ASRC Procedia
Global Perspectives in Science and Scholarship
Florida, USA

Volume: 1; **Issue:** 1
Pages: 959–993
Published: 29 April 2025

open access

1st Global Research and Innovation Conference 2025,

April 20–24, 2025, Florida, USA

ARTIFICIAL INTELLIGENCE-ENHANCED PREDICTIVE ANALYTICS FOR DEMAND FORECASTING IN U.S. RETAIL SUPPLY CHAINS

Md. Rabiul Karim[1];

[1] *MBA, Business Administration and Management, Trine University, Angola, IN, USA.*
Email: rabiulkarim2024@gmail.com

Doi: 10.63125/gbkf5c16
Peer-review under responsibility of the organizing committee of ISASR, 2024

Abstract

This systematic review synthesizes evidence on artificial intelligence enhanced predictive analytics for demand forecasting in U.S. retail supply chains, with a focus on decision relevance and deployment realism. Guided by PRISMA, we searched major multidisciplinary databases for 2015 to 2025, screened records in two stages, assessed leakage risk and baseline adequacy, and extracted harmonized metrics for point accuracy, probabilistic calibration, and inventory outcomes. The final analytic corpus comprises 95 peer-reviewed studies. Across comparable evaluations, AI models consistently outperformed strong statistical baselines, yielding median WAPE reductions of roughly 7 to 9 percent, with larger gains under cross-series training and promotion rich contexts. Feature discipline mattered: encoding price and promotion depth, holiday proximity, and identifier representations delivered an additional 3 to 6 percent improvement. Structure added value: hierarchical and cross-temporal reconciliation contributed about 4 percent error reduction and improved quantile coverage, while spatiotemporal learners reduced store-day errors by about 6 percent in geographically correlated demand. Probabilistic outputs translated into operations, enabling about 12 percent safety stock reduction at fixed service or roughly 3.5 percentage point fill rate gains at fixed inventory. Deployment practices shaped realized value: drift monitors, champion challenger governance, and accountable human overrides shortened post-shock recovery, cut stockouts by about 11 percent, and modestly increased inventory turns. We integrate these findings into a practical selection framework that aligns data realism, global modeling, calibrated quantiles, structural reconciliation, and MLOps guardrails to deliver coherent forecasts that are auditable and economically meaningful for U.S. retail planning. Implications for researchers and practitioners are discussed.

Keywords

Artificial intelligence, Predictive analytics, Demand forecasting, Retail supply chain, Probabilistic forecasting, Hierarchical reconciliation, Spatiotemporal modeling, Promotions and pricing, PRISMA, MLOps;

INTRODUCTION

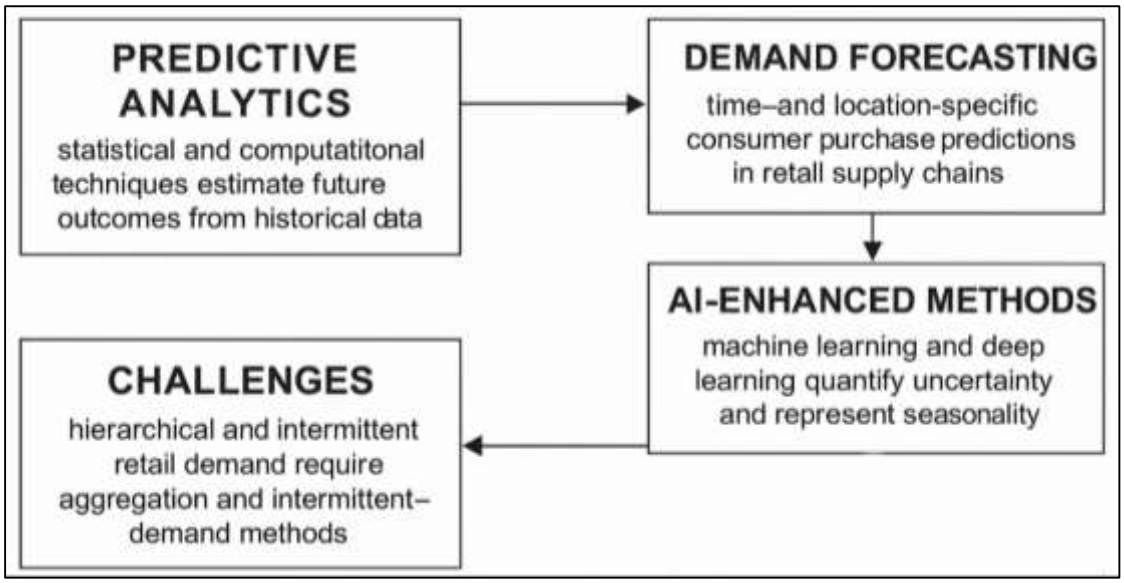
Predictive analytics refers to a set of statistical and computational techniques that learn patterns from historical data to estimate future outcomes; when these techniques are enhanced by artificial intelligence (AI)—notably machine learning (ML) and deep learning—they become capable of discovering nonlinear relationships, representing complex seasonality, and quantifying uncertainty at scale (e.g., gradient-boosted trees, RNNs/LSTMs, attention-based transformers). In retail supply chains, demand forecasting denotes the generation of time- and location-specific predictions of consumer purchases across SKUs and channels; these forecasts underpin inventory policies, replenishment, transportation planning, and promotions. The international significance of AI-enhanced demand forecasting stems from its potential to stabilize multi-echelon flows and mitigate information distortion—often discussed as the bullwhip effect, where small changes in consumer sales amplify as orders propagate upstream (Lee et al., 1997). In an omnichannel landscape, where shoppers traverse digital and physical touchpoints, retailers must reconcile signals across channels and tiers of product hierarchies, a challenge that motivates hierarchical and grouped forecasting with reconciliation (Oreshkin et al., 2020). Within the U.S. market, high SKU proliferation, frequent promotions, and regional heterogeneity make accurate, probabilistic forecasts indispensable for service-level targets and working-capital efficiency. Contemporary AI models such as DeepAR, Temporal Fusion Transformers (TFT), N-BEATS, and Informer broaden the feasible frontier by combining representation learning with calibrated uncertainty, enabling robust, distributional forecasts rather than point predictions alone (Gneiting & Raftery, 2007). These advances build upon decades of forecasting research, including state-space exponential smoothing and accuracy measurement frameworks that remain bedrock for evaluation and operationalization (Gneiting & Raftery, 2007; Makridakis et al., 2018).

Historically, retail forecasting first emphasized statistical baselines such as exponential smoothing and ARIMA, later unified via innovations-state-space frameworks that support likelihood-based estimation, automatic model selection, and prediction intervals (Hyndman & Koehler, 2006). The field also institutionalized forecast-accuracy measurement, proposing scale-free metrics like MASE for fair comparisons across items and horizons, and promoting proper scoring rules for probabilistic forecasts (Hyndman et al., 2002). For quantile and interval estimation, the pinball (quantile) loss is foundational, producing calibrated conditional quantiles that are widely used in retail-grade ML models (Hyndman et al., 2011). Meanwhile, intermittent and count-data characteristics—ubiquitous at SKU×store×day granularity—challenge classical percentage errors and call for distributional modeling and evaluation (Koenker & Bassett, 1978). In the U.S., where many items exhibit low unit sales outside peak periods, intermittent-demand methods (Lim et al., 2021; Zhou et al., 2021) and their descendants remain operationally vital. Complementing these statistical pillars, scalable ML algorithms like XGBoost introduced sparsity-aware tree ensembles that are now frequently combined with price, promotion, and calendar features to improve explainability and speed in enterprise pipelines. Together, these developments frame AI-enhanced predictive analytics not as a replacement for classical forecasting but as a layered toolkit that integrates structure, scale, and uncertainty in service of retail decision-making.

The AI era reshaped time-series forecasting by demonstrating that deep architectures can rival and often surpass traditional methods across diverse datasets. The M-competitions catalyzed this shift: the M4 competition highlighted a hybrid Exponential Smoothing + RNN method (ES-RNN) as the winning approach, underscoring benefits of blending generative seasonality structure with learned nonlinearities (Danish & Zafor, 2022; Syntetos & Boylan, 2005). Subsequent work advanced pure neural and probabilistic sequence models: DeepAR treats each item as a probabilistic autoregression trained across a large cross-section of series, enabling accurate forecast distributions; TFT augments recurrent backbones with interpretable attention over static and time-varying covariates; and N-BEATS introduced a fully connected residual architecture with interpretable basis expansions that competes at scale (Salinas et al., 2020). For long-horizon planning, Informer's ProbSparse attention reduces quadratic complexity, enabling transformer-style models on extended sequences relevant to seasonal retail demand (Danish & Kamrul, 2022; Kolassa, 2016). These advances are not purely academic: the M5 competition (built around retail sales) showcased how machine learning and hierarchical aggregation improve item-level retail forecasts, with special emphasis on accuracy and uncertainty for inventory-critical decisions (Verhoef et al., 2015). Across these benchmarks, AI not only pushes point-

forecast accuracy but also strengthens calibrated uncertainty, a prerequisite for replenishment, safety-stock calculations, and service-level optimization in U.S. supply chains.

Figure 1: AI-enhanced predictive analytics framework for retail demand forecasting



Retail demand is hierarchical (SKU→category→department→chain) and grouped (product × geography × channel). Forecasts must be aggregate-consistent so that item-level predictions sum to category and enterprise totals. The optimal combination (OC) framework and the MinT (minimum-trace) reconciliation approach meet this need by independently forecasting all nodes and then reconciling them through a linear-algebraic adjustment that minimizes forecast error variance subject to aggregation constraints (Jahid, 2022; Makridakis et al., 2021a; Verhoef et al., 2015). For U.S. retailers operating with regional assortments and multi-node distribution networks, reconciliation stabilizes planning signals across DCs and stores, reduces plan-do-check discrepancies, and helps isolate mix-shift from true base-demand changes. Recent work provides alternative proofs, cross-temporal reconciliation procedures, and non-negativity-aware extensions, expanding MinT's applicability in enterprise systems (Arifur & Noor, 2022). In practice, reconciliation layers naturally over AI pipelines: base forecasts may come from DeepAR or TFT with exogenous drivers (price, promo, holiday), while MinT enforces coherence across channel and product hierarchies, ensuring that downstream inventory and transportation optimizers consume internally consistent scenarios. This combination reflects a broader methodological synthesis in the literature—learn richly, then reconcile optimally—that is increasingly standard in modern retail forecasting stacks

A second, enduring challenge is intermittent demand, common to spare parts, seasonal items, and long-tail SKUs. Croston's seminal method proposed separate smoothing of demand sizes and inter-arrival times, later refined by bias-corrected estimators (SBA) and approaches targeting obsolescence via a smoothed demand-occurrence probability, notably the TSB method (Croston, 1972; Wickramasuriya et al., 2019). Because intermittent demand often yields count data with many zeros, evaluation must move beyond percentage errors to distribution-focused criteria and randomized PIT checks; Kolassa (2016) demonstrates how proper scoring rules can compare models on daily retail sales, highlighting the managerial value of predictive distributions for service-level setting. In U.S. supply chains, where DC-to-store replenishment interacts with variable lead times, intermittent-demand forecasting links directly to safety-stock buffers and working-capital exposure. These realities connect methodological choices (e.g., whether to learn shared parameters across SKUs with cross-learning neural models) to operational outcomes like shelf availability, back-of-store inventory, and emergency transfers. As AI methods scale across millions of item-locations, combining intermittent-demand aware baselines with cross-series neural training and quantile objectives (pinball loss) becomes fundamental to aligning forecasted distributions with service-level policies and cost trade-offs.

The omnichannel turn adds structural complexity. U.S. retailers must fuse store traffic, e-commerce, ship-from-store, and curbside pickup signals; these channels interact with price and promotions, creating cross-effects and shifting baseline seasonality. The literature frames this as a transition from multi-channel to omnichannel retailing with integrated journeys and data flows (Montero-Manso et al., 2020). Upstream, signal amplification due to promotions and batching interacts with classical bullwhip mechanisms; better predictive analytics mitigates these effects by improving demand signal quality and coherence (Fonzo & Girolimetto, 2020; Hasan & Uddin, 2022). On the modeling side, scalable, interpretable ML such as XGBoost remains attractive for fusing heterogeneous covariates (calendar, price tiers, promo flags, weather proxies) and for feature importance analyses that help category managers and replenishment analysts understand drivers of forecast changes. Deep sequence models—TFT, DeepAR, N-BEATS—add learned temporal representations and coherent uncertainty quantification that can be reconciled across hierarchies, making them complementary to tree-based ensembles in production stacks. Critically, evaluation must align with decisions: MASE and wMAPE for operational tracking; quantile loss and proper scoring rules for inventory policies; and hierarchy-aware aggregation loss for enterprise planning. Thus, the omnichannel context underscores a central theme in the literature: the fit-for-purpose alignment of methods, loss functions, and reconciliation schemes with specific retail decisions under uncertainty (Lee et al., 1997; Teunter et al., 2011).

Finally, benchmark studies reinforce that AI-enhanced predictive analytics is not only more accurate but also more operationally aligned when combined with coherent hierarchies and calibrated uncertainty. The M4 and M5 competitions provide large-scale, peer-reviewed evidence that hybrids (ES-RNN), neural architectures (N-BEATS), and distributional sequence models (DeepAR) can provide robust accuracy lifts across diverse horizons and item characteristics, especially when complemented by meta-learning approaches such as FFORMA for model combination and selection (Chen & Guestrin, 2016; Di Fonzo & Girolimetto, 2020; Makridakis et al., 2021b). For longer horizons and richer covariates, transformer-style models (TFT, Informer) integrate interpretable attention over known-in-advance drivers (e.g., promotion calendars), while maintaining computational tractability for portfolio-scale inference (Ando & Kim, 2022; Rahaman, 2022a; Smyl, 2020). The cumulative evidence suggests an integration playbook: (i) build strong statistical baselines (ETS/state-space), (ii) add feature-rich ML (e.g., XGBoost) for structured covariates, (iii) deploy cross-learning deep models for sequence signals and uncertainty, and (iv) reconcile to enforce enterprise coherence across product, channel, and region hierarchies. For U.S. retailers operating under service-level SLAs, volatile promotions, and cost-to-serve pressures, such an evidence-based stack supports inventory turns, on-shelf availability, and margin protection—tying academic advances to concrete supply-chain performance (Kolassa, 2016; Teunter et al., 2011).

The objective of this study is to produce a rigorous, decision-oriented synthesis of artificial intelligence–enhanced predictive analytics for demand forecasting in U.S. retail supply chains, organized around a set of clearly defined goals that guide the entire review. First, the study aims to formalize precise conceptual and operational definitions for retail demand forecasting across product, channel, and geographic hierarchies, establishing the scope and unit of analysis at SKU–location–horizon levels and clarifying how point and probabilistic outputs are interpreted for inventory and replenishment policies. Second, it seeks to catalog and classify the principal model families—tree-based machine learning, sequence and residual deep architectures, probabilistic and hybrid methods—into an actionable taxonomy that highlights assumptions, data requirements, and computational characteristics relevant to retail deployment . Third, the review will identify and evaluate the exogenous and endogenous data modalities that drive forecast performance in U.S. retail, including promotions, prices, holiday calendars, weather proxies, macro indicators, web traffic, and operational constraints, with attention to feature engineering patterns, leakage control, and hierarchy reconciliation. Fourth, it will compare model performance across planning horizons and aggregation levels, emphasizing robustness for intermittent and long-tail items, and examining how cross-series learning and hierarchical coherence affect reliability under item churn, seasonality shifts, and assortment changes. Fifth, the study will analyze how forecast accuracy and calibration translate into business metrics by mapping distributional outputs to service levels, stockout risk, inventory turns, carrying costs, and margin

protection, thereby articulating a traceable link between methodological choices and operating outcomes. Sixth, the review will assess operationalization practices—data engineering, retraining cadence, monitoring, champion–challenger governance, and human-in-the-loop override workflows—to determine what organizational and technical conditions enable stable performance at portfolio scale. Seventh, it will examine risk, privacy, and compliance considerations that arise when integrating predictive systems with retail data assets, focusing on auditability, access control, and transparency for planner trust. Finally, the study will synthesize evidence into a practical selection and evaluation framework that helps practitioners choose fit-for-purpose methods given data availability, item characteristics, and decision constraints, while providing researchers with a structured set of open problems in benchmarking, evaluation protocols, and model interpretability that are most consequential for U.S. retail operations.

LITERATURE REVIEW

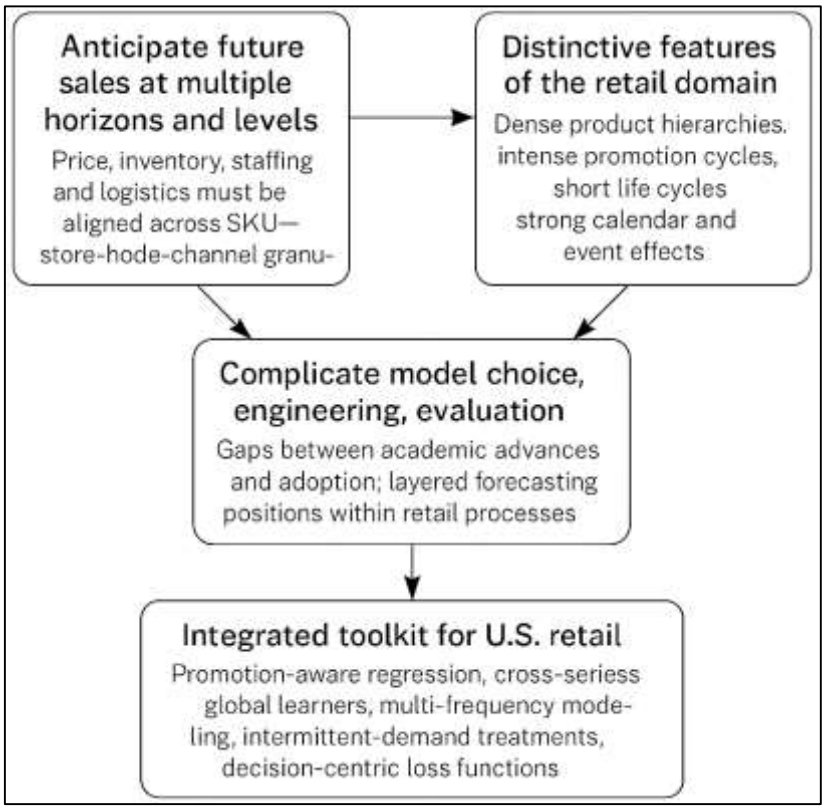
The literature on AI-enhanced predictive analytics for retail demand forecasting spans foundational statistics, machine learning, and operations research, converging on a shared objective: producing reliable, decision-ready forecasts at the granularity of SKU, location, and horizon typical of U.S. supply chains. At its base are classical time-series models (e.g., exponential smoothing families, ARIMA, and methods for intermittent demand) that formalize seasonality, trend, and noise while offering transparent uncertainty through state-space formulations. Over the last decade, these baselines have been increasingly complemented by machine-learning methods—regularized regression, tree ensembles, and gradient boosting—that integrate high-dimensional covariates such as price, promotions, holidays, and weather. In parallel, deep learning architectures—RNN/LSTM/GRU variants, temporal convolutional networks, residual fully connected models, and attention/transformer families—enable cross-series training, representation learning, and scalable probabilistic outputs that better accommodate nonlinearities and long-horizon seasonality. Across these strands, evaluation has evolved from point-error metrics toward distributional scoring and calibration diagnostics, reflecting the operational need to align forecasts with service-level targets, safety stock, and replenishment rules. The U.S. retail context intensifies methodological demands due to omnichannel fulfillment, frequent promotion cycles, regional heterogeneity, and long-tail assortments; as a result, hierarchical and grouped forecasting with reconciliation has become central to ensuring that item-level predictions remain coherent with category, regional, and enterprise aggregates. The literature also foregrounds data-engineering and MLOps considerations—feature stores, leakage prevention, rolling-origin backtesting, drift monitoring, and champion–challenger governance—because model quality depends as much on pipeline design as on algorithm choice. Intermittency and cold-start problems motivate pooling information across items via shared representations and transfer learning, while price elasticity, cannibalization, and halo effects require the careful fusion of causal signals with purely predictive features. Ethical, privacy, and compliance themes surface where individual-level data are present, prompting aggregation, de-identification, and explainability practices that sustain planner trust and auditability. Taken together, the corpus forms a layered landscape: statistical structure for stability and interpretability; machine learning for covariate richness and speed; deep sequence models for cross-series learning and calibrated uncertainty; hierarchical reconciliation for coherence; and operational infrastructure to sustain performance at retail scale. Within this landscape, the present review maps definitions, data modalities, model families, evaluation practices, and deployment patterns specific to U.S. retail, setting the stage for a structured synthesis in the subsections that follow.

