

Artificial Intelligence in Supply Chain Demand Forecasting: A Structured Review of Method Evolution, Context Dependence and Research Gaps

Benchao Zhai

*Information Management, Tiangong University, Tianjin, China
adenzhai@uoem.edu.gr*

Abstract. Demand forecasting is at the starting point of the supply chain planning link, and the forecast deviation is amplified along the chain to form a bullwhip effect, causing the backlog and shortage to occur at the same time. Machine learning and deep learning methods have entered this field in large numbers in recent years, but the literature is scattered across settings as different as retail, medicine, fresh produce and spare parts. This article is based on OpenAlex structured search to review: from 2015 to 2026, there are 2,970 records, and after qualifying the sources of journals and conferences, there are 2,227 records, of which 1,210 contain artificial intelligence method terms, and 18 representative studies are deeply synthesized. The results are threefold. The method genealogy progresses along statistical methods, tree models, deep sequence models, attention and hybrid architecture, and 91.9% of attention literature appears in 2024 and later. The statistical method has not been withdrawn. 71.9% of its literature also comes from the past three years, and it remains the main control benchmark. The accuracy advantage of artificial intelligence depends on the demand form. It is stable in continuous high-density scenarios, while the evidence in intermittent demand and truncated observation scenarios is weak. Cold start, explainability and end-to-end evaluation from forecast to inventory remain three open gaps.

Keywords: supply chain management, demand forecasting, machine learning, deep learning, literature review

1. Introduction

Demand forecasting belongs to the planning link of the supply chain operation reference model, the common upstream input of inventory control, production scheduling and procurement. Once the forecast deviates from real demand, the deviation is amplified upstream along the chain, forming a bullwhip effect whose direct consequence is the coexistence of backlog and stockout on the same chain. Forecast accuracy is therefore not an isolated statistical problem but a precondition constraining the quality of several downstream decisions.

Traditional statistical methods have dominated this link for a long time. Autoregressive moving average, exponential smoothing and its seasonal extensions have clear assumptions, few parameters, interpretable results and stable performance on stationary series. When the demand is driven by

exogenous variables such as promotions, weather, holidays, etc., and results are required simultaneously across multi-level, multi-store and multi-SKU settings, the accuracy headroom of such methods narrows [1]. Machine learning and deep learning methods have been widely tried because they digest nonlinear relationships and multi-source features. The literature has accumulated rapidly over the past five years, but it is spread across conditions as different as retail, e-commerce, pharmaceutical distribution, fresh cold chain and spare parts maintenance, and method classification and application boundaries remain unclarified [2, 3].

Existing empirical studies also share an evaluation problem: most studies take the mean absolute percentage error or the root mean square error as the end point, and the prediction error is not transmitted to the inventory consequences, and the setting of the comparison benchmark is also uneven [4]. This article raises three questions: how to classify and evolve artificial intelligence methods in this field; under what conditions its advantages are established compared with traditional statistical benchmarks; and what methodological gaps exist in existing research. This article crosses the two axes of method genealogy and demand form, giving a conditional judgment under a reproducible search protocol rather than a general ranking across scenarios.

2. Literature review

There have been several reviews in this field. Douaioui et al. have critically combed the machine learning and deep learning models in supply chain demand forecasting, confirming that the relevant research will accelerate growth after 2019 [2]. Vlachos and Reddy divided the application of machine learning in supply chain management into four topics, of which demand forecasting is one, and pointed out that most empirical evidence lacks statistical significance tests for differences between models [3]. Giannopoulos et al. specifically reviewed the performance of machine learning algorithms in intermittent demand prediction [5]. Kim et al. sorted out the development vein and open issues of deep learning time series prediction from the perspective of architecture [6].

These works have established the basic facts, but two points remain undecided. First, demand forecasting, as a sub-process of the planning link, its own applicable boundaries are not pinned down separately: Douaioui et al. discuss the combination of machine learning and deep learning [2], and prediction accounts for only one of the four theme frameworks of Vlachos and Reddy [3]. Second, most of the reviews state the advantages of artificial intelligence relative to the traditional method in a summary way, and do not divide the conditions according to the demand form and data availability, so it is difficult to directly support the selection.

This article differs from the above work in two respects: the retrieval and counting protocol can be fully reproduced, and the search query and the retention of each step are fully reported; the analysis is carried out along the two axes of the method family and the application scenario at the same time, and the conditions for the establishment of the advantage are written into the conclusion itself.

