Taylor & Francis
Taylor & Francis Group

International Journal of Production Research

ISSN: 0020-7543 (Print) 1366-588X (Online) Journal homepage: www.tandfonline.com/journals/tprs20

Machine learning algorithms in intermittent demand forecasting: a review

Panagiotis G. Giannopoulos, Thomas K. Dasaklis, Ioannis Tsantilis & Constantinos Patsakis

To cite this article: Panagiotis G. Giannopoulos, Thomas K. Dasaklis, Ioannis Tsantilis & Constantinos Patsakis (31 Oct 2025): Machine learning algorithms in intermittent demand forecasting: a review, International Journal of Production Research, DOI: 10.1080/00207543.2025.2578701

To link to this article: https://doi.org/10.1080/00207543.2025.2578701

Published online: 31 Oct 2025.

INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH
https://doi.org/10.1080/00207543.2025.2578701

Taylor & Francis
Taylor & Francis Group

Machine learning algorithms in intermittent demand forecasting: a review

Panagiotis G. Giannopoulos[a], Thomas K. Dasaklis[a], Ioannis Tsantilis[b] and Constantinos Patsakis[b]

[a]School of Social Sciences, Hellenic Open University, Patras, Greece; [b]Department of Informatics, University of Piraeus, Piraeus, Greece

ABSTRACT

Forecasting time series with intermittent characteristics poses significant technical challenges. Machine Learning (ML) techniques have the potential to revolutionise the existing state of practice by overcoming the challenges faced by standardised, conventional forecasting methods. In this paper, we shed light on technical details of major interest, such as hyperparameter tuning, data partitioning, training strategies, feature engineering, and mechanisms fostering the incorporation of exogenous variables, critical aspects not adequately covered by previous review approaches. Unlike earlier studies that have narrowly focussed on specific product categories, such as stock keeping units (SKUs) or spare parts, our research adopts a broader, product-agnostic perspective that synthesises findings across diverse industries and sectors. Moreover, our review underscores key elements to enhance the industrialisation of ML techniques by discussing the potential of transfer learning and other prominent methodologies that aim to overcome the practical implementation challenges of ML-oriented forecasting solutions. Finally, we identify research gaps, including the need for consistent benchmarking protocols and standardised evaluation frameworks, and we provide evidence supporting further exploration of emerging techniques (reinforcement learning, more sophisticated deep learning architectures, etc.). Our analysis could prove highly valuable for both researchers and practitioners operating at the intersection of intermittent demand forecasting and ML-based techniques.

ARTICLE HISTORY
Received 26 April 2025
Accepted 22 September 2025

KEYWORDS
Intermittent demand forecasting; machine learning algorithms; classification; tuning techniques; comparative analysis

1. Introduction

Supply chain (SC) networks have been at the forefront of strategic decision-making over the past fifty years (Kouvelis, Chambers, and Wang 2006). Within such networks, practitioners must make several judgmental calls that critically influence the positioning of an organisation against its market competitors. Analysing and forecasting demand patterns over predefined time intervals supports the optimisation of both inventory and production-oriented strategic planning (Gansterer 2015; Widiarta, Viswanathan, and Piplani 2009). Such optimisation approaches are crucial for organisations since they significantly impact cost management per product and service level offered to customers while supporting the assurance of reliability, robustness, and continuity in supply processes, crucial elements for maintaining a competitive advantage (Basten and van Houtum 2014). In this context, developing effective techniques for demand forecasting in various use-case scenarios and business sectors has become crucial for informed, strategic decision-making. To this end, numerous methods and tools have been developed to produce reliable forecasts, especially under demand conditions characterised by specific trends and seasonal patterns. These conditions are influenced by the dynamic development of products, their physical characteristics, and external market forces. While forecasting techniques represent a legacy approach for analysing smooth changes and demand fluctuations, predicting demand time series with irregular patterns remains a challenging task (K. Nikolopoulos 2021).

Among the irregular patterns that could be discerned in a time series, sporadic demand seems to be one of the most challenging to predict. Sporadic demand is a broad descriptor of any irregular demand pattern in which occasional orders punctuate prolonged sequences of zero sales. This term is commonly treated as an umbrella label in the broader literature of forecasting, while in the case of inventory-specific literature, it is frequently used interchangeably with intermittent demand. Intermittent demand reflects a time series whose unpredictability can be formalised through prolonged sequences of zero sales and generally low volumes of sales when demand happens, meaning that any non-zero observation appears as an isolated event (Syntetos et al. 2016). In some cases, scholars tend to further distinguish the intermittent

CONTACT Thomas K. Dasaklis dasaklis@eap.gr School of Social Sciences, Hellenic Open University, 18 Aristotelous St., Patras 26335, Greece

profiles based on erraticness, which reflects the variability in order sizes. In this regard, in cases of highly erratic profiles, the patterns are characterised as lumpy (Boylan and Syntetos 2010), as we discuss later in this section. Due to the challenges imposed in zero-inflated time series, many researchers have focussed on capturing the forecasting difficulties faced by classical, time series-related models (i.e. Autoregressive Integrated Moving Average–ARIMA, exponential moving average–EMA, different exponential smoothing–ES forms, etc.) and developing specialised methods to provide forecasts for time series with these irregular demand patterns. In this paper, we refer to specialised methods as conventional, which include judgmental and context-specific approaches (e.g. Croston's, Teunter–Syntetos–Babai, Syntetos–Boylan Approximation), among others.

Machine Learning (ML) serves as a promising alternative to overcome the challenges faced by conventional methods (Kourentzes 2013; Petropoulos et al. 2022). Recent advances in ML have shifted the research focus towards leveraging ML techniques to address the inherent challenges posed by these irregular demand patterns, with a special area of interest being the management of SCs (Babai et al. 2024). ML algorithms excel at uncovering non-linear relationships and dynamically adapting to complex data, offering significant improvements over conventional methods. However, sparse training datasets often challenge their performance, a hallmark of intermittent demand patterns. Sparsity issues have been continuously reported as phenomena that hinder the training phases since they are related to impurity and high entropy within the processing sets, where incomplete, noisy, or inconsistent information undermines the reliability of the algorithms (Giannopoulos, Dasaklis, and Rachaniotis 2024). Furthermore, the irregularity of such demand patterns complicates the extraction of meaningful trends, making it difficult for ML models to develop robust and generalisable predictions. It is worth noting that while much of the relevant literature has focussed on point forecasts, probabilistic forecasting is equally important. Probabilistic or quantile forecasting explicitly captures forecast uncertainty and is highly relevant for inventory and service-level decisions across SC networks. This is because the estimation uncertainty quantiles could be better incorporated into inventory policy models, thus leading to improved performance, specifically by reducing back-order costs (Sillanpää and Liesiö 2018), and, by extension, advancing service-level quality over specific lead times. In this review, we consider both streams, while also acknowledging, as documented in the literature, that probabilistic forecasting (quantile, density, volatility) remains largely under-investigated in intermittency settings (Fildes, Ma, and Kolassa 2022; Pinçe, Turrini, and Meissner 2021). This gap is also echoed in several ML-centric frameworks (L. Li et al. 2023).

When developing ML-based methods, authors tend to distinguish between intermittent and lumpy demand categories to account for how demand spikes affect the performance of forecasts. Over years of research on this topic, the research community has adopted the classification scheme proposed by Syntetos, Boylan, and Croston (2005), which applies the Average Inter-demand Interval (ADI) and the squared Coefficient of Variation (CV^2), to measure average time between non-zero demand occurrences and the relative variability of order quantities (erraticness), respectively. According to this scheme, demand is typically considered intermittent if $ADI > 1.32$ and $CV^2 \leq 0.49$, and lumpy if both $ADI > 1.32$ and $CV^2 > 0.49$. Importantly, these thresholds were mathematically derived by minimising the theoretical mean-squared-error difference between Croston's original estimator and the Syntetos-Boylan Approximation (SBA) (Croston 1972). Over time, this framework has been generalised and widely applied to classify demand as either intermittent or lumpy, resulting in terminological landscape where the two terms are often used interchangeably. In this paper, we argue that intermittent demand is the most appropriate terminology; thus, we primarily use this term throughout the paper and treat studies referring to lumpy demand patterns as a subset of intermittency, to reflect larger spikes (i.e. erraticness).

Beyond technological contributions, several efforts can be found in the literature that focus on the managerial complexity of inventory systems, facing intermittent demand patterns. For readers interested in delving into the managerial complexity of inventory systems, particularly those that exhibit typical characteristics of spare parts, as well as the interplay between inventory levels, customer service speed, and storage organisation aimed at improving service levels, we recommend visiting the studies by Bacchetti and Saccani (2012) and Basten and van Houtum (2014). These works provide valuable insights into the challenges of managing spare parts and balancing inventory costs with service efficiency. Additionally, a broader spectrum of impactful methodologies in operations research (OR) that focus on inventory control and SC simulation is thoroughly reviewed in Q. Hu et al. (2018). For more technical details on the practical application of forecasting methods in industrial environments, the studies by Syntetos et al. (2016) and Babai, Boylan, and Rostami-Tabar (2021) are highly recommended. Among other critical issues, these works cover the development of modern and dynamic decision-support systems capable of incorporating cutting-edge technologies towards different hierarchies commonly

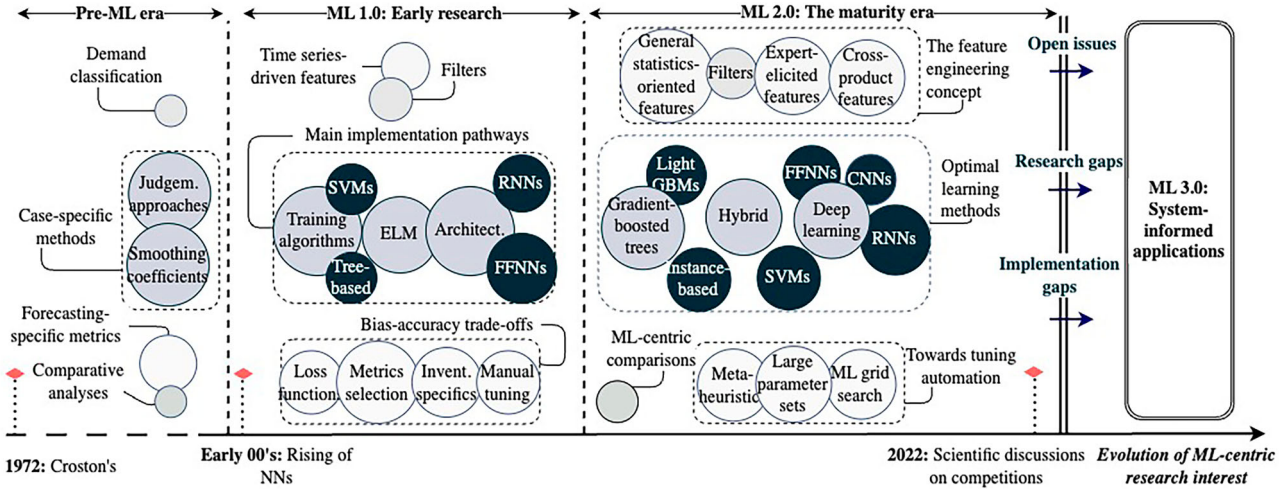


Figure 1. An overview of the evolution of ML-based intermittent demand forecasting methods. *Notation*: ML: Machine Learning, SVMs: Support Vector Machines, RNNs: Recurrent Neural Networks, FFNNs: Feed-forward Neural Networks, CNNs: Convolutional Neural Networks, LightGBM: Light Gradient Boosted Models.

identified in SCs. In this sense, the authors provide directions to bridge the gap between research and practice, themes also explored in Boylan and Syntetos (2010). Finally, significant research efforts have been devoted to the systematic study of both judgmental forecasts in light of installed knowledge bases (Van der Auweraer, Boute, and Syntetos 2019) and regression-based methods leveraging bootstrapping (M. S. Hasni, Babai, and Jemai 2019).

The intensified research efforts on ML-centric forecasting in intermittency settings have led to the emergence of multiple categories of methods with significant potential. Previous studies and forecasting competitions have demonstrated the potential of ML algorithms across various industries, extending beyond spare parts forecasting to sectors such as retail (Fildes, Ma, and Kolassa 2022; Makridakis, Spiliotis, and Assimakopoulos 2022). By critically synthesising the contributions produced over the years in the field of ML-centric intermittent demand forecasting, we conceptually infer that the evolution of ML algorithms could be classified into two eras and a preliminary (Pre-ML) one, as illustrated in Figure 1. Within the Pre-ML era, the primary research interest revolved around case-specific methods utilising smoothing techniques and judgment-based approaches, as well as discussions on the appropriateness of accuracy metrics. All of these contributions laid the groundwork for the development of the first generation of ML applications (Era 1.0). Within this era, the interest shifted towards exploring foundational training structures, including Support Vector Machines (SVMs), tree-based methods, and primarily Neural Network (NN) applications, with significant emphasis on bias-accuracy trade-offs and intensive manual tuning. Building upon the results of the first era's research, the interest in ML Era 2.0 applications has focussed on the maturation of proposed technologies, mainly by exploring the implementation-specific deep learning approaches and extending their applicability by introducing novelties in both tuning and feature engineering processes. Beyond deep learning models, some lightweight, tree-based structures were also studied in-depth. As discussed in this review, all previous contributions had a profound impact on the optimisation of ML-specific techniques that constitute integrated pipelines. Notwithstanding the impact that bootstrapping approaches may have on the evolution of ML-centric techniques, these methods are not included in the proposed illustration, as linking them explicitly to pipeline-based methodologies requires additional considerations beyond the scope of this review.

Despite the promising advancements in ML algorithms for intermittent demand forecasting outlined above, research and implementation gaps remain open across both of the two previous eras of ML-focussed research. Notable efforts aimed at synthesising research findings for two main product categories (spare parts and retailing stock keeping units-SKUs) can be found in the literature (Fildes, Ma, and Kolassa 2022; Pinçe, Turrini, and Meissner 2021). Despite the significant contributions of these papers, we argue that no systematic effort has yet synthesised findings by following a product-agnostic approach while also delving into the technical details of ML-based approaches in intermittent demand forecasting. For example, the seminal work by Pinçe, Turrini, and Meissner (2021), which aligns with our approach, identified key forecasting methods for spare parts and introduced a technical classification

based on parametric, non-parametric, and contextual methods. Similarly, the recent work of Zuvieta, Leevy, and Khoshgoftaar (2025) provides a comprehensive synthesis of forecasting techniques applicable to the aviation industry, identifying that both statistics-oriented and ML approaches may hold promise under certain conditions. Beyond efforts dedicated to spare parts, the work of Fildes, Ma, and Kolassa (2022) provided valuable insights into demand forecasting in the retail sector. Apart from the fact that our approach is not restricted to spare parts or retailing SKUs but analyses both of these two streams, which has not yet been found in the literature, another key distinction between our work and the above-mentioned efforts lies in our more comprehensive focus on ML architectures, training algorithms, implementation pipelines, and evaluation metrics, which constitute significant aspects for advancing the understanding and practical implementation of ML-based forecasting methods.

Elaborating on a product-agnostic approach, our research explores the implementation paradigms shaping the state-of-the-art in ML forecasting by developing a comprehensive taxonomy that encompasses both the methods and the underlying technical intricacies affecting the overall models' performance. Specifically, we critically assess the appropriateness of various accuracy measures recommended by context-specific literature, while exploring the divergent practices observed in current applications. Furthermore, our analysis extends to a detailed discussion of fine-tuning techniques along with systematic examinations of feature engineering strategies. This integrated approach clarifies the landscape of practical ML methodologies and offers clear pathways for implementing the most effective techniques in the context of intermittent demand forecasting. The research contributes to the advancement of ML-centric forecasts in intermittent demand by clearly addressing the following research question:

- **RQ1:** In which industry domains (across various product categories) has intermittent demand been studied using ML-based forecasting–both point and probabilistic–and what cross-domain patterns can be identified?
- **RQ2:** How do ML forecasting architectures, training algorithms, assessment metrics, and feature engineering/selection facilitate intermittent demand forecasting?
- **RQ3:** What are the prevalent research gaps and open challenges, and how could these findings be incorporated into a roadmap for future research, especially for optimiaing ML approaches for intermittent demand forecasting?

The remainder of this paper is structured as follows. Section 2 outlines the study's methodology, Section 3 briefly discusses some bibliometric elements regarding the research activity in the topic, Section 4 analyses and synthesises the literature starting from the products facing intermittent demand patterns and delving into the implementation specifications, Section 5 provides an in-depth analysis regarding the major elements accounting for the tuning and efficacy of the ML models, while in Section 6 we complement our analysis by incorporating insights from organised forecasting competitions, thus building on the context-specific knowledge they provide. Section 7 concludes the survey by summarising the key elements analysed, outlining the current open issues, and offering a roadmap for future research in the field.

2. Research methodology

For conducting our review, we have relied upon a systematic methodological approach. Figure 2 summarises the main phases involved in implementing a fully transparent and reproducible research methodology, which aligns with the methodology proposed by Briner and Denyer (2012) while incorporating elements of the PRISMA statement (Moher et al. 2009) for conducting systematic literature reviews.

Arguably, a review study should be conducted when certain factors are generally met: (i) the research topic attracts significant scholarly interest in the pertinent literature, particularly in recent years, and (ii) the volume of the underlying literature is sufficient to yield adequately generalisable conclusions (Munn et al. 2018). However, there is a notable divergence in scholarly practice regarding the research volume considered sufficient in previous efforts discussing intermittent demand. For example, the study by Van der Auweraer, Boute, and Syntetos (2019) provides an in-depth analysis comprising 44 studies, whereas the work by Babai et al. (2024) adopts a much broader sample size, encompassing 415 studies. Another comprehensive, methodology-oriented study incorporates 56 studies in its core analysis (Pinçe, Turrini, and Meissner 2021). Based on these observations, it can be argued that the scope of the research should be critically evaluated to determine the appropriate sample size. Grounded reviews tend to focus on in-depth analysis and consist of smaller samples, while larger, umbrella-type reviews offer a broader overview

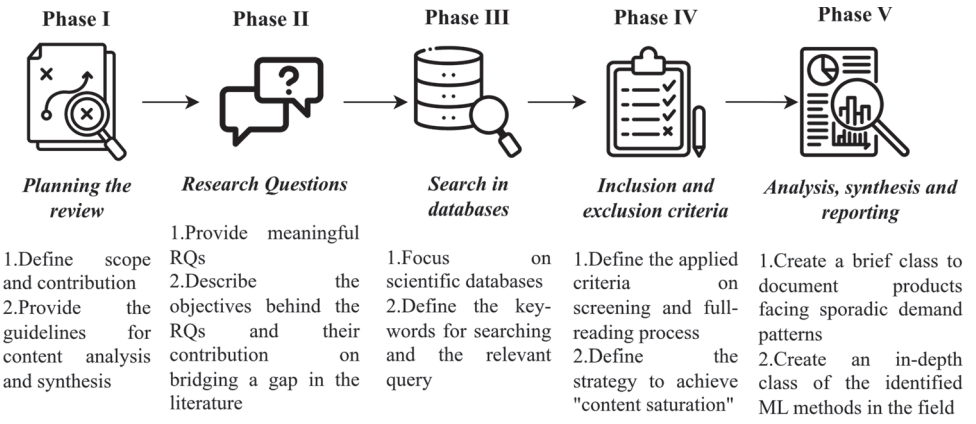


Figure 2. Research methodology.

of specific topics and frequently incorporate significantly larger samples of studies. Since the scope of this review is tailored to analyse ML-specific techniques and highlight the underlying research gaps, we strategically define the main sample to be grounded to enable in-depth analysis. Accordingly, our approach is based on a carefully curated sample of 99 papers. To reach 'content saturation', we incorporate both exhaustive searches in scientific databases (based on our systematic approach) and backward research strategies, as previously employed by Van der Auweraer, Boute, and Syntetos (2019).

2.1. Search criteria

To identify relevant publications, we consulted scientific literature databases, specifically Scopus, Web of Science, and IEEE Xplore. The overall search was conducted without time-frame restrictions. Additionally, we consulted similarly focussed studies and other publications in the field to curate an appropriate set of keywords and construct the most effective search queries. In particular, the study by Van der Auweraer, Boute, and Syntetos (2019) recommends the use of keywords such as 'forecast', 'predict', and 'spare parts' for context-specific literature searches within this domain. By combining these context-specific recommendations with a broader range of terms specialised in the field of ML, such as 'supervised' learning and 'neural network' among others, we developed a comprehensive set of search terms to locate relevant articles. During the data collection phase, we assessed whether a combination of context-specific terms and ML-related keywords appeared in the title, abstract, or literature keywords (using, for instance, the advanced search features of Web of Science and IEEE Xplore). Articles containing at least one such combination formed our sample for analysis. To facilitate the reproducibility of the sampling process, we provide the specifically designed query used for the Scopus database:

```
TITLE-ABS-KEY(("intermittent
demand"OR "lumpy demand"OR
"sporadic demand"OR "spare
part"OR "slow-moving") AND
("machine learning"OR "neural
network"OR ("supervised
learning"OR "unsupervised
learning"OR "reinforcement
learning") OR "intelligent"OR
"artificial intelligence") AND
(forecasting OR prediction)) AND
(LIMIT-TO(LANGUAGE, "English"))
AND (LIMIT-TO(DOCTYPE, "ar",
"cp"))
```

2.2. Inclusion criteria

After concluding the search process across the three scientific databases, we collected a sample of 254 papers, comprising 124 articles and 130 papers published in conference proceedings. It is worth noting that, although we initially identified some of these papers through multiple databases (Web of Science and IEEE Xplore), all selected contributions were also available in the Scopus database. Additionally, we focussed solely on reviewing articles published in peer-reviewed journals and conference proceedings. This approach minimises the introduction of 'noise' in our analysis and aligns with established methodologies employed in prior systematic reviews (Snyder 2019), especially in fields that have traditionally been at the forefront of research trends, such as forecasting within the SC domain. We also excluded all non-English written articles.

The sample of the collected publications was meticulously assessed for eligibility by applying a three-fold, forward approach consisting of screening and full reading. In the first stage of the three-fold selection process, we reviewed the titles and abstracts of all collected articles (initial assessment of the relevance of the articles to the scope of our review). The application of this first-level eligibility criterion resulted in the exclusion of 31 articles and 75 papers from conference proceedings. In the second stage of the screening process, we thoroughly reviewed the introductions and conclusions of the eligible articles. This process led to the exclusion of 26 articles and 16 papers from conference proceedings. The excluded articles were primarily focussed on domains extending beyond the broader context of SC management, such as energy and portfolio management, breakdown estimation in heavy machinery equipment, as well as risk-oriented studies related to medical cases and environmental pollution. Regarding the papers from conference proceedings, in this phase were excluded papers that do not primarily focus on intermittent demand or focussing on the development of tools for facilitating maintenance-oriented, rather than inventory-specific decision making. In the third stage of eligibility assessment, we conducted a full reading of the remaining 106 papers. At this stage, 3 articles and 16 papers from conference proceedings were excluded. These studies focussed on modelling interfaces in B2B (business-to-business) applications, developing online tools that incorporated certain established techniques and other applications that primarily examined the managerial implications of forecasting, without focussing on the evaluation, comparison, or proposal of forecasting techniques using ML methodologies.

Following the above approach, we finalised the collection of 83 relevant papers (60 articles and 23 papers from conference proceedings), completing the forward search process. To achieve content saturation, we extended the search using a backward search process. Specifically, we focussed on studies of similar scope as well as influential works in the field to identify articles that may not have been captured during the forward search. By reviewing other systematic reviews in the field, we examined the references they cited and screened the publications included in those references, following standard screening procedures. The same approach was applied to the five most cited studies, which were included for content analysis, as well as to relevant works published within the last ten years. This process resulted in the identification of 16 additional studies, which were meticulously analysed and incorporated into the core set of relevant studies. Arguably, the drawn process is streamlined with the protocols governing systematic reviews and ensures that any relevant publication has been incorporated into the core sample of analysis and synthesis. To increase the transparency and reproducibility of our approach, the interested reader is referred to Figure 3, which provides the detailed steps we conducted for achieving content saturation.

2.3. Content analysis and synthesis

Thematic content analysis provides a robust methodological framework for systematically examining qualitative data, enabling researchers to identify, scrutinise and interpret patterns of meaning, commonly referred to as 'themes', within qualitative datasets (Dasaklis et al. 2022; Elo and Kyngäs 2008). In this study, we have meticulously adopted a thematic content analysis approach to distill research areas and discern overarching themes from the corpus of the retrieved literature. By employing content analysis techniques, we facilitate the extraction of specific qualitative attributes, which subsequently enables the grouping of articles sharing common characteristics through content synthesis methods. To systematically extract relevant information across the selected instances, the content analysis was conducted within the following coding themes:

(1) Product category, industry and type of demand manifested.
(2) Implementation details, comparative analyses and accuracy measures, assessment of different configurations (if any), fine-tuning techniques, and limitations stated.
(3) Dataset details: number of time series or products, time window (day, week, month, etc.), partitioning strategies and training evaluation measures, data impurity-related (i.e. augmentation) techniques (if any).