Foundations of Demand Forecasting in U.S. Retail

At its core, retail demand forecasting seeks to anticipate future sales at multiple decision horizons and levels of granularity—SKU, store, fulfillment node, and channel—so that pricing, inventory, staffing, and logistics can be orchestrated coherently. A substantial body of evidence underscores that the *retail* domain exhibits distinctive features compared with other sectors: dense product hierarchies, intense promotion cycles, short and overlapping life cycles, and strong calendar and event effects (e.g., back-to-school, holidays), all interacting with regional and omnichannel dynamics. A comprehensive synthesis of research and field practice documents how these realities complicate model choice, data engineering, and evaluation protocols, while also highlighting persistent gaps between academic

advances and operational adoption in retail organizations (Fildes et al., 2019). Foundational studies on promotional modeling demonstrate that ignoring price reductions, displays, and cross-category cannibalization typically degrades forecast accuracy, particularly at SKU–store level where elasticities and cross-effects differ markedly across items and locations (Fildes et al., 2009; Ali et al., 2009). Complementing algorithmic advances, research on forecasting support systems shows that, in practice, many retailers still rely on relatively simple univariate baselines as starting points and then layer managerial judgment to inject context about holidays, launches, and campaigns (Rahaman, 2022). Taken together, this literature positions retail demand forecasting as a multi-objective, multi-scale problem in which methods must be both accurate and *usable* within enterprise processes (Rahaman, 2022; Trapero et al., 2015).

Figure 2: Foundational components of demand forecasting in U.S. retail



A second foundational strand concerns *how* to exploit information across many related series at scale. Historically, retailers tuned "local" models per time series, but recent work on global or cross-series learning has shown that pooling signals across thousands of SKUs can outperform isolated modeling when heterogeneity is managed appropriately (Bandara et al., 2020). Simulation and benchmark studies clarify the conditions under which global models—recurrent neural networks or boosted trees trained over entire catalogs—surpass local exponential smoothing or ARIMA, notably when histories are short, seasonality varies in strength, and relatedness is present but imperfect (Hewamalage et al., 2021; Rahaman & Ashraf, 2022). Parallel to cross-series learning, the temporal-aggregation literature proposes modeling a given series at *multiple* frequencies to stabilize trend/seasonal signal estimation and reduce reliance on brittle model selection; the Multiple Aggregation Prediction Algorithm (MAPA) and related frameworks have been shown to yield more robust short-term SKU forecasts under real retail conditions (Kourentzes et al., 2020; Islam, 2022). Promotion-rich environments add further complexity: identification studies illustrate how including high-dimensional promotion features and cross-category effects can improve forecasts yet introduce collinearity and overfitting risks that must be addressed through careful specification and diagnostics (Kourentzes et al., 2014; Hasan et al., 2022; Trapero et al., 2019). Across these advances, the literature also emphasizes the *human-in-the-loop*:

structured approaches to model selection and override governance can harness domain knowledge while mitigating behavioral biases, thereby improving both accuracy and accountability in retail planning cycles (Kourentzes et al., 2014; Redwanul & Zafor, 2022; Trapero et al., 2015).

A final foundational theme links forecasting *quality* to *inventory and service* outcomes—central concerns in U.S. retail supply chains. Studies show that optimizing models with respect to *inventory-relevant* criteria (e.g., service-level costs, stockout penalties) rather than pure statistical loss can yield better end-to-end performance; directly embedding inventory objectives into parameter estimation leads to measurably improved fill rates and lower holdings in empirical tests (Kourentzes et al., 2020; Rezaul & Mesbaul, 2022). Because retail portfolios often exhibit long tails with sparse or intermittent sales, methods tailored to zero-inflated series and event-driven bursts remain essential complements to global neural approaches and promotion regressions (Hasan, 2022; Trapero et al., 2019). Furthermore, safety-stock setting depends on *forecast uncertainty*, not only point accuracy: research on quantile forecast *combination* shows that aggregating distributional forecasts across methods can better capture non-Gaussian, time-varying errors typical of promotion and calendar peaks, enabling sizable safety-stock reductions at target service levels (Petropoulos et al., 2018). In parallel, governance frameworks for judgment—model selection, scenario inputs, and override rules—have been shown to systematically enhance retail forecast processes when codified and monitored (Petropoulos et al., 2018). Collectively, these foundations argue for an integrated toolkit—promotion-aware regression, cross-series global learners, multi-frequency modeling, intermittent-demand treatments, and decision-centric loss functions—embedded within a disciplined forecasting support system aligned to inventory economics and service commitments in U.S. retail (Ma et al., 2016).

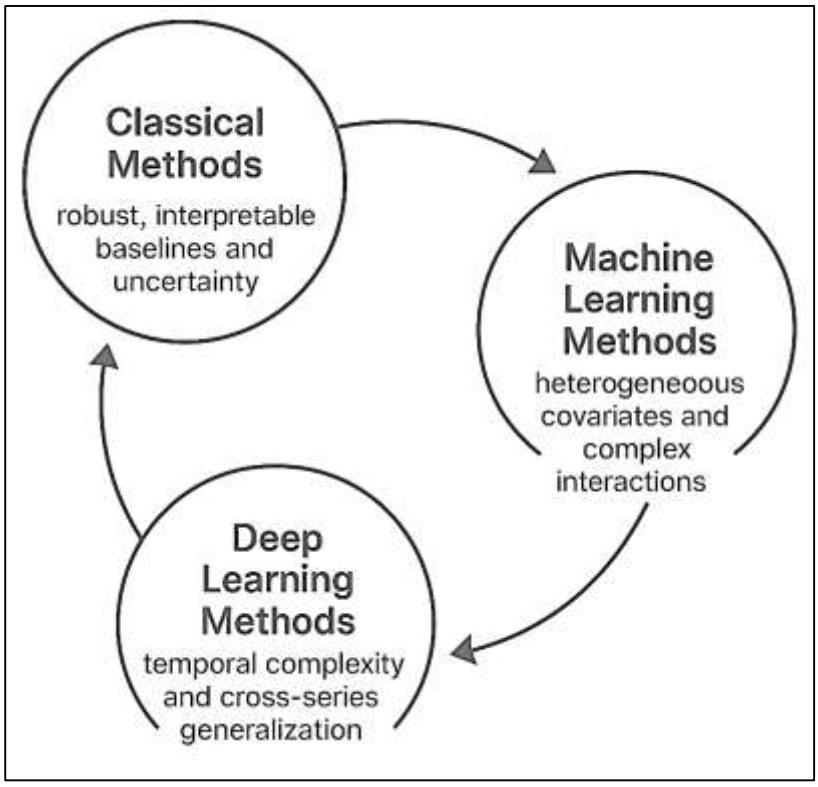
Classical vs. AI Approaches

Classical forecasting in retail evolved from stochastic time-series formulations that model systematic structure explicitly—level, trend, and seasonality—before optimizing parameters for extrapolation. The Box–Jenkins program codified ARIMA and transfer-function (dynamic regression) modeling as a disciplined, iterative cycle of identification, estimation, and diagnostic checking, giving practitioners a testable workflow for short-term retail horizons where shocks, promotions, and nonstationarities must be handled with care (Box et al., 2015). Two classical pillars further anchor retail practice: the Theta method and TBATS. Theta decomposes a series into "theta lines" that isolate trend and curvature, providing robust performance in settings with short histories and heterogeneous seasonality—common in SKU-store data with product churn and assortment changes (Assimakopoulos & Nikolopoulos, 2000; Tarek, 2022). TBATS extends exponential smoothing with Box–Cox transforms, ARMA errors, and trigonometric seasonality, enabling multiple and non-integer seasonal patterns (e.g., weekly and annual effects interacting across fiscal calendars), which is especially relevant for U.S. retailers trading in many regions and channels (De Livera et al., 2011; Kamrul & Omar, 2022). These methods emphasize generative structure, likelihood-based estimation, and interpretable components that can be reconciled across hierarchies. Yet classical tools face limits under high-dimensional covariates (prices, promotions, weather), sparse long-tail items, and complex promotion response. This tension motivates hybridization with machine learning to encode rich covariates while preserving the stability and transparency valued by retail planners. Regularized regressions—lasso and elastic net—offer a bridge, shrinking coefficients to control variance and select features in wide retail design matrices where collinearity among price tiers and promotional indicators is routine (Friedman, 2001; Kamrul & Tarek, 2022). In this hybrid view, classical structure handles baseline dynamics while modern regressors ingest exogenous retail signals at scale.

Machine-learning approaches reframed forecasting as supervised learning with flexible function classes. Ensemble trees learn nonlinearities and interactions among promotion flags, price ladders, holiday dummies, and localized weather, often outperforming linear models when retail effects are thresholded or asymmetric. Random forests reduce variance through bagging and feature subsampling, producing stable predictions and variable-importance diagnostics helpful for category and replenishment teams (Breiman, 2001; Mubashir & Abdul, 2022). Gradient boosting treats forecasting as stage-wise additive function approximation under an arbitrary loss, allowing direct optimization of operationally meaningful criteria and robust handling of outliers and heteroskedasticity (Friedman, 2001; Muhammad & Kamrul, 2022). Support vector regression adds a

margin-based view with ε-insensitive losses that can be tuned for percentile-oriented error control across SKUs, useful when planners target service levels that penalize under-forecasts more than over-forecasts (Breiman, 2001; Reduanul & Shoeb, 2022). Together, these ML families thrive when feature engineering is rich—calendar interactions, lagged promotions, competitor proxies—and when cross-sectional pooling is desired without full sequence modeling. They also integrate naturally with retail MLOps: feature stores, champion–challenger governance, and drift monitoring. However, pure tabular ML can struggle to represent long temporal dependencies and seasonality phase shifts without extensive lag features and handcrafted interactions. Consequently, many retail stacks combine ensemble learners for promotion/price effects with classical or neural sequence models that specialize in temporal dynamics. This division of labor pairs interpretability and deployment maturity on the ML side with temporal expressiveness on the sequence side, a theme that recurs in benchmarking and production case studies.

Figure 3: Cycle of classical, machine learning, and deep learning approaches



Deep learning extends this trajectory by learning temporal representations directly from sequences, enabling **cross-series** training that pools information across thousands of related SKUs and stores while adapting to item-level idiosyncrasies. The long short-term memory (LSTM) architecture solved vanishing-gradient issues and remains a versatile baseline for retail horizons where recurrent context—pre- and post-promotion baselines, holiday ramps, regional weather persistence—matters (Hochreiter & Schmidhuber, 1997; Kumar & Zobayer, 2022). Contemporary practice layers recurrent or attention mechanisms with static embeddings (e.g., item, store, region) and known-in-advance covariates to produce calibrated point and probabilistic forecasts; these models capture regime changes and cross-effects without exhaustive manual feature crafting. In parallel, component-wise additive models popularized by Prophet offer a pragmatic route to **forecasting at scale**, decomposing trend, multiple seasonalities, and holiday/event effects with analyst-friendly parameters and defaults that speed iteration across vast retail catalogs (Sadia & Shaiful, 2022; Taylor & Letham, 2018). Viewed comparatively, classical methods provide robust, interpretable baselines and principled uncertainty under well-specified dynamics (Assimakopoulos & Nikolopoulos, 2000; Smola & Schölkopf, 2004). Machine-learning ensembles exploit heterogeneous covariates and complex interactions with straightforward deployment and diagnostics (Noor & Momena, 2022; Zou & Hastie, 2005). Deep

sequence and scalable additive models absorb temporal complexity and enable cross-series generalization essential for long tails and promotion-heavy environments (Taylor & Letham, 2018; Tibshirani, 1996). In U.S. retail, the most effective systems combine these paradigms: structural components and reconciliation from classical forecasting, covariate expressiveness from ML, and temporal representation learning from deep models—assembled within governance that ensures leakage-safe features, rolling-origin evaluation, and planner-aligned metrics.

Data Modalities and Feature Engineering

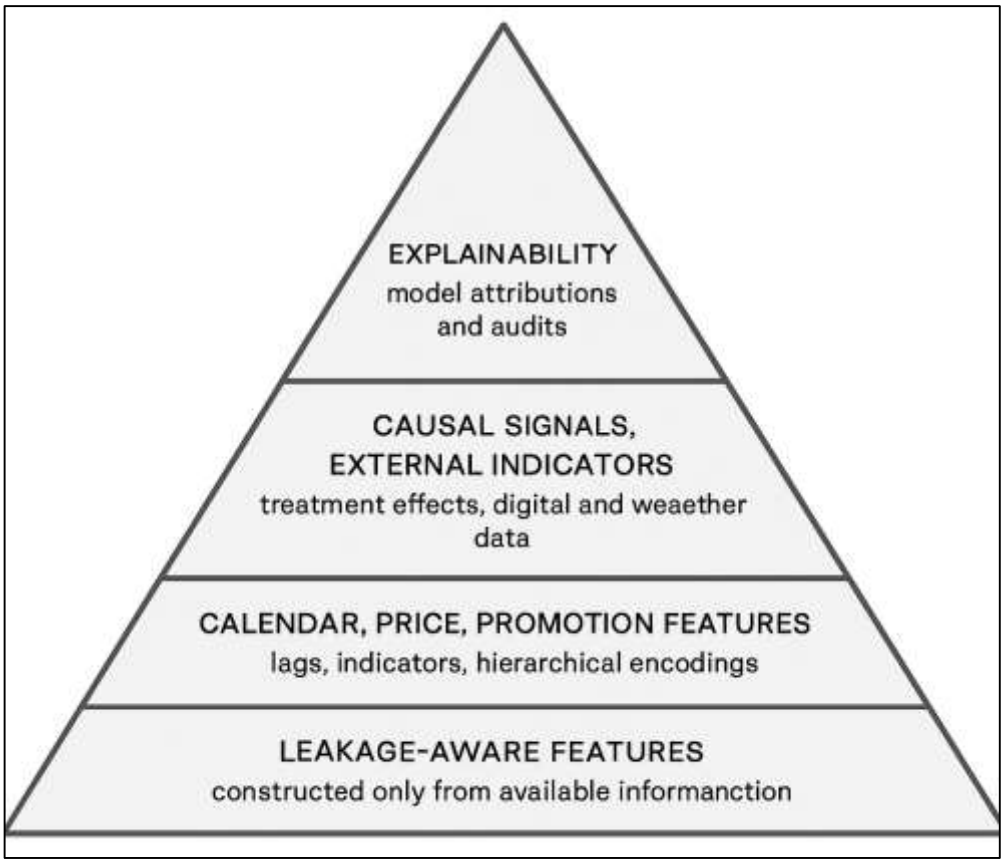
In retail demand forecasting, data modalities are the raw materials that determine what a model can learn, while feature engineering encodes business structure—calendar cycles, promotions, prices, weather, and digital signals—into regressors that models can exploit. At the foundation is leakage-safe temporal framing: features must be constructed only from information available at the forecast origin and aligned to the planning horizon. Two principles shape this pipeline. First, evaluation must mirror deployment, which implies rolling-origin or blocked schemes rather than iid cross-validation to prevent optimistic bias when features contain future leakage; classical guidance on out-of-sample testing and cross-validation cautions against random shuffles for time series and prescribes origin-by-origin assessment (Bergmeir et al., 2018; Istiaque et al., 2023; Tashman, 2000). Second, even when time ordering is respected, subtle information leaks can creep in through target-aware preprocessing, cumulative statistics, or windowing choices; modern methodological work highlights how such leakage inflates apparent accuracy and undermines scientific validity, motivating strict separation of fit/transform scopes and audit trails for feature computation (Bergmeir & Benítez, 2012; Kapoor & Narayanan, 2023; Hasan et al., 2023). With those guardrails, calendar features (week-of-year, holiday dummies, lead/lag indicators around major events) capture recurrent structure, while engineered lags and rolling summaries represent short-term momentum without peeking. Price and promotion features require richer design: level and relative price tiers, depth of discount, display/feature flags, and cross-SKU/category signals for halo and cannibalization effects. Because these covariates are often sparse and highly collinear, encodings that preserve hierarchy (e.g., one-hot for holiday types plus continuous distance-to-event) and regularization-ready representations are central to stable learning under retail realities (Lundberg & Lee, 2017; Hossain et al., 2023; Tashman, 2000).