3. Methodology

3.1. Search strategy and screening process

Literature identification is completed through OpenAlex's public interface, which returns structured records with digital object identifiers, and the search query can be reproduced word by word. The base query requires that the title and abstract hit the demand forecasting and supply chain at the same time, the year is limited to 2015 to 2026, the literature type is limited to journal papers,

reviews and monograph chapters, and 2,970 records are obtained. The second step is to limit the source to journals and conferences, eliminate the duplicate deposition records of repository platforms, and keep 2,227 records. The third step is to add artificial intelligence method word conditions, retaining 1,210 records, accounting for 54.3% of the records after qualifying the source. The fourth step is manual screening according to the eligibility criteria: take demand forecasting in the supply chain context as the main task, and make a quantitative comparison with at least one control benchmark; on this basis, purposeful sampling is made according to the scenario coverage, and 18 studies are comprehensively summarized. In addition to retrieval, Publish or Perish is connected to Google Scholar for supplementary scanning to cover unindexed conference documents.

3.2. Coding framework

The coding unfolds along two axes. The method family axis is divided into five categories: statistical method, tree model integration, deep sequence model, attention mechanism and hybrid architecture. The application scenario axis is divided into five categories: retail and e-commerce, pharmaceutical and healthcare, fresh and perishable, spare parts and maintenance, and manufacturing. The evaluation practice of the 18 studies of in-depth synthesis is separately coded, including the error metric family, the benchmark type, the verification scheme and whether the inventory consequences are connected.

3.3. Counting conventions

The record counts of the method family and the scene are obtained by adding word conditions on the same base query, and each type is not mutually exclusive: a paper comparing autoregressive moving average and long short-term memory networks at the same time will be included in the two categories of statistical methods and deep sequence models. These values reflect the distribution of research attention, which cannot be read as the ranking of method performance, and the categories do not sum to the total.

4. Results

4.1. Literature distribution

In the records after qualifying the source, those containing artificial intelligence method terms have increased from 2 in 2015 to 448 in 2025. The turning point occurred in 2019, which jumped from 5 in the previous year to 20; from 2023 to 2024, it increased from 90 to 271, an increase of 3.01 times, and from 2024 to 2025 the increase fell back to 1.65 times. As of the retrieval date of 2026, there are 248 records, which is an incomplete year and is not included in the trend comparison.

The record counts of the method family are 536 deep sequence models, 231 statistical methods, 191 tree model integrations, 90 hybrid architectures, and 86 attention mechanisms. The ratio of deep sequence models to statistical methods is 2.32 to 1. The difference in concentration in the past three years is obvious: 91.9% of the attention mechanism literature appeared in 2024 and later, the tree model set accounted for 88.5%, the deep sequence model accounted for 77.8%, and the hybrid architecture accounted for 67.8%. The same proportion of statistical methods is 71.9%, which shows that it did not withdraw with the popularization of the deep model, but continued to appear in recent empirical evidence as a control benchmark. In terms of scene distribution there are 582 retail and e-

commerce, 499 manufacturing, 237 pharmaceutical, 230 spare parts and 151 perishable food records (Figure 1).

4.2. The evolution of the method genealogy and the weak points of its argument

The statistical method is stable in stable and strong seasonal sequences, and the meaning of the parameters is clear. Falatouri et al. compared SARIMA with long short-term memory networks with the 37-month data of Austrian retailers [1]; the Fourier time-varying grey model proposed by Ye et al. can be compared with machine learning methods in terms of retail season demand [7]. Therefore, the benchmark method is not a formal companion.

The advantage of tree model integration is to digest table-type characteristics such as price, promotion and store attributes. The deep sequence model is good at cross-sequence learning: Bandara et al. use long short-term memory networks to cross-sequence model the weekly sales of e-commerce multi-SKU products [8]; Punia et al. combine long short-term memory networks with random forests in multi-channel retail, and compare them one by one with the four groups of benchmarks of autoregressive moving averages, neural networks, random forests and long short-term memory networks, and the result is that the single model has a win or defeat and the combination is optimal [9].

The attention mechanism enters late. Oliveira and Ramos evaluated the effectiveness of time series Transformer in retail demand, and its advantages are reflected in the long horizon and multivariate dependence, the price is higher requirements for data volume [10]. The hybrid architecture has since developed into multi-channel fusion networks with interpretation modules [11].