Upon collating the qualitative data, narrative-oriented techniques were applied to synthesise the extracted information and create both context and methodological-specific classes, aiding to facilitate our research scope. Narrative synthesis is a respected approach that focuses on elucidating how studies, which explore different facets of the same phenomenon, can be coherently summarised to construct a holistic understanding of that phenomenon (Campbell et al. 2019). By applying such content-specific analysis and synthesis, our approach focuses on how input variables (features) influence the accuracy of ML-based forecasting techniques tailored for intermittent demand, while examining methodological approaches for incorporating knowledge-based and statistics-engineered features. By assessing these aspects,

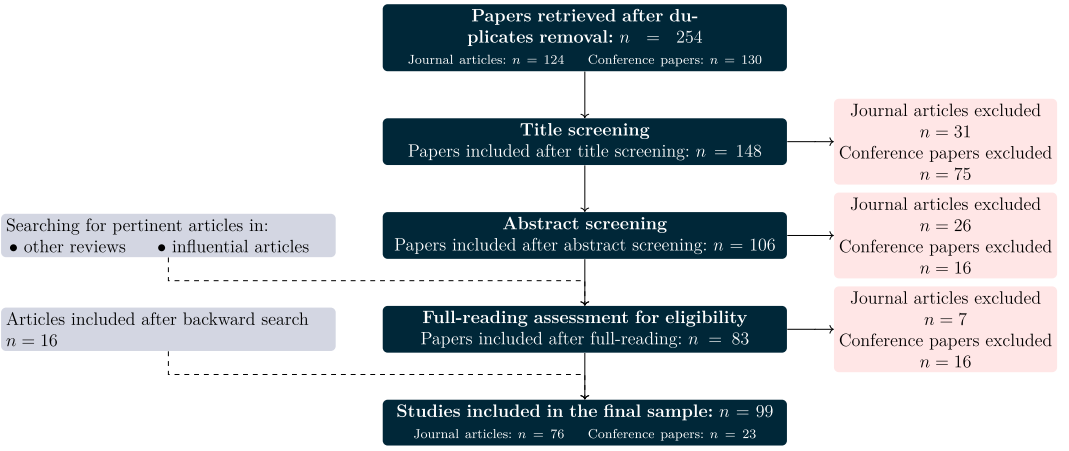


Figure 3. The integrated (forward-backward) approach, followed for achieving content saturation.

including data impurity-related issues, refined training strategies, and the integration of exogenous variables, we synthesise the broader landscape of integrated methodological pipelines that enhance the accuracy potential of ML-based methods in intermittent demand forecasting-technical-oriented aspects not covered so far in the relevant literature.

To ensure maximum consistency in our research and effectiveness in identifying critical themes in the retrieved literature, we use the Holsti coefficient, which calculates the degree of agreement between two or more coders who independently classify or code the same content and is calculated using Equation (1). Following similar approaches presented in the literature (Dasaklis et al. 2024), after collecting the articles and deciding on the thematic analysis dimensions (coding), two authors independently analysed the literature. Subsequently, the degree of convergence in content analysis was calculated using the Holsti index, which was 0.85; thus, reflecting a strong consensus and suggesting that the coding process was quite consistent and reliable.

$$\text{Holsti Coefficient} = \frac{2 \times n_{\text{agreement}}}{N_{\text{coderA}} + N_{\text{coderB}}} \tag{1}$$

3. Bibliometric analysis

In this section, we provide a brief overview of the descriptive statistics for the selected papers, which will be used for content analysis, synthesis, and classification. This descriptive statistical approach complements content analysis by providing a quantitative overview of the literature, thereby enhancing the robustness of our research (Donthu et al. 2021). In particular, complementing qualitative research with statistical artifacts improves the organisation and display of key constructs in the relevant literature, such as publications by year and journal, and highlights prevailing research directions (Aria and Cuccurullo 2017). By focussing on these descriptive statistics, we align with established practices in bibliometric analysis, which aid in understanding the evolution of research topics and identifying significant contributions within a field. Specifically, we analyse the distribution of publications over time, depicted in Figure 4, illustrating the year-wise trend, to identify patterns and growth in research activity related to ML-centric intermittent demand forecasting. The bibliometric analysis presented demonstrates the intensity of research activity concerning the use of ML algorithms for forecasting intermittent time series, especially over the past five years. In addition, while forecasting as a topic has consistently been a central research stimulus, the systematic application of ML algorithms to address this problem appears to have gained significant traction over the last 5 years. This trend could be justified as a result of the advancement of computational systems, providing considerable research flexibility for testing complex, multi-level learning techniques, as well as the broader maturation of ML technologies through interaction with other fields. Blending the insights obtained from both bibliometric and qualitative analyses, as presented in the following sections, we contribute to the literature by presenting key insights for the development of new ML-based pipelines for sporadic demand forecasting.

4. Analysis of the literature

In the broader literature, two main product categories have garnered major research interest: (i) spare parts, and (ii) products designated as retailing SKUs. As previously mentioned, our work is conducted in a product-agnostic manner by consolidating research progression in both of the above research streams. In this regard, our work is grounded in bridging the two research streams by distilling shared principles and transferable design choices for

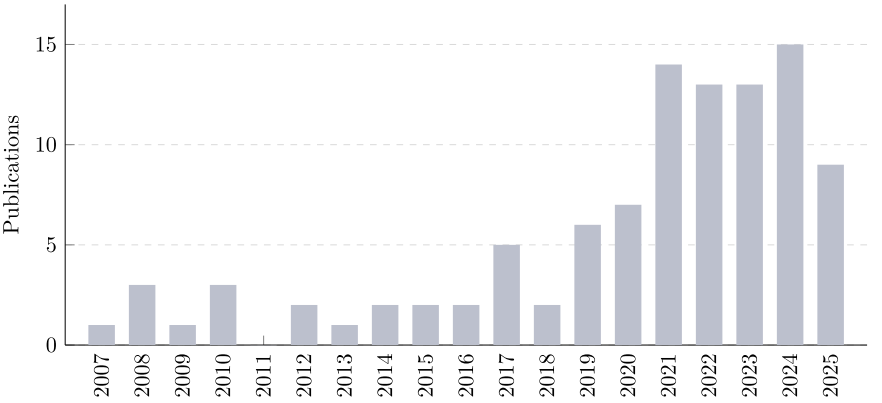


Figure 4. Year-wise distribution of publications.

developing ML pipelines. In this section, we deliberately categorise the ML methods using an industry-specific approach. This approach facilitates the identification of industries facing intermittency and enables the discussion on specific (exogenous) features, having a significant influence on the performance of ML-centric forecasts. Across the industries we identified, demand tends to concentrate in specific products (e.g. spare parts in automotive). Overall, we identify six (6) broader industrial sectors, accumulating a substantial number of research efforts (> 15), along with several additional industries that we include in the broader industrial class 'Miscellaneous', as they concern products with physical characteristics and inventory-related attributes differentiated from the foregoing classes.

To allow the reader to follow the in-depth granularity presented for the identified techniques across each industry, we refer to Figure 5, which represents our proposed taxonomy of the identified methods. Within the scope of our research, we avoid employing the conventional classification of ML methods into supervised, unsupervised, reinforcement, and deep learning, an approach suggested by previous studies (e.g. Procopio et al. 2023), as we believe it is not meaningful in our case. The reasoning behind this argument is that the majority of the identified methods rely on a supervised approach, mainly NNs and decision trees (DTs), either in simple or hybrid forms. As for unsupervised methods, we identified them mainly as components of synthetic techniques rather than standalone methods, while no reinforcement learning approaches were identified at all. In this regard, we developed a tailor-made taxonomy that describes the broader landscape of intermittent demand forecasting. Within the proposed taxonomy, a significant part is the synthetic methods, which are classified into hybrid, ensemble-based, transfer-learning, and meta-learning. Hybrid methods are those developed by combining two or more techniques, at least one of which belongs to the ML domain. Correspondingly, we define ensemble-based methods as those that integrate multiple base learners to improve predictive accuracy and robustness (Ganaie et al. 2022). Transfer- and meta-learning approaches, constitute standalone classes in ML, but we classified them into the broader synthetic class, since the sample of such methods was found to be considerably smaller in comparison to other categories.

Beyond synthetic approaches, the proposed taxonomy encompasses two additional classes: NN-based methods and shallow ML baselines. The latter includes established methods in the field of ML that present the common themes of limited architectural depth and low computational overhead, such as SVMs, tree-based models, and instance-based methods. The term shallow ML baselines was selected to reflect that in the broader ML literature, such methods serve as reference points, especially in cases where efficiency is prioritised. Within this class, instance-based methods refer to algorithms that generate predictions by identifying neighbourhoods of similar historical instances–typically based on distance metrics, without explicitly building on parametric assumptions (i.e. k-Nearest Neighbors (k-NN)).

4.1. ML-centric forecasts in different industrial settings

4.1.1. Automotive industry

The need to maintain multiple vehicle types, aging fleets, and variable failure rates of components creates considerable uncertainty in maintaining the required inventory levels of different types of spare parts, ensuring customer satisfaction, reducing lead times, and minimising costs in the automotive industry (Q. Hu et al. 2018). In this context, the development of reliable forecasting tools has been a long-standing research focus. The first detailed investigation of instance-based approaches for providing forecasts under intermittent conditions was carried out

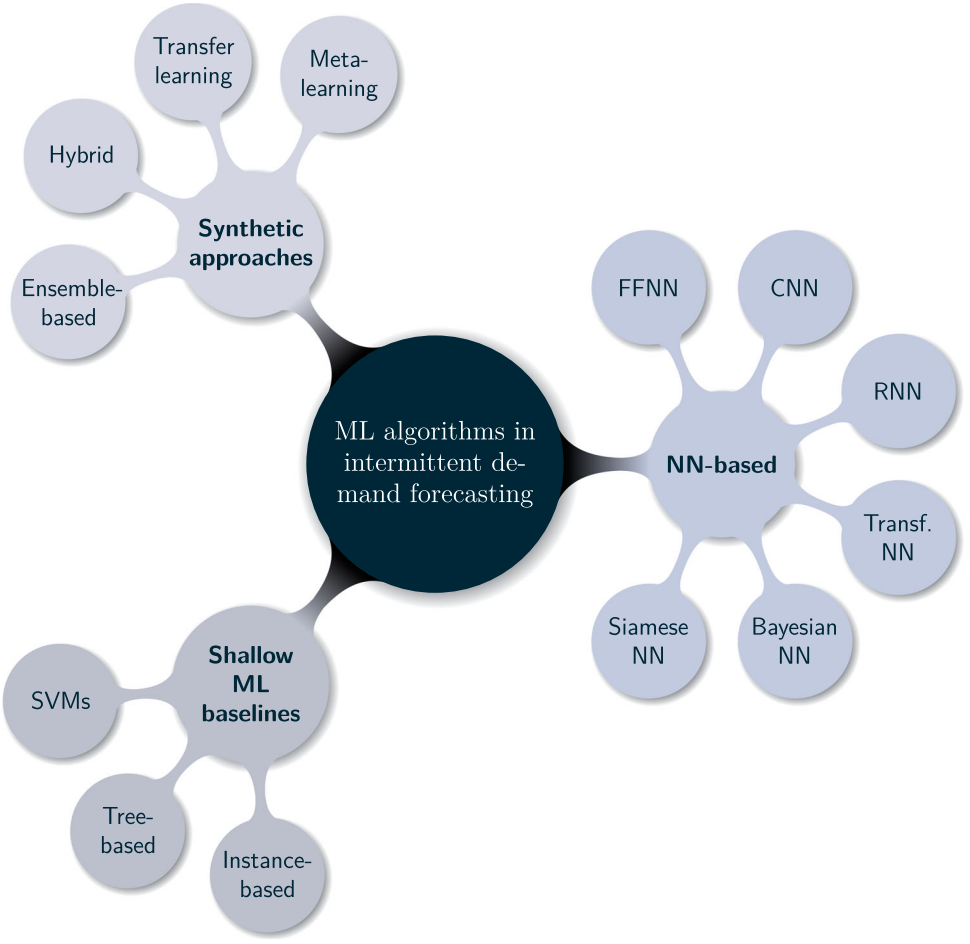


Figure 5. Taxonomy of ML algorithms within the intermittent demand forecasting context. *Notation*: ML: Machine Learning, NN: Neural Network, SVMs: Support Vector Machines, RNNs: Recurrent Neural Networks, FFNNs: Feed-forward Neural Networks, CNNs: Convolutional Neural Networks, Transf. NN: Transformer Neural Networks.

using simulated data related to automotive spare parts demand (K. I. Nikolopoulos, Babai, and Bozos 2016). This study focuses on producing forecasts by employing neighbour distance vectors and analysing how the algorithm performs when assigning higher weights to closer neighbours.

The use of feed-forward neural networks (FFNN) within the automotive industry was proposed in the study by Kourentzes (2013), which introduces two distinct supervised learning variants: (i) 'NN-Dual', which was designed to capture demand size and intervals; and (ii) 'NN-Rate', which was introduced to forecast demand quantities. The study also examines the application of the Levenberg-Marquardt optimiaer with various initialisation schemes. Another study highlighting the ability of FFNNs to provide accurate forecasts, without requiring extensive calibration, focussed on reducing the complexity of automotive SCs in Iran (Jafarzadeh, Rahman, and Wahab 2012). In this case, the approach leverages shallow architectures coupled with back-propagation training algorithms, achieving a satisfactory balance between complexity and accuracy. Further progress in single-layer FFNN applications is evident in the work of Lolli et al. (2017), who argue that while back-propagation methods are effective compared to Extreme Learning Machines (ELMs), they require substantial effort in parameter tuning, which can pose challenges for very large datasets. In the vein of back-propagated training schemes, the incorporation of Bayesian-regularised structures has been found to hold promise, especially for initialising the networks in a fast-generalisable manner (Y. Chen, Liu, and Yu 2010).

Beyond shallow ML baselines and single-layered NN architectures, the study by AlAlaween et al. (2022) examines deep FFNN architectures tailored to both demand and price prediction for hybrid electric vehicle spare parts. Aligned with the focus on deep architectures, İfraz et al. (2023) examine the demand forecasting capabilities for spare parts in a bus fleet. A specialised class of vehicles is designed for military operations (spare parts demand for heavy-duty vehicles is characterised as intermittent). In this context, the reliability of ML algorithms has been

studied through applications focussing on K-X tanks (Kim, Hwang, and Doh 2023; Kim, Kim, and Han 2023). Emphasis is placed on the development of FFNNs based on deep architectures, enhanced with specialised scaling techniques, to improve reliability.

The ability of Recurrent NNs (RNNs) to produce accurate forecasts by capturing complex temporal relationships has been extensively studied. A recent application on a large monthly car sales dataset proposed the development of a six-layer architecture built on an RNN-LSTM (Long Short-Term Memory) framework, enhanced with a modified Adam optimiaer (Chandriah and Naraganahalli 2021). The rationale behind the framework lies in the ability of LSTM units to handle long-term dependencies and the capacity of RNN units to capture sequential patterns. Beyond FFNNs and RNNs, which are considered a conventional class of networks for time series forecasting (Petropoulos et al. 2022), recent novelties in the field of NN-centric forecasting include the application of convolutional NNs (CNNs), typically employed in image analysis (Dhillon and Verma 2020). The innovation of this method lies in transforming tabular data into 3D voxel images for spatial context learning. The framework utilises Neural Architecture Search (NAS) based on the differentiable architecture (DARTS) to optimiae both the input-embedding architecture (which maps tabular features to 3D voxel images) and the 3D CNN architecture (Lee et al. 2022).

For shedding light on which of such deep-learning approaches seem to hold significant promise in the automotive industry, the work of Z. Ma, Wang, and Zhang (2021) compared different architectural schemes of RNNs, FFNNs, CNNs, and Transformer Neural Networks (TNNs), leading to the conclusion that LSTMs still hold significant promise in the field. Some systemic approaches have also been proposed towards preserving the scalability and reducing the re-training of ML schemes. For instance, the work by Sobral et al. (2024) develops an end-to-end MLOps (ML Operations) pipeline for multi-series forecasting by incorporating Temporal Fusion Transformers (TFTs) and coupling the scheme with context-specific, exogenous features.

Significant research interest has also been directed towards developing hybrid methodologies, which leverage improved capabilities by combining at least two algorithms. In this context, the work of S. G. Li and Kuo (2008) stands out as one of the initial efforts to highlight the limitations of conventional methods, particularly their inability to incorporate explanatory features from the broader SC framework. To optimiae operational management in an automobile spare parts warehouse, the study proposes an enhanced fuzzy neural network (EFNN)-based decision support system. Similar efforts could be identified in applications tailored to 4S shops (automobile service centres, which cover operations of Sales, Spare parts provision, Service, and Surveys on customer feedback) (Y. Liu et al. 2019; J. Wang et al. 2020), in which engineering of explanatory, weather-derived features has been found to present a significant influence in the forecasting process, when both ELM and Support Vector Machines (SVM) algorithms were applied.

Apparently, hybrid approaches incorporating simpler supervised techniques, such as trees and SVMs, have been of importance in this field. In the case of spare parts for heavy-duty vehicles, integrating SVMs with nature-inspired meta-heuristics has proven to be a promising approach for enhancing forecasting accuracy (Jiang, Huang, and Liu 2021). This study proposes the application of AUSVM (adaptive univariate SVM), a technique that integrates the SVM and adaptive particle swarm optimisation (APSO) algorithm, to fine-tune the engineered hyperparameters. Other hybrid schemes have focussed on integrating Delphi and fuzzy evaluation techniques to model the relationship between forecast and external factors using SVMs (X. Li, Zhao, and Pu 2019). Another line of research explores hybrid models that couple SVMs with FFNNs (Han et al. 2017). In these synthetic forecasting modules, the SVM is used to model the linear component of the demand signal, while the FFNN captures the nonlinear residuals, as shown in the case of spare parts demand forecasting in urban rail vehicles.

Significant progress has been made in hybrid methods centred on tree-based approaches. Promising avenues include combining CatBoost, a gradient-boosted DT classifier, to predict demand occurrences and LightGBM to estimate the demand quantity (Rožanec, Fortuna, and Mladenić 2022). In the context of business-to-business trading environments for automotive spare parts, the authors in Nguyen, Nguyen, and Nguyen (2025) propose incorporating K-means clustering to group codes with similar sales trends and demand patterns, with the purpose of identifying the optimally-fitted approach in each cluster. This systemic scheme validates that, in clusters with high intermittency, tree-based schemes (i.e. gradient-boosted) hold significant promise. In a similar vein, tree-based techniques were employed as part of a dynamic safety stock optimisation model, in which the forecasting module was developed by coupling a tensor-based LightGBM model with a linear decay correction mechanism to address the obsolescence commonly included in the management of spare parts (Fan et al. 2024).

Ensemble learning protocols have also been proposed in the field of automotive spare parts forecasting. For instance, the study by Chien, Ku, and Lu (2023) introduces a systematic framework in which

base models, including Syntetos-Boylan approximation, Holt-Winters Seasonal (bootstrapping regressor), ARIMA, and XGBoost (eXtreme Gradient Boosting), are aggregated using a stacking module with a linear regression meta-learner. At this point, we note that the term ensemble learning is commonly used interchangeably with tree-based approaches. Despite some architectural similarities, these two schemes differ significantly on the implementation side. In the case of DTs, this could be attributed to the divergent methods that may be applied in a stacking module, which is in contrast to the predefined nature of modules constituting a hierarchy. Since ensembles, especially stack-based, excel in forecasting scenarios of high variability and intermittency, they should be preferred when dynamic aspects are incorporated (Dietterich 2000; Makridakis, Spiliotis, and Assimakopoulos 2020). Beyond ensemble learning approaches, the idea of developing transfer learning mechanisms has recently gained traction. The primary advantage of such mechanisms is their ability to significantly mitigate the greed for data and the time-consuming nature of learning phases in ML models, a phenomenon also explored in the literature (Zhuang, Yu, and Chen 2022). The latter research focuses on developing a tree-based (LightGBM) two-fold model, where each level is used for classification and regression tasks, respectively. By incorporating feature engineering approaches to integrate exogenous parameters influencing demand, the authors significantly enhance the model training process.

4.1.2. Aviation and naval industries

Demand forecasting for spare parts in the aviation industry has shown remarkable progress with applications pertinent to both civil and military operations. In this context, the distinction between repairable and non-repairable spare parts has been a focal point, highlighting the different inventory management requirements for each category (Pogačnik, Duhovnik, and Tavčar 2017). FFNNs have emerged as a promising avenue within the aerospace industry. Expanding on existing architectures, such as those proposed by Gutierrez, Solis, and Mukhopadhyay (2008) and Lolli et al. (2017), the research of Babai, Tsadiras, and Papadopoulos (2020) introduced enhanced FFNN configurations. Their innovation lies in the incorporation of descriptive factors of intermittency, such as the number of periods between non-zero demand occurrences (NZ) and the periods from the last zero-demand occurrence to the target period (FZ), alongside the Bayesian regularisation back-propagation algorithm. The work of de Oliveira, Jorge, and Rocha Filho (2020), further supports that incorporating 'specialised knowledge', reflected in the engineering of context-specific features, adds significant value to the performance of single-layered FFNNs. Similarly, Shafi et al. (2023) focuses on predicting critical spare parts using FFNNs, which are optimiaed for topology and trained on historical demand datasets. On the topology side, the development of Bayesian NNs constitutes an innovative approach to neural modelling, since they offer complementary advantages for intermittent demand forecasting in changing service-logistics contexts by directly encoding conditional dependencies among failure drivers and integrating expert knowledge, as demonstrated for the case of unmanned aerial vehicle support (Boutselis and McNaught 2019). The potential of TNN has been evaluated for the forecasting of the demand for civil aircraft spare parts (Zhang, Xia, and Xie 2024). Key techniques in this research include an encoder-decoder architecture with self-attention mechanisms, which have been proven to effectively capture long-term dependencies. The study demonstrates that TNN-centric approaches hold promise when sparsity levels increase.

Other novelties have focussed on exploiting Croston's method by coupling it with RNNs (i.e. LSTMs) for facilitating the development of two-fold schemes decomposing the forecasting task into interval and quantity prediction (J. Liu et al. 2020). For the case of repairable spare parts for an aircraft fleet, Guo et al. (2017) propose a hybrid, double-level approach that integrates multiple features affecting demand. This research builds upon assessing several forecasting methods (Genetic NN, Linear ES, Secondary ES, Cubic ES, and Grey Model) at the first level while integrating them into a weighted optimisation layer using GA. The GA is also applied to optimiae the neurons' initialisation phase, while back-propagation is employed as the training algorithm. Another notable contribution to the civil aerospace aftermarket is Bombardier's deployed integrated forecasting pipeline, which couples tree-based ML (with rich features, including flight data) to model the intermittent components–demand size and inter-demand interval–with classical time series models, and ensembles the outputs by demand-pattern groups to produce robust spare-parts forecasts (Dodin et al. 2023).

Beyond these paradigms, hybrids coupling ML modules with meta-heuristics are of importance. In alignment with the practices applied in automotive spare parts, the use of particle-based optimiaers emerges as a common practice to identify hyperparameters to fine-tune (i.e. learning rate, batch sizes, etc.) and generalise the results of NN-centric forecasts, as presented in Song et al. (2021). Further exploitation of particle meta-heuristics has been made in applications centred around fuzzy networks. In this vein, Z. Li et al. (2015)

developed a method that utilises the high resolution of the wavelet transform to analyse the time series and then applies a Particle Swarm Optimisation (PSO)-enhanced fuzzy network to the features related to the time series reflecting the demand of spare parts in military aircrafts. In the case of weapon systems, where the demand for unscheduled and non-repairable spare parts is both costly and critical, the accuracy of forecasts has been recognised as vital. Choi and Suh (2020) focussed on detailed comparisons among several leading techniques in the field, including tree-based approaches, back-propagation-trained FFNNs, and SVMs, to identify the optimal method, arguing that data mining techniques facilitate the identification of influencing factors of demand, and concluding that tree-based schemes hold significant promise.

Hybrid forecasts are of importance in the naval industry, especially in the case of critical spare parts. In this vein, the work of Anglou, Ponis, and Spanos (2021) integrates the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method for outlier detection and removal with multiple ML-based regression tools such as Random Forest (RF), Principal Component Regression (PCR), and Generalised Linear Model (GLM); with the results demonstrating that RF consistently outperformed GLM and PCR, particularly after the removal of outliers. Other hybrid forecasting tools include the integration of convolutional layers, such as the work by Huang et al. (2024), which proposes the development of a three-layer combined prediction model using a back-propagation trained NN and showed promising results against statistical benchmarks.

4.1.3. Electronics & electrics industry

Another industry with significant complexity in managing spare parts inventory is the electronics sector, where products often become quickly outdated, necessitating immediate replacement, while others remain in long-term operation. This creates several challenges in inventory management, requiring the maintenance of a large volume of diverse spare parts. To reduce complexity, significant progress has been made in the field, with the majority of applications focussing on the exploration of simple and fuzzy NN forms. The first attempt to apply NNs to forecast intermittency patterns with large demand spikes (i.e. lumpy) was presented by Gutierrez, Solis, and Mukhopadhyay (2008) and later expanded by Mukhopadhyay, Solis, and Gutierrez (2012), with a focus on addressing the management of daily demand faced by an electronics distributor operating in Monterrey, Mexico. These works seem to have catalysed the development of research 'value chains' in the field, implementing extensive comparisons among methods based on research-driven extensions–an approach that, as discussed below, has become an established practice in the domain. The research of Gutierrez, Solis, and Mukhopadhyay (2008) proposes a single-layered MLP trained with back-propagation algorithm, to improve the accuracy achieved by conventional methods, while the work of Mukhopadhyay, Solis, and Gutierrez (2012) extends the comparison framework to include both conventional and time series-related methods. The authors validate the prevailing notion that relates data availability to the performance of FFNNs. In a similar 'value chain' appears to be located the work of Solis, Mukhopadhyay, and Gutierrez (2010), which argues that although significant performance advancements can be achieved when statistics-oriented metrics are applied, this may not be generalisable to inventory systems. This study stresses that forecasting performance based on statistics-based indicators may not be directly translated into inventory systems efficiency (Solis 2015).

