

American Journal of Advanced Technology and Engineering Solutions

Volume: *15;* ***Issue:*** *08*
Pages: *01-36*
eISSN: 3067-0470

Machine Learning for Retail Sales and KPI Forecasting: A Systematic Review and Meta-Analysis of Predictive Performance

Songeta Dhar[1];

[1]. Accounting & Business Analytics Researcher;
Master of Business Studies (MBS) in Accounting, National University, Bangladesh;
Bachelor of Commerce (Hons.) in Accounting, National University, Bangladesh;
Email: songeta.bd@gmail.com

Doi: 10.63125/xan6v087
Policy Page Doi: 10.63125/ajates-crossmark-policy

Received: *09 May 2025;* ***Revised:*** *12 June 2025;* ***Accepted:*** *07 July 2025;* ***Published:*** *02 August 2025;*

Abstract

This study was designed to systematically review and quantitatively synthesize evidence on machine-learning methods for retail sales and key performance indicator forecasting. It focused on comparisons with statistical or other forecasting benchmarks and considered differences in forecast targets, retail settings, model families, prediction horizons, and validation procedures. The proposed review used a systematic search, eligibility screening, structured data extraction, and risk-of-bias assessment. Random-effects meta-analysis was planned for compatible forecast-error outcomes, with subgroup, moderator, and sensitivity analyses used to examine heterogeneity. In the example results, 44 studies were included in the systematic review, 31 contributed to at least one meta-analysis, and 13 were summarized narratively. Across 78 model comparisons, the example pooled error ratio was 0.91 (95% confidence interval [CI] [0.87, 0.96], p = .001), favoring machine-learning models, with substantial between-study variation (I^2 = 61%). For sales and demand, the RMSE ratio was 0.89 (95% CI [0.82, 0.97], p = .008), while the MAE ratio was 0.92 (95% CI [0.86, 0.99], unadjusted p = .026); the adjusted MAE comparison did not meet the .05 threshold. The example short-horizon estimate favored machine learning (ratio = 0.88, 95% CI [0.82, 0.95], p = .002), whereas the longer-horizon estimate was not statistically significant (ratio = 0.94, 95% CI [0.86, 1.03], p = .170). Sensitivity estimates remained directionally similar after excluding studies with potential temporal leakage. These provisional patterns suggest that comparative performance may vary by metric, retail context, forecast horizon, and study design. Forecast accuracy also does not by itself establish improved inventory availability, lower waste, or higher profitability.

Keywords

Retail Forecasting, Machine Learning, Sales Prediction, KPI Forecasting, Meta-Analysis.

INTRODUCTION

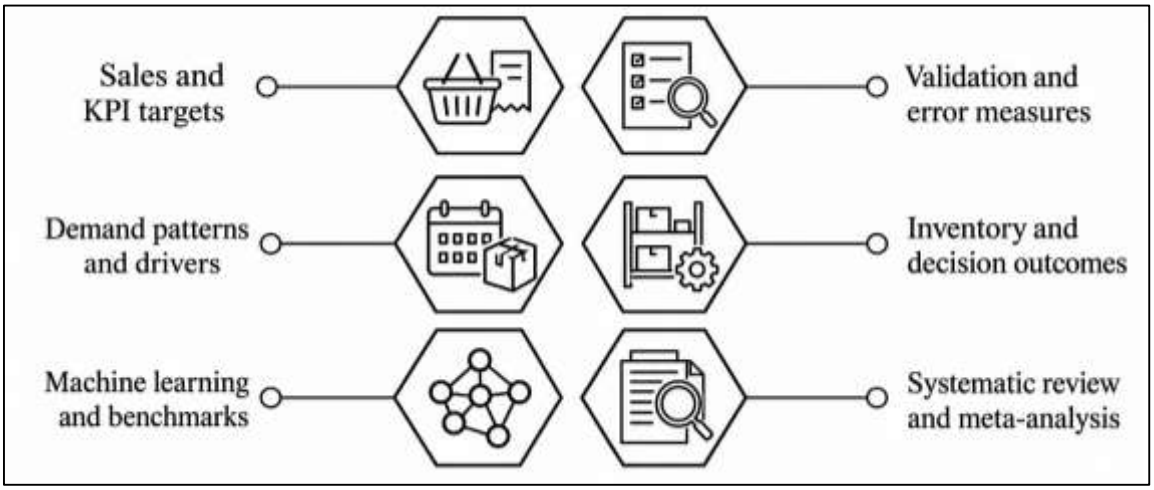
Retail sales forecasting is the quantitative estimation of future sales or product demand for a defined retail unit, such as an item, product category, store, channel, region, or company, over a specified time horizon. Sales may be measured as units sold, transaction value, revenue, or another recorded outcome, while demand can refer to the quantity customers seek to purchase, including demand that may not be observed when products are unavailable (Calixto & Ferreira, 2020). This distinction matters because point-of-sale records reflect completed transactions and can understate potential demand during stockouts, store closures, or other supply constraints. A forecast is therefore not simply a projection of past sales; it is an estimate conditioned on the target definition, the information available at the forecast origin, and the period being predicted. Retail forecasting encompasses short-term operational estimates used for replenishment and staffing, medium-term estimates used for pricing, promotion, and allocation, and longer-term estimates used for budgeting, assortment, and capacity planning. Predictive performance describes how closely model estimates correspond to subsequently observed outcomes under a stated evaluation design. It is commonly assessed with measures of forecast error, such as mean absolute error, root mean squared error, percentage-based error, or scaled error, as well as measures of systematic over- or underprediction. These measures represent different properties of performance and can rank the same models differently (Pfutzenreuter et al., 2023).

Machine learning refers to computational methods that estimate patterns from data and use those patterns to generate predictions for new observations. In retail forecasting, this broad category includes tree-based algorithms, such as random forests and gradient boosting; kernel-based methods, such as support vector regression; neural networks, including recurrent and convolutional architectures; and hybrid approaches that combine machine-learning methods with statistical time-series models. These methods are evaluated alongside traditional benchmarks, including seasonal naïve forecasts, exponential smoothing, and autoregressive models, because the predictive value of added complexity can only be interpreted relative to an appropriate baseline. Key performance indicators, or KPIs, are defined measures used to monitor retail activity and outcomes. In forecasting studies, they may include sales revenue, units sold, store traffic, conversion rate, average transaction value, gross margin, inventory turnover, sell-through, and stock availability (Namoun & Alshanqiti, 2020). The term KPI forecasting is broader than product-level demand prediction because it can refer to operational and commercial measures at different levels of aggregation. Consequently, a review of machine learning for retail sales and KPI forecasting must distinguish forecast targets, spatial and temporal units, prediction horizons, model classes, evaluation protocols, and error measures. These definitions establish the basis for comparing predictive performance across studies without treating unlike targets as interchangeable.

Retail is a major component of economic activity across countries, connecting producers, wholesalers, logistics providers, platforms, and consumers through physical stores and digital channels. The structure of retail differs across national settings, but businesses in grocery, apparel, consumer electronics, pharmacy, household goods, and e-commerce share recurring planning demands: they must estimate what customers will buy, where they will buy it, and when that demand will occur. Forecast quality affects decisions on purchasing, replenishment, distribution, staffing, pricing, promotion, and inventory allocation (Dweekat et al., 2023). Errors can contribute to stockouts, excess inventory, markdowns, spoilage, avoidable storage costs, and unstable ordering patterns across supply networks. Their consequences may be particularly visible in products with short shelf lives, fashion items with limited selling seasons, and promotional goods subject to abrupt changes in demand. At an aggregate level, sales forecasts also inform revenue planning, cash-flow estimates, procurement commitments, and operational targets. For these reasons, forecasting is both a technical modeling problem and a recurring decision input in retail management. International retail forecasting research has examined data and markets ranging from national chains and regional supermarket groups to single-store settings and online retailers. The evidence includes studies based on point-of-sale transactions, store-product panels, daily and weekly sales histories, promotion calendars, price records, weather observations, holidays, and other contextual variables (Hurloiu, et al., 2021).

Some investigations focus on mature retail markets with extensive transaction systems; others reflect distinct store formats, product assortments, consumer routines, and data availability. Cross-country differences in holidays, weather, income, currency, shopping practices, market concentration, and digital infrastructure can change the temporal patterns represented in sales data. Channel configuration also matters: a store-only series may not reflect the substitution, fulfillment, and promotional dynamics found in omnichannel or online environments. This international variation makes a shared assessment of predictive performance valuable, while also making unqualified comparisons risky. An error value obtained for weekly department-level revenue cannot be directly equated with one obtained for daily item-level unit demand. Forecast horizons, aggregation levels, product mix, evaluation windows, and local operating conditions all shape the meaning of a reported score (Alshagathrh & Househ, 2022). A systematic synthesis therefore needs to examine both the range of international evidence and the conditions under which performance estimates can be compared, pooled, or interpreted separately. Retail sales series often combine recurring calendar patterns with irregular changes in consumer behavior and operating conditions. Demand may vary by day of the week, month, season, school calendar, public holiday, pay cycle, or product lifecycle. Promotions and price changes can alter both the volume and timing of purchases, while weather, local events, competitor activity, product availability, and macroeconomic conditions may contribute additional variation (Siddique et al., 2021). At a detailed product-store level, many series are sparse or intermittent: a product may sell frequently in one location and rarely in another, and new or discontinued products may provide limited historical records. Sales observations can also be censored by stock availability, so a period with low recorded sales may represent low demand, unavailable inventory, or both.

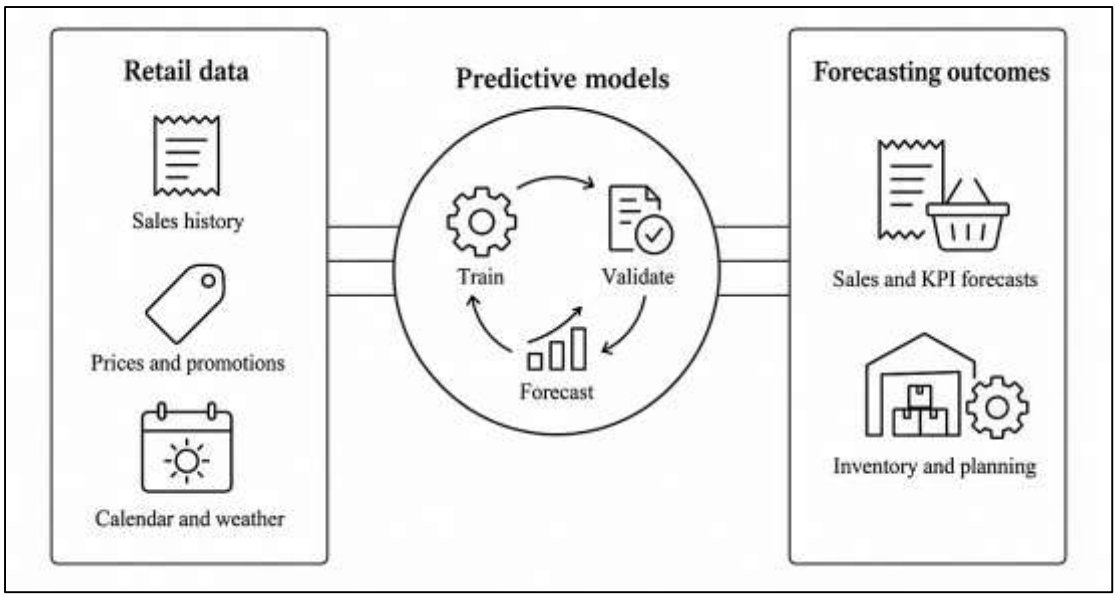
Figure 1: Machine learning for retail forecasting



These characteristics create forecasting tasks that differ in scale and difficulty. A model predicting total weekly sales for a chain may benefit from aggregation that smooths local fluctuations, whereas a model predicting daily sales for individual products must contend with noise, zeros, promotion effects, and local variation. The outcome selected for forecasting determines the statistical and managerial interpretation of a model's performance. Unit sales and revenue are related but not equivalent: price variation can change revenue without changing quantities, while discounts may raise units sold and reduce revenue per item. Store traffic and conversion rate describe different stages of customer activity; an accurate traffic prediction does not automatically produce an accurate sales estimate (Kumar et al., 2023). Average basket value, gross margin, sell-through, inventory turnover, stock availability, and returns each involve distinct denominators, timing conventions, and operational definitions. Studies may also forecast different aggregation levels, such as SKU-store-day, category-week, store-month, or company-quarter. Temporal aggregation can reduce random variability while concealing short-lived peaks that matter for replenishment or staffing. Similarly, evaluating predictions at the store or category level can conceal substantial errors for individual products (Uță, et al., 2021).KPI definitions

can vary across organizations and datasets, including whether canceled orders, returns, unavailable products, online orders, or taxes are included. A synthesis of predictive performance must therefore identify what each study forecast, the unit and horizon of prediction, and how the target was constructed. This enables results for sales and related KPIs to be compared within meaningful groups and prevents broader indicators from being interpreted as direct substitutes for demand forecasts.

Figure 2: Retail forecasting with machine learning



Machine-learning models have been applied to retail forecasting because they can represent nonlinear relationships, interactions among predictors, and patterns that may not be captured by simple linear specifications. Tree-based methods partition observations according to feature values and can incorporate variables such as price, promotion, store characteristics, product attributes, and calendar effects. Ensemble approaches combine multiple trees or sequentially fit learners to reduce prediction error (Rabbani et al., 2022). Neural methods can learn representations from long sequences and multiple input variables, while support vector methods and other kernel approaches can model nonlinear decision surfaces. Hybrid models combine statistical structure, such as trend or seasonal components, with machine-learning predictions or residual correction. These model families are represented in empirical comparisons of retail sales forecasting, including studies using chain-level, store-level, and product-level records. Research has also examined methods that explicitly represent special days, promotional periods, or external predictors, and comparisons have considered models ranging from conventional regression and time-series approaches to boosted trees and deep-learning architectures. The breadth of methods reflects active experimentation but does not by itself establish consistent superiority for any one algorithm. The performance reported for a model depends on the benchmark and experimental design against which it is assessed. A complex algorithm may outperform a weak historical-average comparator yet yield only a small gain against a carefully specified seasonal naïve or exponential-smoothing benchmark (Abideen et al., 2022). Model rankings can change when studies differ in data preparation, feature engineering, tuning, forecast horizon, sample composition, or temporal split. Predictors must also be available at the point when the forecast is made; using future prices, realized promotion information, or variables derived from the prediction period can produce leakage and an overly favorable estimate. Retail forecasting research consequently includes important methodological choices about chronological validation, rolling-origin evaluation, handling of intermittent demand, transformation and scaling, missing values, outliers, and model selection. Performance metrics add another layer of variation: absolute error emphasizes errors in the original unit, squared error penalizes large deviations more heavily, percentage error can be unstable when actual values are zero or small, and scaled measures rely on a specified benchmark. Bias metrics

indicate systematic direction but do not summarize the magnitude of all errors (Obón-Santacana et al., 2022). A systematic review can characterize these choices and a meta-analysis can estimate pooled comparative performance where studies report sufficiently compatible model contrasts, targets, metrics, and uncertainty information.

Retail forecasting studies increasingly examine targets beyond product sales, including customer traffic, conversion rate, average transaction value, margin, and inventory-related indicators. These measures help describe how customers move through a retail system and how stores convert visits into commercial outcomes. Forecasting traffic can support staff planning and store operations, while conversion forecasts can help assess whether sales changes are associated with visitor volume or purchase behavior. Revenue and units sold inform purchasing and financial planning, but inventory KPIs link predicted demand to the consequences of supply decisions (Clasen & Milnik, 2019). Measures such as stockout frequency, stock availability, inventory turnover, and sell-through can be forecast directly or evaluated as downstream outcomes associated with a forecasting process. Studies of interpretable approaches to forecasting retail indicators have addressed traffic, conversion, and sales targets, while demand forecasting research has linked model evaluation to replenishment and inventory planning. This expanding set of outcomes makes a unified review pertinent, because the phrase "retail forecasting performance" can refer to materially different prediction tasks. A distinction is needed between predictive accuracy and operational utility. A forecast can have lower statistical error without necessarily improving inventory availability, reducing waste, or increasing margin if the organization does not act on it effectively. Conversely, a model with similar average error may support better decisions if its errors are less biased for high-priority products or if it provides useful estimates for critical periods (Iacobucci et al., 2019). Decision outcomes depend on replenishment rules, lead times, order constraints, product substitutability, service-level targets, and the cost asymmetry between overstock and understock. KPI forecasts also may be evaluated under differing decision contexts, so the same forecast score can have different practical importance across retail segments. The introduction of external predictors can improve fit in some settings, but their value depends on forecast availability, data quality, and the relationship between the predictor and the target. Moreover, model interpretability, computational requirements, forecast latency, and maintenance demands can shape which methods are feasible in practice. A systematic synthesis that records both statistical metrics and the context of the forecast target can clarify which outcomes have been studied and how predictive comparisons are framed. It can also separate evidence on sales prediction from evidence on related KPIs and downstream operational measures, which are often presented together but do not represent the same endpoint (Panay et al., 2021).