The genealogy itself is clear, but the argument has common weaknesses. Most studies conclude with the victory of the single-point error index. In the absence of statistical significance tests such as Diebold-Mariano, it is impossible to judge whether the improvement of the hybrid architecture relative to the single model is stable. Vlachos and Reddy have pointed out this [3].

4.3. Scenario conditions for the establishment of advantages

After superimposing the scene axis, the advantages of artificial intelligence over traditional methods change with the demand pattern. Pharmaceutical circulation is a continuous demand. After Zhu et al. incorporated the upstream supply chain information of the wholesaler layer into the characteristics, machine learning has been significantly improved relative to the statistical benchmark, and the role is the information availability rather than the model category [12]; Yilmaz et al. further coupled machine learning prediction with robust stochastic programming for medical supply [13]. Retail and e-commerce are continuous and high-density sequences, which are also the most stable category of artificial intelligence [8, 10], but this is the scenario with the most relaxed data conditions, and the conclusion should not be extrapolated to other forms.

There is a truncated observation in the fresh and perishable scene, unsold goods are discarded, and the real demand cannot be recorded. Accordingly, Migueis et al. modeled with the censored data method and superior to SARIMA [14], and Jauhar et al. introduced explainable methods in such chains [15]. Spare parts and maintenance scenarios are intermittent, low frequency and often cold start, and the evidence is the most dispersed: Caserta and D'Angelo process spare parts sequences with a very short history [16], and the review by Giannopoulos et al. judged the machine learning benefits of this scenario as conditional [5].

The measurement method directly changes the conclusion here. The mean absolute percentage error is defined as

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \quad (1)$$

Where y_t is the actual demand of the t period, $\hat{y}_t$ is the corresponding forecast value, and n is the number of evaluation periods. When the demand sequence contains a large number of zero values, the denominator approaches zero, and the indicator loses its meaning. Intermittent demand studies therefore use the weighted form

$$\text{wMAPE} = \frac{\sum_{t=1}^n |y_t - \hat{y}_t|}{\sum_{t=1}^n |y_t|} \quad (2)$$

Li et al. use this kind of measurement and connect the accuracy results to the inventory consequences [4]. Table 1 compares the four scenarios.

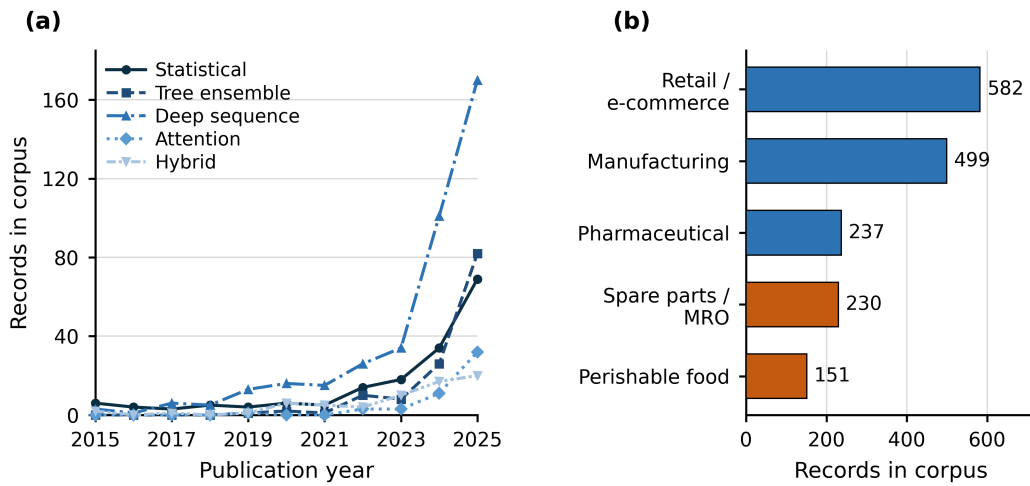


Figure 1. (a) Method-family publication trajectories, 2015-2025. (b) Records per application setting

Table 1. Context-conditioned synthesis of AI demand forecasting across four supply chain settings

Application setting	Demand pattern	Dominant AI methods	Evidence on advantage over statistical benchmarks	Binding constraint
Retail and e-commerce (n = 582)	Continuous, high density, strong promotion and seasonality effects	Cross-series LSTM, tree ensembles, time series transformers	Consistent gains on weekly multi-SKU series; hybrid designs outperform single models [8-10]	Gains concentrate where data are most abundant, so they do not transfer to sparse regimes

