexts?

2. METHODS

2.1 Literature search strategy

The review search covered Scopus, Web of Science, ScienceDirect, and IEEE Xplore using combinations of machine-learning terms, forecasting terms, and supply-chain context terms. The author-reported search was completed in December 2025, restricted to English peer-reviewed studies published from 2015 through 2025, and organized according to PRISMA 2020 [15].

Studies were eligible when they applied at least one ML or deep-learning method to an empirical demand, sales, inventory, or throughput forecasting task in retail, manufacturing, spare-parts, logistics, or distribution settings and reported an evaluation against a baseline or alternative model. Conceptual papers, reviews, unrelated supply-chain applications, duplicate records, and reports without extractable empirical evidence were excluded.

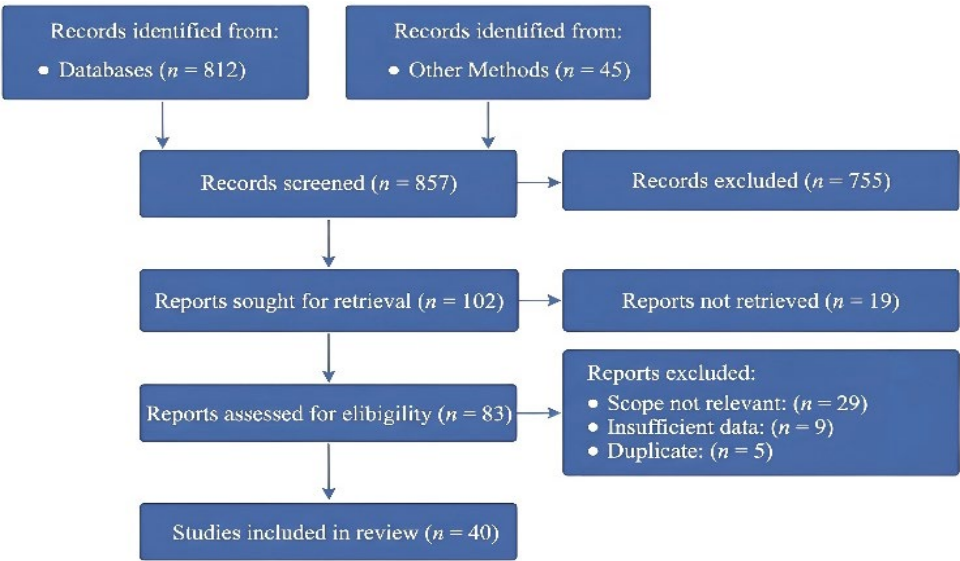


Figure 1. PRISMA 2020 flow diagram based on the author-retained screening log

Quality appraisal considered methodological clarity, data description, the forecast target and horizon, baseline comparability, evaluation metrics, and the transparency of reported limitations, consistent with structured review guidance [16]. The arithmetic in the retained screening log is internally consistent (857 records screened; 40 studies included), as shown in Figure 1.

2.2 Data extraction and analysis procedure

The original record-level exports and extraction spreadsheet were not available during revision. A study-level matrix was therefore reconstructed from DOI-matched primary bibliographic records and accessible publisher abstracts/full texts. For each study, the matrix records the citation and DOI, empirical context and data, target/horizon, model and baseline, metrics, key finding, and limitation. Missing information is

marked Not Reported (NR) rather than inferred. Each study was coded once by predominant model family, sector, and first-author affiliation region; publication year follows the electronic bibliographic record when available. All counts and percentages in Section 3 were recalculated from these 40 rows. Figure 1 reproduces the author-retained screening log and was not independently reconstructed from raw exports. The complete coding is provided in Appendix Table A1.

3. RESULTS

3.1 Descriptive and bibliometric analysis of selected studies

The reconstructed analytical corpus contains 40 DOI-verified empirical studies. The final set of 40 studies reflects the strict eligibility and quality-appraisal criteria applied rather

than an arbitrary cap; this focused corpus enables a fully traceable, study-level synthesis (Appendix Table A1) that larger but shallower samples cannot provide. Descriptive results below are simple study counts; they do not weight studies by sample size or quality and should not be interpreted as a meta-analysis.

3.2 Publication trends

Research activity increased after 2019: 32 of 40 studies (80.0%) were published from 2020 through 2025, including 13 (32.5%) in 2022–2023. Annual counts were 1 (2015), 2 (2017), 1 (2018), 4 (2019), 6 (2020), 5 (2021), 6 (2022), 7 (2023), 4 (2024), and 4 (2025). No eligible 2016 study was retained under the reconstructed coding.

3.3 Geographic distribution

Figure 2 reports the region coded from the first author's affiliation, with the corresponding author's primary affiliation used only when the first-author record was unresolved. This measure describes the location of the published research team, not necessarily the country of the operational dataset.

Europe accounted for 18 studies (45.0%), Asia-Pacific for 17 (42.5%), North America for 4 (10.0%), and other regions for 1 (2.5%). The concentration in Europe and Asia-Pacific, combined with the small North American and other-region samples, limits geographic generalization.



Figure 2. Geographic distribution by first-author affiliation in the reconstructed corpus (n = 40)

3.4 Journals and research fields

The corpus spans forecasting, operations research, industrial engineering, decision support, and applied artificial intelligence (AI) venues. Because disciplinary labels overlap and journal scopes change, the synthesis does not assign mutually exclusive research-field percentages. Instead, the auditable unit is the empirical setting and model-family coding reported in Appendix Table A1.

3.5 Machine learning models used for demand forecasting

Each study was assigned to one predominant model family according to the approach supporting its main reported result. Figure 3 shows 15 hybrid/ensemble studies (37.5%), 14 supervised-learning studies (35.0%), and 11 deep-learning studies (27.5%). These are study counts, not counts of every algorithm mentioned in a paper. Table 1 compares the three families in terms of favorable within-study evidence, boundary conditions, and representative studies.

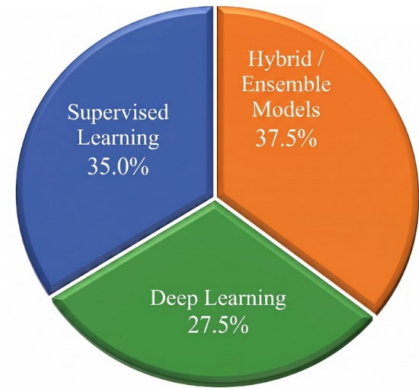


Figure 3. Predominant model-family distribution in the reconstructed corpus (n = 40)

3.5.1 Supervised learning models

Supervised-learning studies represented 14/40 (35.0%). Tree-based and regression methods were especially useful with structured covariates: random forest outperformed K-Nearest Neighbors (KNN) in four apparel datasets [17], Extreme Gradient Boosting (XGBoost) led a four-week e-commerce comparison with a reported computation time of 9 min 13 s [18], simplified tree models remained competitive on 4,523 Belgian products and an M5 robustness set [19], and LightGBM was operationalized across more than 18,000 HP products [20].

