



ISSN ONLINE: 2447-0228



AWARE-HYBRID ULTRA MODEL: A UNIFIED AI FRAMEWORK FOR MULTI-SKU DEMAND FORECASTING AND INTELLIGENT SUPPLY CHAIN OPTIMIZATION

Jaymin V. Soni^{1*} and Chetansinh R. Vaghela²

^{1,2}Department of Mathematics, Faculty of Science, Gokul Global University, Sidhpur-384151, Gujarat, India.

¹<https://orcid.org/0009-0004-8946-7877>, ²<http://orcid.org/0000-0001-5692-5190>

Email: sonijaymin7777@gmail.com, crvaghela16@gmail.com

ARTICLE INFO

Article History

Received: March 28, 2026

Reviewed: August 3, 2026

Accepted: September 3, 2026

Published: September 30, 2026

Keywords:

Retail Demand Forecasting,
Hybrid Ensemble Learning,
Inverse-Error Weighting,
Residual Boosting (CatBoost),
Multi-SKU Multi-Store Analytics.

ABSTRACT

Retail demand forecasting becomes increasingly unstable in multi-SKU, multi-store environments where behaviour is shaped by irregular promotions, heterogeneous store dynamics, and weak temporal persistence, causing conventional time-series models to degrade sharply. Addressing this challenge, the AWARE_HybridUltraModel is formulated as an adaptive, error-driven ensemble designed to operate effectively in settings that violate classical autocorrelation and seasonality assumptions. The framework integrates multiple heterogeneous regressors through an inverse-error weighting mechanism that assigns predictive influence according to each model's empirical loss profile, while a CatBoost-based residual learner captures higher-order deviations left unresolved after ensemble fusion. This layered structure enables the system to remain stable across fragmented signals, nonlinear promotional effects, and SKU-specific variability. Extensive evaluation on a complete two-year multi-store dataset demonstrates that the model achieves MAE ≈ 1.96 , RMSE ≈ 2.45 , and $R^2 \approx 0.974$, consistently outperforming Extra Trees, Gradient Boosting, XGBoost, Random Forests, CatBoost, Ridge, and MLP. In contrast, deep sequence architectures such as LSTM and TCN fail under the dataset's weak sequential continuity, exhibiting catastrophic error amplification with RMSE values above 64. The proposed architecture produces a high-resolution, low-noise demand signal suitable for downstream operational tasks including replenishment planning, safety-stock calibration, and dynamic inventory optimisation. Overall, the AWARE_HybridUltraModel offers a rigorous, data-agnostic forecasting mechanism specifically tailored for modern retail streams where structural volatility and irregular behavioural drivers undermine the assumptions of conventional forecasting models.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Modern retail demand forecasting increasingly involves datasets that depart sharply from the assumptions of classical time-series analysis. In multi-SKU, multi-store environments, demand is typically nonstationary, highly nonlinear, intermittently elevated during promotions, and structurally heterogeneous across both products and locations. Rather than following stable seasonal or trend-driven trajectories, retail demand often manifests as a hybrid of baseline sales, promotional spikes, store-specific behavioural shifts, and noise-dominated fluctuations, forming a composite signal that is difficult to represent with any single statistical or neural forecasting paradigm. These fragmented patterns especially promotion-driven uplift, cross-SKU cannibalisation, short-lived peaks, stock-sensitive dips, and weakly persistent temporal dependencies have been repeatedly identified as major sources of forecasting error in operational retail systems [1], [2], [3].

Such demand streams therefore differ from classical “smooth” or strongly autocorrelated series. Instead, they exhibit the characteristics of irregular, weakly dependent, multi-regime retail demand, where price elasticity shifts over time, promotional events introduce abrupt structural changes, and store-level demographic differences create distinct local patterns. Contemporary empirical studies highlight that deep sequence architectures such as LSTM, GRU, and TCN suffer catastrophic error escalation when applied to datasets lacking meaningful temporal continuity, leading to instability and amplified prediction variance [4]. Similarly, traditional methods ARIMA, ETS, Holt–Winters are constrained by assumptions of trend consistency, stable seasonal decomposition, or monotonic error propagation, limiting their suitability in retail contexts where sudden behaviour shifts dominate the signal structure [5], [6]. Machine learning regressors, including tree ensembles and gradient boosting systems, offer greater representational flexibility and can capture local nonlinearities without explicit temporal modelling. However, their performance remains sensitive to shifting feature relevance, demand shocks, intermittent promotional effects, and latent cross-interactions that evolve across stores and SKUs. Recent research shows that even robust ensemble learners may misestimate contributions when demand structure changes unexpectedly, leading to error oscillation and unstable bias–variance behaviour [7]. Static-weight ensembles inherit these limitations, as they rely on fixed importance profiles that do not adjust to changing error landscapes.[8]