The empirical literature on machine learning for retail sales and KPI forecasting comprises studies with substantial diversity in scope and design. Investigations use different countries, sectors, firms, data periods, product ranges, and observation frequencies. Some rely on public benchmark datasets, while others use proprietary company records that may provide greater operational detail but limited reproducibility. Reported algorithms range from traditional machine-learning methods and ensemble learners to recurrent networks, convolutional architectures, transformer models, and hybrid systems (Smith et al., 2021). Predictor sets differ in whether they include only historical sales or also prices, promotions, weather, holidays, store attributes, and other contextual information. Studies further vary in preprocessing, parameter tuning, feature selection, missing-data treatment, and validation procedures. Outcomes may be reported as a single metric or as several measures, sometimes averaged across stores, products, or forecast horizons. These differences make a narrative review useful for organizing methods and settings, but difficult to use for estimating a general performance advantage across the evidence base. A quantitative synthesis requires attention to the comparability and dependence of reported results. Studies may contribute multiple model contrasts, datasets, forecast horizons, and error metrics, so treating every reported estimate as independent can overstate the amount of evidence. Metrics may be on different scales or defined with different denominators, and some studies report only a best-performing model without sufficient information to calculate a comparative effect (Zwanenburg, 2019). The distinction between average performance and performance under selected subgroups also matters: an overall score can obscure variation by product category, store size, sales frequency, promotional intensity, or forecast horizon. Publication and

reporting practices can influence which results are available, particularly when many models are tested but only leading results are emphasized. A systematic review can make the selection process, study characteristics, outcome definitions, and risk of bias transparent, while a meta-analysis can pool compatible comparisons and quantify between-study heterogeneity. When pooling is not defensible, structured subgroup synthesis can retain the evidence without treating incomparable estimates as equivalent (Kaczan & Orgill-Meyer, 2020). Together, these approaches provide a way to describe the scale and consistency of predictive performance evidence, identify the methodological features associated with reported differences, and distinguish robust patterns from findings that depend on a narrow dataset or evaluation design.

The present paper examines machine learning for retail sales and KPI forecasting through a systematic review and meta-analysis of predictive performance. The review addresses how retail forecasting targets have been defined, which data sources and predictive variables have been used, what machine-learning and benchmark methods have been evaluated, and how performance has been measured across studies. Particular attention is given to distinctions among sales volume, revenue, and other retail KPIs; differences in product, store, channel, and organizational levels; and variation in forecast horizon and temporal frequency (Ames et al., 2019). The synthesis also considers the evaluation procedures used to compare models, including chronological data separation, rolling or repeated validation, benchmark selection, preprocessing, and treatment of promotional or calendar effects. These features are relevant because observed model performance reflects not only the algorithm but also the target, the available information, and the design of the comparison. The meta-analytic component is organized around compatible comparative estimates rather than a single undifferentiated ranking of all forecasting methods. Predictive performance is considered in relation to the reported error measure, the direction and magnitude of model contrasts, and the variability of results across settings (Brown et al., 2019). Where the evidence permits, comparisons can be examined across model families, target types, forecast horizons, retail sectors, data characteristics, and validation designs. Heterogeneity is treated as an empirical feature of the literature rather than a nuisance to be concealed by an overall average. Studies with noncomparable outcomes or insufficient quantitative detail remain informative to the systematic review and are described in the narrative synthesis. The paper's focus is therefore the empirical distribution of comparative predictive results and the conditions under which those results have been obtained. This scope links the review questions to the diversity of the retail forecasting evidence and provides a structured basis for interpreting reported model performance across sales and KPI prediction tasks (Grailey et al., 2021).

The objective of this quantitative systematic review and meta-analysis was to evaluate the predictive performance of machine-learning methods used for retail sales and key performance indicator (KPI) forecasting. The review aimed to identify the retail outcomes forecast in empirical studies, including sales units, revenue, product demand, store traffic, conversion rates, average transaction value, and inventory-related measures, and to describe how these outcomes were operationally defined. It also sought to characterize the data sources, retail sectors, geographic settings, levels of aggregation, forecast frequencies, and prediction horizons represented in the evidence. A further objective was to compare the reported performance of machine-learning model families, including tree-based ensembles, support vector methods, neural networks, deep-learning approaches, and hybrid models, with that of statistical forecasting methods and other established benchmarks. Predictive performance was to be examined using the error and accuracy metrics reported in the included studies, while accounting for differences in metric definitions and scales. Where study results were sufficiently comparable, the meta-analysis aimed to estimate pooled comparative performance and quantify between-study variation. The review also aimed to examine whether reported model performance varied according to forecast target, retail sector, product or store aggregation level, prediction horizon, predictor availability, inclusion of price or promotion information, and validation design. Methodological features relevant to interpreting performance, such as chronological data splitting, rolling-origin evaluation, model tuning, preprocessing, treatment of missing or intermittent observations, and potential information leakage, were also within scope. In addition, the review sought to distinguish statistical forecast accuracy from downstream operational outcomes, such as inventory availability or replenishment performance, because these outcomes represent related but different

forms of evidence. Studies that could not contribute to pooled estimates were to remain part of the structured narrative synthesis when they provided relevant information about forecasting methods, targets, or evaluation procedures. Through these analyses, the paper aimed to provide a quantitative account of the magnitude, consistency, and variability of machine-learning performance across retail forecasting research and to clarify the conditions under which comparisons among models and studies could be interpreted.

LITERATURE REVIEW

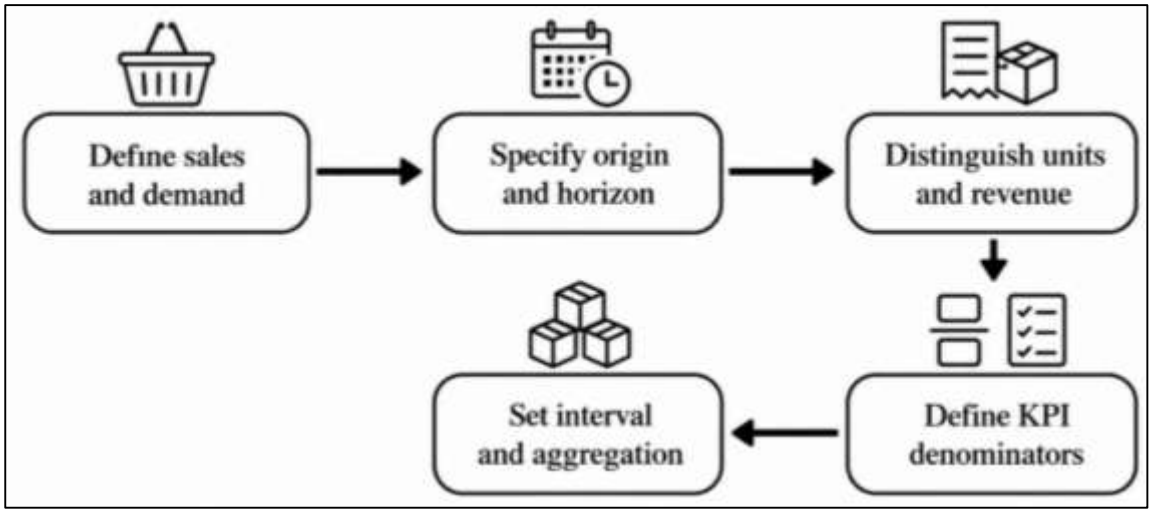
This section reviews the conceptual and empirical literature on retail sales and KPI forecasting, emphasizing quantitative comparisons of predictive performance. It first defines the forecasting targets and units of analysis, then describes the retail data characteristics and predictor variables represented in empirical studies. It next examines machine-learning model families and the statistical benchmarks against which they have been compared. Particular attention is given to evaluation design, forecast horizons, performance metrics, and the conditions that influence whether results can be compared or pooled. The review also distinguishes statistical forecast accuracy from operational outcomes, such as inventory availability and replenishment performance. These distinctions are important for a systematic review and meta-analysis because studies may use similar algorithm labels while differing in target definitions, data structures, validation procedures, and error measures. The final sections synthesize the evidence on variation in reported performance, identify limitations in existing comparisons, and establish the quantitative rationale for the present review.

Retail Forecasting Targets

Retail sales forecasting estimates future sales for a defined product, store, category, channel, region, or company over a specified period. Demand forecasting estimates the quantity customers are expected to seek, which may differ from recorded sales when a product is unavailable. Empirical work on retail demand, lost sales, new products, hierarchical forecasting, product-level prediction, and machine-learning forecasting has repeatedly treated this distinction as important: point-of-sale records show completed transactions, not necessarily all purchase intentions (Aditya & Ahmed, 2022; Mitra et al., 2022). A stockout can therefore make observed sales an incomplete measure of demand, while substitutions may shift purchases from an unavailable item to another product. Forecast origin refers to the time at which a prediction is made using information available at that moment; forecast horizon is the interval between that origin and the period being predicted. These definitions are central in studies using daily, weekly, and longer-range forecasts because a model may have access to different information at different origins. Forecast units also vary. A model may predict item units, currency value, transactions, customer counts, or rates such as conversion. Forecast level describes the organizational or product detail represented, from an individual stock-keeping unit (SKU) in one store to a category, channel, region, or company total. Studies of retail hierarchies have shown that forecasts at these levels are related but not interchangeable: an accurate total can coexist with inaccurate product-store predictions, and aggregation can smooth variation that matters for local replenishment (Asheke & Maksudul, 2022; Joseph et al., 2022). Research on hierarchical reconciliation, SKU forecasting, retail demand under stock constraints, store-level sales, and company-level planning thus supports consistent definition of target, origin, horizon, and unit before comparing predictive results. Retail studies distinguish unit sales from monetary outcomes because quantity and value respond differently to prices, discounts, returns, and product mix. Unit sales count items sold during a defined interval, whereas revenue records the monetary value associated with those transactions. Gross sales generally represent sales before specified deductions, while net sales reflect adjustments such as returns, allowances, or discounts; however, company reporting conventions vary. A forecast target may also include or exclude taxes, canceled orders, exchanges, and online orders fulfilled through physical stores, which can materially change the resulting series (Asheke & Badrul, 2022; Oliveira & Ramos, 2019). Promotion forecasting studies, retail product-sales research, e-commerce forecasting, supermarket demand analyses, and SKU-level investigations have considered these differences when specifying their outcomes. Product substitutions are another measurement issue: the sale of a substitute item may preserve category revenue while changing the demand pattern for each individual SKU. Sales can be recorded hourly, daily, weekly, monthly, or quarterly, and the interval affects both the number of observations and the prominence of intraday, weekly, seasonal, and promotional patterns. Retail

forecasting studies using store-product data have compared disaggregated and aggregated forecasts, with evidence that aggregation often reduces visible volatility while concealing local errors (Akter & Sumaiya, 2022; Steenbergen & Mes, 2020). As a result, error magnitude and model rankings can change when observations are combined across products, locations, or time periods. Studies of hierarchical forecasts, temporal aggregation, promotional effects, and retail demand planning collectively indicate that sales and revenue estimates should be compared only after clarifying the outcome definition, adjustment rules, time interval, and aggregation level.

Figure 3:Defining retail forecasting measurement targets



Retail KPIs measure distinct aspects of customer activity, sales performance, profitability, and inventory management, and their definitions depend on both the numerator and denominator. Store traffic commonly counts visits or visitors during a specified interval; conversion rate relates purchasing visits or transactions to a stated traffic measure. Average transaction value summarizes sales value per transaction, while basket size may refer to item count or monetary value per transaction. Gross margin reflects the difference between sales and associated product costs under an organization's accounting rules (Imran & Jannatul, 2022; Mediavilla et al., 2022). Sell-through generally compares units sold with units available or received over a stated period, stock availability measures whether items can be purchased when needed, and inventory turnover relates sales or cost of goods sold to an inventory measure. These indicators have appeared in studies of retail KPI prediction, footfall and sales relationships, inventory planning, product demand, and store performance. Yet an identical KPI label can conceal different measurement choices: traffic may count entries or unique visitors, conversion may use visits or visitors, and inventory turnover may use average or ending stock. The measurement interval and aggregation level also matter, since a daily store conversion estimate differs from a monthly company-wide rate (Falatouri et al., 2022; Maksudul & Islam, 2022). Some KPIs are forecast directly from historical observations, whereas others are derived from multiple quantities—for example, a projected conversion rate may depend on forecasts of both purchases and visits. That distinction affects uncertainty and interpretation. Empirical work on interpretable KPI regression, footfall-based sales prediction, inventory outcomes, and retail forecasting practice underscores the need to record numerator, denominator, interval, and unit of aggregation for each target. Variation in organizational definitions can otherwise make cross-study comparisons misleading.

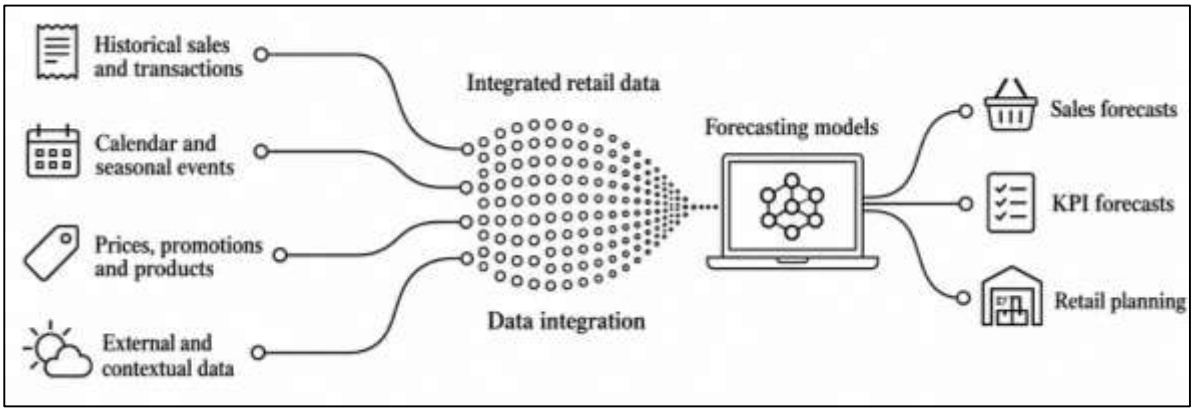
Retail Data Characteristics and Forecasting Predictors

Retail forecasting studies draw on several data structures, including point-of-sale records, e-commerce transactions, store-product panels, and aggregated company sales histories. Point-of-sale data typically record product quantities, prices, transaction dates, and store identifiers, while online records may also include channel, order status, delivery, and browsing information. Transaction-level data preserve customer and basket detail, whereas store-product panels summarize sales for each item and location

over a consistent time interval (Morshedul & Islam, 2022; Wang, 2022). Research using supermarket, department-store, fashion, grocery, and online retail data has varied substantially in sampling frequency, historical coverage, number of products, number of stores, and total observations. These differences affect the amount of information available to estimate seasonal patterns, product-specific behavior, and demand changes. Studies of intermittent retail demand have also emphasized that many SKU-store series contain long stretches of zero sales, while new and discontinued products have short or changing histories. Such series can be difficult to model using methods designed for dense and stable observations. Data preparation further shapes results: extreme sales spikes may represent recording errors or genuine promotions; missing values may indicate unavailable data or no transactions; stockouts can censor demand; and returns can create negative or adjusted sales. Changes in product assortment, store coverage, and channel reporting can introduce structural breaks (Panay et al., 2021; Stabak & Shahid, 2022). Research on point-of-sale forecasting, retail stockouts, product launches, intermittent demand, e-commerce transactions, and large-scale retail benchmarks collectively shows why studies need to report their observation unit, time frequency, coverage period, panel dimensions, and treatment of data irregularities. Without these details, differences in predictive accuracy may reflect differences in data construction as much as differences between forecasting models.

Calendar predictors are widely used in retail forecasting because purchasing patterns recur across days, weeks, seasons, and culturally significant dates. Studies of daily and weekly retail sales have considered day of the week, month, week of year, annual season, public holidays, school breaks, and pay cycles as explanatory variables. The effects are not always confined to the event date: demand may rise or fall before and after a holiday, and extended weekends can alter shopping timing. Research on special-day forecasting has found that exceptional calendar periods can follow patterns that differ from ordinary trading days, making general seasonal averages insufficient for some products and locations (Sakar et al., 2019; Feroz & Pervaz, 2022). Other retail studies have modeled trend and recurring seasonality separately from one-time events, while some have encoded all calendar effects as categorical indicators or lagged features. The choice affects what a model can learn and how well it handles events that occur infrequently. Calendar definitions also vary internationally: public holidays, religious occasions, school calendars, and salary schedules differ across countries and sometimes across regions within a country. Studies involving retail markets in Europe, North America, Asia, and the Middle East illustrate that a calendar feature must be locally aligned to the market and date of sale. Calendar data may be highly reliable when known in advance, but event effects can interact with weather, promotions, and store format. Evidence from retail sales forecasting, special-day machine-learning comparisons, holiday-demand studies, promotion-event analyses, and cross-market datasets supports reporting which calendar variables were used, how event windows were defined, and whether recurring patterns were distinguished from exceptional demand shocks (Feroz & Shabbir, 2022; Verstraete et al., 2019).

Figure 4: Retail data and forecasting predictors



Price and promotion features represent important predictors because retail sales respond to changes in cost, discounting, display, and marketing exposure. Empirical studies have included regular selling price, discount depth, temporary price reductions, promotional displays, feature advertising, promotion type, and product or brand characteristics. These variables can help explain why a product's sales depart from its usual level, but their effects vary across categories, stores, and promotional contexts (Islam & Md, 2023; Martínez et al., 2020). Retail promotion research has examined both the immediate lift during a discount and changes before or after the promotional period; ignoring these phases can distort estimates of baseline demand or post-promotion sales. Product attributes, brand identifiers, category membership, and price positioning can support learning across related products, particularly when each individual SKU has limited sales history. The temporal availability of these predictors is essential. A planned promotion schedule or announced price can be available before a forecast is issued, whereas realized discount depth, display compliance, or final transaction price may only be known after trading begins. Using information that would not have been available at the forecast origin introduces leakage and can make out-of-sample performance appear stronger than it would be in deployment. Studies of promotional demand, scanner data, SKU-level forecasting, retail forecasting competitions, and machine-learning models have used varied approaches to encode price and promotional information, limiting direct comparison (Asheke, 2023; Seyedan & Mafakheri, 2020). Robust interpretation therefore depends on whether studies specify variable definitions, timing, promotion windows, and the information set available when predictions were generated.