3.5.2 Deep learning approaches and LSTM architecture

Deep learning represented 11/40 studies (27.5%). The family included Deep Neural Network (DNN), LSTM/BiLSTM, ConvLSTM, transformer, and mixer architectures. Its strongest evidence appeared where many related series, rich covariates, or complex temporal structure were available [6, 21–23].

Architecture alone did not determine performance. A fashion-retail DNN was not significantly better than random forest on every criterion [7], and a retail comparison found LSTM stronger for stable demand while SARIMA was stronger for seasonal series and SARIMAX for promoted products [10]. These results argue for demand-pattern diagnostics and matched baselines rather than an a priori preference for recurrent or transformer models.

On the M5 benchmark, tested transformers improved Mean Absolute Scaled Error (MASE) by 26–29% and reduced weighted quantile loss by up to 34% relative to seasonal naive [6]. Such results are important but remain benchmark-specific; they cannot be converted into a pooled error-reduction range for proprietary enterprise datasets with different scales, horizons, and metrics.

3.5.3 Hybrid and ensemble models

Hybrid/ensemble approaches formed the largest group (15/40; 37.5%). Favorable within-study evidence included LSTM plus random forest in multi-channel retail [5], stacked ensembles in steel demand [24], Convolutional Neural Network (CNN)-LSTM across 30-, 90-, and 360-day retail horizons [25], an imbalance-aware spare-parts ensemble that reduced Mean Squared Error (MSE) by 47% [26], and a Temporal Convolutional Network (TCN)-transformer with Weighted Mean Absolute Percentage Error (WMAPE) of 4.22% on Favorita data [27]. These gains are case-specific and come with additional tuning, integration, and governance requirements.

3.6 Application areas of machine learning-based demand forecasting

Empirical coverage was highly uneven across sectors. Table 2 reports the mutually exclusive sector coding used in the reconstructed matrix.

3.6.1 Retail and e-commerce

Retail and e-commerce accounted for 29/40 studies (72.5%). This concentration provides rich evidence on promotions, store-item hierarchies, cold starts, and large panels of related series [3, 4, 6, 19, 28], but it also means that the review's aggregate model distribution is largely a retail distribution.

3.6.2 Manufacturing and production planning

Manufacturing and spare-parts settings accounted for 9/40 studies (22.5%). The evidence includes automotive components [29], B2B spare-parts conversion [30], aircraft aftermarket forecasting [31], enterprise print demand [20], and imbalanced spare-parts demand [26]. Intermittency and inventory consequences are more prominent in this group than in most retail benchmarks.

3.6.3 Logistics and inventory management

Only 2/40 studies (5.0%) were coded as logistics/distribution [32, 33]. This small base does not support broad claims about logistics-wide cost or service improvements and identifies a clear gap for field evaluation.

Table 1. Evidence-based comparison of predominant model families

Model Family	Studies, <i>n</i> (%)	Favorable Within-Study Evidence	Boundary Conditions/Counter-Evidence	Examples
Supervised learning	14 (35.0)	Random forest beat KNN in four apparel datasets; XGBoost led a four-week e-commerce case; simplified trees and enterprise LightGBM were competitive.	Strongest with structured covariates. Benefits depend on feature design, pooling and demand pattern; volatile sequences do not guarantee an advantage.	[17-20]
Deep learning	11 (27.5)	On M5, tested transformers reduced MASE by 26–29% and weighted quantile loss by up to 34%; LSTMixer led its 45-store comparison.	DNN was not significantly better than random forest on every metric; LSTM/SARIMA/SARIMAX rankings changed with demand pattern.	[6, 7, 10, 23]
Hybrid/ensemble	15 (37.5)	LSTM-random forest, stacked ensembles, CNN-LSTM, and TCN-transformer won matched comparisons; an imbalance-aware ensemble reduced MSE by 47%.	Integration and tuning burden are higher. Effects are case-specific and cannot be pooled across datasets, horizons, or metrics.	[5, 24-27]

Note: KNN = K-Nearest Neighbors, XGBoost = Extreme Gradient Boosting, CNN = Convolutional Neural Network, DNN = Deep Neural Network, TCN = Temporal Convolutional Network, SARIMA = Seasonal Autoregressive Integrated Moving Average, SARIMAX = Seasonal Autoregressive Integrated Moving Average with eXogenous variables, LSTM = Long Short-Term Memory, MSE = Mean Squared Error, MASE = Mean Absolute Scaled Error.

Table 2. Sector coverage and implications for generalizability

Sector	Studies, <i>n</i> (%)	Empirical Coverage	Generalizability Implication
Retail/e-commerce	29 (72.5)	Promotions, store-item panels, cold starts, fashion, grocery and online retail.	Aggregate findings primarily describe data-rich retail forecasting.
Manufacturing/spare parts	9 (22.5)	Automotive, steel, aerospace, print products and intermittent spare parts.	Operational cost and intermittency are more salient, but most evidence is single-firm.
Logistics/distribution	2 (5.0)	Electronics distribution and intermittent-demand methods.	Too few studies for logistics-wide performance claims.

Table 3. Observed data-scale, computation, and deployment evidence

Study	Data/Scale	Directly Reported Evidence	Deployment Lesson
Oliveira and Ramos [6]	M5 benchmark: 30,490 series across 10 stores.	Transformers reduced MASE by 26–29% and weighted quantile loss by up to 34% vs. seasonal naive.	Strong benchmark evidence, but enterprise integration cost is unobserved.
Spiliotis et al. [8]	~3,300 daily SKU series from a major retailer.	Some ML/cross-learning methods improved accuracy and bias; the advantage was method-dependent.	Global learning helps selected series, not every SKU.
Falatouri et al. [10]	>37 months of Austrian retail sales.	LSTM led for stable demand; SARIMA for seasonal series; SARIMAX for promoted products.	Demand pattern and causal inputs should drive model selection.
Tang et al. [18]	Cross-border e-commerce; four-week inventory horizon.	XGBoost reported as most accurate; computation time 9 min 13 s.	One of the few studies reporting timing; hardware details still limit comparison.
Harshvardhan et al. [20]	>18,000 HP print products in >170 countries.	LightGBM evaluated with bias, RMSE, and WMAPE over paired 12-month comparisons.	Production use required MLOps, governance, and human-in-the-loop review.
Kenaka et al. [26]	Highly imbalanced intermittent spare-parts demand.	Focal-loss/SMOTE ensemble reduced MSE by 47% and increased AUC by 32%.	Occurrence imbalance must be validated without distorting demand distribution.