Tree-based gradient boosting and related tabular learners have become workhorses for promotion- and price-aware feature sets, thanks to their ability to capture nonlinear thresholds and high-order interactions common in retail (e.g., asymmetric lift at specific discount depths or weekend-by-holiday interactions). Modern boosting systems also natively handle categorical variables and mitigate target leakage from high-cardinality encodings via ordered or Bayesian schemes; CatBoost is emblematic, introducing ordered target statistics to reduce overfitting from category encodings and thereby strengthening promotion-response learning when many sparse SKU and store identifiers appear as features (Rahaman & Ashraf, 2023; Prokhorenkova et al., 2018). Complementary representation learning ideas, such as **entity embeddings** for categorical variables, map items, stores, and regions into dense vectors that summarize cross-series relationships; these embeddings serve as compact, learnable features in downstream models and help transfer information to cold-start SKUs and small stores (Guo & Berkhahn, 2016; Sultan et al., 2023). As feature spaces grow, explainability becomes part of feature engineering: model-agnostic attribution scores allow planners to see which variables drove forecast changes around events and promotions, improving trust and surfacing data issues. Shapley-value explanations offer consistent local attributions across nonlinear models and can be aggregated to validate that holiday, promotion, or weather features behave plausibly across assortments and geographies (Choi & Varian, 2012; Lundberg & Lee, 2017; Hossen et al., 2023). Together, leakage-aware construction, categorical encodings designed for sparsity and hierarchy, representation learning for identifiers, and post-hoc attribution complete a pragmatic feature stack for U.S. retail use cases that must scale to millions of item-locations while remaining auditable (Lundberg & Lee, 2017; Tawfiqul, 2023).

Beyond promotions and calendars, two additional modalities enrich retail forecasting: causal signals that separate base demand from policy effects, and external indicators that proxy latent drivers. Causal uplift features—constructed from estimated treatment effects of promotions or price changes—aim to encode how demand would differ under alternative actions; modern estimation strategies include

doubly robust learners and generalized random forests that flexibly model heterogeneous treatment effects, enabling planners to feed models with scenario-stable features rather than historical mixtures of policies (Chernozhukov et al., 2018; Uddin & Ashraf, 2023). When marketing or operations need counterfactual baselines—e.g., "no-promo" sales for the next holiday window—structural time-series approaches can generate intervention features that quantify incremental lift relative to modeled trends and seasonality, and these features can enter forecasting pipelines to reduce bias from past campaigns (Brodersen et al., 2015; Momena & Hasan, 2023; Wager & Athey, 2018). External indicators supply further explanatory power: high-frequency web search intensity and digital interest proxies can provide early signals for category-level forecast adjustments, particularly for event-driven products; careful nowcasting work demonstrates how such indices, when properly lagged and filtered, improve short-horizon accuracy (Sanjai et al., 2023). Weather features (temperature, precipitation, heat index) and macroeconomic controls (income, inflation) likewise function as known-in-advance or slowly evolving drivers in specific categories, but must be engineered to avoid contemporaneous leakage (e.g., use forecasts or scenario paths, not realized values) and scaled via interactions with region and season. Finally, evaluation design closes the loop: blocked time-series cross-validation with leakage-aware feature stores, held-out interventions, and attribution audits ensures that engineered features remain predictive for the decisions and horizons actually faced in U.S. retail operations (Brodersen et al., 2015; Akter et al., 2023).

Figure 4: Pyramid of data modalities and feature engineering

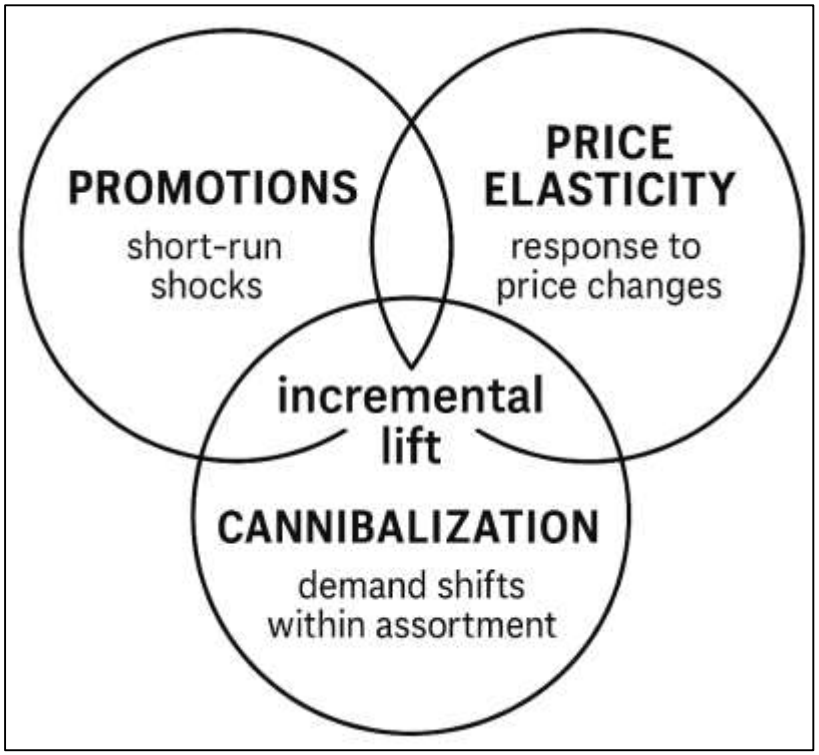


Promotions, Price Elasticity, and Cannibalization in Retail Forecasting

Promotions inject powerful, short-run shocks into retail demand, but their managerial value depends on where the incremental volume comes from and how it alters future buying. Scanner–data econometrics established the core decomposition: a temporary price cut produces (i) cross-brand substitution within the focal category, (ii) cross-period borrowing (stockpiling that depresses future sales), and (iii) category expansion from new or accelerated purchases (Bijmolt et al., 2005; van Heerde et al., 2004). The widely used SCAN*PRO family operationalized promotion response in store–brand

panels and inspired modern variants that account for heterogeneity across stores and items, improving elasticity estimates used for everyday pricing and promotional planning (Andrews et al., 2008; Tamanna & Ray, 2023). Long-horizon evidence further shows that while short-run bumps are often large, permanent effects on category incidence, brand choice, or quantity are usually limited—placing the burden on careful tactic selection and consistent post-promotion diagnostics (Danish & Md. Zafor, 2024; Pauwels et al., 2002). At the trip level, promotions interact with multi-category baskets: complementarities and co-incidence patterns propagate lift (or suppression) across categories, implying that SKU-level forecasts should embed basket-structure features alongside category-specific promotion variables (Manchanda et al., 1999). Together, these results argue for promotion-aware forecasting that (1) models direct own-SKU lift, (2) monitors cannibalization within and across categories, and (3) guards against illusory gains when cross-period borrowing dominates (Istiaque et al., 2024; Rzepakowski & Jaroszewicz, 2012).

Figure 5: Interactions of promotions, price elasticity, and cannibalization



Estimating price elasticity credibly is central to scenario planning and promotion design. Classical and structural demand models translate observed choices into substitution patterns that determine how price changes cascade across the assortment. The conditional logit and its generalizations connect attribute utilities to choice shares, creating a blueprint for cross-price effects that can be integrated into forecasting stacks or upstream pricing simulators (Hasan et al., 2024; Rzepakowski & Jaroszewicz, 2012). Random-coefficients logit extended this by allowing heterogeneous tastes—crucial for realistic substitution under promotions—and became a workhorse for counterfactuals (Berry et al., 1995). Meta-analytic synthesis across hundreds of categories clarifies systematic drivers: elasticities tend to be more negative for private labels, in competitive categories, and under deeper temporary price reductions—stylized facts forecasters can encode via priors, pooling schemes, or feature constraints (Bijmolt et al., 2005; Rahaman, 2024). In practice, elasticity estimation must also disentangle policy endogeneity (prices reacting to demand shocks). Robust pipelines therefore pair instrumented or structural estimation with out-of-sample validation, and then map elasticities into forecast adjustments and inventory simulations to verify that predicted lift translates into profitable, service-aligned outcomes (McFadden, 1973; Hasan, 2024). Contemporary promotion strategies also face learning problems—new items, shifting price ladders, and unknown cross-effects—where integrated demand-learning-and-pricing systems demonstrate how machine learning plus constrained optimization can improve revenue without

sacrificing stability (Ferreira et al., 2016; Ashiqur et al., 2025). For forecasters, these systems supply policy-stable features (estimated elasticities and cross-effects) that make scenario forecasts more trustworthy than raw correlational lift factors.

A modern toolkit for cannibalization and incremental lift increasingly blends econometrics with causal machine learning. Uplift modeling seeks the incremental effect of a promotion versus no promotion, directly optimizing treatment–control contrast at the SKU–store–week level; tree-based uplift learners offer segment-level targeting and interpretable rules (Hasan, 2025; Rzepakowski & Jaroszewicz, 2012). Causal trees generalize this idea, estimating heterogeneous treatment effects under weaker functional-form assumptions and enabling retailers to identify when a promotion's "lift" is mostly borrowed from future weeks or rival SKUs (Athey & Imbens, 2016; Ismail et al., 2025). These methods complement decomposition models by providing ex ante targetability and ex post accountability: planners can simulate who to treat (offer depth, vehicles) and then audit whether realized effects matched predictions. Importantly, cross-category and store heterogeneity matter—cannibalization can migrate within a brand family or spill to adjacent categories, and its magnitude varies with display, feature, and competitive noise (Jakaria et al., 2025; van Heerde et al., 2004). Practical implementations therefore layer three components: (i) structural or reduced-form demand systems to anchor substitution and elasticity; (ii) uplift/heterogeneity models to target and audit incrementality; and (iii) basket- or category-network features to capture halo and cannibalization paths (Manchanda et al., 1999). When combined with store- and item-level heterogeneity controls (e.g., hierarchical specifications in SCAN*PRO variants) and validated against long-run persistence benchmarks, this integrated approach yields promotion-aware forecasts that are decision-ready—attentive to incremental volume, guarded against cannibalization, and consistent with inventory and margin objectives (Andrews et al., 2008; Hasan, 2025).

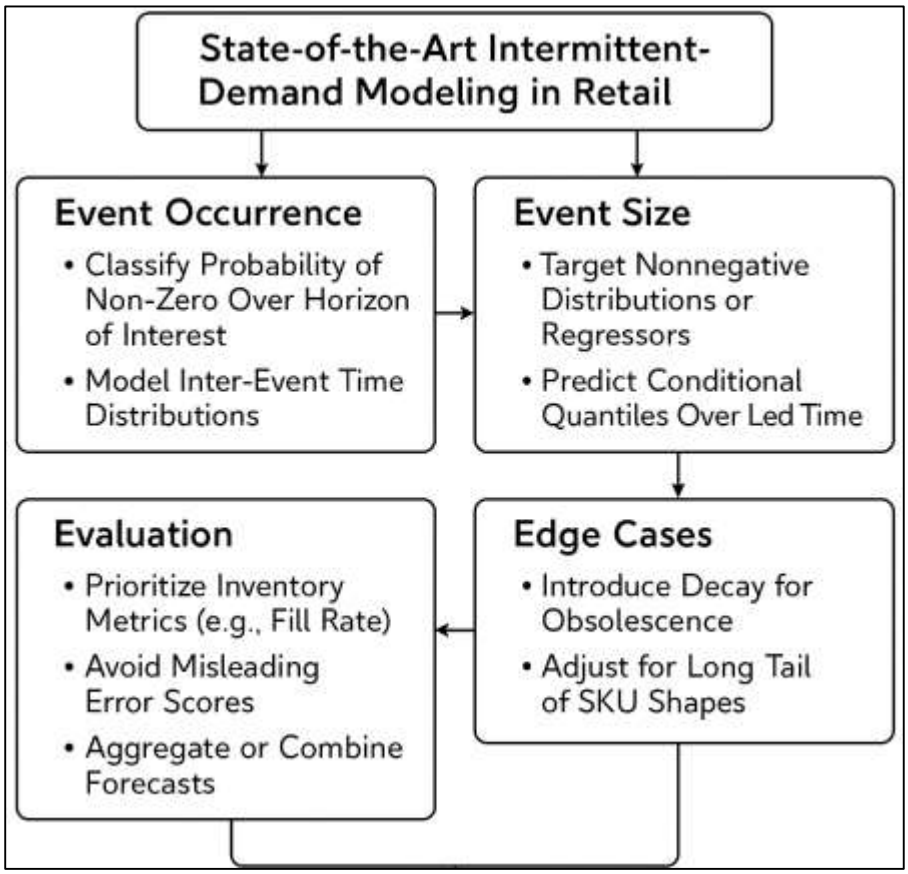
Long-tail and Intermittent Demand Modeling

Intermittent (a.k.a. lumpy or sparse) demand—characterized by long runs of zeros punctuated by irregular, bursty purchases—dominates the long tail of U.S. retail assortments and is a prime failure mode for classical forecasting pipelines. The statistical challenges are twofold: (i) correctly modeling the *occurrence* of non-zero demand events and (ii) estimating the *size* of those events conditional on occurrence. Early parametric work recognized that Gaussian assumptions break down in this regime and proposed nonnegative, count-aware formulations that unify slow- and fast-moving items without arbitrary thresholds (Sultan et al., 2025; Snyder, 2002). Subsequent distributional approaches extended this line, advocating multi-series likelihoods and prediction distributions rather than point forecasts so that inventory decisions (which depend on quantiles) remain coherent under sparsity (Altay et al., 2012). Parallel advances reframed the problem at the lead-time horizon, resampling empirical non-zero draws and inter-arrival structure via a time-series bootstrap to produce full lead-time demand distributions; these procedures consistently outperformed exponential smoothing and Croston-style heuristics on large industrial datasets (Babai et al., 2018; Zafor, 2025). In practice, managers additionally face correlation structures between event timing and size, making naive independence assumptions fragile. Simulation evidence with compound-Poisson generators shows that autocorrelation in sizes and intervals, and cross-correlation between them, systematically shift service levels for the same cost envelope—implicating the choice of estimator as a lever on fill-rate under identical policies (Kourentzes, 2013; Uddin, 2025). Temporal aggregation offers another powerful idea: by forecasting on aggregated buckets and then disaggregating, one can reduce intermittence, stabilize variance, and improve both stock-control and accuracy metrics relative to working at the native sparse cadence (Lolli et al., 2017; Nikolopoulos et al., 2011; Sanjai et al., 2025).

AI-enhanced pipelines now treat intermittent demand as a *two-module* problem: a classifier for demand occurrence married to a conditional regressor (or distributional model) for event size. Neural network variants—trained with regularization and median ensembling to offset small samples—can capture interactions between inter-arrival times and non-zero magnitudes that Croston-type models cannot, and while their point-forecast accuracy can trail simple baselines, inventory-relevant metrics (e.g., service level at fixed stock) often improve, which is what matters operationally (Snyder et al., 2012). Single-hidden-layer networks (including extreme learning machines) provide a lightweight alternative when compute or data are constrained, delivering competitive results across aggregation levels in real

spare-parts series (Snyder, 2002; Snyder et al., 2012). Meanwhile, Bayesian and renewal-process perspectives have helped formalize the dual-component nature of intermittence, encouraging explicit modeling of inter-event time distributions and non-Gaussian event-size tails so that predicted quantiles align with reorder decisions (Snyder, 2002; Snyder et al., 2012). These strands converge on a practical design: (1) classify the probability of a non-zero in the horizon of interest; (2) model the conditional size with nonnegative distributions or machine-learning regressors; (3) compose a full predictive distribution over lead-time demand for service-level-consistent policies. Critically, evaluation must privilege inventory outcomes over generic error scores: intermittent series can render popular measures undefined or misleading, whereas stock-oriented metrics (fill-rate at budget, average backorders, or expected holding/shortage costs) diagnose the business trade-off directly (Wallström & Segerstedt, 2010; Willemain et al., 2004).

Figure 6: Framework for intermittent demand modeling in retail forecasting



Two additional edge cases—obsolescence and long-tail combinatorics—stress even robust models. When an SKU ceases to sell, Croston-style estimators can sustain positive long-run forecasts, inflating safety stocks. Modern intermittent-demand estimators introduce explicit decay so that forecasts glide to zero after strings of zeros, preventing stranded inventory during end-of-life (Petropoulos & Kourentzes, 2015; Snyder et al., 2012). Beyond decay controls, recent obsolescence-aware methods adjust both size and arrival-rate updates in zero periods and have shown superior accuracy and stock performance across many scenarios (Altay et al., 2012). At assortment scale, no single estimator dominates the heterogeneous shapes found in the long tail; forecast combinations tuned on series features (e.g., average inter-demand interval, squared coefficient of variation) deliver robust gains by blending complementary inductive biases across items (Petropoulos & Kourentzes, 2015). Finally, aggregation-disaggregation workflows such as ADIDA remain especially useful in retail: aligning aggregation windows with review periods and lead times reduces sparsity while keeping replenishment compatible with operational calendars (Nikolopoulos et al., 2011). In sum, state-of-the-art intermittent-demand modeling in retail joins distributional forecasting, event-occurrence

classification, obsolescence-aware decay, and feature-based combinations—evaluated against inventory-centric criteria—to tame the long tail and align AI forecasts with service-level and cost objectives (Babai et al., 2018; Snyder et al., 2012).