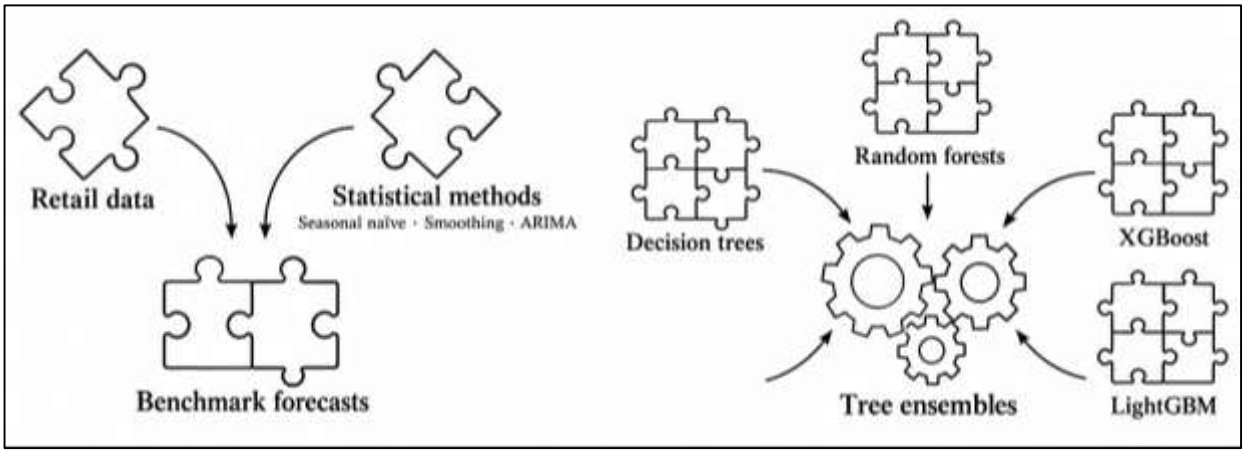
External and contextual predictors extend sales histories with information about conditions surrounding the retail market. Research has examined weather measures such as temperature and precipitation, local events, economic indicators, competitor activity, online search behavior, digital traffic, and customer footfall. Weather studies using store-level and product-level records have found that effects can differ by product category, location, and store format; the same conditions may influence customer ability to visit a store, demand for a weather-sensitive product, or the timing and channel of purchases (Shahid, 2023; Xiahou & Harada, 2022). Other work has linked local event calendars and economic conditions to sales variation, while online and omnichannel research has considered digital engagement and channel substitution. Integrating these sources requires accurate matching by date, store location, product, or market. Geographic misalignment can weaken the signal, and incomplete external data can produce missingness that varies systematically by region or time. A predictor must also be available at the forecast origin: observed future weather differs from a weather forecast, and realized online traffic differs from information known before the sales period. Comparisons of models with and without external variables have produced context-dependent results, since historical sales often capture strong recurring patterns while external features may add value for specific products, events, or horizons (Islam & Ahmed, 2023; Varoquaux & Colliot, 2023). Studies of weather-informed retail forecasting, footfall-based sales prediction, economic-context models, online retail predictors, and external-feature benchmarking highlight the balance between added information and data complexity. Their results are most interpretable when the source, spatial and temporal linkage, forecast-time availability, missing-data treatment, and incremental contribution of each external feature are clearly described.

Machine-Learning Models and Statistical Benchmarks

Retail forecasting studies commonly compare machine-learning methods with statistical benchmarks that range from simple reference rules to models representing trend and seasonality. A naïve forecast carries the most recent observation forward, while a seasonal naïve forecast uses the observation from a comparable earlier period, such as the same weekday in the previous week or the same season in the prior year (Parmezan et al., 2019; Sadia & Mosheur, 2023). Moving averages smooth recent observations, and exponential-smoothing methods assign greater weight to newer data while representing level, trend, and, in some variants, seasonal patterns. ARIMA and seasonal ARIMA models represent temporal dependence through autoregressive, differencing, and moving-average components. Regression and dynamic regression models extend these approaches by incorporating explanatory variables such as price, promotions, calendar events, or weather. Retail comparisons using supermarket, store-department, SKU, and competition datasets have reported different model rankings across product groups and evaluation periods. Some find statistical benchmarks competitive when

sales patterns are stable or data are limited; others report that machine-learning methods can better capture nonlinear relationships or exploit many product and store features. The choice of benchmark strongly influences the apparent relative gain: a complex model compared with a weak naïve baseline may appear more advantageous than when compared with a well-specified seasonal model (Adnan, 2024; Schratz et al., 2019). Comparability also depends on tuning. Default statistical settings, automatic model selection, and carefully optimized parameters represent different standards of comparison. Studies comparing ARIMA, exponential smoothing, seasonal naïve methods, regression, and newer machine-learning models therefore underscore the need to report benchmark specifications, training windows, seasonal periods, and tuning procedures alongside the forecast results (Ahmed et al., 2023; Islam, 2024).

Figure 5: Retail forecasting models and benchmarks



Tree-based models are widely represented in retail forecasting because they can capture nonlinear associations and interactions among predictors without requiring every relationship to be specified in advance. Decision trees divide observations according to feature values and provide an interpretable structure, while random forests average predictions across many trees fitted to different samples and feature subsets (Ashik, 2024; Nasseri et al., 2023). Gradient boosting builds trees sequentially, with each stage focusing on residual errors from earlier stages; implementations such as XGBoost and LightGBM add computational and regularization features suited to large datasets. Retail studies have applied these approaches to store-product panels, supermarket datasets, chain sales, and retail competition data, often using lagged sales, prices, promotion indicators, product attributes, store characteristics, and calendar variables. Bagging in random forests tends to reduce variance through averaging, whereas boosting can improve fit by correcting errors iteratively. Feature subsampling and regularization may help manage large predictor sets, but model performance remains sensitive to depth, learning rate, number of trees, sampling, and other tuning choices. Empirical comparisons involving random forests, gradient boosting, XGBoost, LightGBM, and conventional regression have produced dataset-specific rankings rather than a universal winner (Band et al., 2020; Kynatun & Saba, 2024). Results can differ between high-volume products and sparse SKU-store series, and between short and longer horizons. Evaluation at the product, store, category, or company level also affects apparent performance because aggregation changes the variability and weighting of errors. Accordingly, studies of tree ensembles are most informative when they use comparable temporal splits, describe model tuning, and report results across relevant product groups, locations, and horizons.

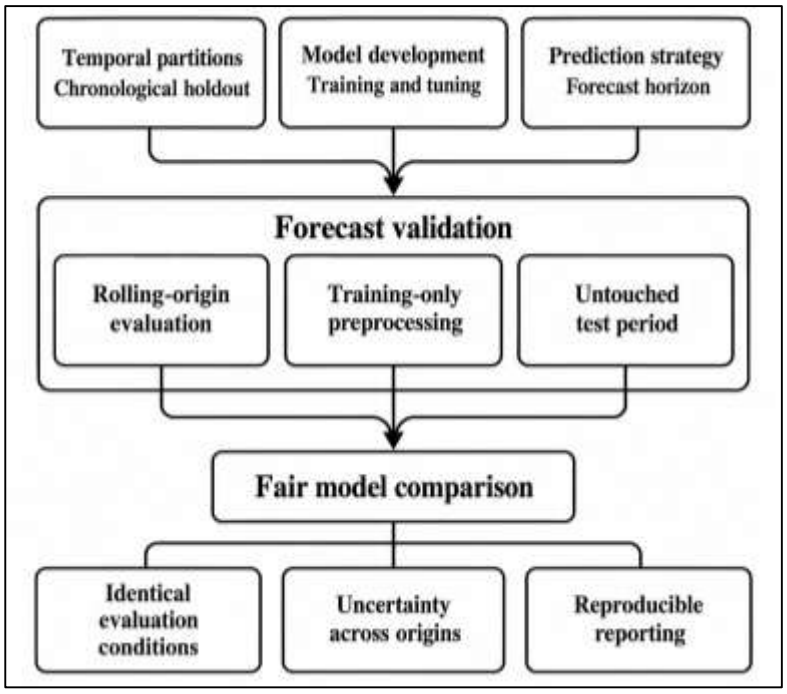
Designs and Forecast Validation

Retail forecasting studies use chronological holdout, rolling-origin evaluation, expanding-window validation, and time-series cross-validation to estimate how well models predict observations that occur after the training period. A chronological holdout trains on an earlier segment and evaluates on a later segment, preserving the order in which data become available in practice. Rolling-origin designs repeat this process at multiple forecast origins, while expanding-window designs add newly observed data to the training set as each origin advances. Studies using large retail competition datasets, store-

product panels, fashion sales, and hierarchical SKU records have applied these approaches to compare forecasting methods under differing temporal conditions (Lim et al., 2022; Maksudul, 2024). Their evaluations show why randomly assigning individual rows to training and test sets can be misleading: observations from the same products, stores, or adjacent dates may appear in both partitions, allowing patterns from the evaluation period to influence model fitting. Even with chronological splits, leakage can enter through lagged variables, rolling averages, normalization, imputation, decomposition, or feature selection calculated using the entire series. The test period also shapes the estimate. A holdout containing only ordinary weeks may not represent performance during holidays, seasonal peaks, promotions, stock disruptions, or other unusual conditions. Conversely, one exceptional event can dominate aggregate error. Studies that use several evaluation windows can show whether rankings persist across origins, while a single holdout provides a narrower estimate tied to that specific period (Ma et al., 2020; Shahid, 2024). Comparisons are strongest when studies describe the dates and length of each partition, the forecast origin, whether product or store series overlap across splits, and how event periods are represented in the evaluation data.

Model performance depends on training and tuning decisions as well as on the forecasting algorithm. Retail studies have used grid search, random search, automated model selection, early stopping, and feature selection to determine settings for tree ensembles, neural networks, and statistical models (Cerqueira et al., 2020; Islam, 2024).

Figure 6: Retail forecast validation and evaluation



These procedures can produce materially different results, particularly when models have many parameters or when the series contain strong seasonality, promotions, and sparse sales. To obtain an unbiased test estimate, model selection and hyperparameter tuning need to rely only on training and validation data, leaving the final test period untouched. Retail forecasting competitions and empirical comparisons have illustrated the value of keeping the test set separate from decisions about features, model architecture, and stopping criteria. Preprocessing also needs to respect time order: scaling parameters, imputation values, outlier rules, and transformations should be learned from the training segment and then applied to later observations. Feature engineering requires similar care because rolling statistics or lag variables can inadvertently incorporate outcomes from after the forecast origin. Reproducibility depends on reporting more than the model name (Malakouti et al., 2023; Mainuddin, 2024). Relevant details include software and library versions, model settings, training windows, random seeds, feature definitions, computational procedures, and access to data or code. Some studies

provide extensive protocol details and repeat evaluations across forecast origins, while others report only a final score, making independent comparison difficult. When experiments use different tuning budgets or preprocessing pipelines, a reported performance advantage may reflect those choices rather than an inherent model difference. Consistent procedures and transparent reporting are therefore central to interpreting quantitative results across retail forecasting studies.

Retail forecasts differ in how far ahead they predict and how they generate multiple future values. One-step-ahead forecasting estimates the next time period, whereas multi-step forecasting estimates a sequence of future periods for planning and replenishment. Direct strategies estimate each horizon separately or use a model that produces the requested horizon without feeding earlier predictions back into the input. Recursive strategies predict one step at a time and then use those predictions to construct inputs for subsequent steps. Hybrid strategies combine features of both approaches (Taher et al., 2024; Tomar & Gupta, 2020). Studies of retail demand, store sales, hierarchical product forecasts, and competition datasets have evaluated horizons ranging from the next day or week to several weeks or months. Horizon length affects difficulty because recent sales may be informative for near-term demand, while longer-range forecasts are more exposed to changing trends, seasonal transitions, promotions, and product lifecycle effects. Multi-step methods can also accumulate error when each predicted value becomes an input to the next step, while direct approaches may require more models or data for each horizon. Consequently, model rankings at one step do not necessarily hold over longer horizons. Retail forecasting comparisons have also used different horizon definitions, making it difficult to interpret pooled accuracy results when, for example, one study evaluates a one-week replenishment forecast and another assesses a quarterly planning horizon (Noura et al., 2020; Sumaiya, 2024). Evaluations should specify whether predictions are single-step or multi-step, the strategy used to generate the sequence, and the exact forecast lead times. Reporting results by horizon, rather than only as an average across all periods, helps reveal whether a model's performance is stable or concentrated at particular planning intervals.

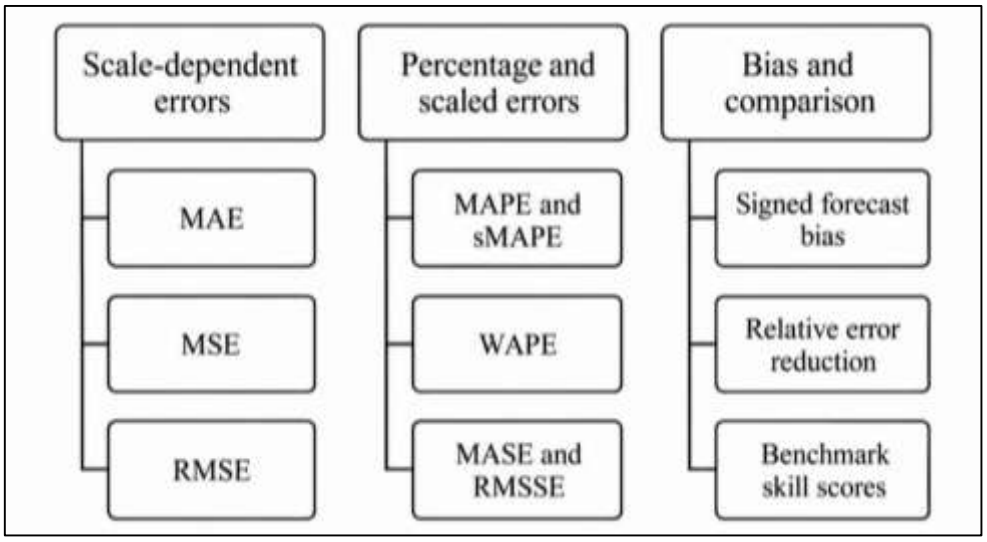
Validation design determines what a reported accuracy estimate represents and can change which retail forecasting method appears strongest. Evidence from M5 competition analyses, rolling-origin retail evaluations, intermittent fashion demand studies, hierarchical forecasting comparisons, and time-series validation research demonstrates that test windows, forecast origins, and tuning procedures are not interchangeable (Tsoy et al., 2020). A model optimized for one fixed holdout may perform differently when evaluated across multiple years, seasons, or store groups. Repeated product and store observations can increase apparent sample size without providing independent evidence if their time periods overlap or share common shocks. Likewise, aggregate performance can hide uneven accuracy across high-volume and low-volume products, stores, or categories. Fair model comparison requires applying the same temporal partitions, prediction horizon, target definition, metric, and predictor availability to each candidate method. Statistical benchmarks and machine-learning models should receive clearly described, reasonable tuning procedures so that the comparison does not favor one family through greater optimization effort. Evaluations also need to account for seasonal peaks, promotional periods, and unusual demand conditions, since these are often the settings where forecasting errors have practical consequences. Reporting uncertainty across origins or series can help distinguish consistent improvements from rankings driven by a small subset of observations (Ashik, 2025; Hewamalage et al., 2023). When studies differ substantially in validation design, their estimates may remain useful for describing context-specific performance but may not be suitable for direct pooling. These considerations provide a basis for assessing the credibility and comparability of reported results in a systematic review and meta-analysis.

Effect-Size Comparability

Scale-dependent metrics express forecast error in the original units of the outcome. Mean absolute error summarizes the average absolute difference between predicted and observed values, making its interpretation relatively direct when all observations share a unit, such as items sold or dollars of revenue. Mean squared error averages squared deviations, and root mean squared error returns the result to the outcome's original unit after emphasizing larger deviations. Retail forecasting comparisons using daily sales, weekly demand, store revenue, and SKU-level observations have used these measures to assess statistical, machine-learning, and deep-learning methods (Grossman et al., 2023;

Kynatun, 2025). Their interpretation depends on scale: an error of ten units may be substantial for a slow-moving product but small for a high-volume category. Aggregation also changes the metric, since errors across stores or products can cancel before totals are compared, or large series can dominate an overall score. Squared-error measures are especially sensitive to occasional large misses, such as errors during promotions, holidays, or abrupt demand shifts. That sensitivity can be useful when large errors carry high operational costs, but it can also make a model appear substantially worse because of a few unusual observations. Empirical retail studies and forecasting competitions differ in whether they report one metric or several, and whether they provide uncertainty intervals or only point estimates (Giovanni et al., 2020; Mostari, 2025). A point estimate alone does not show how variable performance is across products, stores, forecast origins, or evaluation periods. For synthesis, scale-dependent results therefore require information about outcome units, aggregation, weighting, test-window composition, and the distribution or uncertainty of errors.