Rafi et al. [27]	Favorita grocery data with holidays, promotions, oil prices, and transactions.	Hybrid TCN-transformer: MAE 2.01, RMSE 2.81, WMAPE 4.22%.	Competition-style data do not expose full enterprise operating cost.
Rohaani et al. [30]	B2B requests-for-quotation and spare-parts sales records.	~70% recall at ~50% precision; >2.5× the manual process.	Classification-style sales prediction is not comparable with RMSE/MAPE studies.
Kilimci et al. [34]	~875 million records; 5,500 stores; ~1,500 active products; 155 features.	MAPE 0.4238 → 0.2582 with integration and 0.2469 after adding deep learning.	Large gains required a complex retailer-specific pipeline.

Note: XGBoost = Extreme Gradient Boosting, TCN = Temporal Convolutional Network, SARIMA = Seasonal Autoregressive Integrated Moving Average, SARIMAX = Seasonal Autoregressive Integrated Moving Average with eXogenous variables, LSTM = Long Short-Term Memory, MSE = Mean Squared Error, MASE = Mean Absolute Scaled Error, MAE = Mean Absolute Error, WMAPE = Weighted Mean Absolute Percentage Error, ML = machine learning, SMOTE = Synthetic Minority Over-sampling Technique, RMSE = Root Mean Square Error, MAPE = Mean Absolute Percentage Error, AUC = Area Under the Curve.

3.7 Computational trade-offs and deployment considerations

Computational evidence was too heterogeneous for generic labels such as 'minutes', 'hours' or 'days'. Hardware, tuning budgets and preprocessing were seldom reported consistently. Table 3 therefore presents only observed scale, timing and deployment evidence that can be traced to individual studies. Model choice should follow the decision context. Structured, moderate-scale data can favor tree and regression methods; large panels can justify global deep or meta-learning models; intermittent demand may require occurrence/size decomposition or inventory-oriented metrics; and production deployment adds governance, monitoring and human review. Accuracy alone is insufficient when timing, bias, service level or implementation burden are decision-relevant.

4. DISCUSSION

4.1 Performance and benefits of machine learning models

The review supports conditional, not universal, ML advantage. Hybrid models won several matched experiments [5, 24-27], yet no single method dominated all products in a furniture-retail comparison [35]. Deep methods generated large benchmark gains [6] but did not consistently exceed

random forest in a new-fashion case [7], while supervised trees remained competitive in both research and enterprise deployment [19, 20]. Evaluation design can change the conclusion. Croston ranked best under the inventory-oriented Stock-keeping-oriented Prediction Error Costs (SPEC) metric in one intermittent-demand comparison [9], and LSTM, SARIMA, and SARIMAX each led under different retail demand conditions [10]. For this reason, the review does not pool percentage improvements across unlike datasets or metrics; it synthesizes direction and boundary conditions within studies.

4.2 Conceptual synthesis framework for Supply Chain 4.0 demand forecasting

The evidence suggests a four-layer conceptual synthesis for Supply Chain 4.0 demand forecasting. The framework in Figure 4 is a decision architecture distilled from recurring practices in the corpus, not an empirically validated model.

4.2.1 Layer 1: Data sources and acquisition

Layer 1 combines transaction and inventory history with relevant causal inputs such as promotions, calendars, product attributes, weather, and macroeconomic indicators. Cross-product promotions improved high-dimensional retail forecasts [3], while calendar coding [4] and explanatory variables [19] materially shaped accuracy.

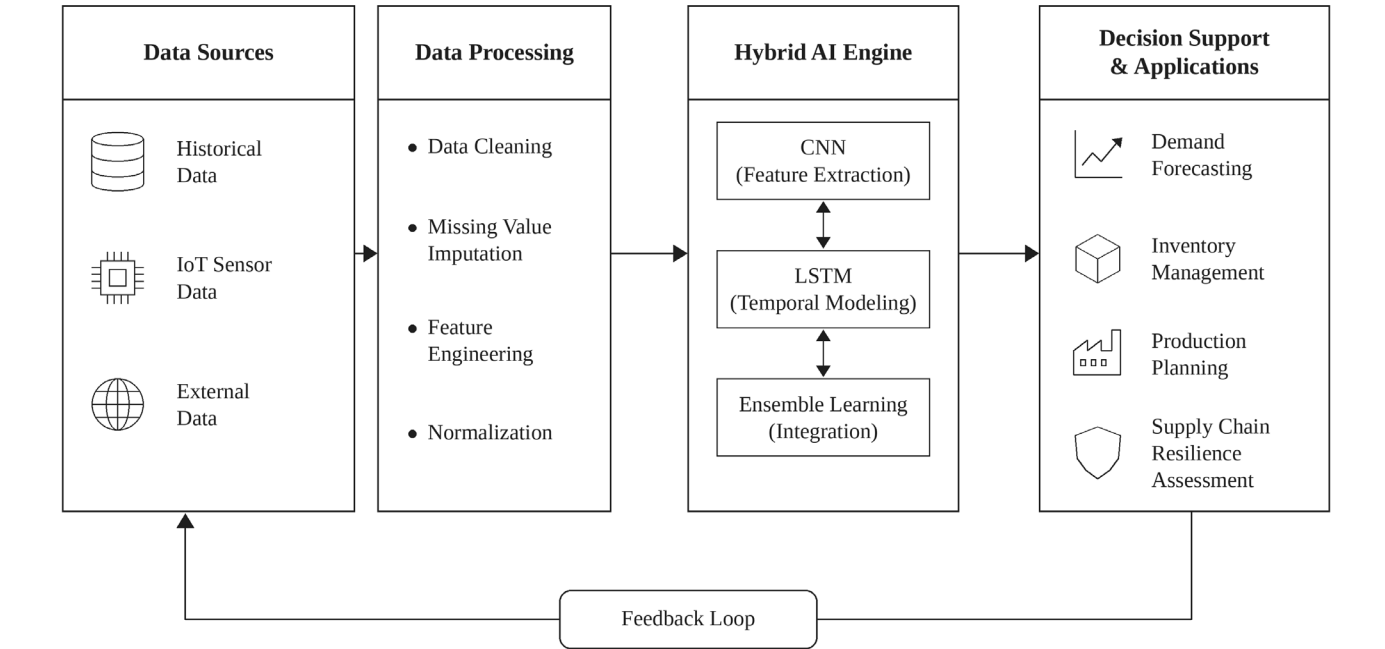


Figure 4. Conceptual synthesis framework for machine learning (ML)-based demand forecasting in Supply Chain 4.0

4.2.2 Layer 2: Data processing and quality management

Layer 2 continuously monitors missingness, stockouts, structural breaks, scale, leakage, and demand intermittency. Gap-aware treatment improved post-disruption forecasts in one study [36], while imbalance-aware learning improved spare-parts occurrence and quantity prediction [26].

4.2.3 Layer 3: Hybrid architecture and model selection

Layer 3 selects or combines models against explicit baselines. It allows supervised trees, recurrent and transformer models, probabilistic methods and ensembles, but requires rolling or time-respecting validation and metrics matched to the decision [6, 9, 10].