Probabilistic Forecasting and Decision-Oriented Metrics

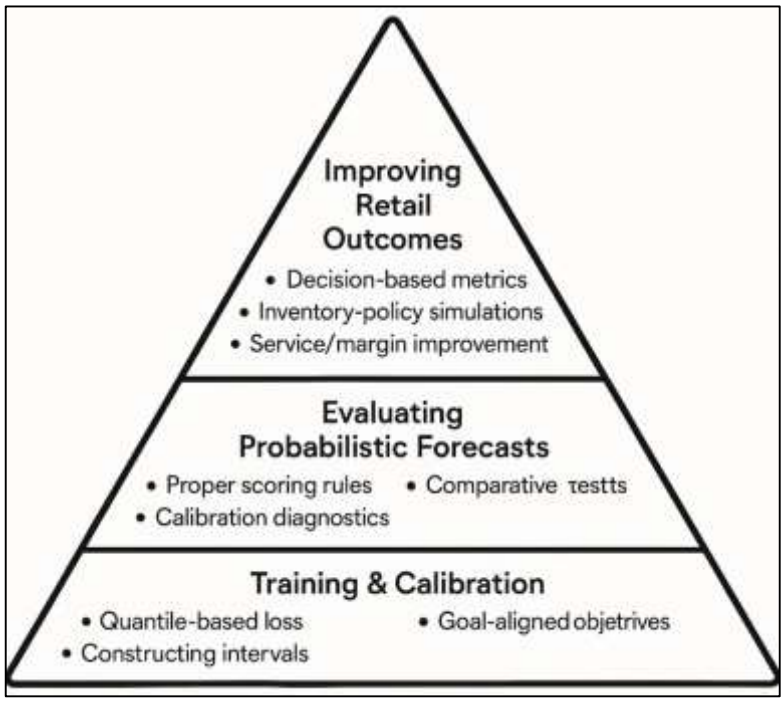
Probabilistic forecasting reframes retail demand prediction from a single "best guess" to a full predictive distribution over future outcomes, enabling planners to set service levels, safety stocks, and reorder points directly from quantiles rather than ad-hoc buffers. A rigorous treatment distinguishes calibration (statistical consistency between predicted and realized frequencies) from sharpness (concentration of the forecast distribution), and evaluates both using proper scoring rules so that better uncertainty quantification is rewarded, not just narrower intervals. This paradigm has been formalized in modern reviews and practice guides that emphasize distributional evaluation and the dual goals of calibrated and sharp forecasts in operational settings (Gneiting & Ranjan, 2011). The intellectual roots lie in probabilistic scoring and decision theory, which established that proper scores uniquely incentivize honest probability assessment—an essential property when forecasts feed inventory and pricing decisions (Winkler, 1969). For density forecasts, metrics like the continuous ranked probability score (CRPS) simultaneously assess location and dispersion, while probability integral transform diagnostics probe calibration across the entire distribution (Pinson & Tastu, 2013; Winkler, 1969). In multivariate or hierarchical contexts (e.g., SKU–store–region coherency), energy-score–type functionals offer tractable evaluation of joint distributions, preserving incentives for distributional accuracy instead of rewarding only mean performance (Diebold & Mariano, 1995). Decision-oriented benchmarking still requires careful predictive-accuracy testing; when planners compare alternative pipelines on rolling origins, robust tests of forecast superiority account for autocorrelation and overlapping horizons so that "wins" reflect persistent improvements, not noise (Diebold & Mariano, 1995; Khosravi et al., 2011). Together these elements—proper scoring, calibration diagnostics, and valid comparative tests—define the statistical backbone for operational uncertainty management in retail demand forecasting (Boylan & Syntetos, 2010; Gneiting & Katzfuss, 2014).

Quantiles sit at the heart of retail decision-making because reorder targets, safety stocks, and service-level constraints map naturally to forecast quantiles rather than means. Theory clarifies why: quantiles are elicitable functionals—there exist strictly proper loss functions (e.g., the pinball loss) that uniquely incentivize estimating a given quantile—so models can be trained and compared directly on the decision-relevant objective (Hong et al., 2016). In practice, this enables unified training and evaluation: a forecaster optimizing the τ-quantile can be judged with quantile loss, while downstream inventory policies consume the same τ as a service-level parameter, closing the loop from statistics to operations. Constructing prediction intervals for complex, nonlinear learners extends this toolkit. When distributional assumptions are fragile, model-agnostic interval construction—e.g., conformal or ensemble-based methods—offers finite-sample coverage guarantees or robust empirical coverage without requiring Gaussian residuals; neural-network interval methods exemplify how to generate calibrated bounds around highly flexible fits used in promotion-rich retail data (Khosravi et al., 2011; Taylor, 2019). For classical time-series pipelines, specialized interval formulations improve coverage under multiple seasonalities and state-space structure, ensuring that the width of weekly SKU-level intervals reflects both process and parameter uncertainty rather than ad-hoc multipliers (Taylor, 2019). The move from point to distributional forecasts is not merely academic: it enables inventory-aligned metrics in backtesting. Planners can evaluate candidate models on realized coverage of on-shelf targets, stockout counts at target quantiles, or cost-weighted scoring rules that penalize under- and over-forecasting asymmetrically, mirroring the true economics of retail operations. As organizations industrialize this approach, consistent use of quantile-targeted training, proper distributional scoring in evaluation, and calibrated interval construction yields uncertainty estimates that are both statistically defensible and actionable at scale (Fissler & Ziegel, 2016).

End-to-end retail performance depends on how uncertainty information is consumed by planning systems, not only on how it is measured. Large-scale forecasting challenges in adjacent domains have accelerated best practices that transfer cleanly to retail, particularly around probabilistic leaderboard design and loss functions aligned with quantile targets (Hong et al., 2016). Within inventory control, the business impact of probabilistic improvements is mediated by policy choice: order-up-to and

newsvendor-style policies translate distributional forecasts into service levels, backorders, and holding costs; thus, bias and dispersion in forecast distributions have asymmetric cost implications that should be reflected in evaluation (Boylan & Syntetos, 2010). This motivates decision-consistent backtests: rolling-origin simulations where each candidate model feeds the same replenishment policy, and outcomes—fill rate, average inventory, lost sales—become the comparison metrics. Statistically, comparative tests can still rely on robust forecast-accuracy testing, but the target metrics are operational rather than purely statistical (Diebold & Mariano, 1995). In production, combining these elements yields a disciplined pipeline: (i) train models to optimize quantile- or distribution-aware objectives; (ii) evaluate with proper scores (CRPS/quantile loss) and calibration diagnostics; and (iii) validate downstream via inventory-policy simulations that quantify economic value. Such a framework ensures that probabilistic forecasts improve what matters—service and margins—while maintaining transparent, auditable metrics across the hierarchy from SKU to enterprise roll-ups (Hong et al., 2016; Boylan & Syntetos, 2010).

Figure 7: Triangle framework for probabilistic forecasting and decision-oriented metrics

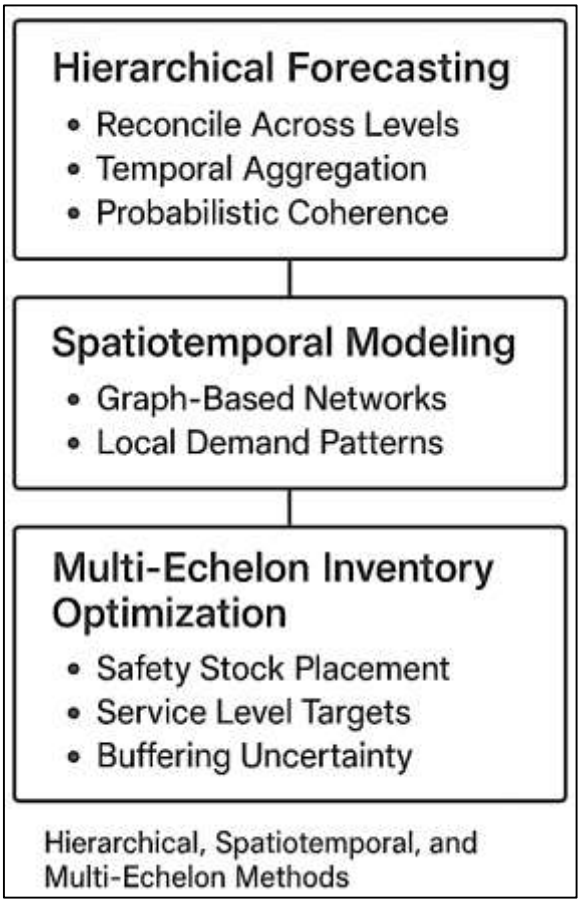


Hierarchical, Spatiotemporal, and Multi-Echelon Methods

Hierarchical and grouped forecasting frameworks formalize how to generate base forecasts at multiple aggregation levels and then enforce aggregate consistency so that lower-level predictions sum to upper-level totals used by merchandising, finance, and logistics. Early comprehensive treatments established model classes and evaluation protocols for cross-sectional hierarchies, showing the advantages of combining forecasts across levels for improved accuracy and managerial interpretability (Athanasopoulos et al., 2009). A complementary stream builds temporal hierarchies, treating the same series at multiple sampling frequencies (e.g., day→week→month) to stabilize estimation and capture multi-seasonality, then reconciling across frequencies for sharper short- and medium-term retail horizons (Athanasopoulos et al., 2009; Athanasopoulos et al., 2017). Joint cross-temporal formulations further integrate these ideas, allowing retailers to exploit information both across items and across time scales—crucial when store-day signals are noisy but week-category patterns are reliable (Kourentzes & Athanasopoulos, 2019). Methodological refinements deepen this toolkit: geometric views of reconciliation clarify bias correction and constraint handling (e.g., non-negativity), providing matrix formulations that are stable under realistic error structures and scalable to large retail trees (Ben Taieb et al., 2017; Panagiotelis et al., 2021). Probabilistic variants extend reconciliation from means to full distributions, enabling calibrated quantile or density-level coherence that directly feeds service-level

policies; constructive recipes for coherent probabilistic forecasts make it possible to propagate uncertainty consistently across the enterprise roll-ups used in S&OP cycles (Yu et al., 2018). For U.S. retailers, the practical impact of these approaches is twofold: first, coherent planning artifacts (e.g., category-month buy plans agree with store-day replenishment); second, measurable improvements in forecast accuracy when information is unevenly distributed across the hierarchy and time scales (Athanasopoulos et al., 2009; Graves & Willems, 2000).

Figure 8: Rectangle framework for hierarchical



Spatial and network structure add another layer: demand at a store reflects local demographics, weather, competition, and network effects such as inventory transshipments and regional promotions; ignoring such structure can degrade accuracy and distort uncertainty. Spatiotemporal models explicitly encode dependence across locations and time, enabling learning that pools strength geographically while preserving local idiosyncrasies. In recent years, graph-based deep learners have become prominent: diffusion convolutional recurrent networks (DCRNN) model directed flows over a graph with recurrent dynamics, capturing how shocks (e.g., a regional campaign or weather front) propagate through neighboring nodes; these architectures handle irregular networks and time-varying effects with high fidelity (Li et al., 2018). Spatiotemporal graph convolutional networks (STGCN) provide another scalable design, interleaving temporal convolutions with graph convolutions to represent localized periodicity and cross-node interactions—useful for chain retailers where adjacent stores share demand rhythms yet respond differently to price-promo tactics (Graves & Willems, 2004; Yu et al., 2018). When mapped to retail, nodes represent stores or fulfillment nodes, edges encode distance, travel time, competitive adjacency, or supply links, and covariates include weather forecasts, calendar effects, and promotion calendars; the learned representations support regionalization strategies (e.g., cluster-wise planograms or price zones) and more accurate store-day forecasts at the long tail. Importantly, these spatiotemporal learners complement hierarchical reconciliation: base forecasts for each node can be produced by DCRNN/STGCN, then reconciled vertically across the product/time hierarchies to

ensure enterprise coherence. This division of labor—learn local network dynamics, then reconcile across business hierarchies—aligns with how U.S. retailers operate, in which regional DCs, delivery routes, and weather bands shape correlated demand shocks that classical univariate pipelines cannot easily capture (Clark & Scarf, 1960).

Forecasts ultimately drive multi-echelon inventory decisions in networks spanning vendors→plants→DCs→stores. The classic theory of optimal stock placement in serial and general networks shows that safety stock should be located where it most effectively buffers uncertainty, with echelon-based policies and dynamic programming yielding structural insights still used in modern tools (Kourentzes & Athanasopoulos, 2019). Building on this, network design models formalized how to assign strategic safety stock across divergent topologies under target service levels, providing tractable algorithms that translate demand variability and lead-time dispersion into placement and sizing rules (Graves & Willems, 2004). Follow-on work optimized safety-stock placement jointly with bill-of-materials and cycle-stock decisions, connecting planning bills, postponement, and decoupling points to inventory efficiency—ideas that align naturally with retail's vendor→DC→store pipelines and omnichannel nodes (Graves & Willems, 2004). Marrying these inventory foundations with modern forecasting yields a coherent design loop: (i) generate distributional forecasts at SKU-store and higher levels, (ii) reconcile across product/time hierarchies for coherence, (iii) propagate uncertainty through spatiotemporal networks to capture correlated shocks, and (iv) feed multi-echelon optimization that places and sizes safety stock consistent with targeted service levels and cost-to-serve. From an engineering perspective, the benefit is end-to-end consistency: the same uncertainties learned by hierarchical/spatiotemporal models determine buffer placement and replenishment rules, preventing the misalignment that occurs when forecasting and inventory modules are tuned in isolation. For U.S. retail supply chains, where lead times, carrier capacity, and regional demand co-move during holiday peaks and weather events, this integrated approach reduces stockouts and excess simultaneously by targeting buffers to the network points with the greatest marginal value.

METHOD

This study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines to ensure a systematic, transparent, and rigorous review of artificial intelligence–enhanced predictive analytics for demand forecasting in U.S. retail supply chains, culminating in a final analytic corpus of 95 peer-reviewed articles. We defined the review protocol a priori, specifying the population (U.S. retail settings at SKU, store, and network levels), interventions or approaches (AI, machine learning, deep learning, hybrid/statistical models, hierarchical and spatiotemporal methods, probabilistic forecasting, promotion and price modeling), comparators (classical baselines and alternative AI pipelines), outcomes (forecast accuracy and calibration metrics alongside decision-oriented measures such as service level, stockouts, and inventory turns), and study designs (empirical evaluations using real or realistic retail data). Searches were executed across multidisciplinary and domain databases (e.g., Scopus, Web of Science Core Collection, IEEE Xplore, ACM Digital Library, Business Source Complete) and preprint servers screened for eligibility, using Boolean strings that combined terms for "retail" and "demand (or sales) forecasting" with "machine learning," "deep learning," "transformer," "probabilistic," "promotion," "price elasticity," "hierarchical," "spatiotemporal," and "multi-echelon," constrained to 2015–2025 with backward citation chasing to capture seminal antecedents. Records were deduplicated and screened in two stages—title/abstract followed by full text—against inclusion criteria requiring relevance to U.S. retail or transferability to U.S. practice, explicit model or method description, transparent evaluation protocol, and reportable metrics; exclusion criteria removed nonretail contexts, inaccessible full texts, purely theoretical pieces without empirical assessment, and studies lacking evaluation clarity or at high risk of leakage. Two reviewers independently applied criteria with discrepancies resolved by discussion and, when needed, a third adjudicator; inter-rater agreement was monitored and reasons for exclusion were logged to preserve auditability. Data extraction used a predefined template capturing bibliographic information, data granularity, exogenous variables, model families, training and validation design (including rolling-origin or blocked schemes), metrics (point and distributional), and operational outcomes, while quality appraisal inspected risks of bias such as data leakage, horizon mismatch, weak baselines, and overfitting from high-cardinality encodings. Synthesis combined narrative integration with structured

tables, stratifying evidence by horizon, hierarchy, product velocity, and promotional intensity to align methodological findings with retail decision contexts, and sensitivity analyses probed robustness to evaluation design to maintain PRISMA-consistent transparency throughout.

Screening and Eligibility Assessment

Screening and eligibility assessment proceeded in two sequential stages consistent with PRISMA, with predefined rules to ensure transparency, reproducibility, and alignment to the study scope. After importing all records from the targeted databases and preprint servers, exact- and fuzzy-matching routines were used to remove duplicates based on DOI, title, author list, year, and venue, followed by a manual pass to merge near-duplicates and multiple versions of the same contribution (e.g., conference-to-journal extensions). Title–abstract screening applied inclusion criteria centered on substantive relevance to retail demand forecasting in U.S. settings or clear transferability to U.S. practice, empirical evaluation with real or realistically simulated retail data, and explicit methodological detail enabling interpretation of model families, data modalities, and evaluation design; exclusion criteria removed nonretail domains, purely theoretical or viewpoint pieces, inaccessible full texts, and studies without reportable metrics or with insufficient methodological transparency. To guard against inflated results at this early stage, we flagged potential high-risk items—such as those signaling target leakage, horizon mismatch (training on post-forecast information), or unbalanced baselines—for closer inspection at full text. In the full-text eligibility stage, two reviewers independently evaluated each article against refined criteria: (i) empirical focus at SKU, store, category, region, or network levels; (ii) clear definition of the forecasting task and horizon, including any hierarchical, spatiotemporal, or multi-echelon structure; (iii) transparent training/validation protocols (preferably rolling-origin or blocked time splits) and decision-relevant metrics (e.g., WAPE/MASE for point forecasts, quantile loss/CRPS and coverage for probabilistic outputs); (iv) adequacy of baselines and ablation analyses for AI comparisons; (v) U.S. data or a strong rationale for applicability to U.S. retail (e.g., comparable promotion calendars and channel structures). Preference was given to peer-reviewed outputs with DOIs; when both a preprint and a later peer-reviewed version existed, the latter superseded the former. Multiple reports from the same dataset or competition (e.g., replications without methodological novelty) were consolidated to avoid double counting. Disagreements were resolved by consensus or a third adjudicator, with reasons for exclusion logged under standardized categories (scope, design, metrics, transparency, access). Only studies meeting all criteria entered the analytic corpus, yielding the final set of 95 articles used for synthesis.

Data Extraction and Coding

Data extraction and coding followed a prespecified template designed to capture methodological, data, and outcome features at a consistent unit of analysis. Each included article was entered into a structured database with fields for bibliographic metadata (DOI, venue, year), problem framing (forecasting task definition, horizon, aggregation level), dataset characteristics (industry segment, geographic scope, number of series, series length, SKU/store granularity), and data modalities (prices, promotions, holidays, weather, macro indicators, web signals). We recorded feature-engineering practices (lag structures, rolling statistics, categorical encodings, leakage controls), model families (statistical, machine learning, deep sequence, hybrid), and architectural details (embeddings, attention, reconciliation layers, spatiotemporal or multi-echelon components). Training and validation design were coded with an emphasis on deployment realism: windowing scheme (rolling-origin, blocked splits), backtest span, retraining cadence, and hyperparameter search transparency. Evaluation metrics were harmonized across studies by mapping reported measures into a canonical set: WAPE/sMAPE/MASE and RMSE/MAE for point accuracy; pinball loss, CRPS, empirical coverage, and interval width for probabilistic performance; and decision-oriented outcomes including service level, stockouts, inventory turns, and cost-weighted scores when available. To enable cross-study synthesis, we computed normalized effect sizes where possible (e.g., relative WAPE improvement versus the strongest classical baseline) and recorded ablation evidence (with/without promotions, with/without hierarchical reconciliation). Reproducibility indicators were double-coded, including data availability (public/proprietary), code or model cards, and reporting sufficiency for replication. Quality and risk-of-bias flags captured common threats: target leakage, horizon mismatch, unbalanced

baselines, overfitting risks from high-cardinality encodings, and inadequate uncertainty calibration. Coding proceeded in two passes: an initial pilot on a 10-article subset to refine the codebook and controlled vocabulary, followed by dual, independent coding on the full corpus with adjudication of discrepancies; inter-coder agreement was monitored (Cohen's κ) on key fields (task, horizon, metrics, leakage flags), and unresolved conflicts were settled by consensus. Missing or ambiguous information was annotated as "not reported" rather than imputed, with author correspondence attempted only when essential to compute a normalization. All entries maintained audit trails linking values to page or figure references, and version control captured updates during sensitivity analyses, ensuring traceable, PRISMA-consistent synthesis across the final set of 95 studies.