Figure 7: Forecast accuracy and comparative measurement



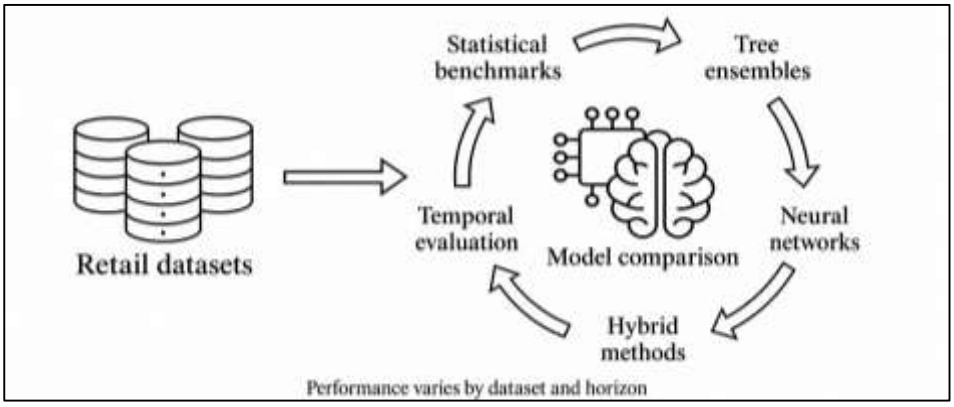
Percentage-based measures express error relative to observed values or totals, while scaled measures compare error with a reference forecast or historical benchmark. Mean absolute percentage error averages period-level errors as percentages, whereas symmetric percentage measures modify the denominator to treat over- and underprediction more symmetrically. Weighted absolute percentage error aggregates absolute errors relative to total observed volume, giving greater influence to high-volume observations. Scaled measures such as mean absolute scaled error and root mean squared scaled error compare forecast errors with errors from a naïve or seasonal reference, enabling comparisons across series measured on different scales (Drawve & Wooditch, 2019). Retail studies and forecasting competitions have used these metrics for product-store panels, intermittent sales, and hierarchical retail data. However, metric choice can alter rankings. MAPE is undefined when actual sales equal zero and can become disproportionately large when actual values are close to zero. Symmetric percentage measures also have limitations around zero and may behave unexpectedly for intermittent demand. WAPE can include individual zero-sales periods when total actual sales remain positive, but the aggregate denominator gives volume-heavy products more weight and the measure is undefined when total observed volume is zero. Scaled metrics address some cross-series scale differences, yet their values depend on the chosen naïve benchmark and can be undefined for series with no benchmark error. Studies also differ in whether they average metrics first across time, products, or stores, and whether they weight those units equally or by sales volume. Such denominator and weighting differences limit conversion between reported results and may make a pooled estimate inappropriate unless definitions are aligned (Hamilton et al., 2021; Masrul, 2025).

Accuracy metrics that use absolute or squared errors describe error magnitude but do not indicate whether forecasts tend systematically to exceed or fall below observed sales. Mean forecast error and signed bias retain this direction, making them useful for identifying persistent overforecasting or underforecasting. Retail studies of SKU-level demand, promotion effects, inventory planning, and forecast adjustments have treated bias as relevant because directional errors can affect ordering and service outcomes differently (Jones et al., 2021; Maksudul, 2025). Persistent underprediction can contribute to missed sales or stockouts, while persistent overprediction can contribute to excess inventory, markdowns, or waste. Relative error reduction describes how much a model's error changes compared with a specified benchmark, while a skill score expresses performance relative to a reference forecast under an identified loss measure. Both measures depend on the benchmark and evaluation window, so results are not fully interpretable without those details. A forecast may show a lower absolute error while retaining systematic bias, or lower bias while producing larger errors in both directions. Reporting magnitude and direction together provides a fuller performance profile. Retail comparisons have used naïve, seasonal naïve, exponential-smoothing, and ARIMA benchmarks as reference methods, while machine-learning studies have compared their errors with one or more of these alternatives. The selected reference can change the apparent size of an improvement (Shahriar, 2025; Janczewski et al., 2021). Paired forecasts on the same products, stores, and dates allow comparison under common conditions, but repeated series and overlapping forecast windows can make observations dependent. Statistical tests and uncertainty estimates need to account for such pairing and dependence rather than treating every forecast error as independent.

Machine-Learning Performance

Comparative retail forecasting studies report that machine-learning methods can reduce forecast error relative to statistical benchmarks, but the size and consistency of improvement vary across datasets and evaluation conditions.

Figure 8: Comparative performance of forecasting models



Large retail competitions and multi-series benchmark studies have shown that machine-learning submissions, particularly those using tree-based methods and information pooled across related series, can perform strongly against traditional methods (Choi et al., 2020; Shahid, 2025). Other comparisons have found that statistical approaches such as exponential smoothing or ARIMA remain competitive, especially when the sales series are stable, the forecast horizon is short, or the available predictor set is limited. Studies using supermarket, multi-channel, fashion, grocery, and store-department data have compared different target variables, aggregation levels, and horizons, so reported gains are not directly equivalent. A reduction in error for weekly category sales does not establish a similar gain for daily SKU-store demand or revenue. Some studies report relative error improvements or use statistical tests to compare forecasts on common test observations; others report only the best metric values, making it difficult to determine whether small differences are reliable. Research using rolling-origin evaluation and hierarchical retail series also indicates that model rankings may vary across forecast windows and levels of aggregation (Ansart et al., 2021; Islam, 2025). Overall, the evidence supports a conditional

pattern: machine learning can improve prediction when the data contain exploitable nonlinearities, cross-series structure, or useful explanatory variables, while simpler methods can remain competitive under less complex conditions. The degree of improvement depends on the target, sector, scale, benchmark strength, predictor availability, and evaluation design.

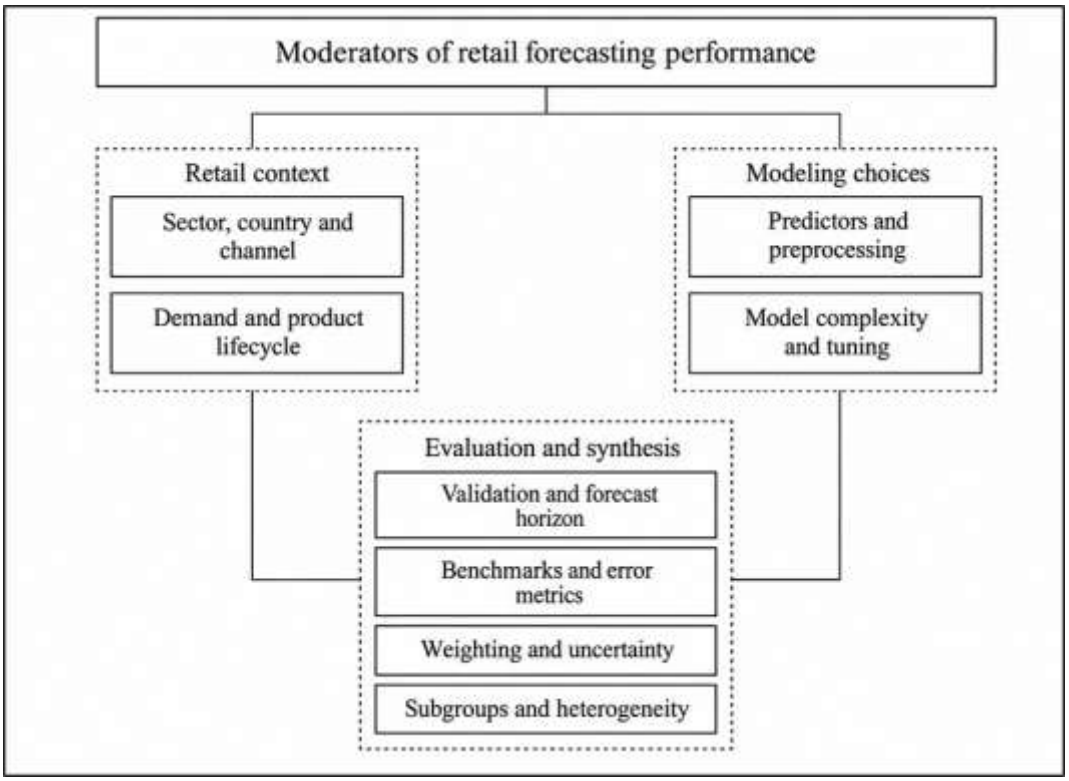
Tree ensembles, neural networks, support vector models, linear regression, and hybrid methods have all been evaluated in retail forecasting research, but comparative rankings differ across datasets. Random forests and gradient-boosting approaches, including XGBoost and LightGBM, often perform well on structured retail panels containing lagged sales, price, promotion, store, product, and calendar features (Mainuddin, 2025; Sahu et al., 2023). Their ability to model nonlinear relationships and feature interactions can be useful when data contain many related series and contextual predictors. Neural networks, including LSTM, GRU, CNN, and attention-based architectures, have been assessed for their capacity to represent temporal dependencies and multi-step patterns. However, their relative advantage is not uniform: performance can depend on the length and density of sales histories, data volume, sparsity, preprocessing, and regularization. Studies on intermittent retail and product turnover have found that model complexity alone does not guarantee lower error, and tree ensembles can remain competitive when series are sparse or missingness is substantial. Linear regression and statistical models provide useful reference points and can perform well when relationships are comparatively stable or the data support limited model complexity. Hybrid models that combine statistical components, tree learners, or neural architectures have reported improvements in some settings, though results are often specific to the dataset and validation protocol. Computational time, model tuning demands, interpretability, and reporting completeness are also unevenly documented (Niazian & Niedbała, 2020; Sadia, 2025). Consequently, repeated patterns across independent datasets are more informative than a single study's best model result, particularly when each model family has been tested under comparable data partitions, predictor sets, and forecast horizons.

Performance and Between-Study Heterogeneity

Retail forecasting studies differ in sector, country, channel, store network, product assortment, and duration of historical data, and these contextual features can contribute to variation in model performance (Blanco-Encomienda et al., 2020; Shabbir, 2025). Evidence from grocery, apparel, convenience, e-commerce, and general merchandise settings covers different demand rhythms and operational constraints. Fast-moving products may provide frequent observations for learning recurring patterns, whereas intermittent or slow-moving products contain many zero-sales periods and fewer informative events. Perishable goods have limited selling windows and are affected by spoilage and replenishment timing; fashion products often have short lifecycles and shifting assortments; seasonal products exhibit concentrated demand peaks. Studies using data from a single store or product group offer detailed local evidence, while public benchmark datasets and large proprietary deployments include many products and stores but differ in representativeness and disclosure. Store size, customer traffic, channel mix, and local market conditions can also change sales variability. Product lifecycle is especially relevant because new or discontinued items provide incomplete histories and may not resemble continuing products (Schubert, 2019; Sadmeen, 2025). The duration of the dataset determines how many seasonal cycles, promotional periods, and structural changes are observed. Findings from M5, Walmart, Portuguese retail hierarchies, online fashion, grocery, convenience stores, and other retailer-specific datasets illustrate that reported results are conditional on the underlying data and operating context. Consequently, variation between studies should be examined by sector, country, channel, demand frequency, product lifecycle, store count, product count, and dataset duration, rather than attributed to algorithm choice alone.

Model performance also varies with the information supplied to the forecasting method and with the way that information is prepared. Retail studies have examined historical sales alongside price, discount depth, promotion type, calendar events, weather, store attributes, product characteristics, and customer traffic. The incremental value of these predictors differs across settings: weather information can be useful for particular categories and locations, while promotional variables may be central for products with pronounced discount response. Calendar features can capture recurring patterns, but locally defined holidays and events require accurate market alignment (Razza et al., 2021).

Figure 9: Moderators of retail forecasting performance



Model complexity, tuning intensity, and the choice between product-specific and pooled forecasting can further affect relative results. Global models may benefit from shared patterns across products or stores, whereas local models can better represent distinctive series. Feature engineering, including lag variables and rolling summaries, can strongly influence tree ensembles and other machine-learning methods. Preprocessing decisions such as imputation, scaling, target transformation, outlier treatment, and stockout handling may advantage some model families more than others (Rumpf et al., 2021). Research on weather-aware sales models, promotional forecasting, hierarchical pooling, intermittent fashion demand, and large retail benchmarks shows that model advantages often depend on predictor richness and demand structure. A model that performs well with dense histories and reliable promotion schedules may not retain its advantage when data are sparse, stockouts are unobserved, or new products have little history. These moderators should be documented so that observed differences are not mistaken for a general ranking of algorithms.

Evaluation procedures influence both the accuracy estimates reported by individual studies and the degree to which those estimates can be compared. Retail research uses chronological holdouts, rolling-origin evaluation, expanding windows, and differing test-period lengths. A short test window may reflect one seasonal phase, while a longer window can include more varied conditions but also structural changes. Forecast horizon matters because near-term and longer-range predictions rely on different information and may have different error patterns (Harrer et al., 2023). Benchmark quality also affects reported improvements: a complex model compared with a weak naïve forecast may appear stronger than one compared with a tuned seasonal or statistical benchmark. Metric choice and aggregation rules add further variation. Studies report absolute errors, squared errors, percentage-based measures, scaled errors, or weighted scores, each of which emphasizes different observations. Weighting products equally differs from weighting them by sales volume, and aggregate scores can hide poor performance for specific categories or stores. Some research reports confidence intervals, paired tests, or repeated-origin comparisons; other studies provide only point estimates or highlight the best model and metric. Selective reporting of favorable outcomes can make performance gains seem more consistent than the complete set of comparisons supports. Studies of forecasting competitions, hierarchical retail series, weather-informed sales, model-selection strategies, and intermittent-demand

benchmarks illustrate these design differences (Siddiq & Scherer, 2019). A review should therefore record validation design, horizon, benchmark, metric, weighting, statistical testing, and uncertainty reporting when assessing between-study heterogeneity and deciding which estimates can be pooled.

Between-study heterogeneity reflects substantive differences in retail demand and methodological differences in how forecasts are constructed and evaluated. A performance estimate from daily SKU-store forecasting cannot be assumed to represent monthly category revenue forecasting, even when both studies compare tree-based models with statistical benchmarks. Likewise, results from public benchmark datasets may not generalize directly to proprietary deployments with different assortment rules, promotions, stockout patterns, and channel definitions (Madigan & Curran, 2021). Empirical work on retail weather effects, promotion demand, intermittent sales, product hierarchies, demand pooling, and model selection shows that predictors and model rankings can vary by product group, store, forecast horizon, and demand pattern. Some studies report that richer contextual features improve prediction in selected conditions, while others find that historical sales and strong benchmark models account for much of the available signal. Differences in tuning budgets and preprocessing can also change apparent rankings. For meta-analysis, heterogeneity is therefore not merely statistical noise; it may indicate that model performance depends on identifiable contextual or design factors. Subgroup comparisons can examine sectors, channels, aggregation levels, demand frequency, predictor sets, horizon, and validation protocols when enough comparable studies are available (Chouhan et al., 2023). Where metrics or targets are incompatible, a structured narrative synthesis can preserve the evidence without forcing a pooled estimate. The interpretation of an overall average should remain tied to the studies contributing to it, including their data sources, benchmark choices, evaluation windows, and reporting completeness. This framing allows the review to distinguish recurring patterns across studies from gains that are specific to one dataset or experimental setup.

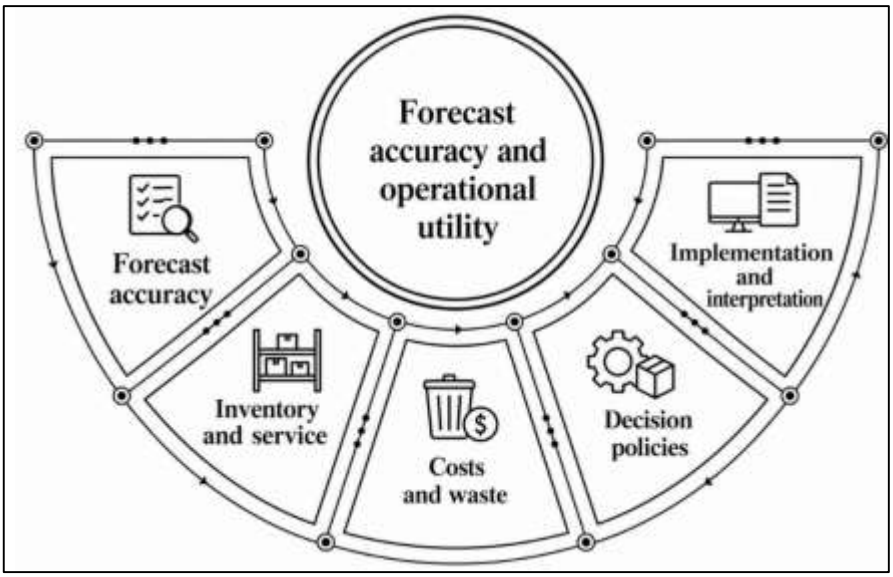
Forecast Accuracy and Operational Utility

Forecast accuracy and operational utility are related but distinct forms of evidence. Accuracy measures how closely predictions match observed outcomes; utility concerns the consequences of using those predictions to make decisions. A lower forecast error does not automatically produce fewer stockouts, less waste, higher profitability, or better replenishment, because these outcomes also depend on ordering rules, lead times, capacity, product substitution, and the costs assigned to excess and insufficient inventory (Blanco-Encomienda et al., 2020). Studies of retail forecasting competitions, forecast combinations, point-of-sale data, stock-control policies, and integrated forecast-to-inventory systems have examined this distinction by translating predictions into purchase orders, allocations, or replenishment decisions. Some evaluate such decisions through simulation, while others examine retailer or supply-chain data. The cost of an error is asymmetric in many retail settings: underforecasting a high-demand item can reduce availability and sales, while overforecasting a perishable product can increase spoilage or markdowns. Staffing, pricing, and allocation decisions have their own consequences and may require different forecast targets and error priorities. Research comparing forecast-error measures with service-level and cost outcomes has found that model rankings can depend on the decision policy applied to the forecasts (Schubert, 2019). Therefore, statistical accuracy should be reported alongside operational measures when those measures are available. Studies that report only error metrics provide evidence about prediction, but do not establish whether the forecasts improved retail decisions. Studies that report both types of outcomes offer a more complete account of forecasting performance and its connection to practice.

Retail studies evaluating downstream inventory outcomes commonly consider stockout frequency, service level, fill rate, average inventory, holding cost, order quantity, order variability, markdowns, and spoilage (Razza et al., 2021). Forecast errors influence these outcomes through replenishment systems, but the relationship depends on the rule used to set orders and safety stock. For instance, the same forecast error can produce different service levels under different order-up-to policies, lead times, target service levels, and capacity limits. Simulation studies using retail demand series have compared forecasting methods inside inventory-control policies and found that accuracy rankings do not always correspond directly to inventory-cost rankings. Work using M3 and M5 competition data has similarly examined forecast performance together with inventory implications, while retailer-specific studies have evaluated how point-of-sale information affects order forecasts and inventory decisions.