4.2.4 Layer 4: Output and decision support

Layer 4 converts forecasts into point and uncertainty outputs, bias/service indicators, explanations, and planning decisions. Model-agnostic explanation tools such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) provide established mechanisms for local and feature-attribution explanations [37, 38], while the HP case illustrates the value of human review and downstream ensemble decisions [20].

Compared with broad Supply Chain 4.0 sustainability frameworks [39], Figure 4 is deliberately scoped to forecasting. Its distinctive element is a feedback loop in which data quality and post-deployment performance are monitored continuously. The framework remains a synthesis proposition requiring prospective validation.

4.3 Limitations and challenges in existing studies

The first limitation is evidence concentration: 29/40 studies concern retail/e-commerce, many use one firm or public competition data, and only two concern logistics/distribution. Proprietary datasets, incomplete reporting of horizons and hardware, and inconsistent baselines also restrict replication and external validity.

The second limitation is evaluation heterogeneity. Studies report Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE)/WMAPE, MASE, quantile loss, classification metrics, and inventory-oriented costs on different scales. Complex deep and hybrid models also introduce tuning, monitoring, and explanation burdens. Explainable artificial intelligence (XAI) can support review [37, 38], but an explanation layer does not by itself establish causal validity or user trust.

4.4 Future research directions

Future work should (1) publish study- and experiment-level extraction data; (2) use time-respecting splits, transparent tuning budgets and strong statistical/ML baselines; (3) report operational metrics such as bias, service level, and inventory cost alongside forecast error; (4) test models across firms, countries and non-retail sectors; (5) quantify compute and energy use; and (6) prospectively validate the Figure 4 framework, including data-quality alarms, uncertainty, XAI and human override.

5. CONCLUSIONS

This review reconstructs and synthesizes 40 DOI-verified empirical studies of ML-based demand forecasting.

Hybrid/ensemble approaches were most common (37.5%), followed by supervised learning (35.0%) and deep learning (27.5%). Evidence supports context-specific gains rather than a universal hierarchy: data scale, covariates, intermittency, horizon, metric, and baseline all affect the result.

The main contribution is traceability. Appendix Table A1 connects every descriptive count and major discussion claim to a study's context, model, baseline, metric, finding and limitation. For managers, the implication is to select models through matched back-testing and operational criteria, while investing in data governance, monitoring and human oversight.

5.1 Limitations

The record-level search exports and original extraction spreadsheet were unavailable during revision; Figure 1 therefore remains the author-reported screening log, whereas Appendix Table A1 is a transparent reconstruction from DOI-matched sources. The review is restricted to English peer-reviewed work from four databases and does not claim exhaustive coverage. Many studies use proprietary or single-company data, and the heterogeneity of metrics and horizons precludes meta-analytic pooling.

ML can materially improve demand forecasting, but credible adoption depends on reproducible evaluation, decision-relevant metrics, and deployment discipline as much as on predictive accuracy.

ACKNOWLEDGMENT

This work was supported by École Mohammadia d'Ingénieurs and École Nationale Supérieure des Mines de Rabat. The authors thank the academic institutions for providing resources and support for this research.

REFERENCES

- [1] Seyedan, M., Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. *Journal of Big Data*, 7: 53. <https://doi.org/10.1186/s40537-020-00329-2>
- [2] Fildes, R., Ma, S.H., Kolassa, S. (2022). Retail forecasting: Research and practice. *International Journal of Forecasting*, 38(4): 1283-1318. <https://doi.org/10.1016/j.ijforecast.2019.06.004>
- [3] Ma, S.H., Fildes, R., Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, 249(1): 245-257. <https://doi.org/10.1016/j.ejor.2015.08.029>
- [4] 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. <https://doi.org/10.1016/j.ijforecast.2020.02.005>
- [5] 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. <https://doi.org/10.1080/00207543.2020.1735666>