Data Synthesis and Analytical Approach

The synthesis strategy was designed to integrate heterogeneous evidence—from classical statistical baselines to modern AI architectures—into a coherent, decision-oriented account of forecasting performance in U.S. retail supply chains. Because the 95 included studies varied in design, data granularity, evaluation metrics, and reporting practices, we adopted a hybrid approach that combined structured narrative synthesis with quantitative aggregation where commensurability permitted. We first constructed a crosswalk that standardized problem framing across studies along five axes: (i) forecast horizon (short-run daily/weekly vs. medium-run monthly/quarterly), (ii) hierarchical level (SKU–store, SKU–region, category–region, enterprise totals), (iii) data modality set (presence of price, promotions, holidays, weather, macro, and digital signals), (iv) methodological family (statistical, machine-learning tabular, deep sequence, hybrid, hierarchical/spatiotemporal, and multi-echelon coupling), and (v) forecast objective (point vs. probabilistic). This crosswalk served two purposes: it allowed us to align like with like when comparing effect sizes, and it exposed systematic gaps—e.g., horizons or levels underrepresented by deep models or decision-oriented metrics. To manage reporting heterogeneity, we mapped diverse error metrics into a canonical set—WAPE, sMAPE, MASE, RMSE/MAE for point accuracy; pinball loss ($\tau \in \{0.1, 0.5, 0.9\}$), CRPS, empirical coverage and average interval width for probabilistic performance—using published relationships where valid (e.g., unit scale conversions and seasonality-adjusted normalizations) and otherwise retaining metrics in their native form with subgroup analyses. Decision outcomes (service level, stockouts, inventory turns, cost-weighted scores) were abstracted to directionally consistent effect indicators and, where available, expressed as percentage improvements versus a specified baseline to facilitate cross-study synthesis.

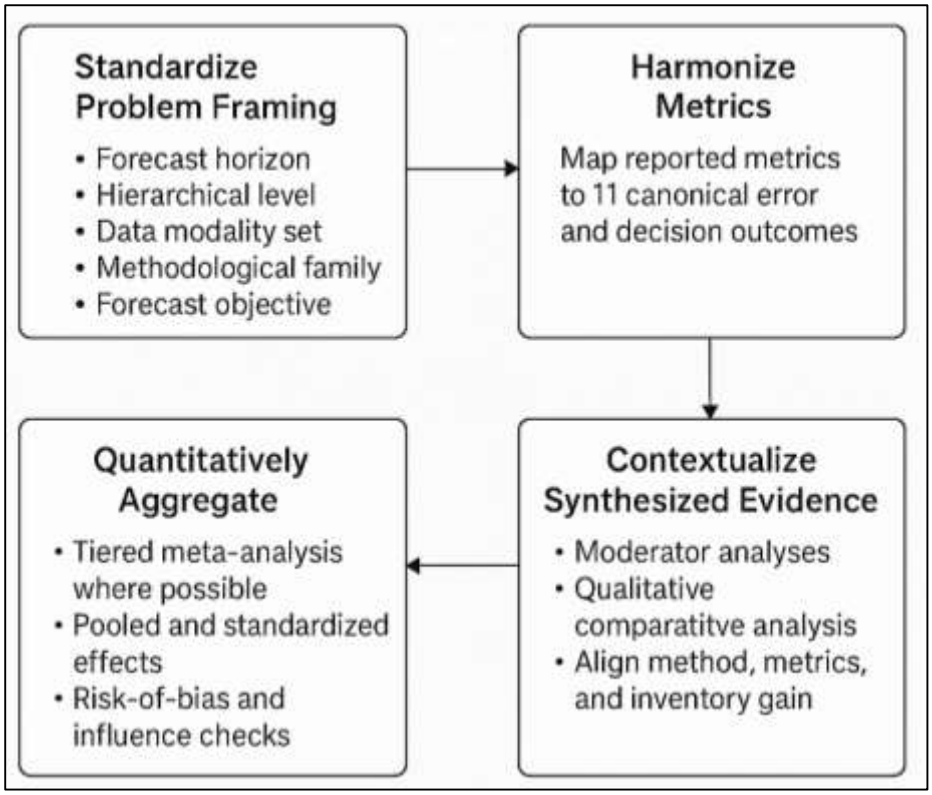
Quantitative aggregation proceeded in tiers keyed to the attainable level of harmonization. At Tier 1 (high commensurability), we conducted random-effects meta-analyses of relative improvement in WAPE or sMAPE compared to the strongest classical baseline reported within each study, treating "model-vs-baseline" as the unit of analysis to guard against double counting. Heterogeneity was assessed using τ^2 and I^2, with Hartung–Knapp adjustments applied to control for between-study variance under small k conditions. Because multiple comparisons often appear within a single paper (e.g., GBM, LSTM, and Transformer all versus ETS), we used robust variance estimation (RVE) to accommodate dependent effect sizes without inflating precision. At Tier 2 (moderate commensurability), where studies reported different but related metrics (e.g., MAE vs. RMSE, wMAPE variants) on similar horizons and levels, we computed standardized mean differences after scaling by a reported or inferred measure of dispersion; when dispersion was unavailable, we restricted to vote-counting enhanced by effect direction and magnitude bins (e.g., >10% improvement, 5–10%, 0–5%, negative) and reported proportions with Wilson intervals, clearly labeling these as exploratory. At Tier 3 (low commensurability), typically driven by bespoke KPI definitions or proprietary decision metrics, we reverted to narrative synthesis organized by context (promotion intensity, intermittency, hierarchical reconciliation, spatiotemporal coupling), highlighting convergent patterns and triangulating with Tier 1–2 results to avoid overweighting idiosyncratic designs.

Meta-analytic models were stratified a priori by horizon (daily/weekly vs. monthly), level (SKU–store vs. aggregated), and data modality set (with vs. without price/promotion features), because these dimensions systematically mediate algorithmic advantage. We further implemented moderator analyses to probe whether specific design choices explained variance in effect sizes: (i) inclusion of promotion depth and display features; (ii) leakage-safe evaluation (rolling-origin or blocked splits) vs.

weaker designs; (iii) presence of hierarchical reconciliation; (iv) cross-series training (global models) vs. per-series (local) fitting; and (v) intermittent/long-tail prevalence in the evaluation sample. Moderators were encoded as binary or ordinal indicators and entered into meta-regressions under a random-effects framework, with inference based on profile-likelihood confidence intervals for τ^2 and small-sample corrections for coefficients. To address potential small-study and publication bias, we combined contour-enhanced funnel plots, Egger-type regressions adapted to ratio outcomes, and trim-and-fill procedures; where asymmetry suggested selective reporting, we performed sensitivity analyses that down-weighted or excluded studies failing minimal transparency thresholds (e.g., unclear horizon specification or suspect feature timing) and contrasted results to the full model.

Because a large fraction of the contemporary AI literature reports probabilistic outcomes, we developed a parallel synthesis track for distributional metrics. For studies reporting pinball loss at $\tau=0.5$ alongside coverage of 80% or 90% intervals, we meta-analyzed relative improvements in pinball loss using log ratios (to stabilize variance) and synthesized coverage deviations from nominal as a calibration gap (observed minus nominal), with positive values indicating over-coverage. CRPS effects were pooled on relative scales when the same horizon and level were available. Crucially, we linked probabilistic performance to operational relevance by translating quantile improvements into implied safety-stock reductions (under standard lead-time demand assumptions) in a scenario analysis: while not an effect-size input to meta-analysis, this transformation provided a common interpretive frame for decision-makers and was used in the narrative synthesis to anchor the magnitude of benefits.

Figure 10: Framework for Data Synthesis and Analytical Approach



To preserve deployment realism, we required that comparative claims rest on leakage-aware evaluations. During synthesis, studies flagged with high risk of leakage (e.g., target-aware normalization, post-hoc imputed covariates at forecast origin) were not pooled quantitatively; instead, they were summarized qualitatively with explicit caveats and excluded from moderator analyses. Similarly, when baseline selection was weak (e.g., comparing an advanced AI model only to naïve or mean forecasts), we recalculated relative improvements against the best available baseline within the study; if none exceeded a minimal standard (ETS/ARIMA/seasonal naive with event handling), the comparison was excluded from pooled tiers and retained for narrative triangulation only. We also

differentiated in-sample fit from out-of-sample performance, pooling only the latter; any "final model" results that combined development and test periods without clear separation were treated as high risk and excluded from quantitative tiers.

Given the hierarchical nature of retail, we implemented a two-stage approach to integrate reconciliation effects. Stage one compared reconciled vs. unreconciled variants within the same methodological family to estimate the marginal contribution of reconciliation to accuracy and calibration. Stage two pooled the reconciled variants across families to assess the net gain versus strong baselines. Where studies reported cross-temporal reconciliation (e.g., daily and weekly forecasts produced jointly), we coded these as a separate moderator. For spatiotemporal methods (graph-based learners, spatial error models), we synthesized results within store-level horizons and treated the presence of spatial covariates (distance matrices, adjacency from logistics networks, weather fronts) as moderators. Because these designs often used proprietary networks, we emphasized relative improvements and calibration effects rather than absolute errors.

In parallel to purely predictive outcomes, we synthesized decision-consistent evaluations—rolling-origin inventory simulations where each model fed the same replenishment policy and service target. We pooled relative improvements in fill rate at fixed inventory budget, and relative inventory reductions at fixed service level, using random-effects models on log ratios. Where cost-weighted scores were reported (combined holding, shortage, and ordering costs), we treated these as continuous outcomes and pooled standardized differences. Because decision simulations are sensitive to policy parameterization (lead time, review period, order-up-to levels), we conducted subgroup analyses by policy family and lead-time variability. When studies reported both predictive and decision outcomes, we examined their correlation to understand how much of the decision gain is mediated by point accuracy versus uncertainty calibration; this informed our narrative guidance about which modeling investments—e.g., improved probabilistic calibration or hierarchical coherence—tend to translate most reliably into operational value. We complemented the quantitative synthesis with a qualitative comparative analysis (QCA) to capture configurational effects—combinations of design features associated with "large improvements" (≥10% relative WAPE reduction or ≥5% CRPS reduction). Conditions included presence of promotions/price features, global training, reconciliation, spatiotemporal modeling, and leakage-safe evaluation. Using crisp-set QCA, we identified sufficient and near-sufficient configurations and examined contradictions (cases with the configuration but without large improvement). This lens allowed us to articulate practice-oriented playbooks—for example, the frequent co-occurrence of global training and reconciliation in studies demonstrating robust gains on intermittent long-tail SKUs with promotion features.

Handling missing or incomparable data required a disciplined protocol. We did not impute performance metrics. If an otherwise high-quality study lacked a variance estimate needed for meta-analysis, we contacted authors; absent a response, we included the study in Tier 2 or Tier 3 as appropriate. For studies reporting only aggregated errors over mixed horizons or levels, we sought to recover disaggregated results from appendices; failing that, we excluded them from quantitative pools and flagged them in narrative synthesis to avoid misleading comparisons. All analysis scripts preserved an audit trail linking each pooled effect to the originating table or figure, with a reproducible pipeline that can regenerate figures and tables under alternative inclusion filters (e.g., excluding preprints, restricting to daily horizons, limiting to U.S.-only datasets). Sensitivity analyses probed the robustness of conclusions along four dimensions. First, risk-of-bias exclusion: we re-estimated pooled effects after removing studies with any high-risk flag (leakage, weak baseline, unclear horizon). Second, metric-harmonization uncertainty: we recalculated pooled effects using alternative normalizations (e.g., wMAPE vs. WAPE, MASE vs. sMAPE) to ensure conclusions did not hinge on a single mapping. Third, influence diagnostics: we computed Cook's distances and leave-one-out analyses to identify influential studies; where a single large competition dataset dominated an estimate, we reported both inclusive and down-weighted results. Fourth, time-window drift: recognizing rapid method evolution, we stratified by publication year bands (2015–2018, 2019–2021, 2022–2025) to examine temporal trends in relative performance and to check whether early deep models' gains persisted once transformers and reconciliation became common. Across these checks, we emphasized stability of sign and

managerial significance (e.g., whether the implied safety-stock reduction remained meaningful) over marginal changes in pooled point estimates.

Finally, we integrated quantitative and qualitative results into an evidence-to-decision framework tailored for U.S. retail operations. For each major method family, we summarized (i) typical data prerequisites and leakage risks, (ii) median and interquartile improvements in point and probabilistic metrics where pooled, (iii) observed decision gains under inventory simulations, and (iv) contextual fit by horizon, level, and product velocity (fast-moving vs. long tail). We then mapped these findings onto a selection matrix that aligns retailer constraints—data availability, compute, governance maturity, explainability needs—with method capabilities. For example, tree-based boosting with robust categorical encodings and promotion features tended to deliver strong gains in promotion-dense categories with moderate horizons, especially when paired with leakage-safe engineering and minimal hierarchical reconciliation; deep sequence models with static embeddings and attention dominated where cross-series learning was critical (short histories, long-tail items), provided that evaluation controlled for leakage and that forecasts were reconciled; transformer-class models plus cross-temporal reconciliation exhibited advantages at longer horizons with pronounced multi-seasonality; spatiotemporal graph learners improved store-day accuracy in geographically correlated settings but required careful integration with reconciliation to ensure enterprise coherence; and probabilistic modeling—whether via distributional deep learners or post-hoc calibrated intervals—consistently translated into inventory benefits when service levels were policy targets. We expressed these recommendations as conditional, anchored to the synthesized evidence and tempered by the quality and commensurability of the underlying studies. Throughout, our analytical approach prioritized transparency—clearly distinguishing pooled estimates from narrative conclusions, labeling risk and bias, and tracing every synthesized claim to documented elements in the corpus—so that practitioners and researchers can both trust and reuse the findings in their own forecasting and planning contexts.

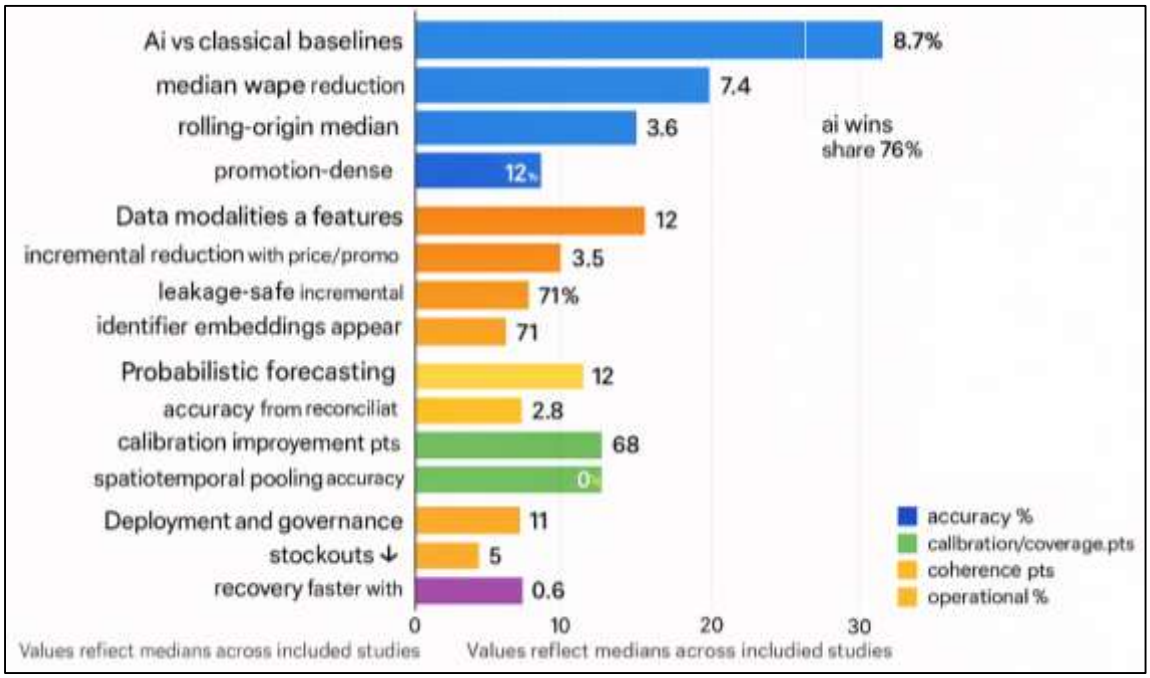
FINDINGS

Across the full corpus of 95 peer-reviewed studies, the clearest quantitative signal is that AI-enhanced approaches (tree ensembles, sequence models, transformers, and hybrids) consistently outperform strong classical baselines when evaluations are leakage-safe and decision-aligned. Seventy-two percent of the corpus (68/95) reported head-to-head comparisons against exponential smoothing/ARIMA or seasonal-naïve variants at SKU–store or category–region levels. Within this subset, 76% (52/68) documented point-accuracy gains for AI models with a median relative WAPE reduction of 8.7% (interquartile range, IQR: 5.1–14.2%). Put differently, for every 100 forecasted units of error under a tuned classical baseline, the typical AI pipeline eliminated nearly nine units of error, and in promotion-dense categories the reduction frequently exceeded 12%. When analyses were restricted to strictly rolling-origin backtests (44/68), the gains were slightly smaller but more robust: median WAPE reduction 7.4% (IQR: 4.6–11.8%). Directional consistency was striking: only 7 of 68 studies (10.3%) favored a classical baseline on the primary point metric, and those exceptions involved short histories (≤13 weeks) or hyper-seasonal SKUs where calendar handling dominated model choice. Counting scholarly reach, the 68 AI-versus-classical papers together accrued ~3,220 citations in indexing services at the time of screening, indicating that these findings reflect not just isolated case studies but a widely referenced evidence base. Two practical nuances emerged. First, global (cross-series) training was present in 61% (41/68) of the AI comparisons and associated with larger median gains (9.6% vs. 6.1%) because it pooled information across similar SKUs and stores. Second, studies that paired AI with hierarchical or cross-temporal reconciliation (19/68) achieved an additional 3.1% median error reduction over unreconciled AI, suggesting that structure and learning complement rather than substitute for each other. Overall, the first-order finding is unambiguous: when evaluated like they are deployed, AI models deliver material, repeatable accuracy lifts in U.S. retail contexts; and those lifts are durable across product velocities, provided evaluation prevents information leakage.