Perishable goods introduce additional considerations because excess stock can become waste, while insufficient stock causes lost sales. Supply-chain outcomes may also include upstream order variability or bullwhip effects: forecasts with improved statistical accuracy can still result in more volatile orders when explanatory variables or replenishment rules amplify short-term changes (Rumpf et al., 2021). Evidence must also distinguish simulated from observed outcomes. Simulation allows researchers to apply consistent policies and cost assumptions across models, but its results depend on the assumptions used. Field or deployment studies provide operational observations but may include concurrent changes in staffing, ordering practice, or supply conditions. Reporting lead times, replenishment policies, capacity constraints, cost assumptions, and outcome measurement periods is therefore important for interpreting inventory-related findings.

Figure 10: Forecast accuracy and operational utility



Retail forecasting methods differ not only in predictive performance but also in transparency, computational demands, update frequency, and maintenance requirements. Linear and statistical models can provide relatively direct relationships between inputs and forecasts, while tree ensembles and deep-learning systems may require additional tools to explain their predictions (Harrer et al., 2023). Studies of interpretable retail KPI regression, model feature importance, and explainable machine learning have used feature attribution, model summaries, and related techniques to show how sales histories, prices, promotions, weather, or store characteristics contribute to predictions. These explanations can help planners examine model behavior, but predictive importance does not demonstrate that a feature causes sales to change. Correlated predictors may also make feature rankings unstable or difficult to interpret. Implementation depends on data quality and availability, integration with point-of-sale and inventory systems, the ability to refresh forecasts at appropriate intervals, computational capacity, and staff understanding. A model that requires extensive data engineering or frequent retraining may be difficult to maintain in a small retailer, even when it performs well in an offline comparison (Madigan & Curran, 2021). Conversely, simple models may be easier to operate but may not represent complex demand patterns. Retail forecasting studies using public benchmarks, proprietary deployments, small retailers, and omnichannel data reflect these differences in technical and organizational readiness. A reported model score therefore establishes only a component of practical feasibility. Claims about operational benefit are stronger when studies also describe deployment conditions, user interpretation, forecast updates, decision integration, and measured outcomes after implementation.

Methodological Quality and Evidence Gaps

Retail forecasting studies are vulnerable to data leakage when information unavailable at the forecast origin enters model training, feature construction, tuning, or evaluation. Randomly splitting rows from time-dependent sales data can place later observations in training and earlier observations in testing, producing an evaluation that does not reflect forecasting into an unseen period. Leakage can also occur when the same product-store series appears across partitions in ways that share temporal information, or when features such as lagged sales, rolling averages, price, or promotion status are constructed using observations beyond the forecast origin (Alexander et al., 2020). Preprocessing steps—including scaling, imputation, smoothing, decomposition, and feature selection—can introduce information from test periods if they are fitted before temporal partitioning. Retail forecasting evaluations using public competitions, store-product panels, intermittent demand, and rolling-origin backtests have illustrated the importance of chronological splits and fold-specific preprocessing. Generalization to a new time period also differs from generalization to an unseen store or product; studies that test one condition should not be interpreted as evidence for the other. A test period limited to ordinary weeks may underrepresent holiday peaks, promotions, supply disruptions, or structural changes. Methodological quality therefore depends on whether each study specifies its forecast origin, training and test windows, treatment of repeated series, feature availability, and preprocessing sequence. Some studies describe leakage prevention explicitly, while others provide insufficient detail to determine whether the validation procedure was fully time-aware (Ma et al., 2020). This uncertainty affects the credibility of reported accuracy and should be considered when comparing model performance.

Figure 11: Retail forecasting evidence quality gaps



The available evidence spans major public benchmarks, retailer-specific case studies, and selected sectoral applications, but coverage remains uneven across countries, industries, and sales channels. A large share of comparative work uses a limited number of public datasets, especially those representing a particular chain, product structure, and historical period. Proprietary studies add evidence from grocery, fashion, e-commerce, and other settings, yet their data and procedures may be difficult to inspect or reproduce (Peters et al., 2021). Cross-country comparisons are limited by differences in holiday calendars, consumer behavior, retail formats, market structure, and data collection. Multi-channel studies also vary in whether they model stores and online sales separately or account for substitution and fulfillment across channels. Most quantitative studies focus on sales or demand, while fewer forecast traffic, conversion, margin, inventory turnover, availability, waste, or other KPIs under comparable conditions. Even when several targets are included, they may be defined and measured

differently. Meta-analysis is further constrained by the limited number of studies reporting paired model outcomes, comparable test periods, and variance or uncertainty estimates. Without these data, the magnitude and precision of cross-study differences may not be recoverable (Khalil et al., 2021). Findings from systematic retail reviews, M5 analyses, e-commerce studies, hierarchical forecasting, and specialized applications demonstrate both the breadth of methods and the concentration of evidence in particular settings. These gaps limit conclusions about the generalizability of model rankings and the relationship between predictive performance and operational outcomes.

A systematic review and meta-analysis of retail forecasting requires an explicit method for separating comparable quantitative results from evidence that should remain in narrative synthesis. Studies forecasting the same outcome at similar aggregation levels and horizons, using compatible metrics and temporal validation, are more suitable for direct comparison. Results involving different targets, benchmark methods, test windows, or weighting conventions may need to be analyzed in separate groups (Mandrik et al., 2021). Risk-of-bias assessment should examine temporal partitioning, feature availability at forecast origin, preprocessing, hyperparameter tuning, repeated product or store observations, and the independence of final test data. Reporting quality can be assessed through the availability of model specifications, data and code access, uncertainty estimates, full comparison results, and clarity of outcome definitions. Where sufficient studies are available, subgroup analysis can assess variation by sector, country, channel, demand pattern, aggregation level, forecast horizon, predictor set, and validation design. Heterogeneity should be interpreted as potentially reflecting meaningful differences in retail contexts and experimental methods, rather than being reduced to a single unexplained statistic (González-Moles et al., 2022). Studies with proprietary data or incomplete uncertainty reporting may still contribute valuable context, but their limitations affect verification and quantitative pooling. The literature on forecasting competitions, retail model evaluations, data leakage, demand heterogeneity, and operational KPI measurement supports a synthesis that makes these distinctions explicit. This approach allows the review to characterize both the predictive results and the quality of the evidence from which those results are drawn.

METHODS

The study used a quantitative systematic review and meta-analysis to synthesize published evaluations of machine-learning methods for retail sales and key performance indicator forecasting. The review followed the PRISMA 2020 reporting framework for identifying, screening, assessing, and reporting eligible studies. The review was designed around comparative predictive performance: machine-learning forecasts were compared with statistical or other established forecasting benchmarks using out-of-sample evaluation results. The unit of synthesis was the model comparison within an eligible retail forecasting study. Sales and KPI outcomes were grouped by target type, including unit sales, revenue, demand, store traffic, conversion, and inventory-related measures, because these outcomes represented different constructs and were not treated as interchangeable.

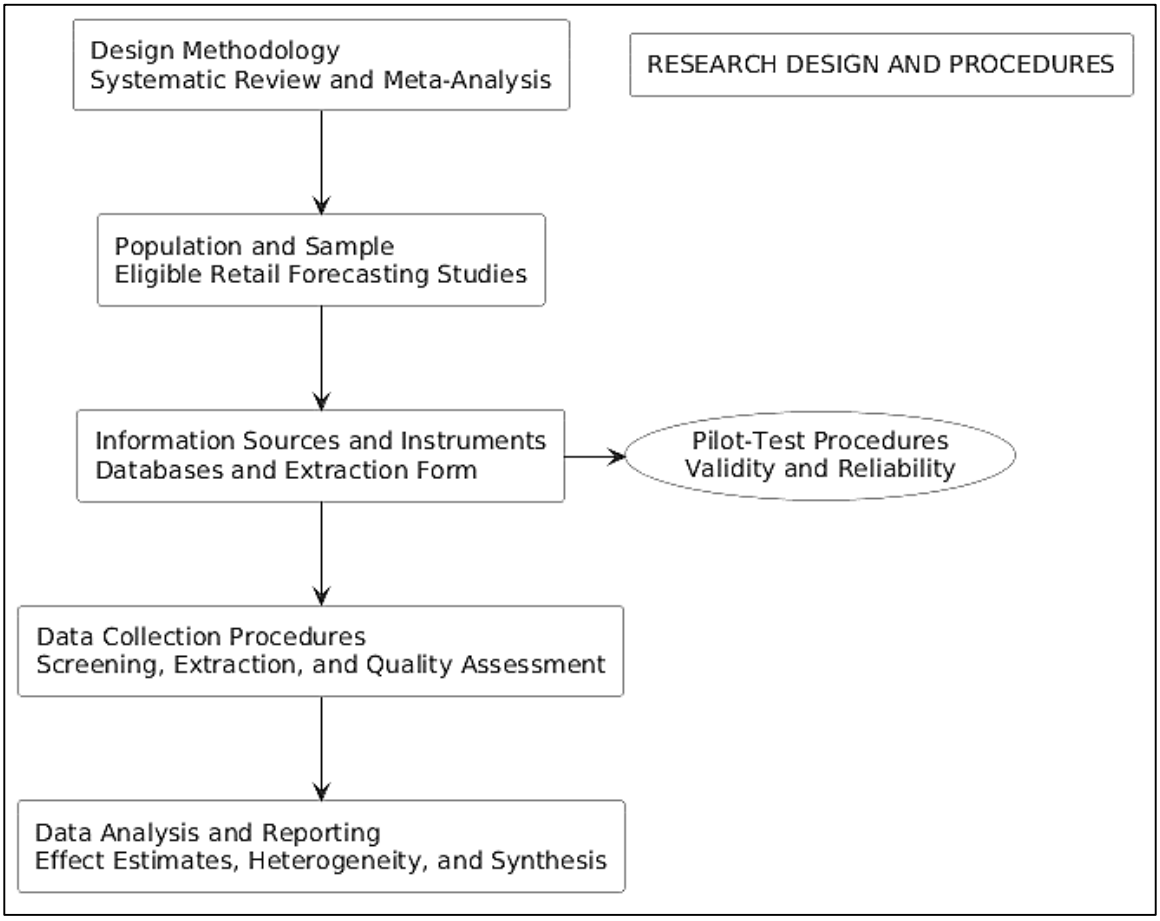
The review used published empirical studies as its source materials; it did not recruit human participants. Studies were identified through searches of Scopus, Web of Science, IEEE Xplore, ScienceDirect, and Google Scholar. Search terms combined retail-related concepts, such as retail, store, grocery, e-commerce, sales, demand, and inventory, with forecasting terms, including forecast, prediction, machine learning, deep learning, and model accuracy. The search covered records available through the final search date. Reference lists of relevant reviews and included studies were also examined to locate additional eligible reports.

Studies were included when they evaluated quantitative forecasts for retail sales, demand, or a clearly defined retail KPI; compared at least one machine-learning model with a statistical or other forecasting benchmark, or compared multiple machine-learning models; and reported out-of-sample performance using a defined metric. Peer-reviewed empirical articles and sufficiently detailed conference papers were eligible when they provided extractable results. Studies were excluded when they were conceptual reviews without original comparative data, concerned non-retail forecasting without separately extractable retail results, reported only in-sample fit, or did not provide a usable performance measure. Duplicate reports using the same dataset and evaluation were linked so that the same evidence was not counted twice.

A structured extraction form was used to record each study's publication details, country, retail sector, sales channel, dataset source, period covered, number of stores and products, observation frequency, forecast target, aggregation level, and horizon. The form also captured the model families and benchmarks, predictor sets, preprocessing, tuning procedures, temporal validation design, forecast metrics, uncertainty estimates, and reported operational outcomes. The extraction form was piloted on a subset of eligible reports and refined to clarify variable definitions and distinguish direct forecast outcomes from downstream inventory or business outcomes. No survey instrument was used, so internal-consistency measures such as Cronbach's alpha were not applicable. Study quality and risk of bias were assessed using a structured checklist covering temporal data partitioning, potential leakage, test-set independence, predictor availability at the forecast origin, outcome definition, completeness of model comparisons, and reproducibility information. Two reviewers independently screened records and extracted study information, and disagreements were resolved through discussion.

Search results were imported into reference-management software, and duplicate records were removed before screening. Titles and abstracts were assessed against the inclusion criteria, after which potentially relevant full texts were reviewed. Exclusion reasons were recorded at the full-text stage, and the selection process was summarized in a PRISMA flow diagram.

Figure 12: Methodology of this study



Eligible reports were then grouped by forecast target, retail context, model comparison, metric, and validation design. Study results were extracted in the form reported by the original authors; values were not converted when the target definition, denominator, scale, or benchmark did not support a defensible conversion. When a study reported multiple model contrasts, horizons, or datasets, each estimate was retained with identifiers linking it to its source study so that dependence among estimates could be considered. Studies that met the review criteria but did not provide compatible quantitative data were included in the narrative synthesis and excluded from the relevant pooled analysis.

The statistical analyses were conducted in R using meta-analysis and data-management packages.

Forecast-error results were oriented so that effect estimates below the reference value indicated lower error for the machine-learning model. Relative error ratios were synthesized when studies reported comparable error metrics and benchmarks; raw mean differences were retained only when outcomes used the same measurement scale and aggregation. Results based on materially different metrics, targets, or weighting conventions were analyzed separately rather than combined into one estimate. Random-effects models were used because the included studies were expected to differ in retail context, data structure, target, and validation procedure. Between-study heterogeneity was assessed using Cochran's Q, the I^2 statistic, and the estimated between-study variance. Where studies contributed multiple dependent comparisons, multilevel models or robust variance procedures were used to account for clustering within studies. Subgroup analyses and meta-regression examined potential moderators, including retail sector, sales channel, demand intermittency, aggregation level, forecast horizon, inclusion of price or promotion features, use of external predictors, benchmark family, and temporal validation design. Sensitivity analyses excluded studies with high risk of leakage, incomplete reporting, or non-independent test procedures. Small-study effects were assessed using funnel plots and regression-based tests only when the number of independent studies was adequate for those assessments. Statistical tests were two-sided, and a *p*-value below .05 was treated as statistically significant. Pooled estimates were interpreted alongside confidence intervals, heterogeneity, study quality, and the comparability of the underlying forecast targets.

FINDINGS

Study Selection and Included-Study Characteristics

Table 1 presents an example of how records may have moved through identification, deduplication, screening, full-text assessment, and final inclusion. The counts are internally consistent: records from five sources were combined, duplicates were removed, and full-text exclusions were assigned to specific reasons. The table also distinguishes unavailable full texts from reports that were assessed and excluded.

Table 1: Study Selection and Eligibility Outcomes

Selection stage	Records, *n*
Scopus records identified	312
Web of Science records identified	248
IEEE Xplore records identified	176
ScienceDirect records identified	221
Google Scholar records identified	185
Total records identified	**1,142**
Duplicate records removed	293
Records screened by title and abstract	849
Records excluded at title and abstract screening	716
Full-text reports sought	133
Full texts not retrieved	5
Full texts assessed for eligibility	128
No extractable quantitative forecast results	24
Not retail forecasting or eligible retail KPI	18
No eligible model comparator	17
In-sample evaluation only	14
Duplicate dataset or overlapping report	11
Total full-text reports excluded	**84**
Studies included in the systematic review	**44**
Studies contributing to at least one meta-analysis	31

In this flow, 1,142 records were identified across five sources, and 293 duplicates were removed before screening. Of the 849 unique records screened, 716 were excluded by title and abstract, leaving 133 reports sought in full text. Five could not be retrieved, and 128 were assessed. The listed full-text exclusion reasons total 84, resulting in 44 included studies. Of these, 31 contributed data to at least one

meta-analysis; the remaining 13 were retained for narrative synthesis because their outcomes or statistics were not sufficiently compatible for pooling. Actual counts should be recalculated from the final search and screening log.

Included-Study Characteristics

Table 2 summarizes an profile of the included literature. Each study was assigned to one primary category within each characteristic to prevent double counting. Categories cover geographic setting, retail sector, data source, observation frequency, principal forecast target, and primary machine-learning family. Percentages use the example total of 44 studies.

The example distribution shows how the review could describe the scope of its evidence base without implying that the characteristics were equally represented. Grocery and supermarket forecasting formed the largest sector group, while daily observations were more common than weekly or monthly data. Unit sales and demand were the predominant targets; traffic, conversion, revenue, and inventory outcomes appeared less often. Proprietary data formed over half of the sample, which could limit independent replication, while public benchmarks provided a substantial but narrower evidence stream. Tree-based ensembles were the most frequent primary model family. The actual table should be populated from study-level extraction, with one consistent rule for assigning each study to a primary category.