- [6] Oliveira, J.M., Ramos, P. (2024). Evaluating the effectiveness of time series transformers for demand forecasting in retail. *Mathematics*, 12(17): 2728. <https://doi.org/10.3390/math12172728>
- [7] Loureiro, A.L.D., Miguéis, V., da Silva, L.F.M. (2018). Exploring the use of deep neural networks for sales forecasting in fashion retail. *Decision Support Systems*, 114: 81-93. <https://doi.org/10.1016/j.dss.2018.08.010>
- [8] Spiliotis, E., Makridakis, S., Semoglou, A.A., Assimakopoulos, V. (2022). Comparison of statistical and machine learning methods for daily SKU demand forecasting. *Operational Research*, 22: 3037-3061. <https://doi.org/10.1007/s12351-020-00605-2>
- [9] Kiefer, D., Grimm, F., Bauer, M., van Dinther, C. (2021). Demand forecasting intermittent and lumpy time series: Comparing statistical, machine learning and deep learning methods. In *Proceedings of the 54th Hawaii International Conference on System Sciences*, pp. 1425-1434. <https://doi.org/10.24251/hicss.2021.172>
- [10] 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. <https://doi.org/10.1016/j.procs.2022.01.298>
- [11] Bouhaddour, S., Sbihi, M., Guerouate, F., Saadi, C. (2024). Assessing tourism prediction models: A comparative study of SARIMA, random forest, and LSTM, considering nonlinear trends and the influence of COVID-19. *Ingénierie des Systèmes d'Information*, 29(6): 2443-2454. <https://doi.org/10.18280/isi.290631>
- [12] Bouhaddour, S., Saadi, C., Bouabdallaoui, I., Sbihi, M., Guerouate, F. (2023). A novel hybrid approach for daily tourism arrival forecasting: The PROPHET-Bayesian Gaussian Process-Forward Neural Network model. *Ingénierie des Systèmes d'Information*, 28(4): 833-842. <https://doi.org/10.18280/isi.280404>
- [13] Aamer, A., Eka Yani, L.P., Alan Priyatna, I.M. (2021). Data analytics in the supply chain management: Review of machine learning applications in demand forecasting. *Operations and Supply Chain Management: An International Journal*, 14(1): 1-13. <https://doi.org/10.31387/oscm0440281>
- [14] Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, 122: 502-517. <https://doi.org/10.1016/j.jbusres.2020.09.009>
- [15] Page, M.J., McKenzie, J.E., Bossuyt, P.M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372: n71. <https://doi.org/10.1136/bmj.n71>
- [16] Paul, J., Criado, A.R. (2020). The art of writing literature review: What do we know and what do we need to know? *International Business Review*, 29(4): 101717. <https://doi.org/10.1016/j.ibusrev.2020.101717>
- [17] Güven, İ., Uygün, Ö., Şimşir, F. (2021). Machine learning algorithms with intermittent demand forecasting: An application in retail apparel with plenty of predictors. *Textile and Apparel*, 31(2): 99-110. <https://doi.org/10.32710/tekstilvekonfeksiyon.809867>
- [18] Tang, Y.M., Chau, K.Y., Lau, Y.Y., Zheng, Z.H. (2023). Data-intensive inventory forecasting with artificial intelligence models for cross-border e-commerce service automation. *Applied Sciences*, 13(5): 3051. <https://doi.org/10.3390/app13053051>
- [19] Wellens, A.P., Boute, R.N., Udenio, M. (2024). Simplifying tree-based methods for retail sales forecasting with explanatory variables. *European Journal of Operational Research*, 314(2): 523-539. <https://doi.org/10.1016/j.ejor.2023.10.039>
- [20] Harshvardhan, M., Curtland, C., Hwang, J., et al. (2025). Print demand forecasting with machine learning at HP Inc. *INFORMS Journal on Applied Analytics*, 55(6): 469-483. <https://doi.org/10.1287/inte.2024.0126>
- [21] Li, J.T., Cui, T.X., Yang, K.W., Yuan, R.P., He, L.Y., Li, M.T. (2021). Demand forecasting of e-commerce enterprises based on horizontal federated learning from the perspective of sustainable development. *Sustainability*, 13(23): 13050. <https://doi.org/10.3390/su132313050>
- [22] Mitra, A., Jain, A., Kishore, A., Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: A novel hybrid machine learning approach. *Operations Research Forum*, 3: 58. <https://doi.org/10.1007/s43069-022-00166-4>
- [23] Theodoridis, G., Tsadiras, A. (2025). Retail demand forecasting: A comparative analysis of deep neural networks and the proposal of LSTMixer, a linear model extension. *Information*, 16(7): 596. <https://doi.org/10.3390/info16070596>
- [24] Raju, S.M.T.U., Sarker, A., Das, A., et al. (2022). An approach for demand forecasting in steel industries using ensemble learning. *Complexity*, 2022(1): 9928836. <https://doi.org/10.1155/2022/9928836>
- [25] Moraes, T.D.C., Yuan, X.M., Chew, E.P. (2024). Hybrid convolutional long short-term memory models for sales forecasting in retail. *Journal of Forecasting*, 43(5): 1278-1293. <https://doi.org/10.1002/for.3073>
- [26] Kenaka, S.P., Cakravastia, A., Ma'ruf, A., Cahyono, R.T. (2025). Enhancing intermittent spare part demand forecasting: A novel ensemble approach with focal loss and SMOTE. *Logistics*, 9(1): 25. <https://doi.org/10.3390/logistics9010025>
- [27] Rafi, M.A., Rodrigues, G.N., Mir, M.N.H., et al. (2025). A hybrid temporal convolutional network and transformer model for accurate and scalable sales forecasting. *IEEE Open Journal of the Computer Society*, 6: 380-391. <https://doi.org/10.1109/ojcs.2025.3538579>
- [28] Ma, S.H., Fildes, R. (2021). Retail sales forecasting with meta-learning. *European Journal of Operational Research*, 288(1): 111-128. <https://doi.org/10.1016/j.ejor.2020.05.038>
- [29] Gonçalves, J.N.C., Cortez, P., Carvalho, M.S., Frazão, N.M. (2021). A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain. *Decision Support Systems*, 142: 113452. <https://doi.org/10.1016/j.dss.2020.113452>
- [30] Rohaan, D., Topan, E., Groothuis-Oudshoorn, C.G.M. (2022). Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider. *Expert Systems with Applications*, 188: 115925. <https://doi.org/10.1016/j.eswa.2021.115925>
- [31] Dodin, P., Xiao, J.Y., Adulyasak, Y., et al. (2023). Bombardier aftermarket demand forecast with machine learning. *INFORMS Journal on Applied Analytics*, 53(6): 425-445. <https://doi.org/10.1287/inte.2023.1164>

- [32] Fu, W.H., Chien, C.F. (2019). UNISON data-driven intermittent demand forecast framework to empower supply chain resilience and an empirical study in electronics distribution. *Computers & Industrial Engineering*, 135: 940-949. <https://doi.org/10.1016/j.cie.2019.07.002>
- [33] Zhang, G.P., Xia, Y.S., Xie, M.H. (2024). Intermittent demand forecasting with transformer neural networks. *Annals of Operations Research*, 339: 1051-1072. <https://doi.org/10.1007/s10479-023-05447-7>
- [34] Kilimci, Z.H., Akyuz, A.O., Uysal, M., et al. (2019). An improved demand forecasting model using deep learning approach and proposed decision integration strategy for supply chain. *Complexity*, 2019(1): 9067367. <https://doi.org/10.1155/2019/9067367>
- [35] Aras, S., Kocakoç, İ.D., Polat, C. (2017). Comparative study on retail sales forecasting between single and combination methods. *Journal of Business Economics and Management*, 18(5): 803-832. <https://doi.org/10.3846/16111699.2017.1367324>
- [36] Riachy, C., He, M.D., Joneidy, S., et al. (2025). Enhancing deep learning for demand forecasting to address large data gaps. *Expert Systems with Applications*, 268: 126200. <https://doi.org/10.1016/j.eswa.2024.126200>
- [37] Ribeiro, M.T., Singh, S., Guestrin, C. (2016). "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, USA, pp. 1135-1144. <https://doi.org/10.1145/2939672.2939778>
- [38] Lundberg, S.M., Lee, S.I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, California, USA, pp. 4768-4777. <https://dl.acm.org/doi/10.5555/3295222.3295230>
- [39] Srhir, S., Jaegler, A., Montoya-Torres, J.R. (2023). Introducing a framework toward sustainability goals in a supply chain 4.0 ecosystem. *Journal of Cleaner Production*, 418: 138111. <https://doi.org/10.1016/j.jclepro.2023.138111>
- [40] Rosienkiewicz, M., Chlebus, E., Detyna, J. (2017). A hybrid spares demand forecasting method dedicated to mining industry. *Applied Mathematical Modelling*, 49: 87-107. <https://doi.org/10.1016/j.apm.2017.04.027>
- [41] Craparotta, G., Thomassey, S., Biolatti, A. (2019). A siamese neural network application for sales forecasting of new fashion products using heterogeneous data. *International Journal of Computational Intelligence Systems*, 12: 1537-1546. <https://doi.org/10.2991/ijcis.d.191122.002>
- [42] Ulrich, M., Jahnke, H., Langrock, R., Pesch, R., Senge, R. (2021). Distributional regression for demand forecasting in e-grocery. *European Journal of Operational Research*, 294(3): 831-842. <https://doi.org/10.1016/j.ejor.2019.11.029>
- [43] Aguilar-Palacios, C., Muñoz-Romero, S., Rojo-Álvarez, J.L. (2020). Cold-start promotional sales forecasting through gradient boosted-based contrastive explanations. *IEEE Access*, 8: 137574-137586. <https://doi.org/10.1109/access.2020.3012032>
- [44] Cai, W.W., Song, Y.P., Wei, Z.G. (2021). Multimodal data guided spatial feature fusion and grouping strategy for e-commerce commodity demand forecasting. *Mobile Information Systems*, 2021(1): 5568208. <https://doi.org/10.1155/2021/5568208>
- [45] 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 and Electrical Engineering*, 103: 108358. <https://doi.org/10.1016/j.compeleceng.2022.108358>
- [46] Giri, C., Chen, Y. (2022). Deep learning for demand forecasting in the fashion and apparel retail industry. *Forecasting*, 4(2): 565-581. <https://doi.org/10.3390/forecast4020031>
- [47] Chaudhuri, K.D., Alkan, B. (2022). A hybrid extreme learning machine model with Harris hawks optimisation algorithm: An optimised model for product demand forecasting applications. *Applied Intelligence*, 52: 11489-11505. <https://doi.org/10.1007/s10489-022-03251-7>
- [48] Wang, J.X., Chong, W.K., Lin, J.Y., Hedenstierna, C.P.T. (2024). Retail demand forecasting using spatial-temporal gradient boosting methods. *Journal of Computer Information Systems*, 64(5): 652-664. <https://doi.org/10.1080/08874417.2023.2240753>
- [49] Li, Q.Y., Yu, M.Y. (2023). Achieving sales forecasting with higher accuracy and efficiency: A new model based on modified transformer. *Journal of Theoretical and Applied Electronic Commerce Research*, 18(4): 1990-2006. <https://doi.org/10.3390/jtaer18040100>
- [50] Kim, S. (2023). Innovating knowledge and information for a firm-level automobile demand forecast system: A machine learning perspective. *Journal of Innovation & Knowledge*, 8(2): 100355. <https://doi.org/10.1016/j.jik.2023.100355>
- [51] Mejía, S., Aguilar, J. (2024). A demand forecasting system of product categories defined by their time series using a hybrid approach of ensemble learning with feature engineering. *Computing*, 106: 3945-3965. <https://doi.org/10.1007/s00607-024-01320-y>