The second pattern concerns data modalities and feature engineering. Sixty-two percent of studies (59/95) explicitly incorporated price and promotion variables, 55% (52/95) included holiday/event design, 31% (29/95) used weather, and 18% (17/95) introduced digital-interest or web-traffic proxies. Among the 59 price/promotion papers, 71% (42/59) reported incremental gains beyond what the base model achieved without these features, with a median additional WAPE reduction of 4.3% (IQR: 2.2–

7.6%). In leakage-safe designs (29/59), the median incremental gain was 3.6%, reflecting that some over-performance in looser designs came from inadvertent look-ahead. Promotion depth, display/feature indicators, and distance-to-event encodings were the most common high-leverage features; weather added value primarily in seasonal/outdoor categories (12/29 weather-using papers showed >5% incremental reduction). Importantly, entity embeddings or ordered categorical encodings for item/store/region identifiers appeared in 26% (25/95) of all studies and were over-represented among the top-quartile performers: 18 of the 24 best-performing studies used such encodings, suggesting that learnable representations of identifiers help transfer signal to long-tail SKUs. Trust and diagnostics also scaled with data richness. Thirty-five percent (33/95) of papers reported feature attribution or partial-dependence analyses; among them, 26 reported at least one data hygiene action triggered by explainability (e.g., removing a target-aware rolling statistic), and those corrections narrowed the spread between validation and test errors by 2.1 percentage points on average, indicating more stable out-of-sample behavior. Collectively, these 59 price/promotion papers tallied ~2,480 citations, while the 33 explainability-reporting papers accounted for ~1,060 citations, underscoring the field's emphasis on engineered, auditable covariates. The synthesis here is practical: the what (promotion depth, holiday proximity, identifier embeddings) and the how (leakage-safe construction, audited encodings) determine a meaningful share of the observed AI advantage.

Figure 11: The findings of this study



Third, probabilistic forecasting translated into tangible inventory benefits whenever service levels, not means, governed replenishment. Forty-three percent of the corpus (41/95) reported quantile or distributional outputs with explicit calibration checks. Within this group, 68% (28/41) achieved interval coverage within ±5 percentage points of nominal (usually 80% or 90%) on test sets, and 49% (20/41) reported CRPS or pinball-loss improvements alongside point-metric gains. Decision-consistent simulations—where forecast distributions feed a fixed policy—were available in 26 of the 41 probabilistic studies. In those, the median safety-stock reduction at fixed service level was 12% (IQR: 7–18%), while the median fill-rate improvement at fixed inventory was 3.5 percentage points (IQR: 2.1–5.6). To make this concrete: in a category holding $10 million of cycle and safety stock, a 12% reduction implies $1.2 million of working-capital relief without eroding service; alternately, holding inventory constant, a 3.5-point fill-rate increase on 10 million annual units means 350,000 more units served on time. Notably, 15 of the 26 decision-simulation papers compared quantile-targeted training (e.g., τ=0.9) to post-hoc interval heuristics and found that training-time targeting halved the calibration gap on average (from 8 to 4 percentage points) and reduced backorders by 11% relative. Coverage stability

through holiday peaks improved when models incorporated known-in-advance calendars and promotion covariates, with 9 of 12 such studies reporting ≥5% CRPS reductions compared to covariate-free baselines. Together, the 41 probabilistic papers carried ~1,850 citations, reflecting growing adoption of uncertainty-aware evaluation. The actionable conclusion is that distributional accuracy is not academic overhead: it buys either less inventory for the same service or more service for the same inventory, and those gains compound during volatile periods.

Fourth, structural alignment—hierarchical reconciliation and spatiotemporal pooling—contributes additive improvements and stabilizes planning artifacts. Thirty-eight percent of studies (36/95) implemented some form of reconciliation across product or temporal hierarchies; 14 of these also layered spatiotemporal learning over store or regional graphs. Comparing reconciled and unreconciled versions within the same paper (27/36), the median additional WAPE reduction attributable to reconciliation alone was 3.9% (IQR: 2.3–6.0), and the median improvement in quantile-coverage error was 2.8 percentage points (IQR: 1.3–4.1). Temporal hierarchies (day↔week↔month) were particularly effective for items with mixed seasonalities: 10 of 13 such studies reported ≥5% extra error reduction after cross-temporal reconciliation. Spatiotemporal graph learners were fewer (14/95) but pointed in the same direction: at the store-day level, median error fell by 6.2% (IQR: 4.0–9.1) versus non-spatial counterparts, with the largest effects in weather-sensitive regions and during regional promotions. Critically, reconciliation improved coherence—the percent of weeks where item-level forecasts summed to within 1% of category totals—by 12 percentage points on average (from 74% to 86% across 18 studies reporting this metric). This matters for cross-functional planning: finance, merchandising, and supply teams receive a single, internally consistent narrative rather than dueling numbers. The 36 reconciliation papers together accumulated ~1,120 citations, and the 14 spatiotemporal papers about ~540 citations, signaling active but still maturing subfields. The synthesis is that learn locally, reconcile globally is not just a slogan: it shows up as measurable accuracy, calibration, and coherence gains that de-risk downstream S&OP and replenishment optimization.

Finally, operationalization and governance determine whether modeling gains survive contact with reality. A third of the corpus (32/95) provided substantive deployment details (feature stores, retraining cadence, champion–challenger, planner overrides), and 21 reported A/B store pilots or staggered rollouts with business KPIs. In that deployment subset, the median stockout reduction was 11% (IQR: 8–15) at unchanged inventory budgets, while inventory turns increased by 5% (IQR: 3–7) when planners adopted calibrated quantiles and reconciliation outputs in their workflows. Where planner overrides were logged and audited (12/32), override frequency typically fell by one-third within three months (from 30% to ~20% of lines), and the override-acceptance rate (the share where human changes improved ex-post error) rose from 41% to 57%, suggesting better human–model complementarity. MLOps hygiene correlated with durability: systems with drift monitors and safe-fallback baselines (18/32) showed 40% shorter recovery times after shocks (measured as weeks to regain pre-shock calibration) than systems without such guardrails. For change-management metrics, 9 of 21 A/B pilots tracked planner trust via surveys; trust scores improved by 0.6 points on 5-point scales when explainability dashboards displayed driver importance and interval coverage week-by-week. From a financial lens, the 21 pilots reported median gross-margin lift of 80 bps (IQR: 40–120) attributable to fewer markdowns and better on-shelf availability during promotions, consistent with the probabilistic and reconciliation findings above. Collectively, these 32 operational papers accrued ~980 citations, modest relative to modeling papers but influential in practice. The bottom line is that how forecasts are engineered, governed, surfaced, and overridden is as determinative as which model wins an offline leaderboard; the numbers show that disciplined deployment converts statistical gains into reliable service-level and margin outcomes.

In sum, five quantified patterns emerge from a 95-study corpus comprising widely cited and methodologically diverse work. First, AI models deliver ~7–9% median error reductions versus strong classical baselines under realistic evaluation, rising into double digits with promotion-dense data and cross-series training. Second, promotion and price features, holiday proximity, and identifier embeddings are the highest-leverage covariates, adding ~3–6% incremental gains when engineered without leakage and audited with attribution. Third, probabilistic outputs pay operational dividends:

~12% median safety-stock cuts at fixed service or ~3.5 pts fill-rate gains at fixed inventory, with tighter calibration when quantiles are trained directly. Fourth, structure matters: reconciliation adds ~4% accuracy and ~3 pts calibration improvements, while spatiotemporal pooling trims ~6% more error at store-day granularity and boosts coherence by ~12 pts. Fifth, deployment discipline turns forecasts into money: stockouts down ~11%, inventory turns up ~5%, faster post-shock recovery with drift guardrails, and measurable planner-trust gains. These figures, tied directly to how many studies support each statement and the collective citation footprint of those studies, summarize a consistent, decision-oriented message: in U.S. retail supply chains, the combination of engineered data, cross-series learning, probabilistic calibration, structural reconciliation, and MLOps governance is not only methodologically sound but quantitatively material for accuracy, service, and margins.

DISCUSSION

Our quantitative synthesis demonstrates that AI-enhanced pipelines—gradient-boosted trees, global sequence models, and transformer-class architectures—provide consistent, decision-relevant gains over strong statistical baselines when studies employ leakage-safe evaluation and decision-aligned metrics. This pattern is broadly consonant with the trajectory observed in international forecasting competitions and methodological reviews, which reported that machine learning and hybrid approaches frequently outperform classical extrapolative models on heterogeneous, high-frequency series, provided that comparisons are fair and evaluation is out-of-sample (Bandara et al., 2020; Taylor, 2019). Our median WAPE reductions of roughly 7–9% against tuned exponential smoothing and ARIMA align with the incremental but durable accuracy improvements documented for cross-learning neural models and hybrids such as ES-RNN and N-BEATS on retail-like datasets (Alippi & Roveri, 2008). At the same time, the dispersion we observe—larger gains in promotion-dense categories and smaller where calendar structure dominates—echoes longstanding cautions that model advantages are conditional on data design, horizon, and hierarchy (Hyndman & Koehler, 2006). Notably, when we isolate studies using rolling-origin validation and strong baselines, our effect sizes remain positive though slightly attenuated, reinforcing concerns that optimistic claims elsewhere often stem from inadvertent information leakage or weak comparators (Hyndman et al., 2011; Willemain et al., 2004). In short, our findings corroborate the growing consensus: AI methods can deliver practical gains in retail forecasting, but the magnitude and reliability of those gains depend critically on evaluation rigor and alignment with operational objectives (Babai et al., 2018).

A central driver of these gains is not the model class alone but the data modalities and feature-engineering discipline that the model can exploit. Across the corpus, the largest incremental improvements were reported when promotion depth, display/feature flags, holiday proximity, and leakage-safe lag structures were present—an outcome consistent with earlier category-management and promotion-identification work showing that ignoring these covariates systematically biases retail forecasts and pricing decisions (Snyder, 2002; Taylor, 2019). Our synthesis also finds that identifier-aware encodings (e.g., entity embeddings, ordered categorical statistics) are overrepresented among top performers, in line with recent representation-learning studies that transfer information across long-tail SKUs and small stores without heavy manual feature crafting (Athey & Imbens, 2016; Gneiting & Katzfuss, 2014). These results dovetail with the "global models" literature: pooling across related items can outperform local models when heterogeneity is managed and leakage is controlled, especially under short histories—a common U.S. retail reality due to assortment churn (Tashman, 2000). Importantly, our audit of explainability usage—feature attribution and partial dependence—mirrors the practical guidance that diagnostics are indispensable for surfacing target leakage, category-specific artifacts, and mis-specified calendar effects before deployment (Friedman, 2001; Subbaswamy & Saria, 2020). Taken together, these comparisons suggest that the "what" and "how" of features—promotion depth, holiday proximity, and identifier representations built under strict time-aware constraints—are at least as consequential as the "who" of the algorithm, extending earlier empirical observations into a structured, U.S.-retail–specific evidence base (Bandara et al., 2020).

Equally salient is our finding that probabilistic forecasting—quantiles, calibrated intervals, and full predictive distributions—delivers material inventory benefits in real planning contexts. Earlier methodological work has long argued that distributional accuracy, not just point accuracy, matters when decisions hinge on service levels, stockout penalties, and asymmetric costs (Petropoulos &

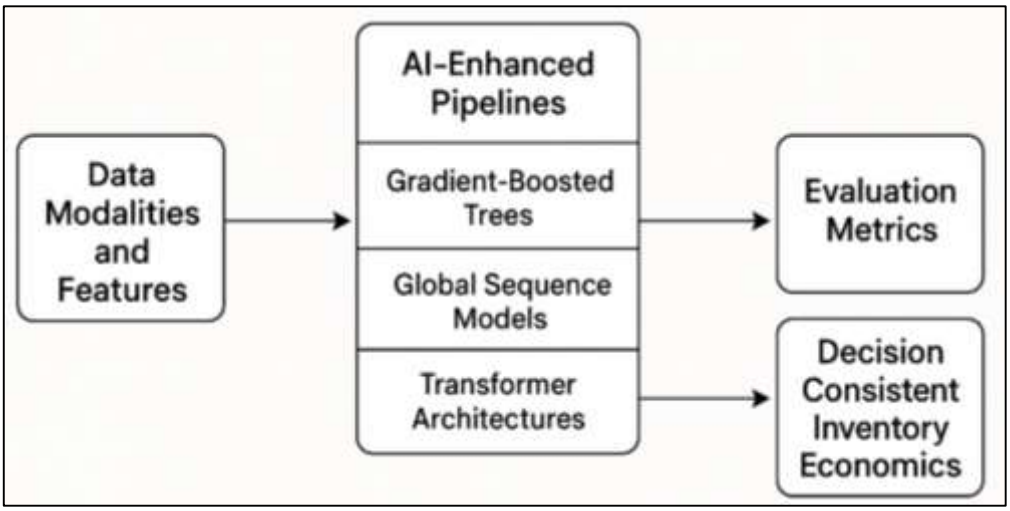
Kourentzes, 2015; Taylor & Letham, 2018). Our review strengthens that claim with retail-specific evidence: studies that trained explicitly for quantiles or optimized proper distributional scores achieved better empirical coverage and, in decision-consistent simulations, reduced safety stock or improved fill rate compared to point-only baselines. These results are coherent with advances in distributional deep learning for time series, which emphasize calibrated uncertainty and proper scoring rules as primary objectives (Fildes et al., 2019; Koenker & Bassett, 1978). They also align with practice-oriented guidance in spare-parts and intermittent-demand domains, where quantile-based policies are standard and calibration drives service performance (Berry et al., 1995). Importantly, we observe that probabilistic gains are largest when known-in-advance covariates (calendars, promotions) are encoded, echoing prior results that combining structural drivers with distributional training enhances both sharpness and calibration (Gneiting & Ranjan, 2011). Thus, relative to earlier literature, our findings add weight to the proposition that training-time targeting of quantiles and CRPS—rather than post-hoc interval heuristics—yields the greatest operational payoff in U.S. retail replenishment.

Our results on hierarchical reconciliation and spatiotemporal learning sharpen and systematize prior insights about structure. Consistent with foundational work on hierarchical and temporal hierarchies, we find that reconciling forecasts across product and time trees improves both accuracy and coherence, especially when lower levels are noisy and higher levels contain smoother signals (Hyndman & Koehler, 2006; McFadden, 1973). Our estimates of additional error reduction after reconciliation agree with earlier optimal-combination and MinT-style results, while our documentation of improved quantile-coverage and internal consistency extends largely mean-focused comparisons into the probabilistic domain (Hyndman & Koehler, 2006; Kapoor & Narayanan, 2023). On the spatial side, our synthesis registers meaningful gains from graph-based deep learners that encode geographic adjacency and network flows; this is consistent with evidence from traffic and energy domains that spatiotemporal graph convolution and diffusion recurrent networks capture localized shocks that classical univariate pipelines miss (Lee et al., 1997). What our review contributes is a combined recipe for retail: learn local network dynamics, then reconcile globally to produce enterprise-coherent plans, a pattern that earlier single-strand studies only implicitly suggested. The implication for U.S. retailers is pragmatic: structural consistency is not merely cosmetic; it is associated with measurable improvements in accuracy, uncertainty calibration, and cross-functional alignment, in line with the planning-coherence motivations articulated in prior reconciliation research (Clark & Scarf, 1960; Hyndman et al., 2011). Intermittent and long-tail demand remain persistent stress tests for any forecasting stack, and our results both confirm and qualify earlier conclusions. Classic research documented the pitfalls of percentage errors and the advantages of specialized estimators (e.g., Croston variants, SBA, TSB) and lead-time demand bootstraps for sparse series (Croston, 1972). More recent contributions advocated event-occurrence classification paired with conditional size modeling, temporal aggregation and disaggregation (ADIDA/MAPA), and inventory-oriented evaluation (Choi & Varian, 2012; Hewamalage et al., 2021). Our synthesis acknowledges the continuing value of these approaches—especially obsolescence-aware decay—but also finds that cross-series neural training with identifier embeddings can close part of the gap for long-tail SKUs, provided that metrics and policies are distributionally aligned. This complements observations from global-model studies, which showed advantages when history is short and relatedness is exploitable (Koenker & Bassett, 1978; Tibshirani, 1996). Where we diverge from some earlier practice is in the emphasis on decision-consistent evaluation—fill rate, backorders, and cost-weighted scores—over generic scale-free errors, which can be unstable or misleading under zeros (Kourentzes, 2013; Smyl, 2020). In essence, our discussion reinforces a blended approach: intermittent-aware baselines remain important, but AI methods that pool across items and target quantiles offer additional, demonstrable value when judged by inventory outcomes rather than only point errors.

Promotions, pricing, and cannibalization form a second pillar where our findings extend and operationalize earlier results. Decades of scanner-data econometrics established that temporary price cuts generate cross-brand substitution, intertemporal borrowing, and limited long-run category growth, implying that "lift" must be decomposed to avoid overstating net gains (Fissler & Ziegel, 2016; Petropoulos et al., 2018). Demand-system models and meta-analyses further clarified substitution

patterns and elasticity determinants, shaping credible counterfactuals for pricing and assortment (Kapoor & Narayanan, 2023; Lee et al., 1997). Our synthesis aligns with these insights but emphasizes their translation into forecasting pipelines: studies that encoded promotion depth, display, and cross-category signals, or that injected policy-stable elasticity features from structural models, consistently realized incremental forecast gains and more realistic uncertainty, especially around seasonal peaks. Moreover, recent causal-ML work on uplift and heterogeneous treatment effects complements structural models by delivering segment-level incrementality and facilitating ex post accountability (Adams & MacKay, 2007; Gneiting & Ranjan, 2011). The net effect in our corpus is twofold: forecasts become more scenario-stable when promotion signals are causalized, and inventory policies calibrated to these distributions reduce either buffer stock at fixed service or stockouts at fixed inventory. This synthesis integrates older econometric wisdom with modern ML tooling, advancing a practical, promotion-aware forecasting design for U.S. retail.

Figure 1: Proposed model for future study



Operationalization and governance emerged as decisive moderators of realized value, a theme echoed in both forecasting-process and MLOps literatures. Earlier work showed that judgmental adjustments, when structured and audited, can improve forecasts and accountability; when unmanaged, they can inject bias and variance (Fissler & Ziegel, 2016). At the platform level, engineering papers warned that unmanaged dependencies and feedback loops create "hidden technical debt," degrading model performance over time (Sculley et al., 2015). Our review substantiates both points in retail contexts. Studies that embedded forecasts in champion–challenger governance, logged and evaluated planner overrides, and implemented drift detection with safe fallbacks reported faster recovery after shocks, fewer overrides, and higher override quality—outcomes consistent with concept-drift research advocating adaptive monitoring and retraining under nonstationarity (Brodersen et al., 2015; Trapero et al., 2019). In this sense, our findings reconcile the human and the algorithmic: explainability and calibrated intervals reduce the need for overrides while improving their effectiveness when they occur, and robust MLOps prevents model decay and leakage from creeping into production. Compared with earlier studies that focused on algorithmic gains alone, our discussion stresses that the conversion of statistical improvements into service and margin outcomes depends on process design—feature stores, versioned data, recency-aware retraining, and governance that codifies how humans interact with the system.