Table 1. (continued)

Pharmaceutical and healthcare (n = 237)	Continuous, shaped by channel structure and regulation	Feature-rich ML using upstream distributor information	Significant gains once upstream channel information enters the feature set [12, 13]	Cross-tier data sharing, not model class, is the binding input
Perishable food (n = 151)	Censored observation: unsold units are discarded, so true demand is unobserved	ML for censored targets, explainable models	Better handling of truncated distributions than SARIMA [14, 15]	Accuracy is measured against a censored target, which biases every benchmark
Spare parts and MRO (n = 230)	Intermittent, lumpy, low frequency, frequent cold start	Croston-type and ML combinations, transformers, kNN variants	Mixed and conditional on series features and error metric [4, 16]	Zero-inflated series break percentage errors; short histories limit learning

Record counts are OpenAlex title-and-abstract matches within the 2,970-record core corpus (retrieved 30 July 2026); categories are not mutually exclusive.

5. Discussion

5.1. Cold start and intermittent demand

The core difficulty of intermittent and low-frequency sequences is not the capacity of the model, but the sample size that can be learned. The application of cross-sequence learning and meta-learning in this scenario is still fragmented. Synthetic data is proposed as a path to alleviate data scarcity [17], and the prediction of spare parts under extremely short historical conditions has also been specially processed [16], but there is no stable solution for new product launch, structural breakpoints and long-term zero-demand sequences. This gap directly corresponds to the constraints of the spare parts scene in Table 1.

5.2. Explainability for planner adoption

To transform the forecast into an order action, the planner needs to approve its basis. The explainable method has appeared in the perishable food supply chain [15] and the multi-channel fusion model [11], but the existing research is mostly limited to the importance of generating characteristics, and has not tested whether the explanation has changed the actual decision-making of the planner. Therefore, the quality of explanation lacks the evaluation standard with decision-making as the endpoint.

5.3. From prediction error to inventory consequences

Most empirical evidence stops at error indicators, and the path of error into inventory decision-making is clear. At a given service level, safety stock can be written as

$$SS = z_{\alpha} \sigma_e \sqrt{L} \quad (3)$$

Among them, z_{α} is the quantile corresponding to the service level, σ_e is the standard deviation of one-step prediction error, and L is the replenishment lead time. The degree of dispersion of the prediction error directly determines the safety stock level, and only comparing the average error is not enough to explain the economic value of the method. A few studies have opened this link: Li et al. reported the linkage between accuracy and inventory consequences [4], Yang et al. put demand forecasting and inventory optimization in the same multi-agent reinforcement learning framework [18], and Yilmaz et al. embed the prediction into robust optimization [13]. In addition, the random division of training sets and test sets destroys the timing structure, and the execution rate of rolling origin verification in this field is still low [6].

5.4. Research limitations

The judgment of this article is limited by three conditions. The search is limited to English records included in OpenAlex, and non-English literature and grey literature are not included, so the coverage is limited. The record counts of the method family and the scene are obtained by keyword matching and the categories are not mutually exclusive. It only reflects the distribution of research attention and does not constitute a measure of the performance of the method. The 18 studies of in-depth synthesis are purposeful sampling. Therefore, this article reports the conditions and mechanisms, and does not give the pooled effect size.

6. Conclusion

This article examines the application of artificial intelligence in supply chain demand forecasting with a reproducible search protocol. The method genealogy progresses along statistical methods, tree models, deep sequence models, attention and hybrid architecture. 91.9% of attention research is concentrated in 2024 and later; statistical methods have not been withdrawn, and the proportion of literature in the past three years is 71.9%, and they remain the main control benchmark. The accuracy advantage of artificial intelligence depends on the demand pattern, which is stable in continuous high-density scenarios, while it is constrained by the sample size and measurement method in intermittent and truncated observation scenarios. Therefore, the selection judgment should first determine the demand form and data availability, and then determine the complexity of the model. Follow-up research can be promoted in three directions: using cross-sequence learning and synthetic data for cold-start sequences; using planner decision-making as the endpoint to test explanation methods; extending the evaluation from error indicators to safety stock and out-of-stock rate, and uniformly using rolling origin verification.

Author contribution

Benchao Zhai conceived the study, designed and executed the literature search protocol, coded and analysed the corpus, and wrote the manuscript.

Declaration of interest

The author declares no conflict of interest.