APPENDIX

List of studies included in the systematic review (period 2015–2025)

In Table A1, we detail the 40 studies included in the systematic review, together with their data, models, metrics, and reported limitations.

Table A1. Study-level extraction matrix of the 40 reviewed studies

ID/Ref.	Study	Context and Data	Target/Horizon	Model and Baseline	Metrics and Key Finding	Reported Limitation
S01 [3]	Ma et al. (2015)	Retail/e-commerce; Asia-Pacific. SKU-level retail sales with intra-	Rolling SKU forecasts; horizon	Multistage LASSO regression Baseline: Models using the	Forecast-accuracy metrics; adding cross-product promotional	Exact sample size and horizon

		and inter-category promotion variables; high-dimensional proprietary data	NR in accessible primary abstract	focal SKU's own predictors	information improved accuracy by 12.6%, mostly from intra-category signals.	require the full text.
S02 [35]	Aras et al. (2017)	Retail/e-commerce; Europe. Weekly sales for 10 products from a global furniture retailer operating in Türkiye	Weekly; exact horizon NR in accessible primary abstract	ETS, ARIMA, ARFIMA, neural networks, ANFIS and forecast combinations Baseline: Single statistical and neural models	Multiple accuracy metrics; no single method dominated all products, and combination performance depended on the series.	Small product sample limits broad generalization.
S03 [40]	Rosienkiewicz et al. (2017)	Manufacturing/spare parts; Europe. Lumpy spare-parts demand in a mining-company setting	NR in accessible primary abstract	Information criteria, regression modelling and artificial neural networks Baseline: Traditional intermittent-demand methods	Forecast-error metrics; the hybrid procedure handled lumpy spare-parts demand better than conventional approaches.	Single-industry case; numerical effect size not available in the abstract.
S04 [7]	Loureiro et al. (2018)	Retail/e-commerce; Europe. Real fashion-retailer data for new products, including product attributes and expert information	Future-season sales for new products	Deep neural network baseline: random forest and other shallow-learning methods	MAPE and other accuracy metrics; deep learning performed well but was not significantly better than random forest on every metric.	Company-specific new-product setting.
S05 [34]	Kilimci et al. (2019)	Retail/e-commerce; Europe. 106 weeks of proprietary retail demand; about 875 million records, 5,500 stores, roughly 1,500 active products and 155 features	Next-week store-product demand	Nine time-series/regression methods, SVR, feed-forward deep network and weighted decision integration Baseline: Individual statistical, SVR and deep-learning models	Cross-validated MAPE; decision integration and deep learning cut MAPE from 0.4238 to 0.2469 ($\approx 41.7\%$ improvement).	Proprietary retailer; complex clustering, feature and blacklist pipeline; MAPE is sensitive to low demand.
S06 [41]	Craparotta et al. (2019)	Retail/e-commerce; Europe. Heterogeneous product-attribute and sales data from a European fashion retailer	Long-term sales of new products	Siamese neural network Baseline: Benchmark new-product forecasting models	Forecast-accuracy metrics; similarity learning produced useful item-level forecasts despite no product sales history.	Evidence comes from one fashion retailer.
S07 [32]	Fu and Chien (2019)	Logistics/distribution; Asia-Pacific. Intermittent component demand from a world-leading semiconductor distributor	Multiple temporal aggregations; exact horizon NR in accessible primary abstract	UNISON framework combining machine learning and temporal aggregation Baseline: Conventional methods and the company's existing practice	Forecast-performance metrics; the deployed framework outperformed conventional approaches and supported operational decisions.	Single electronics-distribution implementation.
S08 [42]	Ulrich et al. (2019)	Retail/e-commerce; Europe. E-grocery demand with covariates relevant to the full conditional distribution	NR in accessible primary abstract	Distributional regression Baseline: Conventional point-forecast specifications	Distributional and point-forecast metrics; modelling the demand distribution added value for inventory decisions beyond a point forecast.	The approach is tailored to an e-grocery context.
S09 [4]	Huber and Stuckenschmidt (2020)	Retail/e-commerce; Europe. Daily product demand from a bakery chain, with explicit calendric special-day features	Daily forecasting	Machine-learning regressors with calendar-event features Baseline: Statistical and feature-reduced alternatives	Daily accuracy metrics; explicitly representing special days markedly improved forecasts for irregular calendar effects.	Findings are sensitive to event coding and one retail format.
S10 [5]	Punia et al. (2020)	Retail/e-commerce; Asia-Pacific. Real multivariate data from online and offline channels of a multi-channel retailer	Shorter and longer horizons across both channels	Hybrid LSTM and random forest Baseline: Neural network, regression, ARIMAX, LSTM and random forest	Bias, accuracy and variance metrics; the hybrid model was statistically superior to its components and robust across horizons and channels.	Computational requirements are not quantified in the abstract.
S11 [8]	Spiliotis et al. (2020)	Retail/e-commerce; Europe. Real daily demand for approximately 3,300 SKUs from a major Greek retailer	Daily SKU forecasts	Regression trees and neural-network methods trained per series and by cross-learning Baseline: Established statistical forecasting methods	Accuracy and bias metrics; several ML methods improved accuracy and bias, and cross-learning helped selected methods.	The advantage was method-dependent rather than universal.