Notable contributions in the domain were also oriented towards incorporating fuzzy logic into FFNN architectures. In this vein, the work of F.-L. Chen, Chen, and Kuo (2010) comparatively assessed a moving back-propagation and moving fuzzy NN (FNN), which was based on the analytical hierarchy process (AHP) to rank the engineered features, encapsulating their impact on advancing the training performance. The introduction of fuzzy layers for the development of FNNs is a significantly differentiated approach to previous hybrids (that is, S. G. Li and Kuo 2008). Further progress in the FNN vein was noticed in the study by Cao and Li (2014), which incorporates a PSO to facilitate robustness and generalisation of FNNs, in the case of critical electrical spare parts. To clarify the landscape for these elements, the authors in Vaitkus, Zylius, and Maskeliunas (2014) compare SVM and FNN along with established conventional approaches to identify under which circumstances each algorithmic class may be preferable. The authors argue that adaptive methodologies, based on residual calibration and SVM-based regression regressors, lead to performance among the assessed combinations.

Combining auto-regressive algorithms with NNs presents another interesting pathway. In this vein, combining ARMA (Auto-Regressive Moving Average) with back-propagation trained FFNN, facilitates the identification of discernible trends and the modelling of non-linear residuals, which was found prominent in the case of spare parts forecasting for hydroelectric equipment (Z. Ma et al. 2022). A similar scheme was followed in Quiñones-Rivera, Rubiano-Ovalle, and Alfonso-Morales (2023), which relies on exogenous features, modelled by the statistical multivariate ARMA with eXogenous (ARMAX), to foster the performance of

an FFNN. Repairable spare parts in healthcare electronics constitute another line of research, in which the potential of ML-based hybrids was illustrated in contrast to both conventional and simpler ML architectures (El Garrab et al. 2023).

Except for standalone applications, ML methods appear to contribute to the development of data-driven frameworks that contextualise the dynamics in SCs, supporting inventory- and production-related decision-making. For instance, the UNISON framework (Fu and Chien 2019) leverages the integration of RNNs with conventional and time series methods by elaborating on the protocol *aggregate-then-forecast-then-disaggregate*. Specifically, the authors employ temporal aggregation to smooth demand patterns. Subsequently, an RNN with four hidden layers is utilised to capture nonlinear relationships, with outputs disaggregated back to their original resolution. On the technical side, the combination of parametric and conventional methods with RNNs is a promising scheme in the broader field of electronics inventories (i.e. semiconductors) (Fu, Chien, and Lin 2018). Ensemble schemes were also used for the development of data-driven frameworks. For instance, the AutoGluon-Tabular Framework, which is tailored to locate the underlying demand pattern, based on established classification protocols, and then applying a diverse set of base learners (NNs, trees, instance-based, etc.) to identify the optimal and best-generalisable method for predicting the demand of each series (Caserta and D'Angelo 2024). An effort to capitalise on Reservoir Computing (RC) frameworks was identified in de Melo Menezes et al. (2015), which argues that RC holds promise over traditional MLPs, ARIMA, and Croston's method, as it can effectively operate under circumstances of limited data structures by relying solely on time series-based features. For contextualisation, we note that RC constitutes a family of recurrent models known for their training efficiency, which is achieved because they do not rely on end-to-end backpropagation, but leverages asymptotically stable recurrent dynamics to minimise the computational cost of training (Ceni and Gallicchio 2024).

Beyond point forecasts, RNN-centric hybrids have recently been implemented for quantile forecasting purposes. For instance, the work of Wu, Fu, and Xia (2025), introduces a production-oriented hybrid that integrates an RNN, for capturing temporal dependencies, with a quantile regression head to estimate the target quantile τ. The proposed scheme also integrates a PSO routine that advances training by tuning weights and key hyperparameters, aiding convergence and mitigating local minima trapping.

4.1.4. Heavy industries

In the heavy industry sector, timely fault prediction and, by extension, effective forecasting of spare parts demand are of critical importance, especially because heavy-duty mechanical equipment is expensive and maintenance costs are substantial. The core of forecasting paradigms in this setting is shaped by FFNN, RNN, and tree-based approaches, both in their simple and hybrid forms. For example, Aktepe, Yanık, and Ersöz (2021) explore the impact of context-specific features on improving the accuracy of SVM, trained with polynomial kernels, and FFNNs developed in a deep architecture with two hidden layers. Similarly, Hoffmann, Lasch, and Meinig (2022) proposed the use of single-layer FFNNs trained with algorithms such as back-propagation and Levenberg-Marquardt. From the RNN side, Amin-Naseri and Tabar (2008) evaluate Elman-type recurrent networks, whereas the authors in X. Hu and Tang (2024) examine LSTM architectures trained solely on time series features. Across their empirical settings (process/petrochemical and enterprise datasets), the authors report that RNN-based schemes can hold significant promise. The main difference between these two schemes lies in memory handling: Elman RNNs use a single recurrent state and are prone to vanishing gradients, whereas LSTMs employ gated cells that retain longer-range dependencies–an advantage during prolonged periods of zero demand.

SVMs present another attractive research tool in several heavy industries (Babaveisi et al. 2023). The first attempt to create a hybrid method tailored to the petrochemical industry was based on the integration of SVM and logistic regression models (Hua and Zhang 2006). Tree-based hybrids present another research stream, with a recent technique developed for after-sales spare parts demand in large engineering manufacturing enterprises (Hong et al. 2023). This approach combines a stacked auto-encoder to extract hidden features, tensor Tucker decomposition to smooth anomalies and capture evolutionary trends, and the LightGBM algorithm for implementing forecasts. In parallel, within the energy industry, a stacked generalisation ensemble has been proposed that combines a conventional intermittent-demand method (i.e. TSB) and a Random Forest as base learners, coupled with an Elastic Net meta-learner, yielding promising results for highly erratic demand patterns (Tsao et al. 2019).

FFNN-centric hybrids present another significant stream, with applications covering petrochemicals (Nasiri Pour, Rostami Tabar, and Rahimzadeh 2009), mining (Rosienkiewicz 2020; Rosienkiewicz, Chlebus, and Detyna 2017), and power production industries

(Xing and Shi 2019). The first ML-centric decomposition scheme has relied on a multi-layered perceptron for predicting non-zero demand intervals and a conventional recursive method to estimate the quantity (Nasiri Pour, Rostami Tabar, and Rahimzadeh 2009). The argument that feature engineering is of paramount importance when developing end-to-end computation frameworks is highlighted in the work by Rosienkiewicz, Chlebus, and Detyna (2017), which applies entropy-oriented criteria (i.e. cross entropy) and Beyesian information criterion (BIC), among others to facilitate the identification of the sets that better inform the training processes in ML. Elaborating on this groundwork, Rosienkiewicz (2020) propose the development of FFNN-centric, SVM-centric, and econometrics-based hybrid structures by constructing meta-models that ingest forecasts from nine classical time series and conventional methods (i.e. ARIMA, smoothing-based, SBA, etc.). By applying the Bayesian information criterion, the authors argue that the ANN-based hybrid emerges as the top performer, ranking first in six cases. Beyond such schemes, Xing and Shi (2019) proposed an integrated approach that combines SVM with a back-propagation–trained FFNN to identify the optimal time interval and enhance forecasting performance.

CNN-centric hybrids have also been emphasised for delivering robust forecasts in intermittent demand. In the case of heavy equipment maintenance, Cui et al. (2024) proposed an eight-level architecture, complementing CNN with information selection enhancement-based attention mechanisms and a Bi-directional LSTM (BiLSTM) layer to enhance the predictive performance of established baselines. Finally, the chaotic nature of spare parts demand time series in the steel industry, driven by the interaction of multiple external factors, has also been discussed through the lens of planning inventories (Sareminia 2023; Wakle et al. 2024). Such complexities are addressed by developing hybrid models built upon time series (i.e.ARIMA) and seasonality detection methods (i.e. STL), which enhance the applicability of NN- and SVM-centric approaches.

Transfer-learning mechanisms have recently been discussed as prominent alternatives to typical, supervised approaches, since they present the advantage of lightening the overall training process, which is crucial for the industrialisation of ML techniques. Recent efforts focus on heavy and complex manufacturing equipment, by proposing an architectural Gated Recurrent Units (GRU) approach that leverages the potential of feature adaptation in different products by learning a weight matrix (Fan et al. 2023). Complementary to transfer learning, a multivariate multi-output LSTM has also been introduced–stacking an LSTM over a dense layer to jointly forecast multiple spare parts–to capture cross-item temporal dependencies while reducing item-level tuning effort (Gabellini et al. 2022).

4.1.5. Retailing industry

Retailing constitutes yet another industry where forecasting SKUs, particularly on a daily basis, is an especially challenging task. Arguably, when aggregating at the daily level, the demand profile becomes irregular, exhibiting intermittent and erratic characteristics. In order to clarify the landscape of the effectiveness of ML in the retail industry, the study by Spiliotis et al. (2022) conducts a comparative evaluation of 18 established methods, including some ML-based approaches. The evaluations are conducted on the M5[1] dataset and data from a Greek retailer to verify and generalise the findings. By conducting a combined assessment that simultaneously considers bias and accuracy, the authors conclude that tree-based approaches, especially RFs and Gradient Boosting Trees (GBTs), are the most robust methods, while NNs derive the greatest benefit from feature-enhanced modelling. Grounded in a comparative assessment of the M5 dataset, the study by Kiefer et al. (2021) underscores that although RFs hold promise in terms of accuracy, they fall short against Croston's method in terms of inventory-specific metrics.

The M5 dataset has been the subject of intense research activity, serving both as a primary validation tool in academic studies and as the basis of the M5 competition, whose results are discussed later in this paper. Applications using it as a testbed naturally fall under the retail industry, as the dataset is based on Walmart's sales data. Although extensive studies and combined analyses on this particular dataset have highlighted tree-based methods as prominent, considerable research interest is also drawn to the development of instance-based and RNN-based applications. Among the class of NN-based applications, the use of RNNs, specifically LSTMs, has been extensively studied. For instance, Oliveira and Ramos (2023) examine cross-learning scenarios driven by the product hierarchy employed in retail planning, enabling global models to capture inter-dependencies across products and regions. Complementing this stream, RNN-based studies in M5 investigate sequence models with lag-based inputs for intermittent SKUs and report competitive performance against classical baselines in both simulated and retail settings (Muhaimin, Prastyo, and Lu 2021). Scalability-oriented issues have also been explored, especially through the lens of transfer-learning protocols, built on deep NNs trained at a dataset-level and fine-tuned on a targeted SKUs, with the results showcasing a

promising avenue for operationalisation (Kiefer, Grimm, and van Dinther 2022).

The potential use of the k-NN algorithm for forecasting intermittent characteristics has been explored in the literature, as seen in the work of Hasan, Ahmed, and Ali (2024). This research is positioned as an extension of earlier efforts to utilise the algorithm (K. I. Nikolopoulos, Babai, and Bozos 2016), enhancing it by incorporating both zero and non-zero values into demand vectors and employing a grid-search optimisation to dynamically determine the optimal demand vector length and the number of nearest neighbours for each dataset.

The idea of developing a unified probabilistic framework, which extends traditional methods in the field, has been proposed by Türkmen et al. (2021). Since the evaluation of this method relied on well-known datasets, but dedicated to different industries, including car parts, Godahewa et al. (2020) retailing products using the M5 dataset, and aviation-related applications throughout the RAF dataset, we intentionally decided to discuss its contributions in a case-agnostic manner. A key advancement is the generalisation of Croston-type models by incorporating renewal processes to better handle demand arrival patterns, clustering of new customers and products, and aging effects (Ducharme et al. 2024). Frameworks, in this vein, could also integrate RNNs, including LSTM networks, to replace the Exponentially Weighted Moving Average (EWMA) with learned representations that model complex temporal dependencies and incorporate auxiliary features. In a similar vein, L. Li et al. (2023) propose a feature-based combination framework (feature-based/diversity-based Intermittent demand forecasting) that is likewise evaluated on RAF and M5, and which improves both point and probabilistic forecasts by mapping a diverse pool of time series-related and conventional methods into combined predictions through the application of an XGBoost-based meta-learner, trained to learn the relationship between features and combination weights. The feature-based module is trained on intermittency-specific features, illustrating the importance of feature engineering in such schemes, while its counterpart relies on diversity-based features from pairwise forecast differences. The use of LightGBM and LSTM modules, may also be prominent in the case of developing densities for describing demand. In this vein, S. Wang, Kang, and Petropoulos (2024) propose an ensemble where LightGBM produces a dense grid of quantile forecasts (via pinball loss), while LSTMs parameterise a Negative-Binomial predictive distribution; the resulting outputs are converted to discrete probability mass functions and linearly pooled to obtain the final density.

Beyond applications related to the M5 dataset, the ML literature also includes tailor-made applications. Specifically, in the case of demand forecasting in apparel stores, the accuracy of k-NN and RF algorithms has been compared, with particular emphasis placed on tuning procedures such as optimiaing the number of trees, the number of predictors per tree, and the optimally fitted k-value for each respective method (Güven, Uygun, and Şimşir 2021). Similarly, the authors in J. Wang et al. (2024) have studied the potential of spatial and cross-sectional data to improve the performance of gradient-boosted DTs. This study showcases that incorporating both time series-related and macroeconomic features, including risk-free indicators, can lead to significant advancements in the typical form of gradient-boosted algorithms compared to established ML-based forecasts.

Other promising methods build on the idea that certain similarities can be captured between different fashion-related products, which, when identified, could facilitate the development of accurate forecasts for new codes. In this case, the application of Siamese NNs (SNNs) for similarity learning purposes has been investigated under the lens of combining descriptive attributes and visual images of fashion products to forecast sales profiles (Craparotta, Thomassey, and Biolatti 2019). The model is benchmarked against traditional tree-based and baseline forecasting methods within the domain of new product sales forecasting. This method introduces a novel integration of unstructured (images) and structured data to automate the challenging process of forecasting sales of new products.

Increasing research interest has also been placed on deep learning applications tailored to replenishment strategies at store-SKU levels (S. Ma 2024). Investigating the online retail settings, the authors in Ahmadov and Helo (2023) propose an end-to-end deep learning approach that builds upon the synthesis of different LSTM-based layers. The implications of this research include the complementarity of compound distribution models and calibration-related discussions, which collectively led to significant accuracy enhancements compared to both conventional and time series-related methods. Other approaches have mainly focussed on modelling the impact of features within the overall training process and validating ML-based methods. For instance, a study has analysed the intermittent characteristics of demand when analysed at the customer level, using the Black Friday case study (Zohdi et al. 2022). In this study, the interest centres on tuning and comparing ML methods from both the NN class and more shallow ML approaches, specifically tree-based and instance-based,

with the results indicating that ELM and MLP models hold promise in the field. The indications that MLP schemes may hold promise in the field of retailing SKU forecasting are further validated by the work of Benhamida et al. (2020), which develops an accurate forecasting tool for the products of an online platform, namely CombTSB, by elaborating on a hybrid selection framework that automatically processes product time series; and selects whether conventional methods or hybrid models built from ARIMA, Theta, and an MLP; are more preferable to produce weekly/monthly forecasts. The results indicate that hybrid schemes are more preferable in long-term forecasting horizons. Aligning with the argument that forecasting volatility can be better captured through engineering of contextual features, Yasir et al. (2024) proposed a two-phased application tailored to a textile apparel firm, which first uses generalised least squares and single-layer perceptron models to estimate coefficients and examine the impact of macroeconomic features.

Except for FFNNs, the application of RNNs has been found to hold promise in the demand forecasting of major e-commerce platforms. In this vein, the study by Praveena and Prasanna Devi (2022) presents an integrated approach, based on a combination of LSTM and time series, as well as parametric methods (i.e. ARIMA), to assist in a new range of management decisions as a vital element of intelligent SC. Extending beyond previous approaches in the field of retail demand forecasting, the study by S. Ma and Fildes (2021) introduces a Double Channel CNN (DCCNN), a meta-learner designed to automatically extract features from sales time series and external influential factors such as promotions, seasonality, and pricing. By conducting experiments on the IRI dataset (as presented by Bronnenberg, Kruger, and Mela 2008), this approach generates a feature representation that links the characteristics of each sales series to the performance of a pool of base forecasters, thereby optimiaing their weighted combination for improved forecasting. LSTMs have also been applied in the case of e-commerce, and recent results show promise against Croston-type baselines, especially in cases where high product variety induces intermittency, thus reinforcing their practical value in disaggregated demand scenarios for both point and quantile forecasts (Khan and Al Hanbali 2025).

4.1.6. Miscellaneous

The miscellaneous class comprises ML-centric applications targeting SKUs with distinctive product characteristics that commonly pertain to the previous classes. In particular, this class comprises research efforts tailored to food production and medical equipment, each represented by fewer than five studies.

Online and on-site food services are commonly faced with intermittent demand. In the context of meal delivery platforms, Hess, Spinler, and Winkenbach (2021) investigate the ability to forecast demand both in the short and long term. In this study, historical order data are aggregated at an hourly level and then classified into horizontal and vertical slices, based on daily and weekly demand trends. By employing both tree-based approaches and SVM regressors, the authors conclude that ML-centric approaches hold promise, even in cases where historical data are limited. In the case of catering operations, Sun et al. (2021) focus on developing a hybrid deep learning architecture aimed at improving the detection of customer preferences. From a technical perspective, the research combines the ability of RNN-based networks, specifically LSTMs, to capture temporal dependencies, with FFNNs implemented as an MLP, thereby creating a combined model with weighted allocations for daily sales. The calibrated model, underpinned by Gaussian mechanisms, was found to be prominent against both time series and simpler ML baselines.

The challenging logistics of groceries have been highlighted in both online and on-site operations. The authors in Ulrich et al. (2021) highlight that different operating settings apply to each scheme, primarily due to the transparency required by e-grocery operators, which in turn limits the flexibility of maintaining high in-stock availability. They propose the Generalised Additive Models for Location, Scale, and Shape (GAMLSS) to model demand distributions, placing it within the distributional regression framework. In contrast, the work of Andrade and Cunha (2023) is tailored to physical grocery operations. Specifically, it centres on tree-based approaches, utilising an enhanced form of the XGBoost algorithm that incorporates structural correction mechanisms to address data discrepancies in inventory records. This work demonstrates that both marketing and cross-product effects may be appropriately translated into feature sets to better inform the training processes. The relevance of encoding marketing activities into features also holds for food and coffee SCs, as illustrated in the work of X. Ma et al. (2024), which proposes an RNN-based implementation to estimate demand by modelling temporal dependencies, while simultaneously capturing cross-SKU promotional interactions through a graph CNN. The CNN-centric layer is developed using the k-NN algorithm by integrating inventory-specific features.

In addition to food services, medical items and consumables have recently faced intermittency. Research insights include the technical challenges of encoding exogenous features due to the interplays between several

actors in the SC. Against this backdrop, the use of dynamic NNs (DNNs) arises as a prominent venue (P. Liu 2020). From an architectural perspective, these networks share similarities with FFNNs, with the main difference lying in their dynamic nature of tuning, due to the implied iterative adjustments in the number of neurons based on predefined thresholds of the Minimum Description Length (MDL) criterion. DNNs seem capable of not only delivering accurate forecasts, but also achieving a good trade-off between bias and variance, an aspect preserving scalability and robust generalisation, especially in cases of limited features' incorporation (Cawley and Talbot 2010). The potential of FFNNs has also been demonstrated in the context of cosmetics and medicine inventories. The authors in X. Hu, Zhang, and Zhang (2021) decompose time series between demand events and then forecast the magnitudes for predicted events by coupling tree-based classifiers with FFNNs. In the case of medical consumables, RNNs have shown promise. M. Hasni, Babai, and Rostami-Tabar (2024) harness the potential of RNN-centric hybrids to minimise both replenishment lead times and holding costs for medical inventories by incorporating LSTM- and SVM-based layers, which operate sequentially and in a manually-tuned fashion under the proposed architecture.

Table 1 consolidates the research avenues that have attracted the most attention, mapping industries facing intermittent demand and providing a granular analysis of the proposed ML techniques (point and probabilistic) across product categories. To preserve clarity and support the representativeness of these results, we exclude applications classified under the miscellaneous class, as the contributions in the respective fields were found to be significantly fewer compared to the other industrial classes. To aid the interpretation of the industry-wise distribution of the 99 studies, we also report the presence of cross-industry assignments based on the experiments conducted.

Based on a product-agnostic approach, we can derive certain conclusions and insights about the most prominent ML-based techniques in the field of intermittent demand forecasting. Arguably, most of the samples we managed to locate relate to the development of forecasting methods for spare parts. However, the applications are not limited to this product category. Significant emphasis has also been identified in applications related to retailing SKUs, where the demand profiles appear to be intermittent at the store/daily level. In both product categories, strong emphasis has been placed on NN-centric, tree-based, and hybrid models. Regarding the use of NNs, a common pattern emerges, indicating that most efforts are oriented towards the development of deep architectures featuring multiple hidden layers, predominantly employing approaches from the broader categories of FFNNs and RNNs.

In the subclass of FFNNs, the most prominent implementation approaches appear to be Multi-Layer Perceptrons. The training of these networks is predominantly performed using the back-propagation algorithm, which is widely regarded as the standard method for weight optimisation in NNs. Moreover, specialised variations such as the Levenberg-Marquardt algorithm are often employed in cases where high accuracy is prioritised, particularly in problems involving small to medium-sized datasets. A common theme underlying research practices in the field is the need to achieve trade-offs between accuracy and training time, with some endeavours indicating that applying ELMs may be a promising approach. In contrast, within the category of RNNs, the extensive use of LSTM networks has been noted. LSTMs are favoured due to their unique capability to capture long-term dependencies and dynamic changes in data, making them highly effective for forecasting applications. Building on sequential modelling of temporal dependencies, RNNs could serve as a promising alternative for implementing cross-learning protocols, which preserve computational efficiency. Their ability has been validated beyond point or quantile forecasts, extending to the development of full-distributions for multiple time series on a dataset level.

Beyond these applications, considerable effort has also been directed towards tuning less complex approaches, with implementations focussing on various types of DTs, evaluating different lightweight and gradient-boosted methods. In this venue, LightGBM and XGBoost have been found to be promising, due to their efficiency and flexibility to be incorporated into composite, ensemble-like structures, acting as base-learners. The findings from the comparative analyses indicate a common practice of assessing ML methods against time-series and conventional approaches. In turn, this supports the prevailing notion that ML-centric methods, in most cases, overcome the challenges faced by traditional forecasts. In our view, however, the broad adoption of ML methods in such comparative analyses not only strengthens their methodological positioning but also contributes to the development of research 'value chains'.

4.2. Point versus probabilistic forecasts: proofs from the ML literature

Notwithstanding the well-contextualised merits of probabilistic forecasts in informing decision-making processes, the development of such schemes is significantly limited in the broader literature (Barrow and Kourentzes 2016;

Table 1. A consolidated overview of ML-centric intermittent demand forecasting.

Industry	Technical class	Technical subclass	Implementation schemes	Comparison
Automotive ($n = 24$)	Shallow ML ($\approx$ 15%)	Instance-based	Centered around weighted distance vectors to improve forecasting accuracy of the k-NN algorithm.	Conventional
		Tree-based	Focused on boosting approaches, by proposing and calibrating algorithms including CatBoost, LightGBM for better occurrence and quantity prediction; and tensor-based GBDT with decay correction for better adjustments.	Conventional, Time series, ML
	NN-based ($\approx$ 45%)	FFNN	Efforts on: (i) better-configuring both shallow and deep architectures, incorporating different learning algorithms (BP, LM) and non-linear activation functions; (ii) developing a Bayesian-regularised structure for initialisation.	Conventional, Time series, ML
		RNN	Elaborating upon the capabilities of LSTM for capturing temporal dependencies in data, by implementing mainly deep architectures and stacked approaches.	Conventional
		CNN	Centers around 3D CNN for advanced data transformation and feature extraction.	Time series, ML
		TNN	MLOps pipelines, centred on the integration of temporal fusion transformers for single and multi-time series forecasting.	ML
	Synthetic ($\approx$ 40%)	Hybrid	Pathways: (i) combinations of fuzzy ANN forms with GA to fine-tune activation function and hyperparameters; (ii) integration schemes of time series-related methods (i.e. ARIMA) with fuzzy ANN forms; (iii) integration of SVM with meta-heuristics (PSO) to fine-tune hyperparameters; (iv) integral schemes composing clustering and tree-based mechanisms.	Time series, ML
		Ensemble	Merges ELM and SVM to account for weather impacts on spare parts demand. Aggregates multiple methods, including SBA, HW, ARIMA, and XGBoost, to enhance accuracy.	Conventional, Time series, ML
		TransferL	Utilises transfer learning with LightGBM for incorporating exogenous factors.	Conventional, ML
Aviation & Naval ($n = 15$)	Shallow ML ($\approx$ 10%)	Tree-based	Proposals for better tuning of both DTs and RFs.	Conventional, Time series, ML
	NN-based ($\approx$ 40%)	FFNN	Development of topology-optimiaed architectures, by incorporating additional lagged inputs and Bayesian regularisation.	Conventional, ML
		TNN	Utilises encoder-decoder with self-attention for long-term dependency modelling.	Conventional, ML
	Synthetic ($\approx$ 50%)	Hybrid	Pathways: (i) integration of conventional methods (i.e. Croston's method) and time series-related with RNNs and tree-based regressors; (ii) development of Genetic NN with weighted optimisation mechanisms; (iii) combination of swarm intelligence and other meta-heuristics with different NN forms (RNNs; Prob. NNs), with inputs in both simple and fuzzy forms.	Conventional, Time series, ML
		TransferL	Utilises transfer learning with LightGBM for incorporating exogenous factors.	Conventional, ML
Industry	Technical class	Technical subclass	Implementation schemes	Comparison
Electronics & Electrics ($n = 15$)	NN-based ($\approx$ 55%)	FFNN	Pathways: (i) Bayesian-optimiaed FFNN topologies; (ii) tuning of traditional MLPs.	Conventional, ML
		RNN	Pathways: (i) UNISON framework integrates RNNs with temporal aggregation and regret-weighted calibration; (ii) application of Reservoir Computing (RC) framework.	Conventional, Time series, ML
	Synthetic ($\approx$ 45%)	Hybrid	Combinations of: (i) time series related (i.e. ARMA(X)) with FFNN; (ii) conventional methods (i.e, SBA) with RNNs.	Time series, ML
		Ensemble	AutoGluon leverages diverse base learners (NN, tree-based, instance-based) with profile-driven evaluation.	Conventional, ML
Heavy industries ($n = 17$)	Shallow ML ($\approx$ 10%)	SVM	Utilises SVM with polynomial kernels and logistic regression for petrochemical and oil industry spare parts.	Conventional, Time series, ML
	NN-based ($\approx$ 30%)	FFNN	Includes single-layer and 2-hidden-layer FFNN architectures optimiaed for construction machinery and feature-specific tasks.	Conventional, ML
		RNN	Pathways: (i) Development of an Elman-type RNNs; (ii) evaluation of LSTM structures in terms of accuracy.	Time series, ML
		CNN	Combines CNN with BiLSTM for advanced maintenance forecasting, incorporating attention mechanisms.	ML
	Synthetic ($\approx$ 60%)	Hybrid	Pathways: (i) combinations of FFNN with time series methods (mainly ARIMA); (ii) tuning through the incorporation of recursive methods and other decomposition techniques (i.e. STL); (iii) combinations of SVM and logistic regression.	Conventional, Time series, ML

(*continued*).