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. <https://doi.org/10.1016/j.procs.2022.01.298>
- [2] Douaoui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine learning and deep learning models for demand forecasting in supply chain management: A critical review. *Applied System Innovation*, 7(5), 93. <https://doi.org/10.3390/asi7050093>
- [3] Vlachos, I., & Reddy, P. G. (2025). Machine learning in supply chain management: Systematic literature review and future research agenda. *International Journal of Production Research*, 63(16), 5987-6016. <https://doi.org/10.1080/00207543.2025.2466062>
- [4] Li, L., Kang, Y., Petropoulos, F., & Li, F. (2023). Feature-based intermittent demand forecast combinations: Accuracy and inventory implications. *International Journal of Production Research*, 61(22), 7557-7572. <https://doi.org/10.1080/00207543.2022.2153941>
- [5] Giannopoulos, P. G., Dasaklis, T. K., Tsantilis, I., & Patsakis, C. (2025). Machine learning algorithms in intermittent demand forecasting: A review. *International Journal of Production Research*. Advance online publication, 1-43. <https://doi.org/10.1080/00207543.2025.2578701>
- [6] Kim, J., Kim, H., Kim, H., Lee, D., & Yoon, S. (2025). A comprehensive survey of deep learning for time series forecasting: Architectural diversity and open challenges. *Artificial Intelligence Review*, 58(7), 216. <https://doi.org/10.1007/s10462-025-11223-9>
- [7] Ye, L., Xie, N., Boylan, J. E., & Shang, Z. (2024). Forecasting seasonal demand for retail: A Fourier time-varying grey model. *International Journal of Forecasting*, 40(4), 1467-1485. <https://doi.org/10.1016/j.ijforecast.2023.12.006>
- [8] Bandara, K., Shi, P., Bergmeir, C., Hewamalage, H., Tran, Q., & Seaman, B. (2019). Sales demand forecast in e-commerce using a long short-term memory neural network methodology. *Lecture Notes in Computer Science*, 462-474. https://doi.org/10.1007/978-3-030-36718-3_39
- [9] 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>
- [10] 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>
- [11] Jahin, M. A., Shahriar, A., & Amin, M. A. (2025). MCDNF: Supply chain demand forecasting via an explainable multi-channel data fusion network model. *Evolutionary Intelligence*, 18(3), 66. <https://doi.org/10.1007/s12065-025-01053-7>
- [12] Zhu, X., Ninh, A., Zhao, H., & Liu, Z. (2021). Demand forecasting with supply-chain information and machine learning: Evidence in the pharmaceutical industry. *Production and Operations Management*, 30(9), 3231-3252. <https://doi.org/10.1111/poms.13426>
- [13] Yilmaz, O. F., Guan, Y., Yilmaz, B. G., Yeni, F. B., & Ozcelik, G. (2025). A comprehensive methodology combining machine learning and unified robust stochastic programming for medical supply chain viability. *Omega*, 133, 103264. <https://doi.org/10.1016/j.omega.2024.103264>
- [14] Migueis, V. L., Pereira, A., Pereira, J., & Figueira, G. (2022). Reducing fresh fish waste while ensuring availability: Demand forecast using censored data and machine learning. *Journal of Cleaner Production*, 359, 131852. <https://doi.org/10.1016/j.jclepro.2022.131852>
- [15] Jauhar, S. K., Harinath, S., Krishnaswamy, V., & Paul, S. K. (2025). Explainable artificial intelligence to improve the resilience of perishable product supply chains by leveraging customer characteristics. *Annals of Operations Research*, 354(1), 103-142. <https://doi.org/10.1007/s10479-024-06348-z>
- [16] Caserta, M., & D'Angelo, L. (2025). Intermittent demand forecasting for spare parts with little historical information. *Journal of the Operational Research Society*, 76(2), 294-309. <https://doi.org/10.1080/01605682.2024.2349734>
- [17] Long, Y., Kroeger, S., Zaeh, M. F., & Brintrup, A. (2025). Leveraging synthetic data to tackle machine learning challenges in supply chains: Challenges, methods, applications, and research opportunities. *International Journal of Production Research*. Advance online publication, 1-22. <https://doi.org/10.1080/00207543.2024.2447927>
- [18] Yang, Y., Wang, M., Wang, J., Li, P., & Zhou, M. (2025). Multi-agent deep reinforcement learning for integrated demand forecasting and inventory optimization in sensor-enabled retail supply chains. *Sensors*, 25(8), 2428. <https://doi.org/10.3390/s25082428>