S12 [28]	Ma and Fildes (2020)	Retail/e-commerce; Asia-Pacific. IRI grocery/drug data: six categories and 100 stores; meta-training on 83,944 series and testing on 36,194 series	Seven-week SKU-store forecasts using rolling 48-week training windows	Deep convolutional meta-learning to weight a pool of base forecasters Baseline: Individual base-forecasting methods and standard combinations	AvgRelMAE and MPE; the meta-learner improved AvgRelMAE by $\approx 3.2\%$ overall and $\approx 5\%$ in promotion weeks.	Requires many related series; the data omit stockouts/inventory and interpretation is limited.
S13 [29]	Gonçalves et al. (2020)	Manufacturing/spare parts; Europe. Real automotive supply-chain data with leading indicators for component demand	Multiple forecast horizons over the component life cycle	Multivariate machine-learning regression models Baseline: ARIMAX	Accuracy and supply-chain cost metrics; leading indicators improved multi-step component-demand forecasts, varying over the product life cycle.	Single automotive case.
S14 [43]	Aguilar-Palacios et al. (2020)	Retail/e-commerce; Europe. Cold-start promotional sales linked to similar historical promotions	Promotion-planning horizon	Gradient-boosted decision trees with contrastive explanations Baseline: Retail promotional-forecasting alternatives	Accuracy plus expert-review usability; the method combined predictive performance with case-based explanations experts could adjust.	Focused on cold-start promotions rather than routine demand.
S15 [17]	Güven et al. (2021)	Retail/e-commerce; Europe. Four intermittent-demand datasets for apparel retail with 28 predictors	NR in accessible primary abstract	Random forest and k-nearest neighbours Baseline: Direct comparison between RF and KNN	RMSE; random forest outperformed KNN on all four datasets.	Only two ML families were evaluated.
S16 [30]	Rohaani et al. (2021)	Manufacturing/spare parts; Europe. Requests-for-quotation and sales records at a large after-sales service provider	B2B spare-parts sales conversion	Supervised learning with natural-language processing Baseline: Existing manual process	Recall and precision; the model identified $\approx 70\%$ of sales at $\approx 50\%$ precision, over $2.5\times$ better than the manual process.	Classification-style sales prediction is not directly comparable with error-based time-series studies.
S17 [9]	Kiefer et al. (2021)	Manufacturing/spare parts; Europe. Intermittent and lumpy demand series	Multi-period inventory-oriented evaluation	Statistical, machine-learning and deep-learning methods Baseline: Croston-family statistical methods	Conventional errors and inventory-oriented SPEC; Croston ranked best on SPEC, showing lower generic error does not guarantee lower operational cost.	A methodological comparison rather than a single deployment.
S18 [21]	Li et al. (2021)	Retail/e-commerce; Asia-Pacific. Multi-party e-commerce demand data distributed across enterprises	Multidimensional time-series demand forecasting	Horizontal federated learning with ConvLSTM Baseline: Centralized RNN/LSTM and non-federated alternatives	Accuracy and privacy-preserving metrics; federated ConvLSTM improved collaboration without directly sharing enterprise records.	Results depend on compatible data schemas across participating firms.
S19 [44]	Cai et al. (2021)	Retail/e-commerce; Asia-Pacific. E-commerce order sequences plus multimodal product and consumer features	Commodity demand; exact horizon NR in accessible primary abstract	Spatial feature fusion and BiLSTM grouping Baseline: Existing neural demand-forecasting solutions	Prediction-error metrics; multimodal fusion reduced information loss relative to order-history-only modelling.	Feature availability may limit transfer to less data-rich firms.
S20 [45]	Joseph et al. (2022)	Retail/e-commerce; Asia-Pacific. Store-item retail demand time series	Store-item forecasting; exact horizon NR in accessible primary abstract	CNN-BiLSTM with Lazy Adam optimization Baseline: Standalone deep-learning and conventional alternatives	Forecast-error metrics; the hybrid architecture improved prediction by combining local feature extraction and bidirectional temporal learning.	Data scale and deployment cost are NR in accessible primary abstract.
S21 [46]	Giri et al. (2022)	Retail/e-commerce; Asia-Pacific. European fashion-retailer sales combined with garment-image attributes	Future demand for fashion products	Multimodal deep-learning forecasting system Baseline: History-only conventional forecasting models	Accuracy metrics; combining visual product attributes with sales data helped, especially for short life cycles and sparse histories.	Single retailer and image-rich product domain.
S22 [10]	Falatouri et al. (2022)	Retail/e-commerce; Europe. More than 37 months of actual sales from an Austrian retailer	Product/store retail forecasts	LSTM Baseline: SARIMA and promotion-augmented SARIMAX	Accuracy by demand pattern; LSTM led for stable demand, SARIMA for seasonal series, and SARIMAX for promoted products.	Strong context interaction prevents a single overall ranking.