Table 1. Continued.

Industry	Technical class	Technical subclass	Implementation schemes	Comparison
		Ensemble	Stacked auto-encoder combined with LightGBM for robust demand forecasting in manufacturing.	Conventional, ML
		TransferL	Implements transfer learning with GRU for heavy equipment demand forecasting; multi-output transfer learning protocol using LSTMs.	Conventional, ML
Retail ($n = 22$)	Shallow ML ($\approx 25\%$)	Instance-based	Pathways: (i) Oriented towards extending the applicability of k-NN algorithm, by proposing a holistic, zero-inclusive approach; (ii) coupling k-NN into graph recurrent layers for producing quantile forecasts.	Conventional, ML
		Tree-based	Approaches tailored to tuning of RF and gradient-boosted algorithms.	Conventional, Time series, ML
	NN-based ($\approx 40\%$)	FFNN	Mainly oriented towards ELM and MLP models, examining several training algorithms (mainly back propagation-trained).	ML
		RNN	Implements cross-learning LSTM-based models to capture inter-dependencies in retail planning.	Time series, ML
		Siam. NN	Exploration of SNNs for similarity learning in new product sales forecasting.	Time series, ML
	Synthetic ($\approx 35\%$)	Hybrid	Pathways: (i) integrative efforts of unified probabilistic framework development, by synthesising conventional models with RNNs; (ii) combination of time series-related methods (i.e. ARIMA) with FFNNs (i.e. MLPs) and RNNs (i.e. LSTMs).	Time series, ML
		Ensemble	Pathways: (i) tuning mechanisms for gradient-boosted algorithms and RFs; (ii) pooling mechanisms for probabil. forecasts.	Conventional, ML
		MetaL	Introduction of DC-CNN method to extract time series-related features and optimiae base forecaster combinations.	ML

Fildes, Ma, and Kolassa 2022; L. Li et al. 2023; Spiliotis et al. 2021). At hand, this is even more evident in the case of ML-centric forecasts. By elaborating upon the above results, it could be argued that the majority of the sample is oriented towards developing point forecasts. Specifically, our analysis indicates that fewer than 15% of the studies develop or incorporate a probabilistic/quantile component when developing forecasting modules. An interpretation of the above finding may stem from the fact that ML modules do not deliver densities automatically, as may be the case in traditional parametric methods. While under-explored, ML-centric applications still hold promise for implementation. To enable the construction of the predictive distribution, ML modules have been trained in a global, dataset-wide fashion. In this vein, the development of RNN modules for fitting a single probabilistic model jointly to all related series, specifically by conditioning on the full historical windows of every series in the dataset, is a very prominent protocol for enabling the model to learn cross-series regularities and covariate–target dependencies that would otherwise require manual specification (Salinas et al. 2020).

Another key takeaway stemming from the core of our analysis is related to the ML topologies. In this vein, we have identified that certain established ML methods, commonly based on RNNs (i.e, LSTMs), trees (i.e.LightGMBs), and combinations of them S. Wang, Kang, and Petropoulos (2024) are of significance for both of the forecasting streams (i.e. point and probabilistic) (Makridakis et al. 2022); which is also supported by the fact that in some cases the same ML modules have been applied to produce both quantile and point forecasts, with minimal adjustments. That said, we could argue that implementation-wise, the current state of practice is streamlined in both point and probabilistic settings. After examining the limited approaches identified, we observed that the scope in most probabilistic forecasts is oriented towards estimating quantiles rather than predicting the total distribution. On the industrial side, most of the efforts have been dedicated to store retailing SKUs, by commonly using as a testbed the M5 dataset. In comparative assessments the landscape differs from the typical procedures followed in point forecasts. For instance, conventional methods are rarely included. This is because most of them need specific modifications in their underlying functions towards operating adequately in probabilistic settings. Lastly, prior knowledge suggests that in different quantiles, the best forecasts may be achieved by different models (Spiliotis et al. 2021), indicating that the selected baselines may vary across different quantiles.

5. Implementation details: on the identification of research value chains

When it comes to the development of ML-based applications, focal points of research revolve around fundamental questions such as (i) how much data is required for these methodologies to operate effectively; (ii) by which means one can validate the improvements in effectiveness they provide; and (iii) how can these methods be further calibrated and tuned to enhance their accuracy. These dimensions represent a common theme across most reviewed studies, with the authors presenting extensive arguments regarding the necessity of pivoting

from time series and conventional techniques to more dynamic models capable of recognising patterns and correlations. In this section, we discuss some proofs gathered by emphasising key dimensions, including the evaluation metrics used, the volume and time basis of the databases analysed, and the types of explanatory features that serve to enhance the predictive power of ML models.

5.1. Accuracy measures

Typically, evaluation protocols differ according to the objective of the forecasting task. Given that the core of our sample is oriented towards point forecasts, in this subsection we first detail the ongoing debate regarding this task and then we discuss the practices found in the case of evaluating probabilistic forecasts. Within the point forecasting literature, two primary evaluation pathways can be identified: one focussing on accuracy improvements and the other on business impact enhancements, particularly in inventory management (Pinçe, Turrini, and Meissner 2021). Both paths share a common denominator in reflecting the overall flexibility of inventory management systems. Improvements in forecast accuracy have been linked to potential benefits in cost reduction and customer service optimisation (Ghobbar and Friend 2003), although in some cases may lead to neutral or even adverse cost impact. To this end, performance may be measured in terms of both statistics-specific and inventory-related metrics (Pinçe, Turrini, and Meissner 2021). Notwithstanding the importance of inventory-specific metrics for performance evaluation, their use remains limited in ML-centric forecasting research. In our sample, only about 10% of studies report at least one inventory-specific measure, and fewer than 5% rely on such measures exclusively. Building on this observation, the core of the discussions presented in this subsection was shaped by the practices followed for selecting statistics-oriented metrics. Inventory-specific metrics were analysed separately from statistics-based metrics, with the corresponding results suggesting that the impact of forecasts on inventory performance is commonly reflected in cost-based (e.g. holding costs, scaled periods-in-stock) and service-based measurements.

For more than forty years, research efforts have been dedicated to mathematically defining KPIs for forecasting accuracy, capturing various aspects and peculiarities of time series. The seminal work of Hyndman and Koehler (2006) made a significant contribution to this field in the early years by presenting a comprehensive overview of metric development activities, categorising existing evaluation metrics into four main classes. This work also outlined the utility and limitations of each class, advising that in problems involving intermittent demand, the use of scaled metrics is generally preferred. Scaled metrics, by benchmarking against naïve (random walk) methods, ensure unit-agnostic comparisons, thereby enhancing the generalisability of methods across different datasets. Aligned with this perspective, the research by Syntetos and Boylan (2005) proposed the scaled metric *sME* (scaled Mean Error), which was later employed in various studies as a criterion for quantifying bias in different intermittent-oriented methods (Petropoulos, Kourentzes, and Nikolopoulos 2016).

To facilitate a thorough mapping of the statistics-oriented metrics used to evaluate ML-based applications, we have drawn the following classification: (i) scale-dependent metrics, (ii) percentage-based, (iii) relative, (iv) scale-free, and (v) miscellaneous. Each of the above classes of metrics has been drawn based on similarities in their respective mathematical foundations and could be further refined by incorporating their mean form, which has been proposed as a solid practice in intermittency contexts (Prestwich et al. 2014). For instance, further analysis could focus on matching the subclasses of central tendency, symmetric, and geometric forms of measures to each of the above classes. Since the scope of our research is oriented towards extracting the common validation patterns of ML-centric techniques, we mainly focus on the first-level classification of metrics. Having delved into the technical specifications of the methods incorporated in this review, we note that approximately one out of two studies (53.5%) have assessed the performance of their method by using at least two different classes of statistics-oriented metrics. The most frequent combination of metric categories applied together was identified as scale-dependent with scale-free and scale-dependent with percentage-based. In 61% of the studies, at least one scale-dependent metric was used for performance evaluation purposes. To better document these findings, we refer to Figure 6, in which we present in a heat-map form, the most frequent combinations of metrics.

The findings validate that combining different metrics reflects a common practice in the field of ML-centric forecasts. Yet, it seems inconsistent with what established works in the field have proposed regarding the evaluation protocols. For instance, the work of Petropoulos et al. (2022) emphasised that different metrics should not be applied for evaluating the performance of a single forecast, a practice aligned with the guidelines given in the recent M5 accuracy competition, in which only one, scale-free metric (i.e. RMSSE-Root Mean Squared Scaled Error) was applied for evaluation purposes (Makridakis, Spiliotis, and Assimakopoulos 2022). Arguably, this presents a significant inconsistency in the existing body of research, which calls for consensus.

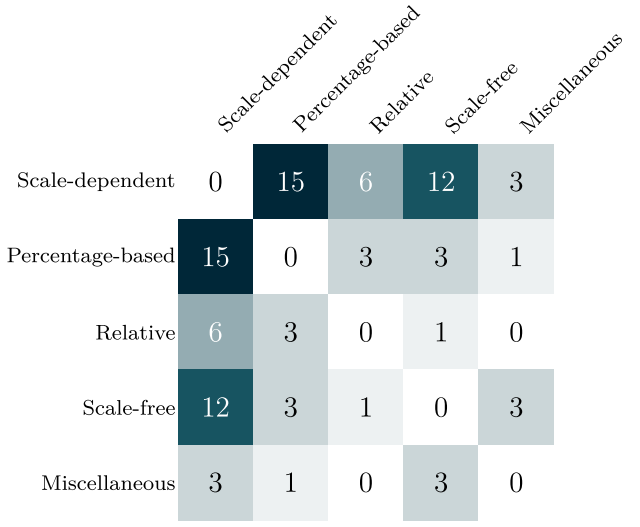


Figure 6. Frequent combinations of metrics used for evaluation purposes.

A partial interpretation of the above finding may come from the scope of the evaluation in each study, an argument inherently tied to the proof that different error metrics may reward different methods (Kolassa 2020). For instance, by utilising absolute-type metrics such as scale-dependent ones, which express deviations based on specific units, authors may aim to quantify forecasting performance and to report, rather than induce, any accuracy improvements achieved by a proposed method. In cases aiming to create a more sensitive basis for model comparison, the specific subclass of squared-error measures (e.g. MSE, RMSE) is often included, since they preserve a downgrade in small performance differences while amplifying larger deviations. Currently, the authors emphasise the generalisability of the training process by exploiting the unit-agnostic characteristics described by scale-free metrics. Additionally, the use of percentage-based and relative metrics, although influenced by the presence of zero values in the evaluation sample, can complement the overall accuracy optimisation when combined with scale-dependent metrics. This approach is particularly effective in cases where appropriately modified versions, such as alternative-MAPE (A-MAPE) or MAAPE (Mean Arctangent Absolute Percentage Error), are applied (Kim and Kim 2016).

By synthesising the above, we could inherently infer that in cases primarily focussing on accuracy advancement for datasets consisting of time series with streamlined units, the authors rely on combinations of scale-dependent and relative or percentage-based approaches. However, if generalisability is also a significant criterion, then combinations including scale-free methods were preferred. To better contextualise the inconsistencies between principled evaluation protocols and the practices identified in the sample, we should note that ML literature has progressed in parallel with mathematically oriented studies that reflect the best evaluation protocols. This means that recent mathematical foundations may not have been taken into account in previous research efforts. Figure 7 presents the most identified metrics in each of the above-mentioned classes developed in this research. In the miscellaneous category, we included inventory-related specifics and certain statistical measures, such as classification-oriented measures that fall outside traditional regression-based forecasting metrics. As previously commented, inventory-specific metrics were found to be limited in the collected sample (just over 10%). That said, we have classified these metrics into the Miscellaneous class, a practice that also aligns with our intention on clearly demonstrating the most frequent inventory measurements.

Among a broad range of metrics discussed in the context of intermittent demand forecasting, significant research interest has been focussed on four key metrics: the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE). Thus, another conclusion can be drawn regarding the practices followed by the authors: if the evaluation session focuses on both scale-free and scale-dependent measures, metrics such as RMSE, MAE, and MASE should be analysed among others. However, if the focus is on assessing percentage-based measures, excluding scale-dependent ones, then the current practice recommends using the MAPE metric. Another finding is that recent works have also shifted interest towards evaluating the Root Mean Squared Scaled Error (RMSSE) metric. This metric has gained traction, as it advances the typical, scale-free metric MASE by allowing for the strict calculation of the forecasting error in cases where the MASE metric, due to its nature, favours median over mean or expectation forecasts (Kolassa 2016, 2020).

An additional instance of analysis could arise from relating the technical classes to the classes of metrics. Such an approach could illuminate how the dynamic characteristics of each method (e.g. stochastic nature, training approach, sensitivity to initial conditions, etc.) interact with the choice of evaluation metrics. To assess whether such patterns can be identified, we present Table 2, which relates the classes embedded in our technical taxonomy with the classes of metrics. For interpretation purposes, we note that when encoding data for the case of hybrids, we have delved into the implementation specifics to identify the core ML pillar and thus relate it to some of the standalone technical classes. In this regard, the following Table 2, except for the classes created before, includes some extra classes in the case

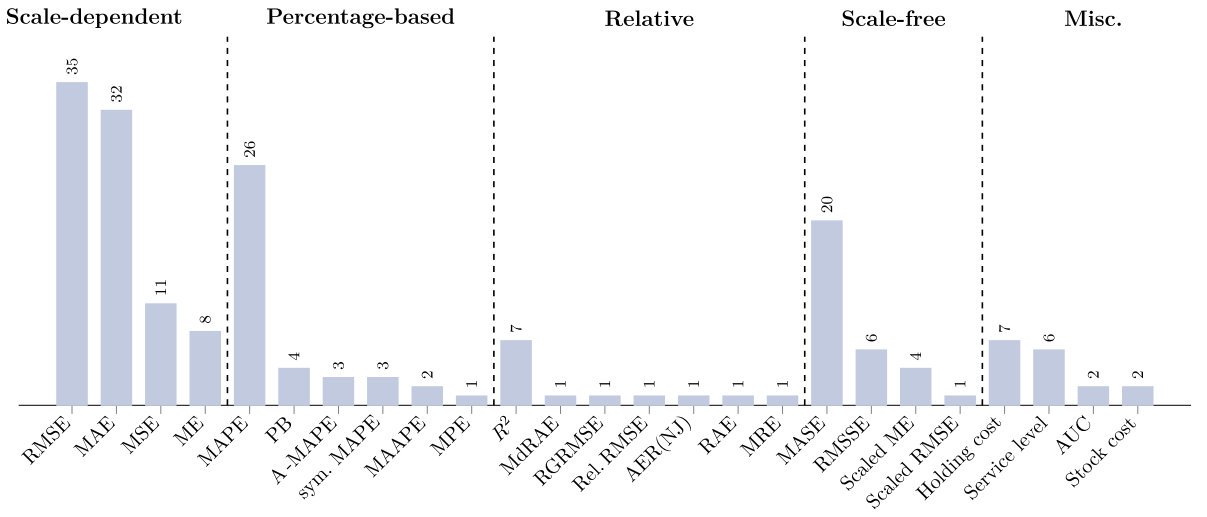


Figure 7. Most frequent metrics of each class (descending within subclasses).

Table 2. Common combination of metrics used for evaluation purposes.

Method class	One set	Pair of metrics	Triplet of metrics	Details on pairs	Details on triplets
Instance-based	57.1%	28.6%	14.3%	Scale-dependent + Relative (50%) | Scale-dependent + Miscellaneous (50%)	Scale-free + Percentage-based + Miscellaneous (100%)
Tree-based	42.1%	47.4%	10.5%	Scale-free + Miscellaneous (22.2%) | Scale-dependent + Scale-free (33.3%)	Scale-dependent + Percentage-based + Miscellaneous (50%) | Scale-dependent + Percentage-based + Scale-free (50%)
SVM-based	30.0%	50.0%	20.0%	Percentage-based + Relative (20%) | Percentage-based + Scale-free (40%) | Scale-dependent + Percentage-based (20%) | Scale-dependent + Scale-free (20%)	Scale-dependent + Percentage-based + Relative (50%) | Percentage-based + Scale-free + Miscellaneous (50%)
FFNN | FFNN-based	39.4%	51.5%	9.1%	Scale-dependent + Relative (24.4%) | Scale-dependent + Percentage-based (24.4%)	Scale-dependent + Percentage-based + Relative (66.7%)
RNN | RNN-based	22.2%	61.1%	16.7%	Scale-dependent + Scale-free (33.3%) | Scale-dependent + Percentage-based (11.1%)	Scale-dependent + Scale-free + Percentage-based (50%) | Scale-dependent + Scale-free + Miscellaneous (50%)
CNN | CNN-based	33.3%	66.7%	0.0%	Scale-dependent + Percentage-based (50%) | Scale-dependent + Relative (50%)	–
Other NNs	60.0%	20.0%	20.0%	Scale-dependent + Percentage-based (100%)	Scale-dependent + Percentage-based + Scale-free (100%)

of NN-based applications. This approach has also been followed for all the shallow ML baselines. This focus is substantiated by evidence from our research, which indicates that hybrid methods are primarily developed to enhance specific categories of methods without altering their core implementation characteristics. This finding suggests that the technical requirements during evaluation are similar between simple and composite algorithms when they share fundamental structures. Hence, we utilise this approach to facilitate a thorough analysis of whether underlying trends and patterns exist in the evaluation of these methods.

Elaborating on our previous findings, which have shed light on forecasting methods centred on trees, SVMs, and NNs (especially FFNNs and RNNs), several conclusions can be drawn based on the proposed analysis. First, in these simple and hybrid subclasses, the researchers have mainly assessed their results using metrics from at least two different classes. In most evaluation assessments of shallow ML approaches (i.e. tree- and SVM-based), the authors have cross-checked their results using scale-free and scale-dependent combinations of metrics. This also holds for the class of RNNs, developed in both simple and hybrid forms. In contrast, authors tend to evaluate FFNN-based applications by mainly using scale-dependent and relative metrics combinations.

By researching the interplays of training and tuning particularities of each of the above methods, some meaningful extensions could be drawn. For instance, hybrid RNN approaches, such as those that integrate LSTMs with manipulation techniques oriented toward captivating exogenous features, often rely on scale-free metrics to account for their inherent sensitivity to vanishing and exploding gradients, while often adopting

hybrid training techniques (e.g. gradient clipping or gated mechanisms) to stabilise long-term dependency capture (Hewamalage, Bergmeir, and Bandara 2021). At the same time, tree-based and SVM-based models tend to display lower sensitivity to initial conditions, but require meticulous hyperparameter tuning (e.g.choosing the kernel in an SVM or the depth of a DT). This is reflected in the widespread practice of evaluating such models with both scale-dependent and relative metrics to confirm robust performance across diverse data distributions and scales (Taieb and Hyndman 2014). On the other hand, FFNNs illustrate the criticality of appropriate weight initialisation in mitigating the pitfalls of gradient-based training, a factor that makes scale-dependent metrics like RMSE particularly informative for diagnosis. Hybrids that integrate FFNNs or RNNs with statistical techniques (e.g. ES or ARIMA), demand composite error metrics that blend scale-free and relative criteria to capture the varying dynamics introduced by each sub-component (Smyl 2020).

To conclude, we comment on a controversial practice identified in the collected sample. In many studies, the authors tend to evaluate the performance of forecasting tools on a dataset rather than a time series level, by applying the same set of metrics uniformly. This practice, aids comparability but can mask series-specific behaviour. In such corpora, the possible presence of smooth-type patterns may inherently pull the evaluation towards well-fitted measures for smooth series, rather than criteria dedicated to intermittency. In this regard, it is possible that such approaches could lead to misinformed conclusions regarding the performance achieved for the intermittency subset. Taken together with the broader point that metric assumptions (scale, noise, loss asymmetry) should match data properties and the forecasting target (point vs. probabilistic), this observation may act as another divergence between the followed practices and the principled pathways. A constructive way forward is dataset-aware validation. For instance, by stratifying by series typologies and pre-declaring specific metrics to each subset; towards preserving comparability while reducing evaluation bias induced by data heterogeneity.

The above discussions were primarily oriented towards clarifying the practices pertinent to the point forecasting literature. In the case of probabilistic forecasts, the practice departs from simply evaluating the performance in terms of accuracy, and it rather focuses on assessing density-oriented criteria. Based on Kolassa (2016), we could argue that common evaluation pathways rely on *calibration* and *sharpness* criteria. Calibration serves as an indicator of how much the produced density fits the real demand events. It is typically calculated by applying the Probability Integral Transform (PIT), defined as $\text{PIT}_t = \hat{F}_t(y_t)$, where $\hat{F}_t$ denotes the forecast cumulative distribution function (CDF) and y_t the real demand event. For discrete counts, a common pathway is the calculation of the randomised PIT, (rPIT) by drawing uniformly on $\left[\hat{F}_t(y_t - 1), \hat{F}_t(y_t)\right]$. Under correct specification, PIT/rPIT values are approximately Uniform$(0, 1)$. On the other hand, sharpness captures the concentration of the predictive distribution – conditional on calibration – and is quantified using proper scoring rules (i.e. loss functions minimised in expectation by the true data-generating process). In practice, several scoring rules could be utilised, such as (i) the logarithmic score, reflecting the negative log-likelihood, $-\log \hat{p}(y_t)$, with $\hat{p}$ the forecast Probability Mass Function or density; (ii) the Continuous Ranked Probability Score for continuous outcomes together with its discrete analogue; (iii) the Discrete Ranked Probability Score; and (iv) the Brier score for binary events.

The above mechanisms are primarily relevant to estimating the total distribution used to describe demand. In the case of quantile forecasting, the landscape is slightly different, since the scope is oriented towards estimating specific quantiles of the distribution (i.e. τ-levels such as 0.50, 0.90, etc.) rather than delivering a full predictive density. This case reflects the common implementation practice in the probabilistic subset of the literature. In this vein, established implementation protocols suggest statistical evaluation using scale-free metrics, such as the scaled-Pinball Loss (SPL), for specific quantiles. Most of the studies seem to adhere to this protocol, yet this is not horizontally aligned with specific work that still applies scale-dependent variants (Türkmen et al. 2021). The last comment about quantile forecasts reflects a common ground regarding the incorporation of both inventory and statistics-oriented metrics in the evaluation phase. This could be attributed to the direct encoding of targeted quantiles into inventory-specific metrics, such as targeted service levels. Based on our preliminary results, we have identified that the common intervals incorporated are based on the right-tail of the distribution (i.e. 80% or 0.8 and beyond), a finding that inherently validates the prevailing notion that in SC decision-making, uncertainty is better quantified by estimating the right-tail (typically $>75\%$) of distributions.