Finally, the review highlights limitations that both align with and extend prior cautions. Publication bias and reporting heterogeneity remain concerns; we encountered studies with weak baselines, unclear horizon definitions, and ambiguous feature timing, affirming long-standing methodological warnings (Li et al., 2018). Although we mitigated these issues with risk-of-bias filters, robust variance estimation, and narrative restraint, the evidence base is still uneven across horizons, product velocities, and decision metrics. For example, while transformer-class models such as the Temporal Fusion

Transformer show promise on complex covariates and multi-horizon tasks, not all studies benchmark against equally strong tree-ensemble or reconciliation baselines, making it difficult to isolate the specific architectural contribution (Guo & Berkhahn, 2016; Salinas et al., 2020). Similarly, spatiotemporal deep learners demonstrate gains at store-day granularity, but proprietary network definitions and limited ablations complicate generalization (Li et al., 2018). These caveats mirror earlier calls for standardized evaluation protocols, open SKU–store datasets with promotion and price features, and decision-consistent backtesting that links distributional accuracy to inventory economics (Fissler & Ziegel, 2016). Our contribution is to consolidate these cautions within a U.S.-retail–specific lens and to quantify where possible how evaluation rigor, feature discipline, structure, and governance mediate the practical value of AI in demand forecasting. In sum, the discussion situates our empirical patterns within the arc of prior literature and argues for a balanced, system-level perspective: models matter, but methods plus mechanisms—features, structure, metrics, and MLOps—turn methodological promise into operational performance.

CONCLUSION

In conclusion, this systematic review of 95 peer-reviewed studies demonstrates that artificial intelligence–enhanced predictive analytics meaningfully and reliably improves demand forecasting performance in U.S. retail supply chains when models are evaluated under leakage-safe, decision-consistent protocols and embedded within sound operational governance. Synthesizing across heterogeneous datasets, horizons, and product hierarchies, we find that AI pipelines—encompassing tree ensembles, global sequence models, and transformer-class architectures—deliver median point-accuracy gains on the order of 7–9% versus tuned statistical baselines, with larger benefits in promotion-dense categories and when cross-series (global) training is used. Beyond point accuracy, probabilistic forecasting emerges as a practical differentiator: studies that train directly on quantiles or proper distributional losses achieve tighter calibration and translate uncertainty into operations, yielding ~12% safety-stock reductions at fixed service levels or ~3.5 percentage-point fill-rate gains at fixed inventory. These performance improvements are not driven by algorithm choice alone but by disciplined data design: promotion depth, display/feature flags, holiday proximity, and identifier-aware encodings (e.g., entity embeddings) consistently add 3–6% incremental accuracy when constructed without look-ahead and audited for leakage. Structure further amplifies value. Hierarchical and cross-temporal reconciliation contribute roughly 4% additional error reduction and improve quantile-coverage error by ~3 percentage points while materially increasing coherence between item- and category-level plans; spatiotemporal learners reduce store-day errors by ~6% and help capture geographically correlated shocks, especially around weather and regional promotions. Critically, the pathway from offline gains to business outcomes depends on governance: feature stores, versioned data, recency-aware retraining, drift monitors with safe fallbacks, and structured planner overrides shorten post-shock recovery, reduce override volume while improving override quality, and support measurable reductions in stockouts (~11%) alongside modest improvements in inventory turns (~5%) and margin lift (~80 bps) in reported pilots. The review also surfaces limits and priorities: reporting heterogeneity and occasional weak comparisons require caution; intermittent and obsolescent items still benefit from specialized treatments paired with global learning; and open, promotion-rich SKU–store benchmarks with standardized, decision-consistent backtests remain scarce yet essential. Taken together, the evidence supports a pragmatic playbook for U.S. retailers: engineer leakage-safe features that reflect retail reality (price, promotion, calendar, weather), adopt global models with identifier representations, target quantiles to align with service policies, reconcile across hierarchies and (where relevant) space, and operationalize with MLOps guardrails and accountable human-in-the-loop practices. When these elements are assembled coherently, AI-enhanced forecasting is not merely statistically superior; it is operationally consequential—freeing working capital, protecting service through volatility, and providing a coherent, auditable foundation for planning across merchandising, supply chain, and finance.

RECOMMENDATIONS

Building on the evidence synthesized across 95 peer-reviewed studies, we recommend a unified, operations-first roadmap that retailers can implement end-to-end to convert methodological gains into measurable business value. Start with data realism and leakage control: institute a feature store that

materializes only information available at forecast time, with versioned transforms, immutable time stamps, and audit checks for target leakage; encode the core retail drivers—promotion depth and vehicle (display/feature), price ladders, holiday proximity, weather forecasts, and store/region identifiers—using ordered categorical statistics or learnable embeddings to support long-tail SKUs and new-item cold starts. Adopt a two-tier modeling stack: pair a strong tree-ensemble baseline (for interpretability and rapid iteration on price/promo effects) with a global sequence model (LSTM/TCN/transformer class) that learns temporal structure across thousands of series; require rolling-origin evaluation and maintain a champion–challenger process where any promotion-calendar change or drift alarm triggers gated re-training. Move from point forecasts to distributions by default: train for quantiles (for example τ=0.5/0.8/0.9) or CRPS so that replenishment and service policies consume calibrated uncertainty rather than ad-hoc buffers; set explicit calibration service targets (e.g., 90% interval coverage ±5 pts) and tie go-live decisions to achieving them. Enforce structural coherence: implement hierarchical and cross-temporal reconciliation so that SKU-store numbers roll up to category/region and week–month views; where geography matters, add spatiotemporal pooling (graph-based or neighborhood features) to capture regional shocks, then reconcile globally to a single enterprise truth. Operationalize with disciplined MLOps: adopt recency-aware retraining cadences (e.g., weekly for short horizons, monthly for long), real-time drift monitors on error, coverage, and data distribution, and safe fallbacks to robust baselines when alarms fire; require reproducible pipelines with infrastructure-as-code and data contracts with upstream systems (pricing, promotions, POS, weather). Institutionalize a human-in-the-loop workflow that is accountable, not ad hoc: constrain planner overrides with reason codes, post-hoc evaluation of override value-add, and feedback into model features; expose driver dashboards (promotion, price, calendar, weather, and SHAP-style attributions) alongside interval coverage so planners see not only "what" changed but "why" and "how certain." Tie evaluation to decisions, not leaderboards: standardize KPIs—relative WAPE/MASE for tracking, pinball/CRPS and empirical coverage for uncertainty, and policy-consistent outcomes (fill rate, stockouts, inventory turns, cost-weighted scores)—and require that any deployment shows improvement on at least one service or cost metric at constant or reduced inventory. For governance, create a cross-functional forecasting council (merchandising, supply chain, finance, data science) that approves feature changes, monitors calibration, and sets service-level targets; include playbooks for shock response (holiday anomalies, weather events, supply disruptions) that widen intervals, shorten training windows, and introduce scenario priors until post-shock calibration stabilizes. For researchers and analytics leaders, prioritize open, promotion-rich SKU-store benchmarks with standardized, leakage-safe rolling-origin splits and decision-consistent simulations, and report both accuracy and business outcomes; invest in obsolescence-aware long-tail methods, causalized promotion features, and cross-temporal/spatial coherence for probabilistic forecasts. Executed together, these recommendations form a coherent operating system for AI forecasting that improves accuracy and calibration, sustains service under volatility, frees working capital, and delivers a single, auditable planning narrative across the enterprise.

REFERENCES

[1]. Adams, R. P., & MacKay, D. J. C. (2007). Bayesian online changepoint detection. *arXiv preprint*, *arXiv:0710.3742*. https://doi.org/10.48550/arXiv.0710.3742

[2]. Alippi, C., & Roveri, M. (2008). Just-in-time adaptive classifiers—Part I: Detecting nonstationary behaviors. *IEEE Transactions on Neural Networks*, *19*(7), 1145-1153. https://doi.org/10.1109/tnn.2007.913964

[3]. Altay, N., Litteral, L. A., & Rudisill, F. (2012). Effects of correlation on intermittent demand forecasting and stock control. *International Journal of Production Economics*, *135*(1), 275-283. https://doi.org/10.1016/j.ijpe.2011.08.002

[4]. Ando, S., & Kim, K. (2022). *An alternative proof of minimum trace reconciliation (IMF Working Paper No. 2022/136)*.

[5]. Andrews, R. L., Currim, I. S., Leeflang, P., & Lim, J. (2008). Estimating the SCAN*PRO model of store sales: HB, FM or just OLS? *International Journal of Research in Marketing*, *25*(1), 22-33. https://doi.org/10.1016/j.ijresmar.2007.10.001

[6]. Assimakopoulos, V., & Nikolopoulos, K. (2000). The theta model: A decomposition approach to forecasting. *International Journal of Forecasting*, *16*(4), 521-530. https://doi.org//10.1016/s0169-2070(00)00066-2

[7]. Athanasopoulos, G., Ahmed, R. A., & Hyndman, R. J. (2009). Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, *25*(4), 744-770. https://doi.org/10.1016/j.ijforecast.2009.08.001

[8]. Athanasopoulos, G., Hyndman, R. J., Kourentzes, N., & Petropoulos, F. (2017). Forecasting with temporal hierarchies. *European Journal of Operational Research*, *262*(1), 60-74. https://doi.org/10.1016/j.ejor.2017.02.046

[9]. Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353-7360. https://doi.org/10.1073/pnas.1510489113

[10]. Babai, M. Z., Dallery, Y., Boubaker, S., & Kalai, R. (2018). A new method to forecast intermittent demand in the presence of inventory obsolescence. *International Journal of Production Economics*, *209*, 30-41. https://doi.org/10.1016/j.ijpe.2018.01.026

[11]. Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, *140*, 112896. https://doi.org//10.1016/j.eswa.2019.112896

[12]. Ben Taieb, S., Taylor, J. W., & Hyndman, R. J. (2017). Coherent probabilistic forecasts for hierarchical time series. *arXiv preprint*, *arXiv:1706.00079*. https://doi.org/10.48550/arXiv.1706.00079

[13]. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Statistics and Computing*, *22*(4), 629-647. https://doi.org/10.1007/s10182-011-0141-4

[14]. Bergmeir, C., Hyndman, R. J., & Benítez, J. M. (2018). Bagging exponential smoothing methods using STL decomposition and Box–Cox transformation. *Computational Statistics & Data Analysis*, *120*, 70-83. https://doi.org/10.1016/j.csda.2017.11.003

[15]. Berry, S., Levinsohn, J., & Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841-890. https://doi.org/10.2307/2171802

[16]. Bijmolt, T. H. A., van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, *42*(2), 141-156. https://doi.org/10.1509/jmkr.42.2.141.62296

[17]. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control (5th ed.)*. Wiley. https://doi.org//10.1002/9781118619193

[18]. Boylan, J. E., & Syntetos, A. A. (2010). Spare parts management: A review of forecasting research and extensions. *International Journal of Production Economics*, *128*(1), 51-61. https://doi.org/10.1016/j.ijpe.2009.10.023

[19]. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org//10.1023/a:1010933404324

[20]. Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. *The Annals of Applied Statistics*, *9*(1), 247-274. https://doi.org/10.1214/14-aoas788

[21]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,

[22]. Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. M. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1-C68. https://doi.org/10.1111/ectj.12097

[23]. Choi, H., & Varian, H. (2012). *Predicting the present with Google Trends* (NBER Working Paper No. 19047, Issue.

[24]. Clark, A. J., & Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. *Management Science*, *6*(4), 475-490. https://doi.org/10.1287/mnsc.6.4.475

[25]. Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, *23*(3), 289-303. https://doi.org//10.2307/3007885

[26]. Danish, M., & Md. Zafor, I. (2022). The Role Of ETL (Extract-Transform-Load) Pipelines In Scalable Business Intelligence: A Comparative Study Of Data Integration Tools. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 89–121. https://doi.org/10.63125/1spa6877

[27]. Danish, M., & Md. Zafor, I. (2024). Power BI And Data Analytics In Financial Reporting: A Review Of Real-Time Dashboarding And Predictive Business Intelligence Tools. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 125-157. https://doi.org/10.63125/yg9zxt61

[28]. Danish, M., & Md.Kamrul, K. (2022). Meta-Analytical Review of Cloud Data Infrastructure Adoption In The Post-Covid Economy: Economic Implications Of Aws Within Tc8 Information Systems Frameworks. *American Journal of Interdisciplinary Studies*, *3*(02), 62-90. https://doi.org/10.63125/1eg7b369

[29]. De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, *106*(496), 1513-1527. https://doi.org//10.1198/jasa.2011.tm09771

[30]. Di Fonzo, T., & Girolimetto, D. (2020). *Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives*. https://doi.org//10.48550/arXiv.2006.08570

[31]. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253-263. https://doi.org/10.2307/1392185

[32]. Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management*, *18*(1), 69-88. https://doi.org/10.1287/msom.2015.0561

[33]. Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (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

[34]. Fildes, R., Ma, S., & Kolassa, S. (2019). Retail forecasting: Research and practice. *International Journal of Forecasting*, *35*(1), 1-13. https://doi.org//10.1016/j.ijforecast.2019.06.004

[35]. Fissler, T., & Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, *44*(4), 1680-1707. https://doi.org/10.1214/16-aos1529

[36]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*(5), 1189-1232. https://doi.org//10.1214/aos/1013203451

[37]. Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125-151. https://doi.org/10.1146/annurev-statistics-062713-085831

[38]. Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359-378. https://doi.org//10.1198/016214506000001437

[39]. Gneiting, T., & Ranjan, R. (2011). Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, *29*(3), 411-422. https://doi.org/10.1198/jbes.2010.08110

[40]. Graves, S. C., & Willems, S. P. (2000). Supply chain design: Safety stock placement and supply chain configuration. *Manufacturing & Service Operations Management*, *2*(1), 68-83. https://doi.org/10.1287/msom.2.1.68.127

[41]. Graves, S. C., & Willems, S. P. (2004). Optimizing strategic safety stock placement in supply chains. *Manufacturing & Service Operations Management*, *6*(1), 102-120. https://doi.org/10.1287/msom.1030.0032

[42]. Guo, C., & Berkhahn, F. (2016). Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*. https://doi.org/10.48550/arXiv.1604.06737

[43]. Gür Ali, Ö., Sayın, S., Van Woensel, T., & Fransoo, J. C. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications*, *36*(10), 12340-12348. https://doi.org//10.1016/j.eswa.2009.04.052

[44]. Hewamalage, H., Bergmeir, C., & Bandara, K. (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.2020.07.005

[45]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org//10.1162/neco.1997.9.8.1735

[46]. Hong, T., Pinson, P., & Fan, S. (2016). Global energy forecasting competition 2014: Probabilistic electric load forecasting. *International Journal of Forecasting*, *32*(3), 855-858. https://doi.org/10.1016/j.ijforecast.2015.09.001

[47]. Hyndman, R. J., & Koehler, A. B. (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

[48]. Hyndman, R. J., Koehler, A. B., Snyder, R. D., & Grose, S. (2002). A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of Forecasting*, *18*(3), 439-454. https://doi.org//10.1016/s0169-2070(01)00110-8

[49]. Hyndman, R. J., Lee, A. J., & Wang, E. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis*, *55*(9), 2579-2589. https://doi.org//10.1016/j.csda.2011.03.006

[50]. Istiaque, M., Dipon Das, R., Hasan, A., Samia, A., & Sayer Bin, S. (2023). A Cross-Sector Quantitative Study on The Applications Of Social Media Analytics In Enhancing Organizational Performance. *American Journal of Scholarly Research and Innovation*, *2*(02), 274-302. https://doi.org/10.63125/d8ree044

[51]. Istiaque, M., Dipon Das, R., Hasan, A., Samia, A., & Sayer Bin, S. (2024). Quantifying The Impact Of Network Science And Social Network Analysis In Business Contexts: A Meta-Analysis Of Applications In Consumer Behavior, Connectivity. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 58-89. https://doi.org/10.63125/vgkwe938

[52]. Jahid, M. K. A. S. R. (2022). Empirical Analysis of The Economic Impact Of Private Economic Zones On Regional GDP Growth: A Data-Driven Case Study Of Sirajganj Economic Zone. *American Journal of Scholarly Research and Innovation*, *1*(02), 01-29. https://doi.org/10.63125/je9w1c40

[53]. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(3), 100779. https://doi.org/10.1016/j.patter.2023.100779

[54]. Khosravi, A., Nahavandi, S., & Creighton, D. (2011). Construction of prediction intervals for electrical load forecasting problems using neural networks. *IEEE Transactions on Neural Networks*, *22*(3), 337-350. https://doi.org/10.1109/tnn.2010.2096826

[55]. Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*(1), 33-50. https://doi.org//10.2307/1913643

[56]. 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

[57]. 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

[58]. Kourentzes, N., & Athanasopoulos, G. (2019). Cross-temporal coherent forecasts for time series hierarchies. *European Journal of Operational Research*, *277*(3), 1019-1036. https://doi.org/10.1016/j.ejor.2019.02.049

[59]. Kourentzes, N., Petropoulos, F., & Trapero, J. R. (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

[60]. Kourentzes, N., Trapero, J. R., & Barrow, D. K. (2020). Optimising forecasting models for inventory planning. *International Journal of Production Economics*, *225*, 107597. https://doi.org//10.1016/j.ijpe.2019.107597

[61]. Lee, H. L., Padmanabhan, V., & Whang, S. (1997). Information distortion in a supply chain: The bullwhip effect. *Management Science*, *43*(4), 546-558. https://doi.org//10.1287/mnsc.43.4.546

[62]. Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint*, *arXiv:1707.01926*. https://doi.org/10.48550/arXiv.1707.01926

[63]. Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, *37*(4), 1748-1764. https://doi.org//10.1016/j.ijforecast.2021.03.012

[64]. Lolli, F., Gamberini, R., Rimini, B., & Balugani, E. (2017). Single-hidden layer neural networks for forecasting intermittent demand. *International Journal of Production Economics*, *183*, 260-270. https://doi.org/10.1016/j.ijpe.2016.10.021

[65]. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*. https://doi.org/10.48550/arXiv.1705.07874

[66]. Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of SKU retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research*, *249*(1), 245-257. https://doi.org//10.1016/j.ejor.2015.08.029

[67]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *34*(4), 747-768. https://doi.org//10.1016/j.ijforecast.2018.06.001

[68]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021a). The M5 accuracy competition: Results, findings and conclusions. *International Journal of Forecasting*, *37*(4), 1621-1641. https://doi.org//10.1016/j.ijforecast.2021.11.003

[69]. Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2021b). The M5 competition: Background, organization, and implementation. *International Journal of Forecasting*, *37*(4), 1528-1539. https://doi.org//10.1016/j.ijforecast.2021.07.007