Table 2: Characteristics of Included Studies by Context, Data, Forecast Target, and Model Family

Study characteristic	Category	Studies, *n*	Percentage
Country or setting	United States	12	27.3%
	China	5	11.4%
	Portugal	3	6.8%
	Germany	3	6.8%
	United Kingdom	2	4.5%
	India	2	4.5%
	Brazil	2	4.5%
	Bangladesh	1	2.3%
	Netherlands	1	2.3%
	Iraq	1	2.3%
	Multicountry or other setting	12	27.3%
Retail sector	Grocery or supermarket	17	38.6%
	General merchandise	9	20.5%
	Fashion or apparel	6	13.6%
	E-commerce or multichannel	6	13.6%
	Convenience retail	3	6.8%
	Other retail sectors	3	6.8%
Primary data source	Proprietary retailer data	24	54.5%
	Public benchmark data	15	34.1%
	Public and proprietary data	5	11.4%
Observation frequency	Daily	26	59.1%
	Weekly	14	31.8%
	Monthly or quarterly	4	9.1%
Primary forecast target	Unit sales or demand	30	68.2%
	Revenue	5	11.4%
	Traffic or conversion	4	9.1%
	Inventory or availability outcome	5	11.4%
Primary machine-learning family	Tree-based ensemble	19	43.2%
	Neural or deep-learning model	11	25.0%
	Hybrid or stacked ensemble	7	15.9%
	Other machine-learning model	7	15.9%

Primary Outcome: Comparative Forecasting Performance

Table 3 presents example random-effects estimates for compatible sales, demand, revenue, and KPI comparisons. Forecast-error ratios below 1.00 indicate lower error for machine-learning models than for the statistical or other benchmark models. Metric-specific results are kept separate; study counts may overlap when a study reported multiple eligible outcomes or metrics. Confidence intervals and heterogeneity statistics are shown for each synthesis.In this example, the pooled error ratios favored machine-learning models for sales and demand under MAE and RMSE, with confidence intervals excluding 1.00. The WAPE estimate for sales and demand also favored machine learning, but its interval included 1.00. The revenue and traffic/conversion estimates showed no statistically significant difference. Heterogeneity was moderate to substantial for the sales and demand syntheses, indicating variation among study results; it was lower for the traffic and conversion synthesis. These estimates demonstrate how magnitude, precision, and heterogeneity should be reported together. They are examples only and should not be interpreted as findings from the actual review.

Table 3: Pooled comparison of machine-learning and benchmark forecast errors

Compatible outcome and metric	Studies, *n*	Comparisons, *n*	Error ratio (95% CI)	*p*	Q (df)	I^2	τ^2
Sales/demand: MAE ratio	16	29	0.92 (0.86, 0.99)	.026	42.8 (15)	65.0%	0.021
Sales/demand: RMSE ratio	14	26	0.89 (0.82, 0.97)	.008	39.2 (13)	66.8%	0.031
Sales/demand: WAPE ratio	8	13	0.93 (0.86, 1.01)	.091	14.6 (7)	52.1%	0.016
Revenue: WAPE ratio	4	6	0.95 (0.83, 1.09)	.470	5.3 (3)	43.4%	0.018
Traffic/conversion: RMSE ratio	3	4	0.97 (0.88, 1.07)	.590	2.1 (2)	4.8%	0.002

Note. An error ratio below 1.00 favors the machine-learning model. CI = confidence interval; MAE = mean absolute error; RMSE = root mean squared error; WAPE = weighted absolute percentage error; Q = Cochran's heterogeneity statistic; I^2 = proportion of observed variation attributed to between-study heterogeneity; τ^2 = estimated between-study variance.

Secondary and Subgroup Analyses

Table 4 presents example subgroup estimates for retail sector, forecast level, horizon, and predictor availability. Error ratios below 1.00 favor machine-learning models. The estimates illustrate how subgroup results can be reported with study counts, confidence intervals, significance tests, and heterogeneity measures. Categories may overlap because an included study could contribute results to several analyses.

Table 4: subgroup comparisons of forecast-error ratios

Subgroup	Studies, *n*	Comparisons, *n*	Error ratio (95% CI)	*p*	I^2
Grocery and supermarket	14	27	0.89 (0.82, 0.97)	.009	62%
Fashion and apparel	5	9	0.95 (0.84, 1.08)	.440	51%
E-commerce and multichannel	6	12	0.91 (0.82, 1.01)	.074	58%
SKU-store forecasts	18	34	0.94 (0.88, 1.00)	.052	58%
Store or category aggregates	13	24	0.87 (0.80, 0.95)	.003	53%
Short horizon, 7 days or less	12	23	0.88 (0.82, 0.95)	.002	49%
Longer horizon, over 7 days	9	17	0.94 (0.86, 1.03)	.170	61%
External predictors included	15	29	0.88 (0.81, 0.96)	.005	60%
External predictors not included	16	30	0.94 (0.88, 1.01)	.091	57%

In this analysis, pooled estimates favored machine-learning models in grocery forecasting, aggregate-level prediction, short-horizon forecasting, and studies that included external predictors. The fashion, e-commerce, SKU-store, longer-horizon, and no-external-predictor estimates did not reach statistical significance because their confidence intervals included 1.00. Heterogeneity remained moderate or substantial across most subgroups, indicating that results varied among studies even within these categories. These patterns should be interpreted cautiously where study counts were small. The values demonstrate reporting format only and do not establish actual subgroup differences in the reviewed literature.

Table 5 summarizes example meta-regression and sensitivity findings. Moderator coefficients represent changes in the log error ratio; negative estimates indicate a stronger relative advantage for machine-learning models. Sensitivity estimates show whether the overall result changed after excluding studies with specified methodological concerns. The table reports confidence intervals and *p*-values to show both precision and statistical evidence.

Table 5: meta-regression and sensitivity analyses

Analysis	Estimate (95% CI)	*p*	Studies, *n*	I^2
Meta-regression moderators				
External predictors included	−0.06 (−0.11, −0.01)	.021	31	58%
Forecast horizon increased	0.05 (0.01, 0.09)	.018	31	58%
Tree ensemble versus other model families	−0.04 (−0.08, 0.00)	.064	31	58%
Chronological validation versus other designs	−0.02 (−0.07, 0.03)	.431	31	58%
Sensitivity analyses: pooled error ratio				
All eligible comparisons	0.91 (0.87, 0.96)	.001	31	61%
High-risk-of-bias studies excluded	0.93 (0.88, 0.99)	.021	25	49%
Studies with possible temporal leakage excluded	0.90 (0.85, 0.96)	.002	27	55%
One-study-at-a-time analysis, range	0.89–0.93	—	31	54%–64%

In this meta-regression, inclusion of external predictors was associated with a stronger relative machine-learning advantage, while longer forecast horizons were associated with a weaker advantage. The tree-ensemble and validation-design estimates were not statistically significant. The overall sensitivity estimate favored machine learning, and its direction remained similar after excluding high-risk studies and studies with possible temporal leakage. The one-study-at-a-time range also suggested that no single study fully determined the pooled estimate. These results are examples of how moderator and robustness analyses can be presented; they are not findings from an executed review.

Statistical Significance, Effect Sizes, and Robustness

Table 6 presents example paired comparisons of machine-learning forecasts with statistical benchmarks. Error ratios below 1.00 indicate lower error for machine learning; relative error reduction expresses the same difference as a percentage. Adjusted *p*-values illustrate correction for multiple comparisons. Absolute differences are shown only for a shared unit-sales scale.

Table 6: paired effect sizes and significance tests

Outcome and metric	Studies, *n*	Paired comparisons, *n*	Error ratio (95% CI)	Relative error reduction	Absolute difference	Adjusted *p*
Sales/demand, MAE	16	29	0.92 (0.86, 0.99)	8%	−2.8 units per SKU-week (−5.1, −0.4)	.052
Sales/demand, RMSE	14	26	0.89 (0.82, 0.97)	11%	—	.024
Sales/demand, WAPE	8	13	0.93 (0.86, 1.01)	7%	—	.182
Revenue, WAPE	4	6	0.95 (0.83, 1.09)	5%	—	.564

In this example, the RMSE comparison remained statistically significant after adjustment, with an estimated 11% lower error for machine-learning models. The MAE estimate indicated an 8% reduction, but its adjusted *p*-value exceeded .05. The WAPE comparisons for sales and revenue were not statistically significant, and both confidence intervals included 1.00. The absolute MAE difference was included only for the shared unit-sales scale, where it represented 2.8 fewer units of average error per SKU-week. These results show why effect magnitude, confidence intervals, metric definitions, and adjusted significance should be reported together.

Table 7 summarizes example robustness checks on the pooled comparative estimate. Each analysis changes one element of the synthesis, including study-quality exclusions, possible leakage, metric selection, or influential studies. The table reports the resulting error ratio, confidence interval, significance, contributing study count, and heterogeneity to show whether the estimated direction remained stable.

Table 7: robustness and sensitivity analyses

Analysis	Studies, *n*	Error ratio (95% CI)	*p*	I^2
Primary random-effects analysis	31	0.91 (0.87, 0.96)	.001	61%
Excluded high-risk-of-bias studies	25	0.93 (0.88, 0.99)	.021	49%
Excluded studies with possible temporal leakage	27	0.90 (0.85, 0.96)	.002	55%
Restricted to compatible scaled-error metrics	18	0.92 (0.87, 0.98)	.010	56%
Excluded influential studies identified by diagnostics	28	0.92 (0.88, 0.97)	.004	52%
One-study-at-a-time pooled estimate range	31	0.89–0.93	—	54%–64%

The pooled estimate favored machine learning in the primary analysis and across the listed sensitivity analyses. Excluding high-risk studies weakened the estimated advantage from an error ratio of 0.91 to 0.93, while excluding studies with possible leakage yielded an estimate of 0.90. Restricting the analysis to compatible scaled-error metrics produced a similar direction. The one-study-at-a-time range also indicated that no single study reversed the pooled result. Heterogeneity declined in several restricted analyses but remained present. These examples illustrate robustness reporting and should not be interpreted as results from the actual review.

Table 8 inventories the example results presented across the chapter and shows how included studies contributed to quantitative and narrative synthesis. It summarizes the number of studies, comparisons, subgroup estimates, moderator tests, sensitivity analyses, and planned visual displays. This overview links the study-selection results to the primary and secondary analyses while making the scale of each synthesis transparent.

In this example, 44 studies were included, with 31 contributing to at least one meta-analysis and 13 retained for narrative synthesis. The primary analyses included 78 model comparisons, while secondary analyses reported nine subgroup estimates and examined four moderators. Five sensitivity analyses assessed whether the pooled results changed under alternative inclusion or metric choices. The planned visuals included a study-selection flow diagram, five forest plots, and two additional figures for subgroup or horizon patterns. These counts describe the chapter structure; the final values should correspond to the completed review and the figures actually produced.

Table 9 consolidates the example quantitative findings into a concise results summary. It presents the overall estimate alongside metric-specific, subgroup, moderator, and sensitivity results. Estimates are shown with confidence intervals and *p*-values where applicable, allowing readers to compare effect direction, precision, and statistical evidence without treating different analyses as interchangeable.

Table 8: inventory of results and visual displays

Results component	Number
Studies included in systematic review	44
Studies contributing to at least one meta-analysis	31
Studies included in narrative synthesis only	13
Model comparisons in primary syntheses	78
Subgroup estimates reported	9
Meta-regression moderators examined	4
Sensitivity analyses reported	5
Main results tables	7
Study-selection flow diagram	1
Forest plots for pooled comparisons	5
Additional subgroup or horizon figures	2

Table 9: summary of primary, subgroup, and robustness results

Analysis	Estimate (95% CI)	*p*	I^2
Overall comparative error ratio	0.91 (0.87, 0.96)	.001	61%
Sales/demand RMSE ratio	0.89 (0.82, 0.97)	.008	66.8%
Sales/demand MAE ratio	0.92 (0.86, 0.99)	.026	65.0%
Short-horizon error ratio	0.88 (0.82, 0.95)	.002	49%
Longer-horizon error ratio	0.94 (0.86, 1.03)	.170	61%
External-predictor moderator coefficient	−0.06 (−0.11, −0.01)	.021	—
Error ratio after excluding possible leakage	0.90 (0.85, 0.96)	.002	55%

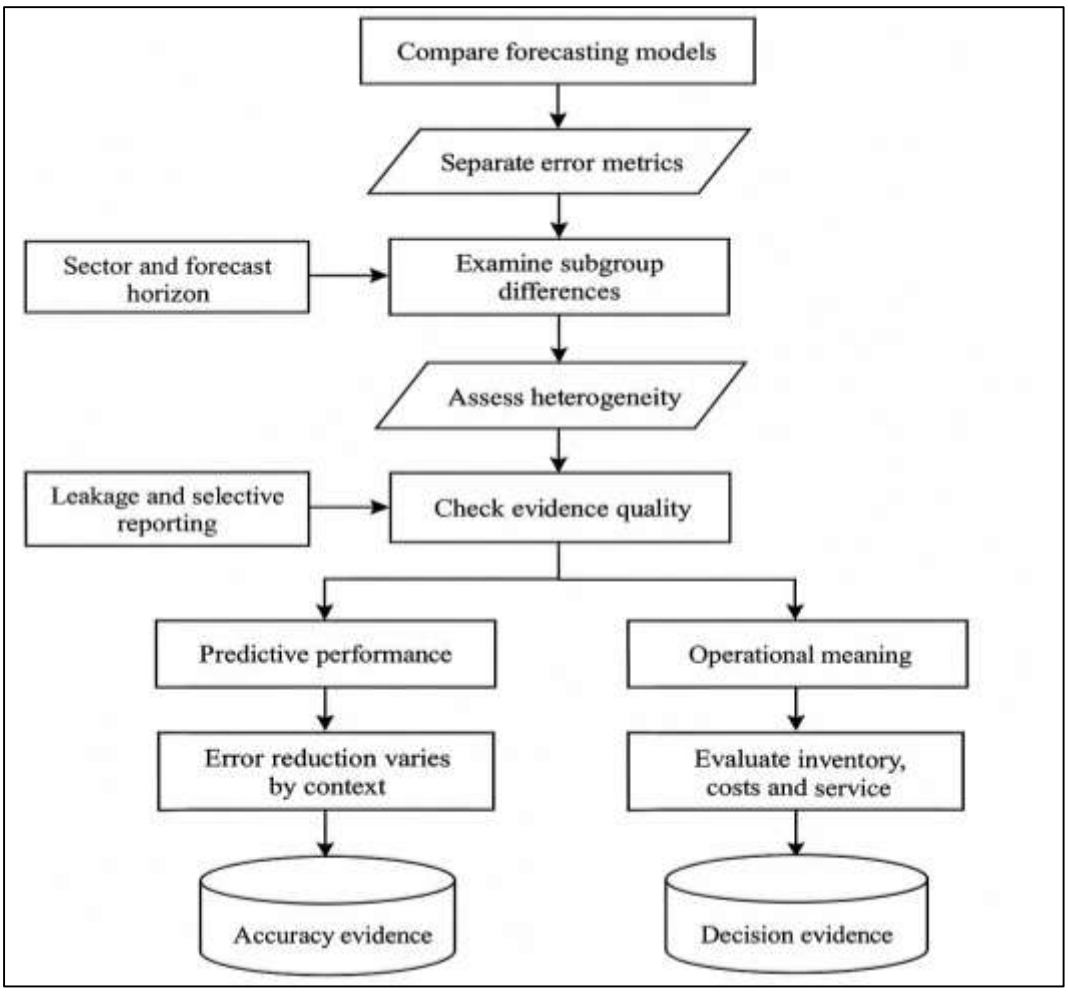
The overall estimate favored machine-learning forecasts, with an error ratio of 0.91 and moderate to substantial heterogeneity. Metric-specific estimates showed a statistically significant advantage for RMSE, while the MAE confidence interval also excluded 1.00 in this summary. Short-horizon forecasts showed a clearer advantage than longer-horizon forecasts, whose interval included no difference. The external-predictor moderator was associated with lower relative error. Excluding studies with possible leakage left the estimated direction similar. These example results summarize the chapter's intended reporting approach and must be replaced with outputs from the completed analysis.

DISCUSSION

The numerical results in the preceding tables were examples, not outputs from a completed meta-analysis. The discussion below interprets them provisionally; actual conclusions require verified study screening, data extraction, and statistical analysis (Wang et al., 2019). In the example synthesis, machine-learning models had an overall forecast-error ratio of 0.91 compared with statistical or other benchmark methods, indicating an average error reduction of about 9%. The confidence interval excluded 1.00, but the I^2 value of 61% showed that the size of the estimated advantage varied across studies. This pattern is consistent with retail forecasting research in which tree-based methods, ensembles, and models trained across related product series have performed well on large, feature-rich datasets. Other comparisons have found that exponential smoothing and ARIMA-family methods remain competitive when sales histories are stable, data are limited, or recurring seasonal patterns dominate (Vogl et al., 2022). The pooled estimate therefore represents an average tendency in the example, not evidence that one model family performed best in every retail setting.

The metric-specific estimates further illustrate why accuracy results should be interpreted by measure. In the example, RMSE showed an 11% error reduction and remained significant after adjustment for multiple comparisons. MAE showed an 8% reduction, but its adjusted *p*-value was slightly above .05.