5.2. Interplays between database properties and implementation specifics

The relationship between the data used to train ML models and the overall accuracy they achieve has been repeatedly documented in the literature (Petropoulos et al. 2022). In fact, the performance of ML models is

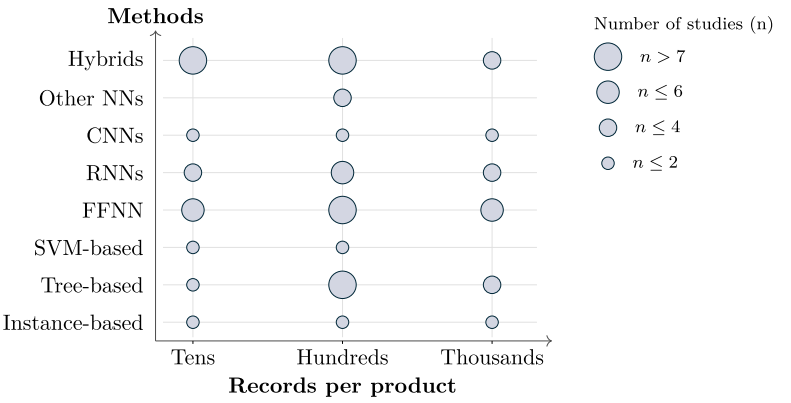
inherently tied to the 'quality' of the data being analysed. The term 'data quality' often refers to the specifications of the datasets on which forecasts are based. Among the dataset specifications that have a significant impact on the accuracy of the models are the number of products (time series), the temporal granularity used to record historical demand data (time basis), and the volume of records per product. These properties, when combined highlight two critical issues in the field of ML models, specifically (i) how much data is necessary to apply a class of ML algorithms effectively; and (ii) how many different products with similar specifications (records and temporal granularity) are required to conduct experiments that substantiate the generalisability of a method. To shed some light on the research practice regarding these critical issues, we synthesise the findings reported in the body of research, excluding the subset of studies that concentrate on the development of meta-learning models, since the corresponding sample size is relatively small for drawing sound conclusions.

The relationship between the number of records and the time basis in a dataset is a recurring topic in the literature on intermittent demand forecasting. Specifically, it is addressed in studies examining the interfaces of temporal aggregation levels with the accuracy of forecasting models (J. Wang et al. 2024). Prior work in the field has demonstrated that the emergence of intermittent demand patterns is influenced by the level of aggregation used for data recording. Strategically, aggregation levels can be developed across three dimensions: product, time, and SC level. However, from a technical perspective, strategic decisions directly impact how and how much data is recorded, creating specific challenges when selecting ML models. Implementation processes commonly include the practice of performing data aggregation in the preprocessing phase to reduce noise and uncertainty during model training. Subsequently, disaggregation techniques or frameworks are applied to shift predictions to the desired temporal granularity as part of post-processing strategies (Andrade and Cunha 2023).

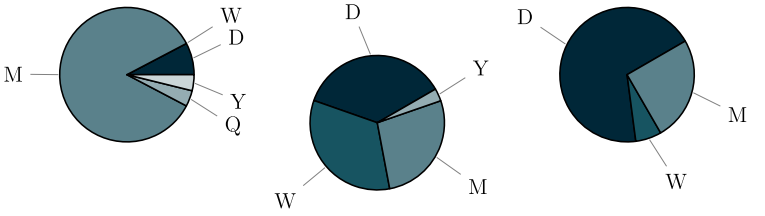
To provide an essential insight on the interplay between data properties (time-bases and total records) and methodologies implemented, we construct Figure 8, which serves a dual purpose. First, it reflects the common methodologies related to the number of data records per time series (part a); and second, it reflects the time bases in relation to the number of records in each time series included in a dataset (part b). To create this figure, we have consolidated the practices mentioned by the authors regarding the total number of records per time series and the time basis of observations. In particular, the recording of counts was conducted at the time series level, rather than at the dataset level. To collect information on how many time series constitute a dataset, we also used the information provided by the authors (e.g. the number of SKUs). For instance, if in a case was reported weekly demand for 925 items facing intermittency, each with observations ranging from 100 to 200; we have accordingly coded these data as weekly, with hundreds of records per series and set the number of series to 925–an element used for the later part of the analysis conducted in this subsection.

From the second part of Figure 8, we can observe that in cases where only limited data are available, intermittency patterns emerge when the data are recorded on a monthly basis. In contrast, when developing richer datasets consisting of hundreds or even thousands of records, the intermittency seems to commonly emerge when daily or weekly time bases are adopted. To complement this, we also note that approximately one out of two datasets reported a monthly time-basis reporting strategy. Combining this observation with the fact that the largest scale involves thousands of records, while there are also cases of model development with only tens of records, it provides indications that the development of ML models indeed requires a significant number of records. However, this volume is finite and categorised into distinct magnitudes, without necessitating an infinite data scale, as might be the case with big data.

Except for the aggregation strategies followed in the collected sample, some meaningful patterns could emerge when analysing the ML methodologies in relation to the number of records per product (time series). In this vein, we could comment that in the case of time series consisting of tens of records, almost all the identified classes have been explored, except for specialised architectural schemes (i.e. transformer NNs). These schemes are commonly developed in multiple layers (deep learning), a factor that leads to the need for quite large data volumes to optimiae their hyperparameter sets towards converging in a generalisable space of forecasts. With only tens of observations per SKU, these preconditions rarely hold; therefore, the practice leans towards shallow MLPs and tree-based learners. Across limited data conditions, hybrid schemes seem to be the primary implementation avenue. Such schemes were commonly identified to build on a statistical component (e.g. Croston/TSB, interval–size decomposition, seasonal filtering) that absorbs intermittency and low-signal structure, while a lightweight ML module learns residuals or ensembles based forecasts–reducing effective parameterisation, stabilising training in small-n settings, and retaining deep-like expressiveness.



(a) Bubble chart of method families in relation to records per series.



(b) Recording time-bases in relation to the total number of records. (Tens/Hundreds/Thousands). *Notation*: (D): Daily, (W): Weekly, (M): Monthly, (Q): Quarterly, (Y): Yearly.

Figure 8. Most-used ML methods in relation to records per series (part a). Time-basis in relation to records in each bucket (part b). (a) Bubble chart of method families in relation to records per series and (b) Recording time-bases in relation to the total number of records. (Tens/Hundreds/Thousands). *Notation*: (D): Daily, (W): Weekly, (M): Monthly, (Q): Quarterly, (Y): Yearly.

For applications based on hundreds of records, a substantial proportion operates at a monthly aggregation level. Within this class of methods, two technical categories can be distinguished: tree-based and NN-based techniques, particularly those employing feed-forward architectural approaches in simple or hybrid configurations. Conversely, for the case of hundreds of records stemming from a weekly recording basis, RNN-centric applications appear to be prominent. This finding could be attributed to a cycle of dependencies at short horizons (e.g. week-of-year, trading-day, and promotion effects) that are preserved under weekly aggregation; gated RNNs (i.e. LSTMs) are expressly designed to capture such sequential autocorrelation and bursty zero-arrival patterns, whereas monthly aggregation collapses these signals into lower-frequency components better handled by feed-forward or tree-based learners (Hewamalage, Bergmeir, and Bandara 2021). Finally, in datasets comprising thousands of records, notable interest is observed in various feed-forward approaches, including MLPs and back-propagation models. In this class of records, there is typically an abundance of data, which offers the flexibility of developing deeper architectural schemes for better capturing dependencies, but on the other hand, brings forward the consideration of trade-offs between model computational cost and training stability. In this vein, an ongoing debate over training schemes emerges: stochastic–gradient back-propagation (e.g. SGD/Adam) scales to larger networks and datasets but demands careful regularisation and scheduling, whereas Levenberg–Marquardt (LM) often converges faster on small–to–mid-sized MLPs at the expense of substantially higher memory use and poorer scalability. Moreover, statistical pre-modelling (e.g. ARMA/ARIMA to capture linear autocorrelation and seasonality) can be leveraged to produce well-informed features or residual targets; by 'pre-whitening' the series, the neural component learns a simpler residual signal, reducing effective parameterisation and, in practice, training time while improving robustness in intermittent settings.

Another concern regarding the use of ML models relates to the number of experiments upon which the authors have attempted to substantiate the generalisability of the developed methods. Since ML-based models are inherently linked to a degree of stochasticity, resulting in certain variations when applied to diverse datasets, some level of fluctuation in outcomes is expected. As a result, the number of series or even datasets to be evaluated for safeguarding the generalisation of a method, is a focal point of research. Focusing on NN-centric applications either simple or hybrid forms, the practices are notably divergent in this vein. Approximately 35% of the studies have evaluated the generalisation capabilities of their models using datasets comprising tens of distinct

products, ranging from ten to eighty. However, a notable proportion of models have been studied in terms of predictive power with datasets involving hundreds (approximately 30%) or even thousands of distinct SKUs (approximately 20%), which are typically maintained in massive datasets related to inventory management in large manufacturing and retail enterprises. A smaller number of applications have been studied using just a few products, usually between three and seven at a minimum, while some implementations validate their results with only one product.

In simpler categories of methods, primarily focussed on tree-based and SVM-based models, the landscape differs slightly. A significant proportion of these methods has been tested on at least 50 different SKUs, with many applications extending their validation to hundreds or thousands of SKUs. These observations highlight that substantial research efforts have been directed towards both tailor-made applications designed to meet the forecasting needs of specific product segments and generalisation studies. In tailor-made approaches, the primary focus is on improving point forecasts for a particular product or group of SKUs (testing on units or tens of products) without necessarily generalising to all SKUs in a dataset. Notably, applications based on some form of NN appear to be preferable for developing tailor-made methods, while simpler approaches are more frequently utilised for methods aimed at delivering generally improved performance across many SKUs or even entire large-scale datasets. This argument holds for specialised NN classes, such as RNN models, in both simple and hybrid forms, where their development aligns with the goals of targeted and generalised methodologies. Ultimately, regarding the provision of insights on the minimum number of products required to evaluate ML methods, our proofs indicate that approximately 9% of studies conduct experiments on six or fewer products. This finding may be considered as a lower threshold for experimentation across different product codes in a series-by-series manner, yet this protocol needs to be further validated in the literature.

5.3. Fine-tuning techniques

In general, fine-tuning techniques can be classified into two categories: pre- and post-training, based on whether they are applied before or after the core training procedure of an ML model (Bojer 2022). In the case of pre-processing techniques, two specific phases can be discerned: the engineering of explanatory features, with the aim of transforming univariate time series into multivariate sets to enhance performance, and the scaling of the records incorporated into the created multivariate data frames. Data scaling practices are commonly based on standardised protocols. In this direction, standardising the data by using min-max or Z-score scalers has gained traction in the broader field of ML-centric forecasting. In the case of intermittency, the application of Z-score with mean absolute deviation (MAD) and adaptive thresholds may be a better solution to overcome the data-specific discrepancies of the traditional Z-score scalers (Xiong et al. 2025). Other approaches such as robust scaling (median per inter-quantile regressors) for outlier-resistant normalisation and power transforms (i.e. following Box–Cox and Yeo–Johnson approaches) might also be pertinent, yet underexplored in the context of intermittency.

A critical factor influencing the effectiveness of ML models is the selection and engineering of explanatory variables, known as features, which enable the capture of various aspects related to the modelling and the temporal evolution of the demand (Makridakis, Spiliotis, and Assimakopoulos 2022). In this vein, the discussions in the literature are broad. In various application domains of ML models, beyond time series forecasting, there are research efforts focussed on developing models based on the entities themselves (e.g. image analysis), where the role of features is minimal or even non-existent. However, in the case of forecasting, the research practice paves different pathways, with numerous authors acknowledging the impact of features in developing accurate and efficient methods. This argument primarily revolves around the notion that the diversity and variability of time series problems constrain the development and widespread adoption of effective pre-trained models, as is common in image and audio analyses. This limitation necessitates feature engineering and selection, which are essential for achieving high levels of accuracy (Hafeez et al. 2021; Lainder and Wolfinger 2022).

In NN-centric applications, the engineering of features seems to have the strongest impact. Nonetheless, their impact has also been discussed in some of the other algorithmic classes; however, these are relatively limited. Table 3 presents the four classes of features mainly used in intermittent contexts. The first is related to the engineering of features aimed at evaluating temporal properties, primarily focussing on the analysis of lags, intervals with peak demand values, fluctuations, and other related elements. The second involves the encoding of domain-specific, contextual knowledge, which arguably has a specific impact on describing demand. However, directly encoding expert insights as features presents technical challenges. Systematic approaches in this vein, may involve incorporating installed knowledge bases integrated with ML components to develop quantitative estimates (Van der Auweraer, Boute, and

Table 3. A mapping of engineered features for ML-centric applications.

Product	Industry	Features [min–max]	Feature classes			
			Temporal	Contextual/Expert-elicited	Market-driven	Misc.
Spare Parts	Automotive	1–10	●	◑	❍	❍
	Aviation & Naval	3–6	◑	◑	❍	◑
	Electronics & Electrics	2–5	●	◑	❍	○
	Heavy Industries	2–7	◑	◑	○	❍
Other SKUs	Retail	3–6	●	◑	◑	❍
	Food production	3–12	◑	❍	◑	❍
	Medical equipment	2–4	❍	◑	○	○

Notes: *Notation*: ● Found in almost all cases; ◑ Found in $\geq$ 50% of cases; ❍ Found in $<$ 25% of cases; ○ Not found

Syntetos 2019). Additionally, market-related variables, including the effects of advertising and promotional activities, as well as calendar-related data such as working days and holidays, influence the identification of trends (S. Ma, Fildes, and Huang 2016). Beyond these targeted features, other categories with potentially significant influence on demand patterns include assessments of weather conditions (Y. Liu et al. 2019) and other indirectly related parameters within the broader SC environment, which are included in the Miscellaneous subclass of features.

Table 3 consolidates the results regarding the engineered features based on both the industrial settings and the underlying product, which is commonly tried to be predicted in each industry. This presentation choice was deliberately made to facilitate the identification of some contextual features, thus providing some systematic guidelines for aligning the research practices. In the broader field of spare parts forecasting, the life-cycles for different product classes may have a significant impact on forecasting, especially when coupled with specific failure rates, usage, and maintenance policies. These observations can be extrapolated to develop contextual features in each of the corresponding industries. For instance, in the case of automotive spare parts features including the vehicle aging, its operating conditions and life-cycle management of different spare parts (critical or not) might hold promise; the same categories could be used for the cases of aviation and naval industries, in which they could be coupled with platform and system taxonomy, planned maintenance cycles, mission and route profiles. In the case of electronics and electrics, contextual features might be positioned in the bill of materials and field-replaceable level, component ratings, and package compatibility, end-of-life notices, and lot or pack constraints; for heavy industries, features might be equipment class and load profile, site severity and access conditions, production and safety criticality, on-site service capability, and spares staging, and the horizon for legacy support. On the other hand, the demand for products associated with retailing SKUs is affected by different factors regarding their merchandising and replenishment practices, which may also vary based on the aggregation level (e.g. store, central warehouses). In this case, features might be better contextualised by the incorporation of expert-driven adjustments reflecting the inventory policies (i.e.service-level targets, etc.), along with promotional calendars and price actions, stock-out and shelf-availability constraints, and other channel-specific rules.

Despite the scientific consensus on the contribution of features to improving forecasting accuracy, discussions on methodologies for assessing the exact contribution of each feature to the improvement of the resulting forecast remain relatively limited. A significant part of the current state of practice focuses on the development of multiple features, with evaluation processes often taking the form of trial-and-error approaches. Meanwhile, research efforts to develop comprehensive methodologies for assessing the impact of each feature individually, after the completion of the training phase, are notably limited. Evaluating the extent to which each feature contributes to the creation of the prediction, by using systematic approaches (i.e. permutation importance, correlation mechanisms), could be considered a problem of quantifying the information entropy introduced by each corresponding variable. Capturing whether each new engineered feature increases or decreases the overall entropy may provide a basis for developing broader directions in the field. This approach could support a systematic study of feature impact, thus reducing the 'artistic' aspect sometimes associated with the engineering of exogenous features.

Except for incorporating expert-elicited features in the pre-processing phase, more straightforward approaches include designing modules in which experts assign weights to the outcomes of forecasting models. In these approaches, to mitigate the introduction of bias, experts may be required to apply weighting adjustments at a secondary level, post-training and validation of the ML components. Another critical issue concerns the level of expert participation needed to carry out informative and impactful interventions. The process of selecting an appropriate aggregation level to reconcile forecasts

is not novel in the forecasting literature. As noted by Makridakis, Spiliotis, and Assimakopoulos (2022), substantial research has explored the determination of the optimal level for integrating expert knowledge. For example, studies such as those by Kourentzes, Petropoulos, and Trapero (2014) highlight the value of expert involvement, particularly in improving long-term forecasts through time-basis aggregation. Similarly, other research efforts suggest that introducing negative adjustments to forecasting outputs can significantly improve accuracy at the SC aggregation level (Fildes et al. 2009). Despite extensive discussions on the topic, the mechanisms for incorporating expert knowledge are still debated, especially within the realm of ML-based approaches.

Beyond the feature engineering pathways on series-by-series ML training fashion, an intriguing extension of features focussed on statistical properties of time series is the application of 'cross-learning' approaches. These methods aim to draw parallels between different time series related to products with correlated demand patterns (Oliveira and Ramos 2023; Semenoglou et al. 2021). By evaluating shared characteristics across various products, the ultimate goal is to validate and optimiae the statistical features, thereby identifying the optimal feature set that enhances predictive accuracy for similar products.

Following the completion of the training phase, extensive validation tests are typically conducted to identify a satisfactory suboptimal set of hyperparameters that significantly enhance the predictive performance of the respective ML model. Consequently, hyperparameter tuning becomes a critical step in developing reliable and high-performing models, as it ensures that each algorithm operates under optimal configurations and avoids over-fitting or under-fitting across diverse datasets. Depending on the model's specifications and the objective function optimisation, different hyperparameters are examined and evaluated after the model has been trained. For instance, in SVM-based approaches, essential hyperparameters such as the penalty parameter C, the kernel type (e.g. linear, RBF), and the kernel coefficient (e.g. γ in the RBF kernel) define the decision boundary and influence the trade-off between margin maximisation and misclassification error, which in forecasting refer to deviations between predicted and actual values. NNs, whether developed in feed-forward, recurrent, or convolutional architectures, typically require tuning the number of layers, the number of neurons per layer, the learning rate, batch size, activation functions, and regularisation settings (e.g. dropout rate), all of which directly impact the network's capacity and training dynamics. For all algorithmic classes, tuning efforts are focussed on finding the set that minimises the loss function. In this step, different metrics can be selected, with practice indicating the selection of scale-dependent metrics (e.g. RMSE) for assessing the training phase more rigorously.

Implementation-wise, both manual and automated approaches could be found in the literature. When adopting manual-type approaches, researchers primarily focus on the detailed analysis of specifications to develop effective, tailor-made models. Conversely, in cases where automated tuning methodologies are employed, the focus shifts toward improving the generalisation capability of the frameworks and underlying ML methods for forecasting. In automated tuning methodologies, research efforts concentrate on employing ML-specific tools or alternatively using sophisticated meta-heuristic algorithms to facilitate convergence to specific points in the grid of matrix-like forms, where the training error is minimised. Some ML-specific tuning tools explored in the literature include Auto-sklearn and Optuna. On the other hand, scikit-learn's GridSearchCV and RandomizedSearchCV are considered as customised or semi-automated approaches (M. Hasni, Babai, and Rostami-Tabar 2024). Finally, a significant subset of hybrid methods involves integrating nature-inspired meta-heuristics into the tuning process. Central implementation schemes are oriented towards utilising GAs with single or multi-objective functions (Guo et al. 2017), such as NSGA-II, as well as swarm intelligence algorithms, to optimiae the tuning process (Jiang, Huang, and Liu 2021; Z. Li et al. 2015; Song et al. 2021).

The above findings suggest that, in intermittent demand forecasting settings, both pre- and post-training processes significantly influence the overall performance of the core ML model. This highlights the need to design and develop integrated ML pipelines, rather than focussing solely on the specifications of the core method. For tailor-made applications, hyperparameter tuning typically involves experimentation within smaller, bounded search spaces. In the case of hybrids built on multiple methods, a careful, commonly hand-crafted calibration is needed, since the interaction of heterogeneous components often creates non-linear dependencies that are difficult to capture through automated search alone. However, system-wide ML pipelines may necessitate the incorporation of cross-learned and platform-based feature sets derived from product and SC hierarchies, which increases the number of hyperparameters, thereby complicating the tuning process. To mitigate these complexity overheads, both metaheuristics and ML-centric approaches can be pivotal role in automating fine-tuning (Yang and Shami 2020). For contextualisation, we note that in system-wide applications, identifying the optimally-fitted hyperparameters is

computationally-intensive. In this vein, coupling heuristics or meta-heuristics in automated, ML-centric tuning protocols may be prominent. Practically, tailor-made heuristics capable of locating promising regions within the search space, can significantly reduce computational demands. In addition, adopting ML-centric protocols for automation could support time-adaptive (i.e. updated state-spaces) runs of metaheuristics, thus contributing to the development of integrated tuning protocols. For safeguarding the scalability, such heuristics and automation schemes should adapt bounded search spaces to the intermittency level of the data and incorporate rolling-origin evaluation schemes, which better capture real-world temporal dynamics.

Another topic of technical interest in the implementation of ML models relates to data partitioning methodologies. As widely discussed in the relevant literature on ML models, the fundamental implementation schemes emphasise that the dataset should be partitioned into three subsets: one for training, one for validation (used to fine-tune hyperparameters), and one for testing, which evaluates the model on unseen data. Many studies favour classical partitioning approaches, by allocating the majority of data for training and smaller portions for validation and testing, while others advocate for k-fold cross-validation. This method divides the dataset into k folds, iteratively training on $k-1$ folds and validating the remaining one. It ensures full utilisation of the dataset, mitigates over-fitting, and offers robust performance evaluation across diverse data patterns, which is crucial for intermittent demand scenarios. Despite its computational cost and challenges with temporal dependencies, k-fold cross-validation remains a prominent choice, particularly when data scarcity or variability is significant. Some interesting extensions to the k-fold training strategy have recently been discussed in the broader context of intermittent demand. The research by Lainder and Wolfinger (2022) argues that both purged and nested strategies significantly enhance the generalisation of results from the training phase. Both of these strategies focus on mitigating the overfitting risk through the strict separation of hyperparameter tuning and validation in nested approaches, and by preventing data leakage via buffer periods in purged strategies.

Since a significant portion of applications relies on percentage-based splits among training, validation, and testing datasets, studying the most common partitioning strategies–and how split choices influence outcomes based on the total available data–is of notable interest. Based on our analysis, a key takeaway is that smaller research datasets, particularly those involving data recorded on a monthly or weekly basis, tend to use relatively smaller partitioning for training sets. Practical implementations in such cases typically range from 65% (Kourentzes 2013) to 75% for creating training sets (Fu and Chien 2019). Conversely, applications related to the analysis of daily data, especially those extending beyond spare parts and primarily focussing on retail-partitioning practices, involve significantly larger proportions for the training sample, often reaching 90% or even higher. This preference may be attributed to the proportion of zero values present in the respective datasets, as well as to the dynamic fluctuations in demand.

6. The impact of competitions

The successful completion of competitions serves as a significant indicator of maturity and intensity of activity within specific research domains, often suggesting that research efforts can be applied to large-scale applications using datasets with specific characteristics (Spiliotis et al. 2020). The qualitative study by Alroomi et al. (2022) suggests that the outcomes of large-scale competitions have the potential to influence real-life forecasting applications, a finding also reported in Seaman and Bowman (2022) and Fildes (2020). Aligning with the latter, in this section, we complement our analysis of the relevant literature by synthesising the results of specific forecasting-related competitions whose outcomes have been published in the literature, often in the form of special issues, highlighting their role in bridging academic research and practical forecasting.

In the broader field of forecasting, notable examples of successful competitions include, but are not limited to, M1-M5, NN3, NN5, and Kaggle (Hyndman 2020). To gather pertinent practices and insights from these competitions, we follow specific eligibility criteria when researching for contributions produced. We concluded that only the M5 competition shares scope similarities with our work, while some discussions about the applicability of advanced ML models could be drawn from the NN3 and M4 competitions (Crone, Hibon, and Nikolopoulos 2011; Makridakis, Spiliotis, and Assimakopoulos 2020). The M5 competition primarily focuses on retail forecasting, whereas we were unable to locate a similar competition for the case of spare parts. These criteria include: (i) the competition should relate to product-based forecasting with intermittent demand patterns; (ii) there should be specific special issues or, at least, articles presenting the corresponding findings; and (iii) the underlying datasets of competitions should be publicly available.

Articles published under the M5 competition special issue cover a broad scope which encompasses diverse thematic areas, ranging from practical applications to higher-level commentary on the maturity of ML

methods, the ongoing relevance of time series-related approaches such as smoothing techniques (Kolassa 2022), the evaluation of methods' applicability in real-life operations (Seaman and Bowman 2022), the representativeness of methods (i.e. whether the feature engineering mechanisms developed for the M5 competition dataset are transferable to other domains) (Theodorou et al. 2022), and strategies for assessing the robustness of bottom-up aggregated methods (S. Ma and Fildes 2022). Additionally, it includes works focussed on disseminating best practices for ML pipelines based on practical, acquired knowledge (Bojer 2022). On the implementation side, we observe that in both accuracy subsets, the techniques fall into five categories: tree-based, ensemble, hybrid, RNNs, and transfer learning approaches; while in the case of uncertainty forecasting the top-ranked submission were focussed on tree-based and LSTM-centric schemes. The observation that tree-based methods, particularly those relying on gradient-boosted algorithms, represent a widely utilised category aligns with our findings on the research sample and reflects a logical extension of research interest, which has recently shifted from NN-based approaches to tree-based ones. The primary implementation techniques revolve around LightGBM algorithms (Januschowski et al. 2022; Lainder and Wolfinger 2022), while more complex forms include the development of recursive tree-building models based on the same algorithm (In and Jung 2022). Although ensemble structures, particularly stacked models, and specific NN architectures have been criticised for their computational demands, significant research efforts have emerged within the M5 competition framework. Notably, the work of Bandara et al. (2022) focuses on developing an ensemble model that combines Pooled Regression (PR; shallow ML) with LightGBM (tree-based) for forecasting. Similarly, NNs remain a significant component of research efforts, especially RNN-based architectures (Jeon and Seong 2022) and other hybrid forms for generating point forecasts. Indicative examples include combinations of FFNN (MLP architectures) with tree-based models (Nasios and Vogklis 2022); combinations of deep RNN-based models for higher aggregation levels with tree-based models for intermittent-type patterns at lower aggregation levels (Anderer and Li 2022); and combinations of autoregressive RNNs with gradient-boosted tree models (Chiew and Choong 2022). Moreover, even in hybrid methods, NNs are utilised in conjunction to enhance gradient-boosted trees further. Moreover, transfer learning methodologies are particularly noteworthy, as they potentially reduce the computational cost associated with ML methods, thereby paving the way for significant progress towards operationalisation and industrialisation, as previously discussed (Wellens, Udenio, and Boute 2022).