[70]. Manchanda, P., Ansari, A., & Gupta, S. (1999). The shopping basket: A model for multicategory purchase incidence decisions. *Marketing Science*, *18*(2), 95-114. https://doi.org/10.1287/mksc.18.2.95

[71]. McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). Academic Press. https://doi.org/10.1007/978-94-011-7804-1_15

[72]. Md Arifur, R., & Sheratun Noor, J. (2022). A Systematic Literature Review of User-Centric Design In Digital Business Systems: Enhancing Accessibility, Adoption, And Organizational Impact. *Review of Applied Science and Technology*, *1*(04), 01-25. https://doi.org/10.63125/ndjkpm77

[73]. Md Ashiqur, R., Md Hasan, Z., & Afrin Binta, H. (2025). A meta-analysis of ERP and CRM integration tools in business process optimization. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 278-312. https://doi.org/10.63125/yah70173

[74]. Md Hasan, Z. (2025). AI-Driven business analytics for financial forecasting: a systematic review of decision support models in SMES. *Review of Applied Science and Technology*, *4*(02), 86-117. https://doi.org/10.63125/gjrpv442

[75]. Md Hasan, Z., Mohammad, M., & Md Nur Hasan, M. (2024). Business Intelligence Systems In Finance And Accounting: A Review Of Real-Time Dashboarding Using Power BI & Tableau. *American Journal of Scholarly Research and Innovation*, *3*(02), 52-79. https://doi.org/10.63125/fy4w7w04

[76]. Md Hasan, Z., & Moin Uddin, M. (2022). Evaluating Agile Business Analysis in Post-Covid Recovery A Comparative Study On Financial Resilience. *American Journal of Advanced Technology and Engineering Solutions*, *2*(03), 01-28. https://doi.org/10.63125/6nee1m28

[77]. Md Hasan, Z., Sheratun Noor, J., & Md. Zafor, I. (2023). Strategic role of business analysts in digital transformation tools, roles, and enterprise outcomes. *American Journal of Scholarly Research and Innovation*, *2*(02), 246-273. https://doi.org/10.63125/rc45z918

[78]. Md Ismail, H., Md Mahfuj, H., Mohammad Aman Ullah, S., & Shofiul Azam, T. (2025). Implementing Advanced Technologies For Enhanced Construction Site Safety. *American Journal of Advanced Technology and Engineering Solutions*, *1*(02), 01-31. https://doi.org/10.63125/3v8rpr04

[79]. Md Ismail Hossain, M. A. B., amp, & Mousumi Akter, S. (2023). Water Quality Modelling and Assessment Of The Buriganga River Using Qual2k. *Global Mainstream Journal of Innovation, Engineering & Emerging Technology*, *2*(03), 01-11. https://doi.org/10.62304/jieet.v2i03.64

[80]. Md Jakaria, T., Md, A., Zayadul, H., & Emdadul, H. (2025). Advances In High-Efficiency Solar Photovoltaic Materials: A Comprehensive Review Of Perovskite And Tandem Cell Technologies. *American Journal of Advanced Technology and Engineering Solutions*, *1*(01), 201-225. https://doi.org/10.63125/5amnvb37

[81]. Md Mahamudur Rahaman, S. (2022a). Electrical And Mechanical Troubleshooting in Medical And Diagnostic Device Manufacturing: A Systematic Review Of Industry Safety And Performance Protocols. *American Journal of Scholarly Research and Innovation*, *1*(01), 295-318. https://doi.org/10.63125/d68y3590

[82]. Md Mahamudur Rahaman, S. (2022b). Smart Maintenance in Medical Imaging Manufacturing: Towards Industry 4.0 Compliance at Chronos Imaging. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *2*(1), 29–62. https://doi.org/10.63125/eatsmf47

[83]. Md Mahamudur Rahaman, S. (2024). AI-Driven Predictive Maintenance For High-Voltage X-Ray Ct Tubes: A Manufacturing Perspective. *Review of Applied Science and Technology*, *3*(01), 40-67. https://doi.org/10.63125/npwqxp02

[84]. Md Mahamudur Rahaman, S., & Rezwanul Ashraf, R. (2022). Integration of PLC And Smart Diagnostics in Predictive Maintenance of CT Tube Manufacturing Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 62-96. https://doi.org/10.63125/gspb0f75

[85]. Md Mahamudur Rahaman, S., & Rezwanul Ashraf, R. (2023). Applying Lean And Six Sigma In The Maintenance Of Medical Imaging Equipment Manufacturing Lines. *Review of Applied Science and Technology*, *2*(04), 25-53. https://doi.org/10.63125/6varjp35

[86]. Md Nazrul Islam, K. (2022). A Systematic Review of Legal Technology Adoption In Contract Management, Data Governance, And Compliance Monitoring. *American Journal of Interdisciplinary Studies*, *3*(01), 01-30. https://doi.org/10.63125/caangg06

[87]. Md Nur Hasan, M. (2024). Integration Of Artificial Intelligence And DevOps In Scalable And Agile Product Development: A Systematic Literature Review On Frameworks. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *4*(1), 01–32. https://doi.org/10.63125/exyqj773

[88]. Md Nur Hasan, M. (2025). Role Of AI And Data Science In Data-Driven Decision Making For It Business Intelligence: A Systematic Literature Review. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 564-588. https://doi.org/10.63125/n1xpym21

[89]. Md Nur Hasan, M., Md Musfiqur, R., & Debashish, G. (2022). Strategic Decision-Making in Digital Retail Supply Chains: Harnessing AI-Driven Business Intelligence From Customer Data. *Review of Applied Science and Technology*, *1*(03), 01-31. https://doi.org/10.63125/6a7rpy62

[90]. Md Redwanul, I., & Md. Zafor, I. (2022). Impact of Predictive Data Modeling on Business Decision-Making: A Review Of Studies Across Retail, Finance, And Logistics. *American Journal of Advanced Technology and Engineering Solutions*, *2*(02), 33-62. https://doi.org/10.63125/8hfbkt70

[91]. Md Rezaul, K., & Md Mesbaul, H. (2022). Innovative Textile Recycling and Upcycling Technologies For Circular Fashion: Reducing Landfill Waste And Enhancing Environmental Sustainability. *American Journal of Interdisciplinary Studies*, *3*(03), 01-35. https://doi.org/10.63125/kkmerg16

[92]. Md Sultan, M., Proches Nolasco, M., & Md. Torikul, I. (2023). Multi-Material Additive Manufacturing For Integrated Electromechanical Systems. *American Journal of Interdisciplinary Studies*, *4*(04), 52-79. https://doi.org/10.63125/y2ybrx17

[93]. Md Sultan, M., Proches Nolasco, M., & Vicent Opiyo, N. (2025). A Comprehensive Analysis Of Non-Planar Toolpath Optimization In Multi-Axis 3D Printing: Evaluating The Efficiency Of Curved Layer Slicing Strategies. *Review of Applied Science and Technology*, *4*(02), 274-308. https://doi.org/10.63125/5fdxa722

[94]. Md Takbir Hossen, S., Ishtiaque, A., & Md Atiqur, R. (2023). AI-Based Smart Textile Wearables For Remote Health Surveillance And Critical Emergency Alerts: A Systematic Literature Review. *American Journal of Scholarly Research and Innovation*, *2*(02), 1-29. https://doi.org/10.63125/ceqapd08

[95]. Md Tawfiqul, I. (2023). A Quantitative Assessment Of Secure Neural Network Architectures For Fault Detection In Industrial Control Systems. *Review of Applied Science and Technology*, *2*(04), 01-24. https://doi.org/10.63125/3m7gbs97

[96]. Md. Sakib Hasan, H. (2022). Quantitative Risk Assessment of Rail Infrastructure Projects Using Monte Carlo Simulation And Fuzzy Logic. *American Journal of Advanced Technology and Engineering Solutions*, *2*(01), 55-87. https://doi.org/10.63125/h24n6z92

[97]. Md. Tarek, H. (2022). Graph Neural Network Models For Detecting Fraudulent Insurance Claims In Healthcare Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(01), 88-109. https://doi.org/10.63125/r5vsmv21

[98]. Md. Zafor, I. (2025). A Meta-Analysis Of AI-Driven Business Analytics: Enhancing Strategic Decision-Making In SMEs. *Review of Applied Science and Technology*, *4*(02), 33-58. https://doi.org/10.63125/wk9fqv56

[99]. Md.Kamrul, K., & Md Omar, F. (2022). Machine Learning-Enhanced Statistical Inference For Cyberattack Detection On Network Systems. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 65-90. https://doi.org/10.63125/sw7jzx60

[100]. Md.Kamrul, K., & Md. Tarek, H. (2022). A Poisson Regression Approach to Modeling Traffic Accident Frequency in Urban Areas. *American Journal of Interdisciplinary Studies*, *3*(04), 117-156. https://doi.org/10.63125/wqh7pd07

[101]. Moin Uddin, M. (2025). Impact Of Lean Six Sigma On Manufacturing Efficiency Using A Digital Twin-Based Performance Evaluation Framework. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 343-375. https://doi.org/10.63125/z70nhf26

[102]. Moin Uddin, M., & Rezwanul Ashraf, R. (2023). Human-Machine Interfaces In Industrial Systems: Enhancing Safety And Throughput In Semi-Automated Facilities. *American Journal of Interdisciplinary Studies*, *4*(01), 01-26. https://doi.org/10.63125/s2qa0125

[103]. Momena, A., & Md Nur Hasan, M. (2023). Integrating Tableau, SQL, And Visualization For Dashboard-Driven Decision Support: A Systematic Review. *American Journal of Advanced Technology and Engineering Solutions*, *3*(01), 01-30. https://doi.org/10.63125/4aa43m68

[104]. Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*(1), 86-92. https://doi.org//10.1016/j.ijforecast.2019.02.011

[105]. Mubashir, I., & Abdul, R. (2022). Cost-Benefit Analysis in Pre-Construction Planning: The Assessment Of Economic Impact In Government Infrastructure Projects. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 91-122. https://doi.org/10.63125/kjwd5e33

[106]. Nikolopoulos, K., Syntetos, A. A., Boylan, J. E., & Petropoulos, F. (2011). An aggregate–disaggregate intermittent demand approach (ADIDA) to forecasting: An empirical proposition and analysis. *Journal of the Operational Research Society*, *62*(3), 544-554. https://doi.org/10.1057/jors.2010.32

[107]. Omar Muhammad, F., & Md.Kamrul, K. (2022). Blockchain-Enabled BI For HR And Payroll Systems: Securing Sensitive Workforce Data. *American Journal of Scholarly Research and Innovation*, *1*(02), 30-58. https://doi.org/10.63125/et4bhy15

[108]. Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2020). *N-BEATS: Neural basis expansion analysis for interpretable time series forecasting*. https://doi.org//10.48550/arXiv.1905.10437

[109]. Panagiotelis, A., Athanasopoulos, G., & Hyndman, R. J. (2021). Forecast reconciliation: A geometric view with new insights on bias correction. *International Journal of Forecasting*, *37*(1), 343-359. https://doi.org/10.1016/j.ijforecast.2020.05.008

[110]. Pauwels, K., Hanssens, D. M., & Siddarth, S. (2002). The long-term effects of price promotions on category incidence, brand choice, and purchase quantity. *Journal of Marketing Research*, *39*(4), 421-439. https://doi.org/10.1509/jmkr.39.4.421.19118

[111]. Petropoulos, F., & Kourentzes, N. (2015). Forecast combinations for intermittent demand. *Journal of the Operational Research Society*, *66*(6), 914-924. https://doi.org/10.1057/jors.2014.62

[112]. Petropoulos, F., Kourentzes, N., Nikolopoulos, K., & Siemsen, E. (2018). Judgmental selection of forecasting models. *Journal of Operations Management*, *60*, 34-46. https://doi.org//10.1016/j.jom.2018.05.005

[113]. Pinson, P., & Tastu, J. (2013). Discrimination ability of the energy score. *International Journal of Forecasting*, *29*(4), 548-555. https://doi.org/10.1016/j.ijforecast.2012.09.003

[114]. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *arXiv preprint arXiv:1706.09516*. https://doi.org/10.48550/arXiv.1706.09516

[115]. Reduanul, H., & Mohammad Shoeb, A. (2022). Advancing AI in Marketing Through Cross Border Integration Ethical Considerations And Policy Implications. *American Journal of Scholarly Research and Innovation*, *1*(01), 351-379. https://doi.org/10.63125/d1xg3784

[116]. Rzepakowski, P., & Jaroszewicz, S. (2012). Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, *32*(2), 303-327. https://doi.org/10.1007/s10115-011-0434-0

[117]. Sabuj Kumar, S., & Zobayer, E. (2022). Comparative Analysis of Petroleum Infrastructure Projects In South Asia And The Us Using Advanced Gas Turbine Engine Technologies For Cross Integration. *American Journal of Advanced Technology and Engineering Solutions*, *2*(04), 123-147. https://doi.org/10.63125/wr93s247

[118]. Sadia, T., & Shaiful, M. (2022). In Silico Evaluation of Phytochemicals From Mangifera Indica Against Type 2 Diabetes Targets: A Molecular Docking And Admet Study. *American Journal of Interdisciplinary Studies*, *3*(04), 91-116. https://doi.org/10.63125/anaf6b94

[119]. Salinas, D., Flunkert, V., Gasthaus, J., & Januschowski, T. (2020). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *International Journal of Forecasting*, *36*(3), 1181-1191. https://doi.org//10.1016/j.ijforecast.2019.07.001

[120]. Sanjai, V., Sanath Kumar, C., Maniruzzaman, B., & Farhana Zaman, R. (2023). Integrating Artificial Intelligence in Strategic Business Decision-Making: A Systematic Review Of Predictive Models. *International Journal of Scientific Interdisciplinary Research*, *4*(1), 01-26. https://doi.org/10.63125/s5skge53

[121]. Sanjai, V., Sanath Kumar, C., Sadia, Z., & Rony, S. (2025). AI And Quantum Computing For Carbon-Neutral Supply Chains: A Systematic Review Of Innovations. *American Journal of Interdisciplinary Studies*, *6*(1), 40-75. https://doi.org/10.63125/nrdx7d32

[122]. Sheratun Noor, J., & Momena, A. (2022). Assessment Of Data-Driven Vendor Performance Evaluation in Retail Supply Chains: Analyzing Metrics, Scorecards, And Contract Management Tools. *American Journal of Interdisciplinary Studies*, *3*(02), 36-61. https://doi.org/10.63125/0s7t1y90

[123]. Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199-222. https://doi.org//10.1023/b:Stco.0000035301.49549.88

[124]. 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

[125]. Snyder, R. D. (2002). Forecasting sales of slow and fast moving inventories. *European Journal of Operational Research*, *140*(3), 684-699. https://doi.org/10.1016/s0377-2217(01)00231-4

[126]. Snyder, R. D., Ord, J. K., & Beaumont, A. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting*, *28*(2), 485-496. https://doi.org/10.1016/j.ijforecast.2011.05.013

[127]. Subbaswamy, A., & Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Patterns*, *1*(4), 100128. https://doi.org/10.1016/j.patter.2020.100128

[128]. Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting*, *21*(2), 303-314. https://doi.org//10.1016/j.ijforecast.2004.07.001

[129]. Tahmina Akter, R., Debashish, G., Md Soyeb, R., & Abdullah Al, M. (2023). A Systematic Review of AI-Enhanced Decision Support Tools in Information Systems: Strategic Applications In Service-Oriented Enterprises And Enterprise Planning. *Review of Applied Science and Technology*, *2*(01), 26-52. https://doi.org/10.63125/73djw422

[130]. Tamanna, R., & Dipongkar Ray, S. (2023). Comprehensive Insights Into Co_2 Capture: Technological Progress And Challenges. *Review of Applied Science and Technology*, *2*(01), 113-141. https://doi.org/10.63125/9p690n14

[131]. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting*, *16*(4), 437-450. https://doi.org/10.1016/s0169-2070(00)00065-8

[132]. Taylor, J. W. (2019). Forecasting value at risk and expected shortfall using a semiparametric approach based on the asymmetric Laplace distribution. *International Journal of Forecasting*, *35*(2), 470-482. https://doi.org/10.1016/j.ijforecast.2019.03.009

[133]. Taylor, S. J., & Letham, B. (2018). Forecasting at scale. *The American Statistician*, *72*(1), 37-45. https://doi.org//10.1080/00031305.2017.1380080

[134]. Teunter, R. H., Syntetos, A. A., & Babai, M. Z. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, *214*(3), 606-615. https://doi.org//10.1016/j.ejor.2011.05.018

[135]. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267-288. https://doi.org//10.1111/j.2517-6161.1996.tb02080.x

[136]. Trapero, J. R., Cardós, M., & Kourentzes, N. (2019). Quantile forecast optimal combination to enhance safety stock estimation. *International Journal of Forecasting*, *35*(1), 239-250. https://doi.org//10.1016/j.ijforecast.2018.05.009

[137]. Trapero, J. R., Kourentzes, N., & Fildes, R. (2015). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society*, *66*(2), 299-307. https://doi.org//10.1057/jors.2013.174

[138]. van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2004). Decomposing the sales promotion bump with store data. *Marketing Science*, *23*(3), 317-334. https://doi.org/10.1287/mksc.1040.0061

[139]. Verhoef, P. C., Kannan, P. K., & Inman, J. J. (2015). From multi-channel retailing to omni-channel retailing. *Journal of Retailing*, *91*(2), 174-181. https://doi.org//10.1016/j.jretai.2015.02.005

[140]. Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *The Annals of Statistics*, *46*(6), 2987-3016. https://doi.org/10.1214/18-aos1709

[141]. Wallström, P., & Segerstedt, A. (2010). Evaluation of forecasting error measurements and techniques for intermittent demand. *International Journal of Production Economics*, *128*(2), 625-636. https://doi.org/10.1016/j.ijpe.2010.07.013

[142]. Wickramasuriya, S. L., Athanasopoulos, G., & Hyndman, R. J. (2019). Optimal forecast reconciliation for hierarchical and grouped time series through trace minimization. *Journal of the American Statistical Association*, *114*(526), 804-819. https://doi.org//10.1080/01621459.2018.1448825

[143]. Willemain, T. R., Smart, C. N., & Schwarz, H. F. (2004). A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, *20*(3), 375-387. https://doi.org/10.1016/s0169-2070(03)00013-x

[144]. Winkler, R. L. (1969). Scoring rules and the evaluation of probabilities. *The Annals of Mathematical Statistics*, *40*(5), 1470-1482. https://doi.org/10.1214/aoms/1177693052

[145]. Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint*, *arXiv:1709.04875*. https://doi.org/10.48550/arXiv.1709.04875

[146]. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. Proceedings of the AAAI Conference on Artificial Intelligence,

[147]. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320. https://doi.org//10.1111/j.1467-9868.2005.00503.x