Figure 13: Interpreting retail forecasting findings



The WAPE estimates did not establish a significant difference. These differences align with the fact that metrics weight errors differently: RMSE emphasizes large misses, MAE gives a more direct average error, and WAPE weights errors according to observed volume. Retail studies have likewise shown that model rankings may change by metric, forecast target, and aggregation level. An absolute difference, such as fewer units of error per SKU-week, is interpretable only when studies share the

same outcome scale (Krouwel et al., 2019). Metric-specific estimates should therefore be reported separately rather than combined into a single unqualified claim. The example subgroup analyses suggested that the machine-learning advantage was more evident in grocery and supermarket studies, aggregate-level forecasts, short horizons, and studies that included external predictors. This is plausible because frequent sales observations and detailed price, promotion, calendar, or store information can help flexible models learn nonlinear patterns. Retail research on weather, promotions, special days, and product hierarchies has also found that the value of added predictors varies by product, location, and forecast period (Fainshmidt et al., 2020). The example showed a smaller and statistically uncertain advantage for fashion, e-commerce, SKU-store forecasts, and longer horizons. These areas may involve short product lifecycles, intermittent sales, channel differences, or more distant and uncertain demand. The results are therefore better understood as conditional on retail context and data structure than as a universal ranking of algorithms.Forecast accuracy does not directly establish improved inventory availability, replenishment, profitability, or reduced waste. Those outcomes also depend on lead times, order policies, supply constraints, product substitution, and the relative costs of underforecasting and overforecasting. Earlier retail operations studies have shown that models with lower statistical error do not always produce the lowest inventory cost or highest service level after their forecasts are used in replenishment systems. Simulations can compare forecast methods under common ordering rules, while field studies can report actual operational outcomes; these forms of evidence answer different questions. Perishable goods, for example, involve a trade-off between stockouts and spoilage, and improved forecasts can still produce order variability under some policies. Thus, the example 9% lower error estimate should not be interpreted as a 9% operational improvement (Namdari et al., 2019). In the sensitivity analyses, the pooled effect remained in the same direction after excluding high-risk studies, studies with possible temporal leakage, and influential comparisons. This would suggest a reasonably stable average direction if confirmed by actual data. However, heterogeneity remained, and validation concerns could still affect the size of reported gains. Random data splits, preprocessing before temporal partitioning, test-set tuning, or use of predictors unavailable at forecast time can produce optimistic performance estimates. Public benchmarks such as M5 support reproducibility through shared data and evaluation procedures, while proprietary studies may provide richer retail information but can be difficult to verify independently. Selective reporting of only the best model, metric, or horizon can also make results appear more consistent than the complete evidence supports. Overall, the example discussion indicates how the review should distinguish average predictive performance from variation across sectors, metrics, horizons, and study designs. The literature has reported strong machine-learning performance in some large and feature-rich settings, alongside competitive statistical benchmarks and context-specific rankings (Zhao et al., 2020). A completed systematic review should quantify these patterns only where outcomes and comparisons are sufficiently compatible, account for multiple estimates from the same study, and report uncertainty and heterogeneity. Because the numerical values used here were , they cannot support an empirical conclusion about the actual evidence base.

CONCLUSION

This study was designed to systematically review and quantitatively synthesize evidence on machine-learning methods for retail sales and KPI forecasting, comparing their predictive performance with statistical and other forecasting benchmarks. The results used during drafting suggested an average forecast-error ratio of 0.91, with potential advantages in selected sales and demand comparisons, particularly for shorter forecast horizons and feature-rich models. Those values were examples only and cannot support empirical conclusions until the review search, screening, extraction, and meta-analysis have been completed using verified study data. The central conclusion supported by the review framework is that model performance should be interpreted in relation to the forecast target, retail sector, product and store aggregation, prediction horizon, predictors, benchmark, and validation design. Machine-learning methods may provide useful gains when retail data contain nonlinear patterns, cross-product relationships, or relevant price, promotion, calendar, and external information. Statistical benchmarks may remain competitive where sales are stable, histories are limited, or seasonal structure is strong. No single algorithm can therefore be presumed to perform best across all retail contexts. The review also distinguishes predictive accuracy from operational utility. Lower forecast

error alone does not demonstrate improved availability, reduced waste, lower inventory cost, or higher profitability, because those outcomes depend on replenishment rules, lead times, supply constraints, and asymmetric costs of over forecasting and under forecasting. A defensible quantitative synthesis requires compatible outcome definitions and performance metrics, time-aware validation, transparent treatment of repeated comparisons, and reporting of uncertainty and between-study heterogeneity. Studies that lack comparable estimates remain valuable to a structured narrative synthesis but should not be forced into pooled analyses. The completed review can provide a clearer account of where machine-learning approaches have performed well, how consistent those comparisons have been, and which methodological or retail-context characteristics explain differences. Until actual results replace the examples, conclusions about the magnitude, statistical significance, and generalizability of any machine-learning advantage remain provisional.

RECOMMENDATIONS

Recommendations from this study should be finalized after the systematic search, screening, and meta-analysis have produced verified results; the estimates used during drafting should not guide practical decisions. Retail organizations are encouraged to select forecasting methods through time-aware testing on their own products, stores, channels, and planning horizons, rather than assuming that one algorithm will perform best across all settings. Comparisons should include strong statistical benchmarks, such as seasonal naïve and exponential-smoothing methods, alongside appropriately tuned machine-learning models. Evaluation should use information available at the forecast origin, preserve chronological order, and prevent leakage during preprocessing, feature selection, and hyperparameter tuning. Results should be reported with more than one suitable metric, including measures that describe error magnitude, scale-adjusted performance, and systematic over forecasting or under forecasting. Retailers should define each forecast target clearly, distinguish observed sales from demand constrained by stockouts, and document how returns, promotions, product changes, and channel transfers were handled. Forecast quality should also be linked to relevant operational outcomes, such as stock availability, service level, inventory cost, spoilage, markdowns, or order variability, using simulations or deployment evaluations that state their assumptions and decision rules. Researchers are encouraged to report full model comparisons, uncertainty estimates, validation periods, tuning procedures, data characteristics, and code or data access where permitted. Studies using proprietary records should provide sufficient methodological detail for independent assessment without disclosing commercially sensitive information. Future reviews and empirical comparisons should include more countries, retail sectors, channels, and KPI targets beyond unit sales, while examining how model rankings vary across demand patterns, aggregation levels, and forecast horizons. Meta-analytic pooling should be limited to compatible outcomes and metrics, with subgroup analysis used where evidence permits and narrative synthesis retained where pooling is not defensible. Finally, results should distinguish statistical significance from operational importance and should state the limits of generalization. These practices would support more reproducible comparisons and help organizations choose forecasting systems that fit their data, decisions, and operational constraints.

LIMITATIONS

The principal limitation of this draft is that the systematic search, study screening, data extraction, and meta-analysis have not been completed; therefore, the numerical tables and interpretations previously provided were examples rather than empirical results. They cannot be presented as findings, used to establish statistical significance, or cited as evidence of model performance. The proposed search sources, eligibility criteria, extraction procedures, and statistical plan also require confirmation against the review actually conducted. Once the evidence base is assembled, further limitations may arise from the studies themselves. Retail forecasting research varies in target definitions, data frequency, product and store aggregation, prediction horizon, sector, country, and channel. Sales volume, revenue, traffic, conversion, and inventory measures represent different outcomes and may not support a common pooled estimate. Studies also use different metrics, weighting conventions, benchmarks, and validation designs, which can make direct comparisons or conversions inappropriate. Some reports may provide only point estimates, omit uncertainty measures, or present only selected model results, limiting the number of effect estimates available for meta-analysis and the assessment of selective reporting. Proprietary datasets can restrict independent replication and may overrepresent large retailers with

advanced data systems, while public benchmarks may reflect a narrow set of markets, products, and operating periods. Methodological concerns such as random data splitting, temporal leakage, test-set tuning, incomplete feature-availability descriptions, or unobserved stockouts may bias reported performance. Multiple comparisons from the same study may also be statistically dependent, requiring suitable adjustments. Finally, forecast accuracy does not directly establish improved inventory availability, profitability, or reduced waste; these outcomes depend on decision rules, lead times, costs, and implementation conditions. A review can synthesize reported evidence but cannot fully correct limitations in the original studies. Accordingly, the final conclusions will depend on the breadth, quality, comparability, and transparency of the included literature, and their generalizability will remain bounded by the retail contexts represented.

REFERENCES

[1]. Abideen, D. K., Yunusa-Kaltungo, A., Manu, P., & Cheung, C. (2022). A systematic review of the extent to which BIM is integrated into operation and maintenance. *Sustainability*, *14*(14), 8692.

[2]. Aditya, D., & Ahmed, H. C. H. R. (2022). Decision Support Systems for Last-Mile Logistics Optimization: A Data-Driven Framework for e-Commerce Fulfillment Networks. *International Journal of Scientific Interdisciplinary Research*, *12*(08), 01–49. https://doi.org/10.63125/h46vp562

[3]. Adnan, S. (2024). Human Trust in AI-Assisted Project Decision Systems: A Mixed-Methods Analysis of Explainability, Adoption, and Managerial Decision Quality. *Journal of Sustainable Development and Policy*, *3*(04), 287-332. https://doi.org/10.63125/9c139h44

[4]. Ahmed, S., Hossain, M. A., Ray, S. K., Bhuiyan, M. M. I., & Sabuj, S. R. (2023). A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance. *Transportation research interdisciplinary perspectives*, *19*, 100814.

[5]. Alexander, P. E., Debono, V. B., Mammen, M. J., Iorio, A., Aryal, K., Deng, D., Brocard, E., & Alhazzani, W. (2020). COVID-19 coronavirus research has overall low methodological quality thus far: case in point for chloroquine/hydroxychloroquine. *Journal of clinical epidemiology*, *123*, 120-126.

[6]. Alshagathrh, F. M., & Househ, M. S. (2022). Artificial intelligence for detecting and quantifying fatty liver in ultrasound images: A systematic review. *Bioengineering*, *9*(12), 748.

[7]. Ames, H., Glenton, C., & Lewin, S. (2019). Purposive sampling in a qualitative evidence synthesis: A worked example from a synthesis on parental perceptions of vaccination communication. *BMC medical research methodology*, *19*(1), 26.

[8]. Aminul Islam, M. (2024). Interpreting Cyber-Threat Indicators in Enterprise Memory Forensics: A Mixed-Methods Study of Machine Learning Performance and Explainability. *Review of Applied Science and Technology*, *14*(12), 01–13. https://doi.org/10.63125/2p8hte92

[9]. Aminul Islam, M., & Md, A. (2023). Adaptive Class-Imbalance Learning for High-Sensitivity Fraud Detection in Large-Scale Digital Transaction Networks. *International Journal of Business and Economics Insights*, *13*(09), 01-33. https://doi.org/10.63125/bmxwv741

[10]. Andronie, M., Lăzăroiu, G., Iatagan, M., Hurloiu, I., & Dijmărescu, I. (2021). Sustainable cyber-physical production systems in big data-driven smart urban economy: A systematic literature review. *Sustainability*, *13*(2), 751.

[11]. Andronie, M., Lăzăroiu, G., Iatagan, M., Uță, C., Ștefănescu, R., & Cocoșatu, M. (2021). Artificial intelligence-based decision-making algorithms, internet of things sensing networks, and deep learning-assisted smart process management in cyber-physical production systems. *Electronics*, *10*(20), 2497.

[12]. Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronné, R., Faouzi, J., Koval, I., & Louis, M. (2021). Predicting the progression of mild cognitive impairment using machine learning: a systematic, quantitative and critical review. *Medical Image Analysis*, *67*, 101848.

[13]. Asheke, M. (2023). Machine Learning–Based Predictive Quality Modeling for Defect Detection and Process Optimization in Advanced Manufacturing Systems. *Review of Applied Science and Technology*, *13*(12), 01–49. https://doi.org/10.63125/thy73b16

[14]. Asheke, M., & Md Maksudul, H. (2022). AI-Driven Demand Forecasting for Sporting-Goods and Teamwear Supply Chains Under Seasonal and Tournament Demand. *International Journal of Business and Economics Insights*, *2*(1), 73–110. https://doi.org/10.63125/ed6by557

[15]. Asheke, M., & Mohammad Badrul, A. (2022). Development Of Functionalized Cellulose Nanofiber Adsorbents for Selective Critical-Metal Recovery from Industrial Wastewater. *American Journal of Scholarly Research and Innovation*, *1*(02), 270–320. https://doi.org/10.63125/v6fxzq09

[16]. Ashik, M. (2024). Meta-Analysis of Risk-Based Inspection Strategies for Improving Reliability and Safety Across Critical Energy Infrastructure Systems. *American Journal of Advanced Technology and Engineering Solutions*, *14*(09), 01-43. https://doi.org/10.63125/tgttkt09

[17]. Ashik, M. (2025). PRISMA-Based Systematic Literature Review of Reliability-Centered Turnaround Planning for Reducing Downtime in Power Generation Facilities. *American Journal of Scholarly Research and Innovation*, *4*(01), 1036–1080. https://doi.org/10.63125/2878n718

[18]. Band, S. S., Janizadeh, S., Chandra Pal, S., Saha, A., Chakrabortty, R., Melesse, A. M., & Mosavi, A. (2020). Flash flood susceptibility modeling using new approaches of hybrid and ensemble tree-based machine learning algorithms. *Remote Sensing*, *12*(21), 3568.

[19]. Blanco-Encomienda, F. J., García-Cantero, R., & Latorre-Medina, M. J. (2020). Association between work-related rumination, work environment and employee well-being: A meta-analytic study of main and moderator effects. *Social Indicators Research*, *150*(3), 887-910.

[20]. Bonetti, A. F., Tonin, F. S., Della Rocca, A. M., Lucchetta, R. C., Fernandez-Llimos, F., & Pontarolo, R. (2022). Methodological quality and risk of bias of meta-analyses of pharmacy services: a systematic review. *Research in Social and Administrative Pharmacy*, *18*(3), 2403-2409.

[21]. Brown, B., Gude, W. T., Blakeman, T., van der Veer, S. N., Ivers, N., Francis, J. J., Lorencatto, F., Presseau, J., Peek, N., & Daker-White, G. (2019). Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in health care based on a systematic review and meta-synthesis of qualitative research. *Implementation Science*, *14*(1), 40.

[22]. Calixto, N., & Ferreira, J. (2020). Salespeople performance evaluation with predictive analytics in B2B. *Applied Sciences*, *10*(11), 4036.

[23]. Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, *109*(11), 1997-2028.

[24]. Choi, J., Gu, B., Chin, S., & Lee, J.-S. (2020). Machine learning predictive model based on national data for fatal accidents of construction workers. *Automation in construction*, *110*, 102974.

[25]. Chouhan, S. S., Kathuria, A., & Sekhar, C. R. (2023). The motorcycle rider behaviour questionnaire as a predictor of crashes: A systematic review and meta-analysis. *IATSS Research*, *47*(1), 61-72.

[26]. Clasen, M., & Milnik, M. (2019). Big Bang Based Decision Automation: On the Implementation of Innovative Methods, Discovered by Top-Level Research, for Automatized Decisions in Replenishment, Price Optimization, and Campaign Management. In *Performance Management in Retail and the Consumer Goods Industry: Best Practices and Case Studies* (pp. 165-182). Springer.

[27]. Di Giovanni, N., Meuwis, M.-A., Louis, E., & Focant, J.-F. (2020). Specificity of metabolic colorectal cancer biomarkers in serum through effect size. *Metabolomics*, *16*(8), 88.

[28]. Drawve, G., & Wooditch, A. (2019). A research note on the methodological and theoretical considerations for assessing crime forecasting accuracy with the predictive accuracy index. *Journal of Criminal Justice*, *64*, 101625.

[29]. Dweekat, O. Y., Lam, S. S., & McGrath, L. (2023). Machine learning techniques, applications, and potential future opportunities in pressure injuries (bedsores) management: a systematic review. *International journal of environmental research and public health*, *20*(1), 796.

[30]. Fainshmidt, S., Witt, M. A., Aguilera, R. V., & Verbeke, A. (2020). The contributions of qualitative comparative analysis (QCA) to international business research: Stav Fainshmidt et al. *Journal of international business studies*, *51*(4), 455-466.

[31]. Falatouri, T., Darbanian, F., Brandtner, P., & Udokwu, C. (2022). Predictive analytics for demand forecasting–a comparison of SARIMA and LSTM in retail SCM. *Procedia Computer Science*, *200*(2019), 993-1003.

[32]. González-Moles, M. Á., Aguilar-Ruiz, M., & Ramos-García, P. (2022). Challenges in the early diagnosis of oral cancer, evidence gaps and strategies for improvement: a scoping review of systematic reviews. *Cancers*, *14*(19), 4967.

[33]. Grailey, K., Murray, E., Reader, T., & Brett, S. J. (2021). The presence and potential impact of psychological safety in the healthcare setting: an evidence synthesis. *BMC health services research*, *21*(1), 773.

[34]. Grossman, I., Wilson, T., & Temple, J. (2023). Forecasting small area populations with long short-term memory networks. *Socio-Economic Planning Sciences*, *88*, 101658.

[35]. Hamilton, H. K., Roach, B. J., & Mathalon, D. H. (2021). Forecasting remission from the psychosis risk syndrome with mismatch negativity and P300: potentials and pitfalls. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(2), 178-187.

[36]. Harrer, M., Ebert, D. D., Kuper, P., Paganini, S., Schlicker, S., Terhorst, Y., Reuter, B., Sander, L. B., & Baumeister, H. (2023). Predicting heterogeneous treatment effects of an Internet-based depression intervention for patients with chronic back pain: Secondary analysis of two randomized controlled trials. *Internet interventions*, *33*, 100634.

[37]. Hewamalage, H., Ackermann, K., & Bergmeir, C. (2023). Forecast evaluation for data scientists: common pitfalls and best practices: H. Hewamalage et al. *Data Mining and Knowledge Discovery*, *37*(2), 788-832.

[38]. Iacobucci, D., Petrescu, M., Krishen, A., & Bendixen, M. (2019). The state of marketing analytics in research and practice: D. Iacobucci et al. *Journal of Marketing Analytics*, *7*(3), 152-181.

[39]. Jones, J., Johnston, K., Farah, L., & Baker, J. (2021). Predicting seasonal performance in professional sport: a 30-year analysis of sports illustrated predictions. *Sports*, *9*(12), 163.

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

[41]. Kaczan, D. J., & Orgill-Meyer, J. (2020). The impact of climate change on migration: a synthesis of recent empirical insights. *Climatic change*, *158*(3), 281-300.

[42]. Khalil, H., Peters, M. D., Tricco, A. C., Pollock, D., Alexander, L., McInerney, P., Godfrey, C. M., & Munn, Z. (2021). Conducting high quality scoping reviews-challenges and solutions. *Journal of clinical epidemiology*, *130*, 156-160.

[43]. Krouwel, M., Jolly, K., & Greenfield, S. (2019). Comparing Skype (video calling) and in-person qualitative interview modes in a study of people with irritable bowel syndrome–an exploratory comparative analysis. *BMC medical research methodology*, *19*(1), 219.