Beyond these elements, discussions on the results of competitions could provide valuable insights regarding the benchmarking protocols and baseline methods. In practice, determining which methods are considered competitors (performance baselines) is a matter of domain-specific knowledge, which significantly influences conclusions regarding the relative superiority of one method over another. An analysis of the findings from the competition reveals that only a small share of the proposed methods surpass the bottom-up ES approaches used as basic benchmarks in the case of point forecasts, which also holds for the uncertainty track, where three out of four submissions failed to outperform the seasonal ARIMA. Furthermore, while some methods outperformed the benchmark in terms of accuracy, the improvements were, in some instances, only incremental (Kolassa 2022). These findings are particularly noteworthy, highlighting that although ML techniques demonstrate new potential compared to simpler, plug-and-play methods, their use should not be regarded as the only avenue. The artifacts gathered from the analysis conducted in the retrieved literature sample further support this observation.

As indicated in Table 1, the authors typically follow the evaluation strategy of comparing proposed methods against at least two other categories of methods in the case of point forecasts. The majority of these categories consist of conventional (dedicated to intermittent demand) and time series methods (i.e. smoothing approaches). Collectively, these observations emphasise that the decision to develop an ML model, which involves creating an entire pipeline for data management and processing, should be made with a focus on the optimal trade-off between accuracy and computational cost. Within this context, we propose that demonstrating the reliability and strengthening the positioning of an ML-centric method should involve three successive stages of comparative evaluation: (i) against conventional methods; (ii) against time series methods; and (iii) against ML methods that have been previously validated in the literature.

7. Discussion

Our thorough analysis encompasses various aspects of technical interest in the field of ML-centric intermittent demand forecasting and synthesises the interdisciplinary body of practice to support the research questions set in this study. Commencing with the products that gathered the major interest, the research findings suggest that ML-centric efforts could be clustered into two

main categories: spare parts and retailing SKUs. Spare parts are typically associated with industries such as automotive, aviation, naval, electronics, and heavy industries. The retailing counterpart of intermittency is commonly encountered in store-level hierarchies, while other applications include food services and medical consumables. The common ground in the identified efforts is reflected in the development of integrated ML pipelines that encompass feature engineering, partitioning protocols, architectural schemes, and hyperparameter tuning approaches, each aligned to address the distinct challenges of intermittent demand.

ML pipelines typically begin with the construction of feature-rich datasets, incorporating temporal lags, intermittency indicators, and exogenous inputs, including both contextual and expert-elicited features, all of which are consistently linked to improved predictive accuracy. Arguably, the choice of data partitioning strategy is informed by the frequency, size, and volatility of the dataset. The identified pattern regarding partitioning protocols suggests that, in monthly or weekly datasets, authors tend to use smaller training samples, whereas in daily-level datasets, they often allocate larger training percentages. Additionally, we have found that cross-validation strategies such as k-fold remain largely underused, despite being widely recommended in the broader ML literature for improving generalisation and mitigating overfitting risks. Regarding model architecture, the literature is dominated by NNs (FFNNs, LSTM-type RNNs) and tree-based learners like LightGBM and XGBoost, selected for their balance between accuracy and computational efficiency. Hybrid and ensemble approaches are frequently adopted to combine statistical structure with ML flexibility, while transfer learning is gaining traction as a means to reduce data hunger and enable knowledge reuse across product families. Training routines commonly involve extensive hyperparameter tuning, using manual grid search, AutoML tools, or metaheuristic algorithms such as GA, PSO, or NSGA-II. Finally, evaluation practices in point forecasts often rely on combinations of different metrics (i.e. scale-dependent with percentage-based or scale-free), which seems in contrast to what established protocols suggest. In the case of quantile forecasts, evaluation practices seem to be streamlined by commonly relying on the the SPL metric.

Although the above validates the significant research progress that has been made in ML-centric intermittent demand forecasting, we believe certain aspects require further investigation, focussing on addressing specific research gaps and fostering consensus on particular technical issues. Within this context, we structure the final section of our work along three key dimensions. First, we outline the research gaps identified through a critical evaluation of the findings. Subsequently, we present specific open issues that emerge from a comparative assessment of findings from existing research efforts in SC and spare parts and SKUs forecasting, alongside other application domains of ML-based models. The emphasis here is primarily on the strengths of methods that, although under-explored, could potentially significantly enhance existing models or serve as a foundation for reconciling the dynamic nature and adaptiveness of ML models. Finally, by synthesising insights from these two areas of interest, we provide a concise roadmap for future research in the field, aiming to advance the current state of knowledge and, by extension, address the third research question set in this research.

7.1. Research gaps and open issues

Despite extensive research into ML algorithms for intermittent-demand forecasting, research practices have not converged in several areas of technical interest. A notable limitation refers to the limited development of probabilistic forecasts, despite their significance in inventory-related decision-making. While preprocessing, feature engineering, and other calibration procedures described in the implementation details could be pertinent to both probabilistic and point-forecasting schemes, because they primarily concern data representation and model, topologythis equivalence does not hold for evaluation protocols. Notwithstanding, the discussions made for reflecting the current evaluation protocols in this field and the preliminary identification of streamlined practices regarding the incorporation of both inventory- and statistics-related metrics, the limited number of ML-centric probabilistic forecasts hinders the ability to draw broader conclusions regarding the alignment of practices reflecting topics such as the set of metrics and baselines. This observation opens a fruitful area for future efforts aiming at bridging the theoretical gap of how uncertainty could be formalised in ML pipelines to support decision-making.

Beyond the above-mentioned research gap, several issues relevant to point forecasts remain open. Starting from the fundamental question of what constitutes the optimal ML technique for producing intermittent-type forecasts, it is evident that there is no definitive answer. The research findings show that the selection of algorithms is significantly dependent on the properties of the underlying dataset. For datasets with relatively large sample sizes, research has primarily focussed on the development of FFNN architectures and hybrid approaches that integrate both neural networks (FFNN, RNN) and tree-based models. In the

case of larger datasets, recent novelties illustrate the potential of MLP-centric architectures such as N-BEATS (Neural Basis Expansion Analysis for Time Series) and N-HiTS (Neural Hierarchical Interpolation for Time Series). Yet, they fall short against gradient-boosted trees (e.g. LightGBM, XGBoost), which can be broadly attributed to the operationalisation capabilities, validated through performance-computation cost trade-offs (Hobor et al. 2025). Conversely, when the sample size is smaller, comprising only tens or hundreds of records, research practices exhibit greater variability, with notable efforts directed towards developing specialised hybrid architectures. In the class of RNNs, LSTM-infused implementations are generally considered effective, whereas other network types, such as GRUs, remain underexplored.

On the other hand, the terrain remains highly debatable regarding the selection of the optimal technique for implementing FFNNs. Comparative assessments often highlight the potential advantages of MLP approaches while raising critical questions, such as the trade-offs between accuracy and training time, with arguments suggesting that ELMs may serve as more powerful tools in specific scenarios. As previously highlighted, research findings indicate that while ML approaches represent the next step in research practices, they are not an exclusive solution at the operational level. Specific techniques can achieve high accuracy levels by combining simple statistical methods (Fiorucci and Louzada 2020; Pawlikowski and Chorowska 2020; Shaub 2020). From a comparative perspective, this observation fuels discussions that remain underexplored, particularly regarding the methodologies required to benchmark the effectiveness of a proposed method. In this regard, we believe that following the comparison pathway of three successive stages (against conventional, time series, and bootstrapping methods), as previously discussed, may add value to assessing the utility, business contribution, and novelty of the ML methods, yet this protocol needs to be further assessed in the literature.

Beyond open issues oriented to methods, a notable gap was identified in evaluation metrics. Despite the contribution of specialised methods for selecting evaluation metrics–many of which recommend scale-agnostic (e.g. RMSSE) or percentage-based symmetric metrics (e.g. symmetric MAPE)–this practice has not been fully adopted in research. The findings indicate that the authors use evaluation metrics from two or more classes, often including scale-free or percentage-based measures. The central question here revolves around the necessity of scale-dependent methods. This appears to be related to the need for strict metrics based on absolute error values, which may be reasonable during training loss evaluation but less so during blind forecasting. Another issue is related to the limited inclusion of inventory-specific metrics in the evaluation of point forecasts. Previous findings suggest that in the broader field of intermittent demand forecasting, the incorporation of such metrics is limited (Pinçe, Turrini, and Meissner 2021; S. Wang, Kang, and Petropoulos 2024), which is even more evident in the field of ML-centric forecasting. Since forecasts primarily act to support inventory decision-making, the incorporation of inventory-specific measures should be prioritised in future efforts.

Regarding ML model tuning strategies, the issue of partitioning remains open. Notably, research highlights cross-validation methods, such as k-fold and its variations, as prominent approaches to enhancing generalisation to the test sample by avoiding over-fitting hyperparameters to the training sample. However, a significant portion of the research continues to employ percentage splitting logic, with some studies utilising validation samples significantly smaller than the training samples. This raises a critical question about how the training sample size affects the generalisability of the methods. Recent discussions in the field suggest that some solutions to the training process may be oriented towards using larger samples that could be created by utilising synthetic data and augmentation methods.

Feature engineering is an established process in ML pipelines, tailored to transforming time series from univariate to multivariate. However, the current practice proves that many of these applications are tailor-made in this implementation phase. In the context of both hyperparameter tuning and general feature selection, the integration of nature-inspired algorithms into core ML approaches remains under-investigated. This presents a discrepancy, given the well-established ability of such techniques to efficiently reach optimal or near-optimal solutions in large search spaces–an asset that could help streamline these otherwise time-consuming processes.

Another issue stems from the limited protocols for assessing the performance of an ML method at a dataset-wide level. In many cases, the authors tend to evaluate a method by applying a horizontal setup, specifically by using the same set of features and evaluation metrics. This is in contrast to the common ground in ML, based on which the same algorithms may perform differently when applied in series with different properties. In this vein, we suggest that the evaluation of an ML forecast would be more informative if unfolded in a two-dimensional manner. Firstly, in a series-by-series approach, by focussing on locating the set of features that consistently work well under varying data properties in intermittent-demand patterns (recognising that different

features may be most pertinent in different cases). That is, by analysing performance across predefined states over a selected set of time series; in this case, the lower threshold of six products, identified by common practice in the sample, should be further evaluated and clarified. Secondly, by employing transfer mechanisms for evaluating the method's performance in specific regions of the dataset that exhibit data properties similar to those analysed in the initial phase. Given the impact of features on a model's performance, this procedure may rely on incorporating feature pooling mechanisms; that is, keeping the method fixed while, for each series, selecting an appropriate subset from a set of features computed across the dataset.

Beyond feature engineering and generalisation protocols, some data-centric gaps also exist. Initially, the availability and application of data augmentation techniques tailored specifically to intermittent-type time series remain limited, resulting in a reduced capacity to increase the representativeness and robustness of such datasets artificially. This is in contrast to the practices applied in the broader field of univariate time series forecasting (Semenoglou, Spiliotis, and Assimakopoulos 2023), thus creating a significant gap in the literature. Furthermore, the computing paradigms and frameworks capable of reconciling heterogeneous datasets, particularly those characterised by irregular sampling intervals and differing granularity, are still restricted in both number and functionality. Some recent efforts in this direction have relied on the flexibility of Echo State Network (ESN)-infused paradigms (Tsantilis et al. 2025). Other prominent avenues may include the coupling of distribution-aware training with point estimation, towards jointly optimiaing quantile (pinball) losses for calibrated uncertainty and point-wise objectives (Wu, Fu, and Xia 2025; Jeong and Lee 2025). In both directions, there is still significant room for improvement, creating a critical need for the development of specialised computational methods and tools designed explicitly to address the inherent diversity and complexity encountered within forecasting datasets.

7.2. A roadmap for further research

In what follows, we summarise specific dimensions that, although particularly prominent in our view, have not yet been explored to their fullest extent or accorded the attention they deserve. By critically presenting a road map for further research in the following topics, we aim to pave the way for discussions in the relevant domains, thereby facilitating evidence-based documentation regarding the techniques and means of implementation, which could support the industrial-scale adoption of such tools.

(1) Since the effectiveness of forecasting methods varies across time horizons, a promising direction is the development of protocols to identify the most suitable methods for short-, medium-, and long-term planning. Future research should also aim to develop integrated forecasting models that combine multiple techniques and metrics aligned with specific time intervals, enhancing accuracy and applicability. Such combinatorial approaches can support context-specific forecasting and improve the strategic use of ML methods in diverse business settings.

(2) Another area of further research lies in the integration of reinforcement learning (RL) techniques with existing forecasting models, particularly to address the absence of expert opinions while also considering the unique operating conditions of SC networks. This includes exploring how RL can dynamically optimiae forecasting strategies in response to the volatile and unpredictable nature of intermittent demand.

(3) Future research should extend the use of meta- and transfer learning mechanisms to enhance forecasting under intermittent demand, where sparsity hinders the potential of ML methods. Despite promising early results on developing transfer learning mechanisms built upon GRU- and CNN-based architectures, the full potential of these approaches remains underexplored, highlighting a key opportunity for advancing industrial-scale forecasting solutions.

(4) Both exogenous and expert-elicited features were found to have a significant impact on ML models. We argue that defining engineering processes for incorporating such variables within the training sample represents a prominent research avenue in the development of automated ML pipelines. Towards this direction, we consider that leveraging mature research activities from other forecasting-oriented domains (i.e. meteorology), could serve as a starting point for enriching feature sets and extending them beyond the typical time series-related features.

(5) Although discussions surrounding the maturation of ML pipelines have been identified, streamlined protocols for embedding ML models within legacy ERP systems are under-delivered. Future research should tailor capabilities to ERP-specific needs, including cost–benefit analyses of integrations and data-fusion schemes that enable ensemble modelling in intermittent-demand settings. In this thematic area, future directions may centre on operationalising upstream layers that automatically identify demand patterns–combining stock-out detection with level correction and demand-size

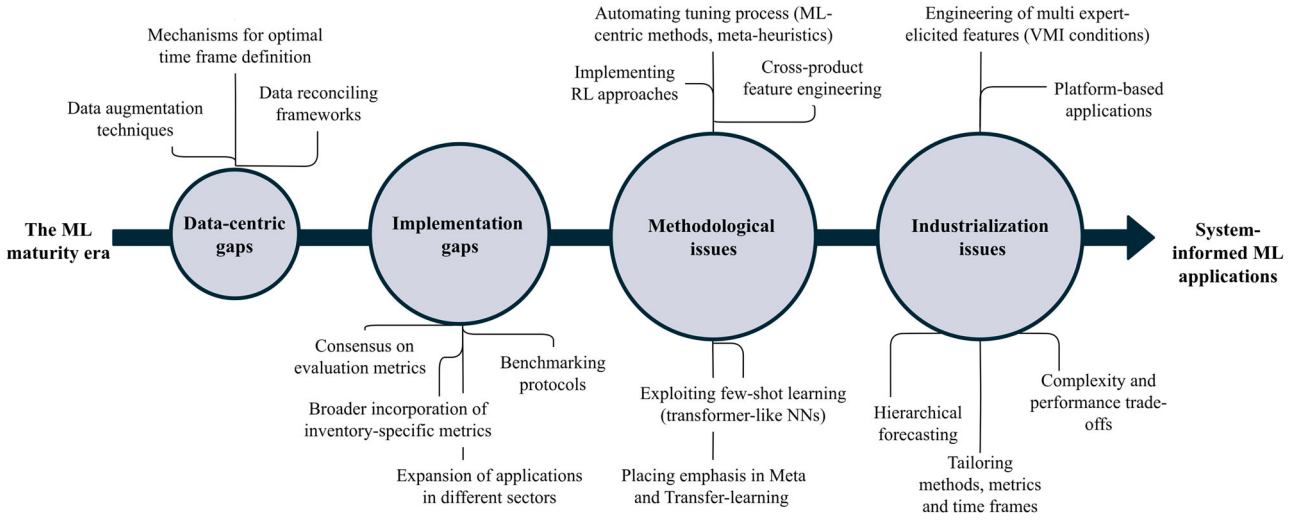


Figure 9. A roadmap towards industrialisation of ML methods in intermittent demand forecasting contexts.

decomposition–thereby enabling ERP workflows to persist stockout flags and regime labels alongside forecasts (Svetunkov and Sroginis 2025). In parallel, advancing vendor-managed forecasting tools requires robust hybrid model-assignment frameworks informed by hierarchical structures or clustering logic, alongside scalable reconciliation mechanisms that ensure forecast coherence across SC levels.

By critically synthesising the research and implementation gaps identified with the future research avenues proposed, we conceptually infer that four major research themes require further consideration. Bridging these gaps requires a research roadmap towards the industrialisation of ML-centric methodologies (see Figure 9). In our view, this roadmap could be pivotal for future work in both point and probabilistic forecasting. ML-specific forecasts may build on the knowledge generated by point forecast schemes and tailor these insights, as described in detail earlier in the manuscript, to probabilistic objectives. Since both approaches share architectural and data-centric similarities, we believe the focus should shift to other pressing issues that require systematic research documentation, such as developing robust methods for reconciling data distributions through weighted combinations of forecasts from different quantiles. Another important avenue is the establishment of benchmarking protocols for probabilistic forecasting, as current practice often relies on simple comparisons against naive methods.

7.3. Concluding remarks and limitations

ML techniques constitute a prominent, yet demanding avenue for improving the accuracy and effectiveness of forecasting systems. In this review, we consolidate the state of research on ML-based forecasting methods for intermittent demand by offering a product-agnostic, technically detailed synthesis across diverse industries, thereby providing insights for the development of comprehensive ML pipelines. Some key conclusions derived from our analysis:

- The ML-centric research streams focussing on probabilistic forecasting are significantly under-explored in comparison to point-forecasting approaches.
- Intermittent demand manifests itself in various product types, beyond spare parts. ML applications cover a broad spectrum of industries, including automotive, aerospace, retail, electronics, and heavy industries.
- There is a clear dominance of NN-based and tree-based methods, along with growing relevance of meta- and transfer learning mechanisms, which signifies the industrialisation potential of lightweight schemes. Most applications focus on FFNNs and RNNs (esp. LSTMs), while newer architectures (e.g. Transformers) remain underexplored.
- Feature engineering–especially with exogenous variables (e.g.promotions, weather)–significantly improves model performance.
- Transfer learning and meta-learning are promising for scalability and reducing training data requirements.
- Most ML models require careful hyperparameter tuning and data partitioning strategies, which are not consistently reported across studies.
- Real-world datasets are underutilised; many studies rely on synthetic or curated datasets that may not reflect operational complexity. In addition, parse, noisy, and irregular data patterns are major barriers to ML performance in intermittent demand.

- There is no universally accepted benchmarking protocol for comparing ML models in the context of intermittent demand. In the case of point forecasts, there is no consensus on which accuracy metrics (i.e.RMSSE, RMSE, MAPE, MASE) should be used, making cross-study comparisons difficult.
- Evaluation frameworks often lack consideration of operational outcomes (e.g. inventory cost, targeted service level), which could be attributed to the limited development of probabilistic forecasts.
- Interpretability and integration into decision-support systems are essential but often neglected in research. This is particularly true since many high-performing models are complex and computationally intensive, which limits their real-world adoption.
- Operational needs (e.g. real-time updating, explainability, low-resource environments) require attention in model design.

Some limitations relevant to conducting our review should be kept in mind. During the search process, we may not have achieved total content saturation, although we have included papers published in peer-reviewed journals and conference proceedings. This focus might have inadvertently excluded relevant works published as technical reports (especially for the grey literature). However, our goal was to provide a comprehensive analysis and deliver reliable conclusions based on robust underlying data from the scientific literature, ensured by quality control, credibility, rigor, and reproducibility. Some process-oriented and theoretical issues inherent to content analysis should also be reported. In particular, content analysis can be overly reductive and/or subjective, leading to increased abstraction and/or bias and potentially impacting the credibility and trustworthiness of the overall analysis and classification of the underlying data. To mitigate these inherent limitations, we have undergone extensive consultation sessions within the research team to reach a consensus on the key topics discussed, which has also been validated by calculating the Holsti coefficient.

Note

1. Kaggle: https://www.kaggle.com/c/m5-forecasting-accuracy

Acknowledgements

The authors would like to sincerely thank the Editor, Associate Editor, and two anonymous reviewers of the manuscript for their constructive comments on a previous version, which significantly improved its quality.

Disclosure statement

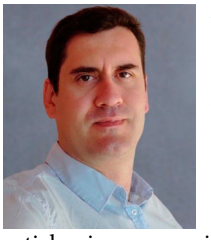
No potential conflict of interest was reported by the author(s).

Funding

This work has been partly supported by the University of Piraeus Research Center.

Notes on contributors

Panagiotis G. Giannopoulos holds an integrated diploma in Mechanical Engineering (MSc equivalent) and an MSc in Industrial Management and Technology. He is currently a PhD student at the Hellenic Open University. His main research interests include: Operations Research, Supply Chain management, Industry 4.0 technologies, Applied & Computational Intelligence algorithms. He has contributed to more than 15 papers published in peer-reviewed journals and proceedings of international conferences.



Dr. Thomas K. Dasaklis is an Assistant Professor at the Hellenic Open University, Greece. His research interests focus on Artificial Intelligence, blockchain-related applications and Industry 4.0 technologies in business administration and supply chain management. Dr. Dasaklis has authored and co-authored more than 65 articles in peer-reviewed journals and international conferences, with publications in outlets such as International Journal of Production Research, International Journal of Production Economics, European Journal of Operational Research, and Informatics & Telematics, among others. He actively collaborates with academia and industry on research and innovation projects related to digital transformation in business, supply chain management and usage of emerging technologies.

Ioannis Tsantilis holds a B.Sc. in Computer Systems Engineering & Telecommunications, an M.Sc. in Advanced Information Systems, and is currently a PhD candidate in Informatics at the University of Piraeus, Greece. He has extensive experience as a Senior DevOps Engineer and Researcher, contributing to large-scale industry projects and European R&D initiatives. He has co-authored publications in peer-reviewed journals and conferences. His research explores the application of chaos theory and machine learning to improve demand forecasting, with particular interest in sporadic demand prediction and the optimisation of predictive models for inventory management.

Constantinos Patsakis received a B.Sc. in mathematics from the University of Athens, Greece, an M.Sc. degree in information security from Royal Holloway, University of London, and a PhD in cryptography and malware from the University of Piraeus. In the past, he has worked as a researcher at the UNESCO Chair in

Data Privacy, at Rovira i Virgili, at Trinity College, Dublin, and the Luxembourg Institute of Science and Technology. He is a Professor at the University of Piraeus and an Adjunct Researcher at the Athena Research and Innovation Center. He has authored more than 150 publications in prestigious peer-reviewed international conferences and journals and participated in several national and European Research and Development projects. His primary areas of research include cryptography, security, privacy, blockchain technology, and cybercrime.

Data availability statement

The data that support the findings of this study are available from the corresponding author, [Thomas K. Dasaklis], upon reasonable request.

References

Ahmadov, Y., and P. Helo. 2023. "Deep Learning-Based Approach for Forecasting Intermittent Online Sales." *Discover Artificial Intelligence* 3 (1): 45. https://doi.org/10.1007/s44163-023-00085-1.

Aktepe, A., E. Yanık, and S. Ersöz. 2021. "Demand Forecasting Application with Regression and Artificial Intelligence Methods in a Construction Machinery Company." *Journal of Intelligent Manufacturing* 32 (6): 1587–1604. https://doi.org/10.1007/s10845-021-01737-8.

AlAlaween, W. H., O. A. Abueed, A. H. AlAlawin, O. H. Abdallah, N. T. Albashabsheh, E. S. AbdelAll, and Y. A. Al-Abdallat. 2022. "Artificial Neural Networks for Predicting the Demand and Price of the Hybrid Electric Vehicle Spare Parts." *Cogent Engineering* 9 (1): 2075075. https://doi.org/10.1080/23311916.2022.2075075.

Alroomi, A., G. Karamatzanis, K. Nikolopoulos, A. Tilba, and S. Xiao. 2022. "Fathoming Empirical Forecasting Competitions' Winners." *International Journal of Forecasting* 38 (4): 1519–1525. https://doi.org/10.1016/j.ijforecast.2022.03.010.

Amin-Naseri, M. R., and B. R. Tabar. 2008. "Neural Network Approach to Lumpy Demand Forecasting for Spare Parts in Process Industries." In *Proceedings of the 2008 International Conference on Computer and Communication Engineering (ICCCE)*, 1378–1382. Kuala Lumpur, Malaysia. https://doi.org/10.1109/ICCCE.2008.4580831.