S23 [22]	Mitra et al. (2022)	Retail/e-commerce; Asia-Pacific. Multi-channel retail-company demand data	Multiple retail channels; exact horizon NR in accessible primary abstract	Novel hybrid machine-learning forecast combination Baseline: Conventional and advanced single forecasting models Random forest, gradient boosting, XGBoost and stacked ensembles Baseline: SVR, ELM and multilayer perceptron Extreme learning machine optimized by Harris hawks optimization Baseline: ARIMA, ELM, Bayesian-optimized ELM, GRU and LSTM	AE, MSE, MLSE and normalized MAE; the hybrid design was assessed across several error criteria, not a single headline metric.	Company-specific data and heterogeneous metrics limit cross-study pooling.
S24 [24]	Raju et al. (2022)	Manufacturing/spare parts; Asia-Pacific. Steel-industry demand dataset	NR in accessible primary abstract		R-squared and RMSE (cross-validated); stacked ensembles outperformed the individual reference models under identical conditions.	One industrial dataset; operational deployment not evaluated.
S25 [47]	Chaudhuri and Alkan (2022)	Retail/e-commerce; Asia-Pacific. Historical product-demand data from an e-commerce company	Real-time product-demand forecasting		Forecast-error and efficiency metrics; the optimized ELM beat ARIMA and neural alternatives while staying computationally fast.	External validity beyond the case company is untested.
S26 [18]	Tang et al. (2023)	Retail/e-commerce; Asia-Pacific. Intensive inventory records from a cross-border e-commerce company	Four-week inventory forecast	XGBoost and other AI models Baseline: Alternative AI forecasting algorithms	RMSE and computation time; XGBoost gave the best reported accuracy at a computation time of 9 min 13 s.	Single enterprise; the reported RMSE scale is company-specific and not comparable across datasets.
S27 [48]	Wang et al. (2023)	Retail/e-commerce; Asia-Pacific. Large e-commerce transaction records covering many SKUs and explanatory variables	Short-horizon SKU-level retail demand	Spatial-temporal gradient boosting tree Baseline: Standard gradient-boosting and alternative retail forecast models	Forecast-accuracy metrics; combining cross-sectional and temporal information improved large-scale SKU forecasting.	One large e-commerce context; code and computing cost are NR in accessible primary abstract.
S28 [33]	Zhang et al. (2023)	Logistics/distribution; North America. Intermittent-demand time series evaluated against established inventory-demand methods	Intermittent multi-period forecasting	Transformer neural network Baseline: Croston-family and other intermittent-demand methods	Intermittent-demand errors; Transformers can exploit cross-series structure, but gains depend on sparsity and evaluation design.	Formal issue year is 2024; coded by 2023 online-first date.
S29 [31]	Dodin et al. (2023)	Manufacturing/spare parts; North America. Bombardier aircraft aftermarket demand, including flight-derived features	Intermittent spare-parts demand sizes and inter-demand intervals	Integrated tree-based ML and traditional time-series pipeline Baseline: Existing planning forecasts and traditional models	Accuracy and inventory-planning outcomes; separately modelling demand occurrence and size improved aftermarket forecasting.	Results reflect one aerospace supply chain.
S30 [49]	Li and Yu (2023)	Retail/e-commerce; Asia-Pacific. Historical e-commerce sales with large fluctuations	Sales forecasting; exact horizon NR in accessible primary abstract	Modified encoder-decoder Transformer Baseline: Standard Transformer and established forecasting models	Accuracy and efficiency metrics; architecture changes improved the use of volatile sales data over the tested comparators.	Dataset description and deployment cost are limited in the abstract.
S31 [50]	Kim (2023)	Manufacturing/spare parts; Asia-Pacific. Monthly firm-level South Korean vehicle sales and endogenous/exogenous factors, 2011-2020	Monthly automobile demand	Linear regression, neural network, random forest, stochastic gradient descent and ensemble learning Baseline: Single-source and reduced-input models	Accuracy metrics with RReliefF analysis; combining firm-level and macroeconomic inputs improved prediction and interpretability.	One national automobile market.
S32 [19]	Wellens et al. (2023)	Retail/e-commerce; Europe. 4,523 products from a Belgian retailer plus an external M5 robustness dataset	Retail sales forecasting	Simplified tree-based models with feature engineering and SHAP Baseline: More complex tree-based specifications and standard retail benchmarks	Accuracy metrics with SHAP interpretation; simpler tree methods stayed competitive when explanatory variables and feature engineering were available.	Formal issue year is 2024; coded by 2023 online-first date.

S33 [6]	Oliveira and Ramos (2024)	Retail/e-commerce; Europe. M5 benchmark with 30,490 series from 10 stores	Short-term retail demand	Transformer, Informer, Autoformer, PatchTST and temporal fusion Transformer Baseline: AutoARIMA, AutoETS and seasonal naive	MASE and weighted quantile loss; leading Transformers improved MASE by 26–29% and cut weighted quantile loss by up to 34% versus seasonal naive.	Public-competition results may not reflect enterprise deployment constraints.
S34 [25]	Moraes et al. (2024)	Retail/e-commerce; Asia-Pacific. Public retail data: 50 products across 10 stores, 500 daily series with 1,826 observations each (2013-2017)	30-, 90- and 360-day sales forecasts	Stacked and parallel CNN-LSTM architectures Baseline: Simple neural networks and standard autoregressive methods	MAPE and SMAPE; parallel CNN-LSTM led at 90 and 360 days (14.00%, 13.42%) and the stacked model at 30 days (14.15%).	One public dataset with calendar-only exogenous data; averaged metrics mask SKU heterogeneity.
S35 [36]	Riachy et al. (2024)	Retail/e-commerce; Europe. Retail sales affected by pandemic-era gaps and structural breaks	Post-disruption demand forecasting	Deep-learning framework with gap-aware feature engineering Baseline: Existing historical-data-driven models	Forecast-error metrics before/after gap treatment; explicit handling of large historical gaps improved adaptation to post-pandemic demand.	Findings are tied to disruption-induced missingness.
S36 [51]	Mejía and Aguilar (2024)	Retail/e-commerce; Other. Retail product-category time series grouped by extracted demand characteristics	Product-category demand; exact horizon NR in accessible primary abstract	Clustering, feature engineering, model selection and ensemble learning Baseline: One-model-for-all-products strategies	Cluster-level accuracy metrics; selecting and combining models by demand-pattern cluster improved adaptation to product heterogeneity.	Transferability depends on stable clustering features.
S37 [20]	Harshvardhan et al. (2025)	Manufacturing/spare parts; North America. More than 18,000 HP print products across over 170 countries	Direct multi-horizon enterprise forecasts	LightGBM within a human-in-the-loop MLOps pipeline Baseline: Statistical and consensus forecasts	Bias, RMSE and weighted MAPE (12-month paired); enterprise ML improved accuracy and fed downstream ensemble decisions at HP.	Implementation required substantial data engineering, governance and expert review.
S38 [26]	Kenaka et al. (2025)	Manufacturing/spare parts; Asia-Pacific. Highly imbalanced intermittent spare-parts demand	Demand occurrence and quantity forecasting	Ensemble deep learning with SMOTE and focal loss Baseline: Unbalanced-data forecasting alternatives	MSE and AUC; the ensemble cut MSE by 47% and raised AUC by 32% by handling zero-demand imbalance explicitly.	Oversampling may alter the empirical demand distribution if not carefully validated.
S39 [23]	Theodoridis and Tsadiras (2025)	Retail/e-commerce; Europe. Historical sales and covariates from 45 retail stores	Retail demand time series	LSTMixer, temporal fusion Transformer, temporal convolutional network and Time-Series Mixer Baseline: Common statistical and machine-learning methods	Accuracy metrics plus ablation; LSTMixer was the strongest predictor, and ablation confirmed the added LSTM block's contribution.	Evidence is restricted to the selected store dataset.
S40 [27]	Rafi et al. (2025)	Retail/e-commerce; North America. Favorita grocery sales with holidays, promotions, oil prices and transaction variables	Multi-step retail sales forecasting	Hybrid temporal convolutional network and Transformer Baseline: LSTM, GRU and temporal fusion Transformer	MAE, RMSE and weighted MAPE (cross-validated); the hybrid reached MAE 2.01, RMSE 2.81 and wMAPE 4.22%, beating the tested deep-learning baselines.	Competition-style grocery data may understate enterprise integration costs.