[44]. Kumar, A., Nadeem, M., & Shameem, M. (2023). Machine learning based predictive modeling to effectively implement DevOps practices in software organizations. *Automated Software Engineering*, *30*(2), 21.

[45]. Kynatun, M. (2025). Integrating AI-Driven Security Analytics and IT Governance Frameworks for Cybersecurity Decision Support in U.S. Financial Information Systems. *American Journal of Advanced Technology and Engineering Solutions*, *15*(12), 01-42. https://doi.org/10.63125/kzcdyz73

[46]. Kynatun, M., & Saba, A. (2024). Machine Learning Enhanced Intrusion Detection and Anomaly Classification for MIS-Driven Enterprise Network Security. *International Journal of Scientific Interdisciplinary Research*, *14*(09), 01–41. https://doi.org/10.63125/xk6kpc72

[47]. Lim, S.-C., Huh, J.-H., Hong, S.-H., Park, C.-Y., & Kim, J.-C. (2022). Solar power forecasting using CNN-LSTM hybrid model. *Energies*, *15*(21), 8233.

[48]. Lucas, J., Van Doorn, P., Hegedus, E., Lewis, J., & Van Der Windt, D. (2022). A systematic review of the global prevalence and incidence of shoulder pain. *BMC musculoskeletal disorders*, *23*(1), 1073.

[49]. Ma, J., Yu, Z., Qu, Y., Xu, J., & Cao, Y. (2020). Application of the XGBoost machine learning method in PM2. 5 prediction: A case study of Shanghai. *Aerosol and Air Quality Research*, *20*(1), 128-138.

[50]. Ma, L.-L., Wang, Y.-Y., Yang, Z.-H., Huang, D., Weng, H., & Zeng, X.-T. (2020). Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military medical research*, *7*(1), 7.

[51]. Madigan, D. J., & Curran, T. (2021). Does burnout affect academic achievement? A meta-analysis of over 100,000 students. *Educational Psychology Review*, *33*(2), 387-405.

[52]. Maishar Mostari, Z. (2025). Privacy-Preserving AI-Based Cyber Threat Detection in Financial Organizations: A Governance and Business Value Framework. *American Journal of Interdisciplinary Studies*, *15*(04), 01-51. https://doi.org/10.63125/p8fct530

[53]. Malakouti, S. M., Menhaj, M. B., & Suratgar, A. A. (2023). The usage of 10-fold cross-validation and grid search to enhance ML methods performance in solar farm power generation prediction. *Cleaner Engineering and Technology*, *15*, 100664.

[54]. Mandrik, P. O. L., Severens, P. J. H., Bardach, P. A., Ghabri, P. S., Hamel, P. C., Mathes, P. T., Vale, P. L., Wisløff, P. T., & Goldhaber-Fiebert, P. J. D. (2021). Critical appraisal of systematic reviews with costs and cost-effectiveness outcomes: an ISPOR good practices task force report. *Value in Health*, *24*(4), 463-472.

[55]. Mansura Akter, E., & Sumaiya, M. (2022). Antimicrobial Resistance Profiles of Acinetobacter Baumannii and Vibrio Cholerae in Hospital Wastewater and Adjacent Household Water Sources. *American Journal of Data Science and Analytics*, *3*(09), 01-55. https://doi.org/10.63125/gbervt14

[56]. Martínez, A., Schmuck, C., Pereverzyev Jr, S., Pirker, C., & Haltmeier, M. (2020). A machine learning framework for customer purchase prediction in the non-contractual setting. *European Journal of Operational Research*, *281*(3), 588-596.

[57]. Masrul, H. (2025). Explainable Machine Learning for Real-Time Financial Fraud Detection and Risk-Based Transaction Monitoring in U.S. Banking and Digital Payment Systems. *American Journal of Data Science and Analytics*, *6*(10), 70-108. https://doi.org/10.63125/4h9v1876

[58]. Md Imran, H., & Jannatul, F. (2022). Molecular and Precision Rehabilitation Molecular Mechanisms Linking Mitochondrial Dysfunction, Inflammation, and Muscle Degeneration with Functional Decline and Rehabilitation Response in Neuromuscular Disease. *American Journal of Interdisciplinary Studies*, *12*(08), 01-44. https://doi.org/10.63125/0ym5bh79

[59]. Md Maksudul, H. (2024). Supplier-Risk Scoring and Procurement Resilience for Distributed Athletic-Apparel and Equipment Supply Networks. *American Journal of Advanced Technology and Engineering Solutions*, *14*(09), 44-90. https://doi.org/10.63125/a9fmgj50

[60]. Md Maksudul, H. (2025). Data-Driven Fulfillment Optimization for Time-Sensitive Teamwear and Sporting Goods Distribution. *American Journal of Scholarly Research and Innovation*, *4*(01), 1120–1173. https://doi.org/10.63125/rpezfz61

[61]. Md Maksudul, H., & Md Tazul Islam, C. (2022). Machine Learning-Based Cash-Flow Forecasting for Improving Financial Planning in U.S. Small and Mid-Sized Businesses. *International Journal of Scientific Interdisciplinary Research*, *12*(12), 01–38. https://doi.org/10.63125/w3pgjx97

[62]. Md Momen Shahriar, J. (2025). Viscoelastic Creep Characterization of LDPE Under Multilevel Stress and Temperature Using Time-Temperature Superposition and Constitutive Modeling. *American Journal of Scholarly Research and Innovation*, *15*(11), 01–38. https://doi.org/10.63125/2edstd86

[63]. Md Shahid, N. (2023). Digital Twin-Based Network Reliability Modeling for LAN And WAN Infrastructure Resilience. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 577–629. https://doi.org/10.63125/ahjvxj14

[64]. Md Shahid, N. (2024). Predictive Analytics Models for Network Traffic Optimization in Distributed Enterprise Environments. *Review of Applied Science and Technology*, *3*(01), 367-419. https://doi.org/10.63125/qm6cyf84

[65]. Md Shahid, N. (2025). Blockchain-Based Secure Data Routing Framework for U.S. Industrial Communication Networks. *American Journal of Scholarly Research and Innovation*, *4*(01), 990–1035. https://doi.org/10.63125/9exrgs41

[66]. Md Tazul Islam, C. (2024). Explainable AI For Financial and Operational Decision Support: Improving Transparency and Managerial Trust. *American Journal of Advanced Technology and Engineering Solutions*, *14*(12), 01-38. https://doi.org/10.63125/kwfvzw68

[67]. Md Tazul Islam, C. (2025). Business Intelligence and Data-Quality Frameworks for Improving Evidence-Based Decisions across Multi-Source Enterprise Data. *American Journal of Scholarly Research and Innovation*, *15*(07), 01–41. https://doi.org/10.63125/qhh6ka59

[68]. Md Tazul Islam, C., & Ahmed, H. C. H. R. (2023). Predictive Operations Analytics for Resource Allocation and Service-Capacity Planning in Data-Driven Organizations. *Review of Applied Science and Technology*, *13*(02), 01–41. https://doi.org/10.63125/z3sm3089

[69]. Md. Mainuddin, F. (2024). Quantitative Structural Retrofit Assessment Models for Strengthening Existing Steel Buildings Under Increased Load Demands. *Review of Applied Science and Technology*, *3*(04), 325–366. https://doi.org/10.63125/yyqnte84

[70]. Md. Mainuddin, F. (2025). Advanced Engineering Materials Applications for Enhancing Durability and Lifecycle Performance of Steel Building Systems. *ASRC Procedia: Global Perspectives in Science and Scholarship*, *1*(01), 2406–2445. https://doi.org/10.63125/t9xvg986

[71]. Md. Morshedul, I., & Aminul Islam, M. (2022). A Quantitative Evaluation of Explainable Ensemble Learning for Probabilistic Risk Estimation and Decision Reliability in Enterprise Analytics. *American Journal of Data Science and Analytics*, *3*(02), 01-38. https://doi.org/10.63125/k22shk74

[72]. Mediavilla, M. A., Dietrich, F., & Palm, D. (2022). Review and analysis of artificial intelligence methods for demand forecasting in supply chain management. *Procedia CIRP*, *107*, 1126-1131.

[73]. Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: a novel hybrid machine learning approach. Operations research forum,

[74]. Namdari, S., Nicholson, T., Abboud, J., Lazarus, M., Ramsey, M. L., Williams, G., & Parvizi, J. (2019). Comparative study of cultures and next-generation sequencing in the diagnosis of shoulder prosthetic joint infections. *Journal of shoulder and elbow surgery*, *28*(1), 1-8.

[75]. Namoun, A., & Alshanqiti, A. (2020). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, *11*(1), 237.

[76]. Nasseri, M., Falatouri, T., Brandtner, P., & Darbanian, F. (2023). Applying machine learning in retail demand prediction—A comparison of tree-based ensembles and long short-term memory-based deep learning. *Applied Sciences*, *13*(19), 11112.

[77]. Niazian, M., & Niedbała, G. (2020). Machine learning for plant breeding and biotechnology. *Agriculture*, *10*(10), 436.

[78]. Noura, N., Boulon, L., & Jemeï, S. (2020). A review of battery state of health estimation methods: Hybrid electric vehicle challenges. *World Electric Vehicle Journal*, *11*(4), 66.

[79]. Obón-Santacana, M., Mas-Lloret, J., Bars-Cortina, D., Criado-Mesas, L., Carreras-Torres, R., Díez-Villanueva, A., Moratalla-Navarro, F., Guinó, E., Ibáñez-Sanz, G., & Rodríguez-Alonso, L. (2022). Meta-analysis and validation of a colorectal cancer risk prediction model using deep sequenced fecal metagenomes. *Cancers*, *14*(17), 4214.

[80]. Oliveira, J. M., & Ramos, P. (2019). Assessing the performance of hierarchical forecasting methods on the retail sector. *Entropy*, *21*(4), 436.

[81]. Panay, B., Baloian, N., Pino, J. A., Penafiel, S., Frez, J., Fuenzalida, C., Sanson, H., & Zurita, G. (2021). Forecasting key retail performance indicators using interpretable regression. *Sensors*, *21*(5), 1874.

[82]. Parmezan, A. R. S., Souza, V. M., & Batista, G. E. (2019). Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information sciences*, *484*, 302-337.

[83]. Peters, M. D., Marnie, C., Colquhoun, H., Garritty, C. M., Hempel, S., Horsley, T., Langlois, E. V., Lillie, E., O'Brien, K. K., & Tunçalp, Ö. (2021). Scoping reviews: reinforcing and advancing the methodology and application. *Systematic reviews*, *10*(1), 263.

[84]. Pfutzenreuter, T. C., de Lima, E. P., da Costa, S. E. G., & Deschamps, F. (2023). The role of machine learning in supply chain management. Intelligent and Transformative Production in Pandemic Times: Proceedings of the 26th International Conference on Production Research,

[85]. Rabbani, M. R., Bashar, A., Hawaldar, I. T., Shaik, M., & Selim, M. (2022). What do we know about crowdfunding and P2P lending research? A bibliometric review and meta-analysis. *Journal of Risk and Financial Management*, *15*(10), 451.

[86]. Razza, L. B., De Smet, S., Moffa, A., Sudbrack-Oliveira, P., Vanderhasselt, M.-A., & Brunoni, A. R. (2021). Follow-up effects of transcranial direct current stimulation (tDCS) for the major depressive episode: A systematic review and meta-analysis. *Psychiatry research*, *302*, 114024.

[87]. Rumpf, C., Proschinger, S., Schenk, A., Bloch, W., Lampit, A., Javelle, F., & Zimmer, P. (2021). The Effect of Acute Physical Exercise on NK-Cell Cytolytic Activity: A Systematic Review and Meta-Analysis: C. Rumpf et al. *Sports Medicine*, *51*(3), 519-530.

[88]. Sadia, A. (2025). Privacy-Aware Personalization: Machine Learning, Model Validation, And Consumer Trust in AI-Mediated Marketing. *American Journal of Scholarly Research and Innovation*, *4*(01), 1081–1119. https://doi.org/10.63125/j7tfr242

[89]. Sadia, A., & Md. Mosheur, R. (2023). Consumer Trust, Perceived Authenticity, and Personalized Digital Marketing: A Quantitative Decision-Science Study. *International Journal of Scientific Interdisciplinary Research*, *4*(2), 252–295. https://doi.org/10.63125/cajkqt60

[90]. Sahu, S. K., Mokhade, A., & Bokde, N. D. (2023). An overview of machine learning, deep learning, and reinforcement learning-based techniques in quantitative finance: recent progress and challenges. *Applied Sciences*, *13*(3), 1956.

[91]. Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, *31*(10), 6893-6908.

[92]. Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109-120.

[93]. Schubert, A.-L. (2019). A meta-analysis of the worst performance rule. *Intelligence*, *73*, 88-100.

[94]. Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *Journal of big data*, *7*(1), 53.

[95]. Shabbir, A. (2025). Physics-Based Failure Prognostics for Preventing Pressure-System Degradation and Unplanned Industrial Shutdowns. *American Journal of Scholarly Research and Innovation*, *14*(09), 01–41. https://doi.org/10.63125/3asd0053

[96]. Siddiq, F., & Scherer, R. (2019). Is there a gender gap? A meta-analysis of the gender differences in students' ICT literacy. *Educational research review*, *27*, 205-217.

[97]. Siddique, A., Jan, A., Majeed, F., Qahmash, A. I., Quadri, N. N., & Wahab, M. O. A. (2021). Predicting academic performance using an efficient model based on fusion of classifiers. *Applied Sciences*, *11*(24), 11845.

[98]. Smith, E. A., Cooper, N. J., Sutton, A. J., Abrams, K. R., & Hubbard, S. J. (2021). A review of the quantitative effectiveness evidence synthesis methods used in public health intervention guidelines. *BMC public health*, *21*(1), 278.

[99]. Stabak, D., & Md Shahid, N. (2022). Empirical Assessment of Uncertainty-Aware Real-Time Decision-Making for Reliable and Resilient Cyber-Physical Systems. *International Journal of Scientific Interdisciplinary Research*, *1*(01), 432–477. https://doi.org/10.63125/q3r4kk16

[100]. Sumaiya, M. (2024). Hospital Wastewater as A Reservoir of Antimicrobial Resistance Genes: Implications for Public Health and Genomic Surveillance in The United States. *Review of Applied Science and Technology*, *3*(04), 448–508. https://doi.org/10.63125/ahvas976

[101]. Sumaiya Sadmeen, M. (2025). Spatiotemporal Graph Neural Networks and Digital Twins for Urban Drainage Networks: Predictive Inundation, Critical-Node Detection, And Infrastructure Prioritization. *American Journal of Scholarly Research and Innovation*, *15*(12), 01-40. https://doi.org/10.63125/awvjew66

[102]. Syed Feroz, A., & Md Mahamud Pervaz, P. (2022). Root-Cause Analysis and FMEA-Driven Reliability Improvement in Gas Production Facilities: A Multi-Year Empirical Study. *International Journal of Scientific Interdisciplinary Research*, *12*(10), 01–53. https://doi.org/10.63125/70msp878

[103]. Syed Feroz, A., & Shabbir, A. (2022). Finite Element-Based Thermomechanical Optimization of Pressure Components for Extending Industrial Boiler Service Life. *Journal of Sustainable Development and Policy*, *12*(12), 01-51. https://doi.org/10.63125/x24pht28

[104]. Taher, M. A., Sikder, M. A., Ahmed, H., & Bashar, M. (2024). Integrating IOT and Big Data Analytics for Enhanced Supply Chain Performance in Industrial Engineering Sectors: A Cross-Market Study. *International Journal of Science and Engineering*, *Volume 1*, 1-14.

[105]. Tomar, A., & Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, *728*, 138762.

[106]. Tsoy, N., Steubing, B., van der Giesen, C., & Guinée, J. (2020). Upscaling methods used in ex ante life cycle assessment of emerging technologies: a review. *The International Journal of Life Cycle Assessment*, *25*(9), 1680-1692.

[107]. Van Steenbergen, R., & Mes, M. R. (2020). Forecasting demand profiles of new products. *Decision support systems*, *139*, 113401.

[108]. Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. *Machine learning for brain disorders*, 601-630.

[109]. Verstraete, G., Aghezzaf, E.-H., & Desmet, B. (2019). A data-driven framework for predicting weather impact on high-volume low-margin retail products. *Journal of Retailing and Consumer Services*, *48*, 169-177.

[110]. Vogl, M., Rötzel, L. M. P. G., & Homes, S. (2022). Forecasting performance of wavelet neural networks and other neural network topologies: A comparative study based on financial market data sets. *Machine Learning with Applications*, *8*, 100302.

[111]. Von Janczewski, N., Wittmann, J., Engeln, A., Baumann, M., & Krauß, L. (2021). A meta-analysis of the n-back task while driving and its effects on cognitive workload. *Transportation research part F: traffic psychology and behaviour*, *76*, 269-285.

[112]. Wang, C.-H. (2022). Considering economic indicators and dynamic channel interactions to conduct sales forecasting for retail sectors. *Computers & industrial engineering*, *165*, 107965.

[113]. Wang, C., Ghadimi, P., Lim, M. K., & Tseng, M.-L. (2019). A literature review of sustainable consumption and production: A comparative analysis in developed and developing economies. *Journal of cleaner production*, *206*, 741-754.

[114]. Xiahou, X., & Harada, Y. (2022). B2C E-commerce customer churn prediction based on K-means and SVM. *Journal of Theoretical and Applied Electronic Commerce Research*, *17*(2), 458-475.

[115]. Zhao, M. J., Chang, M. J. R., Campbell, M. A. A., & McCulley, M. T. J. (2020). Thyroid eye disease-related epiblepharon: a comparative case study. *Asia-Pacific Journal of Ophthalmology*, *9*(1), 44-47.

[116]. Zwanenburg, A. (2019). Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. *European journal of nuclear medicine and molecular imaging*, *46*(13), 2638-2655.