Anderer, M., and F. Li. 2022. "Hierarchical Forecasting with a Top-down Alignment of Independent-Level Forecasts." *International Journal of Forecasting* 38 (4): 1405–1414. https://doi.org/10.1016/j.ijforecast.2021.12.015.

Andrade, L. A. C. G., and C. B. Cunha. 2023. "Disaggregated Retail Forecasting: A Gradient Boosting Approach." *Applied Soft Computing* 141: 110283. https://doi.org/10.1016/j.asoc.2023.110283.

Anglou, F. Z., S. Ponis, and A. Spanos. 2021. "A Machine Learning Approach to Enable Bulk Orders of Critical Spare-Parts in the Shipping Industry." *Journal of Industrial Engineering and Management* 14 (3): 604–621. https://doi.org/10.3926/jiem.3446.

Aria, M., and C. Cuccurullo. 2017. "bibliometrix: An R-tool for Comprehensive Science Mapping Analysis." *Journal of Informetrics* 11 (4): 959–975. https://doi.org/10.1016/j.joi.2017.08.007.

Babai, M. Z., M. Arampatzis, M. Hasni, F. Lolli, and A. Tsadiras. 2024. "On the Use of Machine Learning in Supply Chain Management: A Systematic Review." *IMA Journal of Management Mathematics* 36 (1): 21–49. https://doi.org/10.1093/imaman/dpae029.

Babai, M. Z., J. E. Boylan, and B. Rostami-Tabar. 2021. "Demand Forecasting in Supply Chains: A Review of Aggregation and Hierarchical Approaches." *International Journal of Production Research* 60 (1): 324–348. https://doi.org/10.1080/00207543.2021.2005268.

Babai, M. Z., A. Tsadiras, and C. Papadopoulos. 2020. "On the Empirical Performance of Some New Neural Network Methods for Forecasting Intermittent Demand." *IMA Journal of Management Mathematics* 31 (3): 281–305. https://doi.org/10.1093/imaman/dpaa003.

Babaveisi, V., E. Teimoury, M. R. Gholamian, and B. Rostami-Tabar. 2023. "Integrated Demand Forecasting and Planning Model for Repairable Spare Part: An Empirical Investigation." *International Journal of Production Research* 61 (20): 6791–6807. https://doi.org/10.1080/00207543.2022.2137596.

Bacchetti, A., and N. Saccani. 2012. "Spare Parts Classification and Demand Forecasting for Stock Control: Investigating the Gap between Research and Practice." *Omega* 40 (6): 722–737. https://doi.org/10.1016/j.omega.2011.06.008.

Bandara, K., H. Hewamalage, R. Godahewa, and P. Gamakumara. 2022. "A Fast and Scalable Ensemble of Global Models with Long Memory and Data Partitioning for the M5 Forecasting Competition." *International Journal of Forecasting* 38 (4): 1400–1404. https://doi.org/10.1016/j.ijforecast.2021.11.004.

Barrow, D. K., and N. Kourentzes. 2016. "Distributions of Forecasting Errors of Forecast Combinations: Implications for Inventory Management." *International Journal of Production Economics* 177:24–33. https://doi.org/10.1016/j.ijpe.2016.03.017.

Basten, R. J. I., and G. J. van Houtum. 2014. "System-Oriented Inventory Models for Spare Parts." *Surveys in Operations Research and Management Science* 19 (1): 34–55. https://doi.org/10.1016/j.sorms.2014.05.002.

Benhamida, F. Z., O. Kaddouri, T. Ouhrouche, M. Benaichouche, D. Casado-Mansilla, and D. López-de-Ipiña. 2020. "Stock&Buy: A New Demand Forecasting Tool for Inventory Control." In *Proceedings of the 2020 5th International Conference on Smart and Sustainable Technologies (SpliTech)*, 1–6. Split, Croatia. https://doi.org/10.23919/SpliTech49282.2020.9243824.

Bojer, C. S. 2022. "Understanding Machine Learning-Based Forecasting Methods: A Decomposition Framework and Research Opportunities." *International Journal of Forecasting* 38 (4): 1555–1561. https://doi.org/10.1016/j.ijforecast.2021.11.003.

Boutselis, P., and K. McNaught. 2019. "Using Bayesian Networks to Forecast Spares Demand from Equipment Failures in a Changing Service Logistics Context." *International Journal of Production Economics* 209:325–333. https://doi.org/10.1016/j.ijpe.2018.06.017.

Boylan, J. E., and A. A. Syntetos. 2010. "Spare Parts Management: A Review of Forecasting Research and Extensions." *IMA Journal of Management Mathematics* 21 (3): 227–237. https://doi.org/10.1093/imaman/dpp016.

Briner, R. B., and D. Denyer. 2012. "Systematic Review and Evidence Synthesis as a Practice and Scholarship Tool." In *The*

Oxford Handbook of Evidence-Based Management, edited by D. M. Rousseau, 112–129. Oxford University Press.

Bronnenberg, B. J., M. W. Kruger, and C. F. Mela. 2008. "Database Paper–the IRI Marketing Data Set." *Marketing Science* 27 (4): 745–748. https://doi.org/10.1287/mksc.1070.0330.

Campbell, M., S. V. Katikireddi, A. Sowden, and H. Thomson. 2019. "Lack of Transparency in Reporting Narrative Synthesis of Quantitative Data: A Methodological Assessment of Systematic Reviews." *Journal of Clinical Epidemiology* 105:1–9. https://doi.org/10.1016/j.jclinepi.2018.08.019.

Cao, Y., and Y. Li. 2014. "A Two-Stage Approach of Forecasting Spare Parts Demand Using Particle Swarm Optimization and Fuzzy Neural Network." *Journal of Computational Information Systems* 10 (15): 6785–6793. https://doi.org/10.12733/jcis11577.

Caserta, M., and L. D'Angelo. 2024. "Intermittent Demand Forecasting for Spare Parts with Little Historical Information." *Journal of the Operational Research Society* 294–309. https://doi.org/10.1080/01605682.2024.2349734.

Cawley, G. C., and N. L. C. Talbot. 2010. "On over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation." *The Journal of Machine Learning Research* 11:2079–2107.

Ceni, A., and C. Gallicchio. 2024. "Residual Echo State Networks: Residual Recurrent Neural Networks with Stable Dynamics and Fast Learning." *Neurocomputing* 597:127966. https://doi.org/10.1016/j.neucom.2024.127966.

Chandriah, K. K., and R. V. Naraganahalli. 2021. "RNN / LSTM with Modified Adam Optimizer in Deep Learning Approach for Automobile Spare Parts Demand Forecasting." *Multimedia Tools and Applications* 80 (17): 26145–26159. https://doi.org/10.1007/s11042-021-10913-0.

Chen, F.-L., Y.-C. Chen, and J.-Y. Kuo. 2010. "Applying Moving Back-Propagation Neural Network and Moving Fuzzy Neuron Network to Predict the Requirement of Critical Spare Parts." *Expert Systems with Applications* 37 (6): 4358–4367. https://doi.org/10.1016/j.eswa.2009.11.092.

Chen, Y., P. Liu, and L. Yu. 2010. "Aftermarket Demands Forecasting with a Regression–Bayesian–BPNN Model." In *Proceedings of the 2010 IEEE International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, 52–55. Hangzhou, China. https://doi.org/10.1109/ISKE.2010.5680793.

Chien, C.-F., C.-C. Ku, and Y.-Y. Lu. 2023. "Ensemble Learning for Demand Forecast of after-market Spare Parts to Empower Data-Driven Value Chain and an Empirical Study." *Computers and Industrial Engineering* 185: 109670. https://doi.org/10.1016/j.cie.2023.109670.

Chiew, E., and S. S. Choong. 2022. "A Solution for M5 Forecasting – Uncertainty: Hybrid Gradient Boosting and Autoregressive Recurrent Neural Network for Quantile Estimation." *International Journal of Forecasting* 38 (4): 1442–1447. https://doi.org/10.1016/j.ijforecast.2022.01.009.

Choi, B., and J. H. Suh. 2020. "Forecasting Spare Parts Demand of Military Aircraft: Comparisons of Data Mining Techniques and Managerial Features from the Case of South Korea." *Sustainability (Switzerland)* 12 (15): 6045. https://doi.org/10.3390/su12156045.

Craparotta, G., S. Thomassey, and A. Biolatti. 2019. "A Siamese Neural Network Application for Sales Forecasting of New Fashion Products Using Heterogeneous Data." *International Journal of Computational Intelligence Systems* 12 (2): 1537–1546. https://doi.org/10.2991/ijcis.d.191122.002.

Crone, S. F., M. Hibon, and K. Nikolopoulos. 2011. "Advances in Forecasting with Neural Networks? Empirical Evidence from the NN3 Competition on Time Series Prediction." *International Journal of Forecasting* 27 (3): 635–660. https://doi.org/10.1016/j.ijforecast.2011.04.001.

Croston, J. D. 1972. "Forecasting and Stock Control for Intermittent Demands." *Journal of the Operational Research Society* 23 (3): 289–303. https://doi.org/10.1057/jors.1972.50.

Cui, Z., H. Jia, Q. Gao, and H. Song. 2024. "Maintenance Spare Parts Prediction Based on Multilevel Migration Learning CNN-ISE-Attention-BiLSTM." *IEEE Access* 12:15208–15221. https://doi.org/10.1109/ACCESS.2024.3357994.

Dasaklis, T. K., E. Kopanaki, P. T. Chountalas, N. P. Rachaniotis, T. G. Voutsinas, K. Giannakis, and G. Chondrokoukis. 2024. "Exploring the Implementation Challenges of the Electronic Freight Transport Information (eFTI) Regulation: An Empirical Perspective from Greece." *Logistics* 8 (1): 30. https://doi.org/10.3390/logistics8010030.

Dasaklis, T. K., T. G. Voutsinas, G. T. Tsoulfas, and F. Casino. 2022. "A Systematic Literature Review of Blockchain-Enabled Supply Chain Traceability Implementations." *Sustainability* 14 (4): 2439. https://doi.org/10.3390/su14042439.

de Melo Menezes, B. A., D. de Siqueira Braga, B. Hellingrath, and F. B. de Lima Neto. 2015. "An Evaluation of Forecasting Methods for Anticipating Spare Parts Demand." In *Proceedings of the 2015 Latin America Congress on Computational Intelligence (LA-CCI)*, 1–6. Curitiba, Brazil. https://doi.org/10.1109/LA-CCI.2015.7435980.

de Oliveira, A. C. A., J. M. Jorge, and G. P. Rocha Filho. 2020. "Neural Network with Specialized Knowledge for Forecasting Intermittent Demand." In *Transdisciplinary Engineering for Complex Socio-Technical Systems – Real-Life Applications. Advances in Transdisciplinary Engineering*, Vol 12, edited by J. Pokojski et al., 524–533. Amsterdam: IOS Press. https://doi.org/10.3233/ATDE200113.

Dhillon, A., and G. K. Verma. 2020. "Convolutional Neural Network: A Review of Models, Methodologies and Applications to Object Detection." *Progress in Artificial Intelligence* 9 (2): 85–112. https://doi.org/10.1007/s13748-019-00203-0.

Dietterich, T. G. 2000. "Ensemble Methods in Machine Learning." In *Multiple Classifier Systems*, 1–15. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-45014-9_1.

Dodin, P., J. Xiao, Y. Adulyasak, N. Etebari Alamdari, L. Gauthier, P. Grangier, P. Lemaitre, and W. L. Hamilton. 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.

Donthu, N., S. Kumar, D. Mukherjee, N. Pandey, and W. M. Lim. 2021. "How to Conduct a Bibliometric Analysis: An Overview and Guidelines." *Journal of Business Research* 133:285–296. https://doi.org/10.1016/j.jbusres.2021.04.070.

Ducharme, C., B. Agard, and M. Trépanier. 2024. "Improving Demand Forecasting for Customers with Missing Downstream Data in Intermittent Demand Supply Chains with Supervised Multivariate Clustering." *Journal of Forecasting* 43 (5): 1661–1681. https://doi.org/10.1002/for.3095.

El Garrab, H., D. Lemoine, A. Lazrak, R. Heidsieck, and B. Castanier. 2023. "Predicting the Reverse Flow of Spare Parts in a Complex Supply Chain: Contribution of Hybrid Machine Learning Methods in an Industrial Context." *International Journal of Logistics Systems and Management* 45 (2): 131–158. https://doi.org/10.1504/ijlsm.2023.131425.

Elo, S., and H. Kyngäs. 2008. "The Qualitative Content Analysis Process." *Journal of Advanced Nursing* 62 (1): 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x.

Fan, L., X. Liu, W. Mao, K. Yang, and Z. Song. 2023. "Spare Parts Demand Forecasting Method Based on Intermittent Feature Adaptation." *Entropy* 25 (5): 764. https://doi.org/10.3390/e25050764.

Fan, L., Z. Song, W. Mao, T. Luo, W. Wang, K. Yang, and F. Cao. 2024. "Change Is Safer: A Dynamic Safety Stock Model for Inventory Management of Large Manufacturing Enterprise Based on Intermittent Time Series Forecasting." *Journal of Intelligent Manufacturing* 36: 3983–4003. https://doi.org/10.1007/s10845-024-02442-y.

Fildes, R. 2020. "Learning from Forecasting Competitions." *International Journal of Forecasting* 36 (1): 186–188. https://doi.org/10.1016/j.ijforecast.2019.04.012.

Fildes, R., P. Goodwin, M. Lawrence, and K. Nikolopoulos. 2009. "Effective Forecasting and Judgmental Adjustments: An Empirical Evaluation and Strategies for Improvement in Supply-Chain Planning." *International Journal of Forecasting* 25 (1): 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010.

Fildes, R., S. Ma, and S. Kolassa. 2022. "Retail Forecasting: Research and Practice." *International Journal of Forecasting* 38 (4): 1283–1318. https://doi.org/10.1016/j.ijforecast.2019.06.004.

Fiorucci, J. A., and F. Louzada. 2020. "GROEC: Combination Method via Generalized Rolling Origin Evaluation." *International Journal of Forecasting*36 (1): 105–109. https://doi.org/10.1016/j.ijforecast.2019.04.013.

Fu, W., C. F. Chien, and Z. H. Lin. 2018. "A Hybrid Forecasting Framework with Neural Network and Time Series Method for Intermittent Demand in Semiconductor Supply Chain." In *Advances in Production Management Systems. Smart Manufacturing for Industry 4.0*. APMS 2018. *IFIP Advances in Information and Communication Technology*, Vol 536, edited by I. Moon, G. Lee, J. Park, D. Kiritsis, and G. von Cieminski. Cham: Springer. https://doi.org/10.1007/978-3-319-99707-0_9.

Fu, W., and C.-F. Chien. 2019. "UNISON Data-Driven Intermittent Demand Forecast Framework to Empower Supply Chain Resilience and an Empirical Study in Electronics Distribution." *Computers and Industrial Engineering* 135:940–949. https://doi.org/10.1016/j.cie.2019.07.002.

Gabellini, M., F. Calabrese, A. Regattieri, and E. Ferrari. 2022. "Multivariate Multi-output LSTM for Time Series Forecasting with Intermittent Demand Patterns." In *Proceedings of the 27th Summer School Francesco Turco*, 1–7. Sanremo, Italy.

Ganaie, M. A., M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan. 2022. "Ensemble Deep Learning: A Review." *Engineering Applications of Artificial Intelligence* 115:105151. https://doi.org/10.1016/j.engappai.2022.105151.

Gansterer, M. 2015. "Aggregate Planning and Forecasting in Make-to-order Production Systems." *International Journal of Production Economics* 170:521–528. https://doi.org/10.1016/j.ijpe.2015.06.001.

Ghobbar, A. A., and C. H. Friend. 2003. "Evaluation of Forecasting Methods for Intermittent Parts Demand in the Field of Aviation: A Predictive Model." *Computers and Operations Research* 30 (14): 2097–2114. https://doi.org/10.1016/S0305-0548(02)00125-9.

Giannopoulos, P. G., T. K. Dasaklis, and N. Rachaniotis. 2024. "Development and Evaluation of a Novel Framework to Enhance K-NN Algorithm's Accuracy in Data Sparsity Contexts." *Scientific Reports* 14:28210. https://doi.org/10.1038/s41598-024-79198-1.

Godahewa, R., C. Bergmeir, G. Webb, R. Hyndman, and P. Montero-Manso. 2020. "Car Parts Dataset (with Missing Values) [Data Set]." Zenodo. Version 2. https://doi.org/10.5281/zenodo.4656022.

Guo, F., J. Diao, Q. Zhao, D. Wang, and Q. Sun. 2017. "A Double-Level Combination Approach for Demand Forecasting of Repairable Airplane Spare Parts Based on Turnover Data." *Computers and Industrial Engineering* 110:92–108. https://doi.org/10.1016/j.cie.2017.05.002.

Gutierrez, R. S., A. O. Solis, and S. Mukhopadhyay. 2008. "Lumpy Demand Forecasting Using Neural Networks." *International Journal of Production Economics* 111 (2): 409–420. https://doi.org/10.1016/j.ijpe.2007.01.007.

Güven, İ., Ö. Uygun, and F. Şimşir. 2021. "Machine Learning Algorithms with Intermittent Demand Forecasting: An Application in Retail Apparel with Plenty of Predictors." *Tekstil ve Konfeksiyon* 31 (2): 99–110. https://doi.org/10.32710/tekstilvekonfeksiyon.809867.

Hafeez, G., I. Khan, S. Jan, I. A. Shah, F. A. Khan, and A. Derhab. 2021. "A Novel Hybrid Load Forecasting Framework with Intelligent Feature Engineering and Optimization Algorithm in Smart Grid." *Applied Energy* 299:117178. https://doi.org/10.1016/j.apenergy.2021.117178.

Han, Y., L. Wang, J. Gao, Z. Xing, and T. Tao. 2017. "Combination Forecasting Based on SVM and Neural Network for Urban Rail Vehicle Spare Parts Demand." In *Proceedings of the 2017 36th Chinese Control Conference (CCC)*, 4660–4665. Dalian, China. https://doi.org/10.23919/ChiCC.2017.8028090.

Hasan, N., N. Ahmed, and S. M. Ali. 2024. "Improving Sporadic Demand Forecasting Using a Modified K-Nearest Neighbor Framework." *Engineering Applications of Artificial Intelligence* 129: 107633. https://doi.org/10.1016/j.engappai.2023.107633.

Hasni, M., M. Z. Babai, and B. Rostami-Tabar. 2024. "A Hybrid LSTM Method for Forecasting Demands of Medical Items in Humanitarian Operations." *International Journal of Production Research* 62 (17): 6046–6063. https://doi.org/10.1080/00207543.2024.2306904.

Hasni, M. S., M. Z. Babai, and Z. Jemai. 2019. "Spare Parts Demand Forecasting: A Review on Bootstrapping Methods." *International Journal of Production Research* 57 (15–16): 4791–4804. https://doi.org/10.1080/00207543.2018.1424375.

Hess, A., S. Spinler, and M. Winkenbach. 2021. "Real-Time Demand Forecasting for an Urban Delivery Platform." *Transportation Research Part E: Logistics and Transportation Review* 145:102147. https://doi.org/10.1016/j.tre.2020.102147.

Hewamalage, H., C. Bergmeir, and K. Bandara. 2021. "Recurrent Neural Networks for Time Series Forecasting: Current Status and Future Directions." *International Journal of Forecasting* 37 (1): 388–427. https://doi.org/10.1016/j.ijforecast.2020.06.008.

Hoffmann, M. A., R. Lasch, and J. Meinig. 2022. "Forecasting Irregular Demand Using Single Hidden Layer Neural Networks." *Logistics Research* 15 (1): 1–13. https://doi.org/10.23773/2022_6.

Hong, K., Y. Ren, F. Li, W. Mao, and X. Gao. 2023. "Robust Interval Prediction of Intermittent Demand for Spare Parts Based on Tensor Optimization." *Sensors* 23 (16): 7182. https://doi.org/10.3390/s23167182.

Hu, Q., J. E. Boylan, H. Chen, and A. Labib. 2018. "OR in Spare Parts Management: A Review." *European Journal of Operational Research* 266 (2): 395–414. https://doi.org/10.1016/j.ejor.2017.07.058.

Hu, X., and M. Tang. 2024. "A Study of LSTM-based Intermittent Demand Forecasting for Enterprises." In *Proceedings of the 2024 8th Asian Conference on Artificial Intelligence Technology (ACAIT)*, 207–210. Fuzhou, China. https://doi.org/10.1109/ACAIT63902.2024.11022316.

Hu, X., X. Zhang, and D. Zhang. 2021. "A Novel Approach of Intermittent Demand Prediction in Industrial Domain." *Journal of Physics: Conference Series* 1756:012009. https://doi.org/10.1088/1742-6596/1756/1/012009.

Hua, Z., and B. Zhang. 2006. "A Hybrid Support Vector Machines and Logistic Regression Approach for Forecasting Intermittent Demand of Spare Parts." *Applied Mathematics and Computation* 181 (2): 1035–1048. https://doi.org/10.1016/j.amc.2006.01.064.

Huang, G., Y. Yang, W. Li, X. Cao, and Z. Yang. 2024. "A Convolutional Neural Network-Back Propagation Based Three-Layer Combined Forecasting Method for Spare Part Demand." *RAIRO – Operations Research* 58 (5): 4181–4195. https://doi.org/10.1051/ro/2024159.

Hyndman, R. J. 2020. "A Brief History of Forecasting Competitions." *International Journal of Forecasting* 36 (1): 7–14. https://doi.org/10.1016/j.ijforecast.2019.03.015.

Hyndman, R. J., and A. B. Koehler. 2006. "Another Look at Measures of Forecast Accuracy." *International Journal of Forecasting* 22 (4): 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

İfraz, M., A. Aktepe, S. Ersöz, and T. Çetinyokuş. 2023. "Demand Forecasting of Spare Parts with Regression and Machine Learning Methods: Application in a Bus Fleet." *Journal of Engineering Research (Kuwait)* 11 (2):100057. https://doi.org/10.1016/j.jer.2023.100057.

In, Y., and J.-Y. Jung. 2022. "Simple Averaging of Direct and Recursive Forecasts via Partial Pooling Using Machine Learning." *International Journal of Forecasting* 38 (4): 1386–1399. https://doi.org/10.1016/j.ijforecast.2021.11.007.

Jafarzadeh, S. G., M. N. A. Rahman, and D. A. Wahab. 2012. "Forecasting Capabilities of Spare Part Production with Artificial Neural Networks Model in a Supply Chain." *World Applied Sciences Journal* 20 (5): 674–678. https://doi.org/10.5829/idosi.wasj.2012.20.05.2413.

Januschowski, T., Y. Wang, K. Torkkola, T. Erkkilä, H. Hasson, and J. Gasthaus. 2022. "Forecasting with Trees." *International Journal of Forecasting* 38 (4): 1473–1481. https://doi.org/10.1016/j.ijforecast.2021.10.004.

Jeon, Y., and S. Seong. 2022. "Robust Recurrent Network Model for Intermittent Time Series Forecasting." *International Journal of Forecasting* 38 (4): 1415–1425. https://doi.org/10.1016/j.ijforecast.2021.07.004.

Jiang, P., Y. Huang, and X. Liu. 2021. "Intermittent Demand Forecasting for Spare Parts in the Heavy-Duty Vehicle Industry: A Support Vector Machine Model." *International Journal of Production Research* 59 (24): 7423–7440. https://doi.org/10.1080/00207543.2020.1842936.

Khan, N. T., and A. Al Hanbali. 2025. "Machine Learning Approaches for Disaggregated and Intermittent Demand Forecasting for Last-Mile Logistics." *Transportation Research Procedia* 84:307–314. https://doi.org/10.1016/j.trpro.2025.03.077.

Kiefer, D., F. Grimm, M. Bauer, and C. van Dinther. 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 (HICSS 2021)*, 1425–1434. Maui, HI, USA. https://www.hdl.handle.net/10125/70784.

Kiefer, D., F. Grimm, and C. van Dinther. 2022. "Artificial Intelligence in Supply Chain Management: Investigation of Transfer Learning to Improve Demand Forecasting of Intermittent Time Series with Deep Learning." In *Proceedings of the 55th Hawaii International Conference on System Sciences (HICSS-55)*, 1–10. Maui, HI, USA (virtual). https://www.aisel.aisnet.org/hicss-55/da/decision_support_for_scm/3/.

Kim, J.-D., J.-H. Hwang, and H.-H. Doh. 2023. "A Predictive Model with Data Scaling Methodologies for Forecasting Spare Parts Demand in Military Logistics." *Defence Science Journal* 73 (6): 666–674. https://doi.org/10.14429/DSJ.73.19129.

Kim, J.-D., T.-H. Kim, and S. W. Han. 2023. "Demand Forecasting of Spare Parts Using Artificial Intelligence: A Case Study of K-X Tanks." *Mathematics* 11 (3): 501. https://doi.org/10.3390/math11030501.

Kim, S., and H. Kim. 2016. "A New Metric of Absolute Percentage Error for Intermittent Demand Forecasts." *International Journal of Forecasting* 32 (3): 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003.

Kolassa, S. 2016. "Evaluating Predictive Count Data Distributions in Retail Sales Forecasting." *International Journal of Forecasting* 32 (3): 788–803. https://doi.org/10.1016/j.ijforecast.2015.12.004.

Kolassa, S. 2020. "Why the 'best' Point Forecast Depends on the Error or Accuracy Measure." *International Journal of Forecasting* 36 (1): 208–211. https://doi.org/10.1016/j.ijforecast.2019.02.017.

Kolassa, S. 2022. "Commentary on the M5 Forecasting Competition." *International Journal of Forecasting* 38 (4): 1562–1568. https://doi.org/10.1016/j.ijforecast.2021.08.006.

Kourentzes, N. 2013. "Intermittent Demand Forecasts with Neural Networks." *International Journal of Production Economics* 143 (1): 198–206. https://doi.org/10.1016/j.ijpe.2013.01.009.

Kourentzes, N., F. Petropoulos, and J. R. Trapero. 2014. "Improving Forecasting by Estimating Time Series Structural Components across Multiple Frequencies." *International Journal of Forecasting* 30 (2): 291–302. https://doi.org/10.1016/j.ijforecast.2013.09.006.

Kouvelis, P., C. Chambers, and H. Wang. 2006. "Supply Chain Management Research and Production and Operations Management: Review, Trends, and Opportunities." *Production and Operations Management* 15 (3): 449–469. https://doi.org/10.1111/j.1937-5956.2006.tb00257.x.

Lainder, A. D., and R. D. Wolfinger. 2022. "Forecasting with Gradient Boosted Trees: Augmentation, Tuning, and Cross-Validation Strategies: Winning Solution to the M5 Uncertainty Competition." *International Journal of Forecasting* 38 (4): 1426–1433. https://doi.org/10.1016/j.ijforecast.2021.12.003.

Lee, E., M. Nam, and H. Lee. 2022. "Tab2vox: CNN-Based Multivariate Multilevel Demand Forecasting Framework by Tabular-to-Voxel Image Conversion." *Sustainability (Switzerland)* 14 (18): 11745. https://doi.org/10.3390/su141811745.

Li, L., Y. Kang, F. Petropoulos, and F. Li. 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.

Li, S. G., and X. Kuo. 2008. "The Inventory Management System for Automobile Spare Parts in a Central Warehouse." *Expert Systems with Applications* 34 (2): 1144–1153. https://doi.org/10.1016/j.eswa.2006.12.003.

Li, X., X. Zhao, and W. Pu. 2019. "Battle Damage-Oriented Spare Parts Forecasting Method Based on Wartime Influencing Factors Analysis and ε-support Vector Regression." *International Journal of Production Research* 58 (4): 1178–1198. https://doi.org/10.1080/00207543.2019.1614691.

Li, Z., Y. Zhang, X. Yan, and Z. Peng. 2015. "A Novel Prediction Model for Aircraft Spare Part Intermittent Demand in Aviation Transportation Logistics Using Multi-components Accumulation and High Resolution Analysis." *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering* 229 (2): 384–395. https://doi.org/10.1177/0954410014531742.

Liu, J., L. Lin, Z. Li, H. Guo, and Y. Lv. 2020. "Spare Aeroengine Demand Prediction Model Based on Deep Croston Method." *Journal of Aerospace Information Systems* 17 (2): 125–133. https://doi.org/10.2514/1.I010707.

Liu, P. 2020. "Intermittent Demand Forecasting for Medical Consumables with short Life Cycle Using a Dynamic Neural Network during the COVID-19 Epidemic." *Health Informatics Journal* 26 (4): 3106–3122. https://doi.org/10.1177/1460458220954730.

Liu, Y., Q. Zhang, Z.-P. Fan, T.-H. You, and L.-X. Wang. 2019. "Maintenance Spare Parts Demand Forecasting for Automobile 4S Shop considering Weather Data." *IEEE Transactions on Fuzzy Systems* 27 (5): 943–955. https://doi.org/10.1109/TFUZZ.2018.2831637.

Lolli, F., R. Gamberini, A. Regattieri, E. Balugani, T. Gatos, and S. Gucci. 2017. "Single-Hidden Layer Neural Networks for Forecasting Intermittent Demand." *International Journal of Production Economics* 183:116–128. https://doi.org/10.1016/j.ijpe.2016.10.021.

Ma, S. 2024. "Retail Store-SKU Level Replenishment Planning with Attribute-Space Graph Recurrent Neural Networks." *Expert Systems with Applications* 249: 123727. https://doi.org/10.1016/j.eswa.2024.123727.

Ma, S., and R. Fildes. 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.

Ma, S., and R. Fildes. 2022. "The Performance of the Global Bottom-up Approach in the M5 Accuracy Competition: A Robustness Check." *International Journal of Forecasting* 38 (4): 1492–1499. https://doi.org/10.1016/j.ijforecast.2021.09.002.

Ma, S., R. Fildes, and T. Huang. 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.

Ma, X., R. Han, Y. Chen, Q. Qiu, R. Yan, and L. Yang. 2024. "Intelligent Spare Ordering and Replacement Optimisation Leveraging Adaptive Prediction Information." *Reliability Engineering and System Safety* 252:110420. https://doi.org/10.1016/j.ress.2024.110420.

Ma, Z., B. Tang, K. Zhang, Y. Huang, D. Cao, J. Luo, and J. Zhang. 2022. "Predicting Spare Parts Inventory of Hydropower Stations and Substations Based on Combined Model." *Mathematical Problems in Engineering* 2022: 1643807. https://doi.org/10.1155/2022/1643807.

Ma, Z., C. Wang, and Z. Zhang. 2021. "Deep Learning Algorithms for Automotive Spare Parts Demand Forecasting." In *Proceedings of the 2021 International Conference on Computer Information Science and Artificial Intelligence (CISAI)*, 358–361. Kunming, China. https://doi.org/10.1109/CISAI54367.2021.00075.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos. 2020. "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods." *International Journal of Forecasting* 36 (1): 54–74. https://doi.org/10.1016/j.ijforecast.2019.04.014.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos. 2022. "M5 Accuracy Competition: Results, Findings, and Conclusions." *International Journal of Forecasting* 38 (4): 1346–1364. https://doi.org/10.1016/j.ijforecast.2021.11.013.

Makridakis, S., E. Spiliotis, V. Assimakopoulos, S. Chen, A. Gaba, I. Tsetlin, and R. L. Winkler. 2022. "The M5 Uncertainty Competition: Results, Findings and Conclusions." *International Journal of Forecasting* 38:1365–1385. https://doi.org/10.1016/j.ijforecast.2021.10.009.

Moher, D., A. Liberati, J. Tetzlaff, and D. G. Altman. The PRISMA Group. 2009. "Preferred Reporting Items for Systematic Reviews and Meta-analyses: The PRISMA Statement." *PLoS Medicine* 6 (7): e1000097. https://doi.org/10.1371/journal.pmed.1000097.

Muhaimin, A., D. D. Prastyo, and H. H.-S. Lu. 2021. "Forecasting with Recurrent Neural Network in Intermittent Demand Data." In *Proceedings of the 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 802–809. Noida, India. https://doi.org/10.1109/Confluence51648.2021.9376880.

Mukhopadhyay, S., A. O. Solis, and R. S. Gutierrez. 2012. "The Accuracy of Non-traditional versus Traditional Methods of Forecasting Lumpy Demand." *Journal of Forecasting* 31 (8): 721–735. https://doi.org/10.1002/for.1242.

Munn, Z., M. D. J. Peters, C. Stern, C. Tufanaru, A. McArthur, and E. Aromataris. 2018. "Systematic Review or Scoping Review? Guidance for Authors When Choosing between a Systematic or Scoping Review Approach." *BMC Medical*

Research Methodology 18 (1): 143. https://doi.org/10.1186/s12874-018-0611-x.

Nasios, I., and K. Vogklis. 2022. "Blending Gradient Boosted Trees and Neural Networks for Point and Probabilistic Forecasting of Hierarchical Time Series." *International Journal of Forecasting* 38 (4): 1448–1459. https://doi.org/10.1016/j.ijforecast.2022.01.001.

Nasiri Pour, A., B. Rostami Tabar, and A. Rahimzadeh. 2009. "A Hybrid Neural Network and Traditional Approach for Forecasting Lumpy Demand." *World Academy of Science, Engineering and Technology* 40:384–390.

Nguyen, D. V. H., T. X. H. Nguyen, and H. G. Nguyen. 2025. "K-Means Clustering and Machine Learning-Based Forecasting Model for the Automotive Spare Parts Industry." In *Information Systems for Intelligent Systems*. ISBM 2024. *Lecture Notes in Networks and Systems*, Vol 1255, edited by A. Iglesias, J. Shin, B. Patel, and A. Joshi. Singapore: Springer. https://doi.org/10.1007/978-981-96-1747-0_39.

Nikolopoulos, K. 2021. "We Need to Talk about Intermittent Demand Forecasting." *European Journal of Operational Research* 291 (2): 549–559. https://doi.org/10.1016/j.ejor.2019.12.046.

Nikolopoulos, K. I., M. Z. Babai, and K. Bozos. 2016. "Forecasting Supply Chain Sporadic Demand with Nearest Neighbor Approaches." *International Journal of Production Economics* 177:139–148. https://doi.org/10.1016/j.ijpe.2016.04.013.

Oliveira, J. M., and P. Ramos. 2023. "Investigating the Accuracy of Autoregressive Recurrent Networks Using Hierarchical Aggregation Structure-Based Data Partitioning." *Big Data and Cognitive Computing* 7 (2): 100. https://doi.org/10.3390/bdcc7020100.

Pawlikowski, M., and A. Chorowska. 2020. "Weighted Ensemble of Statistical Models." *International Journal of Forecasting* 36 (1): 93–97. https://doi.org/10.1016/j.ijforecast.2019.03.019.

Petropoulos, F., D. Apiletti, V. Assimakopoulos, M. Z. Babai, D. K. Barrow, S. Ben Taieb, C. Bergmeir, et al. 2022. "Forecasting: Theory and Practice." *International Journal of Forecasting* 38 (3): 705–871. https://doi.org/10.1016/j.ijforecast.2021.11.001.

Petropoulos, F., N. Kourentzes, and K. Nikolopoulos. 2016. "Another Look at Estimators for Intermittent Demand." *International Journal of Production Economics* 181:154–161. https://doi.org/10.1016/j.ijpe.2016.04.017.

Pinçe, Ç., L. Turrini, and J. Meissner. 2021. "Intermittent Demand Forecasting for Spare Parts: A Critical Review." *Omega* 105:102513. https://doi.org/10.1016/j.omega.2021.102513.

Pogačnik, B., J. Duhovnik, and J. Tavčar. 2017. "Aircraft Fault Forecasting at Maintenance Service on the Basis of Historic Data and Aircraft Parameters." *Eksploatacja i Niezawodnosc* 19 (4): 624–633. https://doi.org/10.17531/ein.2017.4.17.

Praveena, S., and S. Prasanna Devi. 2022. "A Hybrid Demand Forecasting for Intermittent Demand Patterns Using Machine Learning Techniques." In *Proceedings of the 2022 1st International Conference on Computational Science and Technology (ICCST)*, 557–561. Chennai, India. https://doi.org/10.1109/ICCST55948.2022.10040407.

Prestwich, S., R. Rossi, S. A. Tarim, and B. Hnich. 2014. "Mean-Based Error Measures for Intermittent Demand Forecasting." *International Journal of Production Research* 52 (22): 6782–6791. https://doi.org/10.1080/00207543.2014.917771.

Procopio, A., G. Cesarelli, L. Donisi, A. Merola, F. Amato, and C. Cosentino. 2023. "Combined Mechanistic Modeling and Machine-Learning Approaches in Systems Biology – A Systematic Literature Review." *Computer Methods and Programs in Biomedicine* 240:107681. https://doi.org/10.1016/j.cmpb.2023.107681.

Quiñones-Rivera, H., O. Rubiano-Ovalle, and W. Alfonso-Morales. 2023. "Demand Forecasting Using a Hybrid Model Based on Artificial Neural Networks: A Study Case on Electrical Products." *Journal of Industrial Engineering and Management* 16 (2): 363–381. https://doi.org/10.3926/jiem.3928.

Rosienkiewicz, M. 2020. "Accuracy Assessment of Artificial Intelligence-Based Hybrid Models for Spare Parts Demand Forecasting in Mining Industry." In *Information Systems Architecture and Technology: Proceedings of 40th Anniversary International Conference on Information Systems Architecture and Technology – ISAT 2019. Advances in Intelligent Systems and Computing*, Vol 1052, edited by Z. Wilimowska, L. Borzemski, and J. Świątek. Cham: Springer. https://doi.org/10.1007/978-3-030-30443-0_16.

Rosienkiewicz, M., E. Chlebus, and J. Detyna. 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.

Rožanec, J. M., B. Fortuna, and D. Mladenić. 2022. "Reframing Demand Forecasting: A Two-Fold Approach for Lumpy and Intermittent Demand." *Sustainability (Switzerland)* 14 (15): 9295. https://doi.org/10.3390/su14159295.

Salinas, D., V. Flunkert, J. Gasthaus, and T. Januschowski. 2020. "DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks." *International Journal of Forecasting* 36:1181–1191. https://doi.org/10.1016/j.ijforecast.2019.07.001.

Sareminia, S. 2023. "A Support Vector Based Hybrid Forecasting Model for Chaotic Time Series: Spare Part Consumption Prediction." *Neural Processing Letters* 55 (3): 2825–2841. https://doi.org/10.1007/s11063-022-10986-4.

Seaman, B., and J. Bowman. 2022. "Applicability of the M5 to Forecasting at Walmart." *International Journal of Forecasting* 38 (4): 1468–1472. https://doi.org/10.1016/j.ijforecast.2021.06.002.

Semenoglou, A.-A., E. Spiliotis, and V. Assimakopoulos. 2023. "Data Augmentation for Univariate Time Series Forecasting with Neural Networks." *Pattern Recognition* 134:109132. https://doi.org/10.1016/j.patcog.2022.109132.

Semenoglou, A.-A., E. Spiliotis, S. Makridakis, and V. Assimakopoulos. 2021. "Investigating the Accuracy of Cross-Learning Time Series Forecasting Methods." *International Journal of Forecasting* 37 (3): 1072–1084. https://doi.org/10.1016/j.ijforecast.2020.11.009.

Shafi, I., A. Sohail, J. Ahmad, J. C. Martínez Espinosa, L. A. Dzul López, E. B. Thompson, and I. Ashraf. 2023. "Spare Parts Forecasting and Lumpiness Classification Using Neural Network Model and Its Impact on Aviation Safety." *Applied Sciences (Switzerland)* 13 (9): 5475. https://doi.org/10.3390/app13095475.

Shaub, D. 2020. "Fast and Accurate Yearly Time Series Forecasting with Forecast Combinations." *International Journal of Forecasting* 36 (1): 116–120. https://doi.org/10.1016/j.ijforecast.2019.03.032.

Sillanpää, V., and J. Liesiö. 2018. "Forecasting Replenishment Orders in Retail: Value of Modelling Low and Intermittent Consumer Demand with Distributions." *International Journal of Production Research* 56 (12): 4168–4185. https://doi.org/10.1080/00207543.2018.1431413.

Smyl, S. 2020. "A Hybrid Method of Exponential Smoothing and Recurrent Neural Networks for Time Series Forecasting." *International Journal of Forecasting* 36 (1): 75–85. https://doi.org/10.1016/j.ijforecast.2019.03.017.

Snyder, H. 2019. "Literature Review as a Research Methodology: An Overview and Guidelines." *Journal of Business Research* 104:333–339. https://doi.org/10.1016/j.jbusres.2019.07.039.

Sobral, P., R. Teixeira, R. Marques, N. Figueiredo, M. Antunes, and D. Gomes. 2024. "Improving Automotive Aftermarket Forecasting with MLOps." In *Proceedings of the 2024 11th International Conference on Future Internet of Things and Cloud (FiCloud)*, 382–389. Vienna, Austria. https://doi.org/10.1109/FiCloud62933.2024.00065.

Solis, A. O. 2015. "Better Statistical Forecast Accuracy Does Not Always Lead to Better Inventory Control Efficiency: The Case of Lumpy Demand." In *Proceedings of the International Conference on Modeling and Applied Simulation (MAS 2015)*, edited by F. Bruzzone, A. De Felice, C. Frydman, M. Massei, I. Merkuryev, and A. O. Solis, 211–217. Bergeggi, Italy.

Solis, A. O., S. Mukhopadhyay, and R. S. Gutierrez. 2010. "Inventory Control Performance of Various Forecasting Methods When Demand Is Lumpy." In *Proceedings of the 9th International Conference on Modeling and Applied Simulation (MAS 2010)*, 1–10. Fes, Morocco.

Song, W., J. Wu, J. Kang, and J. Zhang. 2021. "Research on Maintenance Spare Parts Requirement Prediction Based on LSTM Recurrent Neural Network." *Open Physics* 19 (1): 618–627. https://doi.org/10.1515/phys-2021-0072.

Spiliotis, E., A. Kouloumos, V. Assimakopoulos, and S. Makridakis. 2020. "Are Forecasting Competitions Data Representative of the Reality?" *International Journal of Forecasting* 36 (1): 37–53. https://doi.org/10.1016/j.ijforecast.2018.12.007.

Spiliotis, E., S. Makridakis, A. Kaltsounis, and V. Assimakopoulos. 2021. "Product Sales Probabilistic Forecasting: An Empirical Evaluation Using the M5 Competition Data." *International Journal of Production Economics* 240:108237. https://doi.org/10.1016/j.ijpe.2021.108237.

Spiliotis, E., S. Makridakis, A.-A. Semenoglou, and V. Assimakopoulos. 2022. "Comparison of Statistical and Machine Learning Methods for Daily SKU Demand Forecasting." *Operational Research* 22 (3): 3037–3061. https://doi.org/10.1007/s12351-020-00605-2.

Sun, Q., X. Feng, S. Zhao, H. Cao, S. Li, and Y. Yao. 2021. "Deep Learning Based Customer Preferences Analysis in Industry 4.0 Environment." *Mobile Networks and Applications* 26 (6): 2329–2340. https://doi.org/10.1007/s11036-021-01830-5.

Syntetos, A. A., Z. Babai, J. E. Boylan, S. Kolassa, and K. Nikolopoulos. 2016. "Supply Chain Forecasting: Theory, Practice, Their Gap and the Future." *European Journal of Operational Research* 252 (1): 1–26. https://doi.org/10.1016/j.ejor.2015.11.010.

Syntetos, A. A., and J. E. Boylan. 2005. "The Accuracy of Intermittent Demand Estimates." *International Journal of Forecasting* 21 (2): 303–314. https://doi.org/10.1016/j.ijforecast.2004.10.001.

Syntetos, A. A., J. E. Boylan, and J. D. Croston. 2005. "On the Categorization of Demand Patterns." *Journal of the Operational Research Society* 56 (5): 495–503. https://doi.org/10.1057/palgrave.jors.2601841.

Taieb, S. B., and R. J. Hyndman. 2014. "A Gradient Boosting Approach to the Kaggle Load Forecasting Competition." *International Journal of Forecasting* 30 (2): 382–394. https://doi.org/10.1016/j.ijforecast.2013.07.005.

Theodorou, E., S. Wang, Y. Kang, E. Spiliotis, S. Makridakis, and V. Assimakopoulos. 2022. "Exploring the Representativeness of the M5 Competition Data." *International Journal of Forecasting* 38 (4): 1500–1506. https://doi.org/10.1016/j.ijforecast.2021.07.006.

Tsao, Y.-C., I. N. Pujawan, N. Kurniati, and A. M. A. Yaqin. 2019. "Spare Parts Demand Forecasting in Energy Industry: A Stacked Generalization-Based Approach." In *Proceedings of the 2019 International Conference on Management Science and Industrial Engineering (MSIE 2019)*, 163–167. Phuket, Thailand. https://doi.org/10.1145/3335550.3335573.

Türkmen, A. C., T. Januschowski, Y. Wang, and A. T. Cemgil. 2021. "Forecasting Intermittent and Sparse Time Series: A Unified Probabilistic Framework via Deep Renewal Processes." *PLoS One* 16 (11): e0259764. https://doi.org/10.1371/journal.pone.0259764.

Ulrich, M., H. Jahnke, R. Langrock, R. Pesch, and R. Senge. 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.

Vaitkus, V., G. Zylius, and R. Maskeliunas. 2014. "Electrical Spare Parts Demand Forecasting." *Elektronika ir Elektrotechnika* 20 (10): 7–10. https://doi.org/10.5755/j01.eee.20.10.8870.

Van der Auweraer, S., R. N. Boute, and A. A. Syntetos. 2019. "Forecasting Spare Part Demand with Installed Base Information: A Review." *International Journal of Forecasting* 35 (1): 181–196. https://doi.org/10.1016/j.ijforecast.2018.09.002.

Wakle, S. P., V. P. Toshniwal, R. Jain, G. Soni, and B. Ramtiyal. 2024. "A Data-Driven Approach for Planning Stock Keeping Unit (SKU) in a Steel Supply Chain." *International Journal of Mathematical, Engineering and Management Sciences* 9 (2): 283–304. https://doi.org/10.33889/IJMEMS.2024.9.2.015.

Wang, J., W. K. Chong, J. Lin, and C. P. T. Hedenstierna. 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.

Wang, J., X. Lai, S. Zhang, W. M. Wang, and J. Chen. 2020. "Predicting Customer Absence for Automobile 4S Shops: A Lifecycle Perspective." *Engineering Applications of Artificial Intelligence* 89: 103405. https://doi.org/10.1016/j.engappai.2019.103405.

Wang, S., Y. Kang, and F. Petropoulos. 2024. "Combining Probabilistic Forecasts of Intermittent Demand." *European Journal of Operational Research* 315 (3): 1038–1048. https://doi.org/10.1016/j.ejor.2024.01.032.

Wellens, A. P., M. Udenio, and R. N. Boute. 2022. "Transfer Learning for Hierarchical Forecasting: Reducing Computational Efforts of M5 Winning Methods." *International Journal of Forecasting* 38 (4): 1482–1491. https://doi.org/10.1016/j.ijforecast.2021.09.011.

Widiarta, H., S. Viswanathan, and R. Piplani. 2009. "Forecasting Aggregate Demand: An Analytical Evaluation of Top-down versus Bottom-up Forecasting in a Production Planning Framework." *International Journal of Production Economics* 118 (1): 87–94. https://doi.org/10.1016/j.ijpe.2008.08.013.

Wu, C.-g., X. Fu, and Y. Xia. 2025. "Spare Part Demand Forecasting Using PSO Trained Quantile Regression Neural Network." *Computers & Industrial Engineering* 200:110841. https://doi.org/10.1016/j.cie.2024.110841.

Xing, R., and X. Shi. 2019. "A BP–SVM Combined Model for Intermittent Spare Parts Demand Prediction." In *Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 1085–1090. Bari, Italy. https://doi.org/10.1109/SMC.2019.8914609.

Xiong, P., L. Zhang, M. Ruan, Z. Zhang, and Y. Li. 2025. "Agent-Based Collaborative Model for Forecasting Large-Scale Intermittent Spare Parts in Smart Manufacturing Industry." *Computers & Industrial Engineering* 209:111479. https://doi.org/10.1016/j.cie.2025.111479.

Yang, L., and A. Shami. 2020. "On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice." *Neurocomputing* 415:295–316. https://doi.org/10.1016/j.neucom.2020.07.061.

Yasir, M., Y. Ansari, K. Latif, H. Maqsood, A. Habib, J. Moon, and S. Rho. 2024. "Machine Learning–assisted Efficient Demand Forecasting Using Endogenous and Exogenous Indicators for the Textile Industry." *International Journal of Logistics Research and Applications* 27 (12): 2867–2886. https://doi.org/10.1080/13675567.2022.2100334.

Zhang, G. P., Y. Xia, and M. Xie. 2024. "Intermittent Demand Forecasting with Transformer Neural Networks." *Annals of Operations Research* 339 (1): 1051–1072. https://doi.org/10.1007/s10479-023-05447-7.

Zhuang, X., Y. Yu, and A. Chen. 2022. "A Combined Forecasting Method for Intermittent Demand Using the Automotive Aftermarket Data." *Data Science and Management* 5 (2): 43–56. https://doi.org/10.1016/j.dsm.2022.04.001.

Zohdi, M., M. Rafiee, V. Kayvanfar, and A. Salamiraad. 2022. "Demand Forecasting Based Machine Learning Algorithms on Customer Information: An Applied Approach." *International Journal of Information Technology (Singapore)* 14 (4): 1937–1947. https://doi.org/10.1007/s41870-022-00875-3.

Zuvieta, C. M., J. L. Leevy, and T. M. Khoshgoftaar. 2025. "A Survey on Statistical and ML-Based Demand Forecasting Methods for Spare Parts in Aviation." *IEEE Access* 13:44800–44816. https://doi.org/10.1109/ACCESS.2025.3550091.

References

Hobor, L., M. Brcic, L. Polutnik, and A. Kapetanovic. 2025. "Comparative Analysis of Modern Machine Learning Models for Retail Sales Forecasting." arXiv preprint. https://arxiv.org/abs/2506.05941.

Jeong, H., and S. Lee. 2025. "Probabilistic–Robust Loss with Uncertainty-Aware Learning for Diverse Demand Forecasting Patterns." SSRN Preprint. https://ssrn.com/abstract=5370683.

Svetunkov, I., and A. Sroginis. 2025. "Why Do Zeroes Happen? A Model-Based Approach for Demand Classification." arXiv preprint. https://arxiv.org/abs/2504.05894.

Tsantilis, I., P. G. Giannopoulos, T. K. Dasaklis, and C. Patsakis. 2025. "An Analytics-Driven Approach to Sporadic Demand Forecasting Using Reservoir Computing." SSRN preprint. https://ssrn.com/abstract=5209158.