To meet these challenges, contemporary literature increasingly supports the use of error-adaptive hybrid architectures, where diverse models collaborate dynamically and residual behaviour is explicitly learned rather than left unmodelled [9]. Yet many proposed hybrid systems continue to rely on heuristic weighting rules, simplistic meta-learners, or partial correction strategies that fail under conditions of intermittent, promotion-sensitive, heterogeneous demand [10]. Responding to this gap, the present study introduces the *AWARE_HybridUltraModel*, a forecasting framework expressly designed for chaotic, low-autocorrelation, multi-regime retail demand. The model is constructed around a layered, error-driven architecture that integrates heterogeneous regressors including Extra Trees, Gradient Boosting, XGBoost, Random Forests, CatBoost, Ridge, and MLP using an inverse-error weighting mechanism that dynamically allocates influence according to each model’s empirical loss characteristics. This adaptive weighting makes the ensemble robust to variability in model competency across differing SKU-store combinations and across demand regimes, including baseline periods, high-variance promotional surges, and store-level behavioural divergences [11], [12]. The first-layer ensemble is followed by a CatBoost-based residual learner trained to capture higher-order nonlinear patterns and systematic deviations left unresolved by the model fusion stage. This residual correction is critical in environments marked by promotion-triggered structural breaks, long-tail SKU behaviour, and localised store-level anomalies all of which generate error artefacts that base predictors struggle to learn [13].

A structurally realistic two-year multi-store dataset incorporating daily baseline demand, promotion-induced demand uplift, holiday-driven spikes, store-specific multipliers, pricing variation, and stochastic demand noise serves as the evaluation platform. Results demonstrate that the proposed architecture achieves $MAE \approx 1.96$, $RMSE \approx 2.45$, $R^2 \approx 0.974$, outperforming all benchmark models. Conversely, LSTM and TCN collapse due to the dataset’s inherently weak temporal continuity, producing RMSE values exceeding 64, consistent with theoretical findings on deep sequence instability in low-dependence environments [14], [15], [16]. The *AWARE_HybridUltraModel* therefore offers a rigorous, data-agnostic forecasting system for complex retail demand types, particularly those characterised by: Nonstationary demand, short-lived promotional spikes, Weak autocorrelation, Intermittent or irregular patterns, Cross-SKU heterogeneity, Store-level behavioural divergence. This framework produces a high-resolution, low-noise demand signal suitable for decision-making pipelines including replenishment planning, safety-stock allocation, store-level inventory optimisation, and algorithmic retail control systems. Overall, the model provides a generalizable forecasting mechanism built for modern retail environments where temporal regularity is the exception rather than the norm.

II. LITERATURE REVIEW

Retail demand forecasting research has evolved across statistical, machine-learning, and deep-learning paradigms, each contributing unique modelling perspectives while revealing crucial limitations in real-world, multi-SKU environments. Early forecasting foundations were established through classical models such as ARIMA, SARIMA, and exponential smoothing, which provided structured frameworks for stationary or seasonally stable series [17]. However, numerous empirical studies highlight that retail data often behaves as short, noisy, weakly correlated micro-series that violate these assumptions, causing traditional temporal models to underperform under promotional irregularities, price shocks, and store-specific behavioural variability [18], [19]. To overcome these constraints, regression-based retail forecasting approaches began incorporating pricing, promotional calendars, and cross-category effects into multivariate models, demonstrating improved practical accuracy in both academic and industry settings [20]. With the rise of data-rich retail systems, tree-based ensemble models including Random Forests, Gradient Boosting, XGBoost, and Extra Trees emerged as dominant techniques due to their robustness to nonlinear interactions, hierarchical structure, and heterogeneous feature contributions [21], [22]. Subsequent research consistently observed their superiority over classical benchmarks, particularly in contexts involving promotional uplift, SKU-level heterogeneity, and multi-store interactions [23].

Innovations such as CatBoost further enhanced forecasting accuracy through native handling of categorical variables and robust regularisation strategies [24]. Parallel advances in deep learning introduced LSTM, GRU, and TCN architectures capable of modelling complex temporal dependencies [25]. These networks achieved strong results in domains where temporal coherence is high, but multiple studies demonstrate their instability when applied to retail series with weak autocorrelation, discontinuous demand segments, or irregular promotional structures [26], [27]. Retail datasets often contain sudden spikes, cross-SKU cannibalisation, stock-dependent behavioural shifts, and multi-store divergences, all of which disrupt sequential learning mechanisms and result in overfitting, prediction drift, and inflated error variance in deep models [28]. Hybrid and stacking-based ensembles have therefore gained traction, aiming to integrate the strengths of various models through feature-driven combination rules, meta-learners, and weighted fusion mechanisms [29]. While these approaches outperform individual base models, most rely on static weighting, heuristic blending, or partial residual modelling, leaving systematic error components unaddressed particularly those arising from promotional shocks, price-sensitive behaviour, and SKU–store fragmentation. Few studies explicitly combine inverse-error adaptive weighting with a dedicated nonlinear residual learner capable of correcting structural deviations left after ensemble fusion.

Given the increasing complexity of modern retail demand characterised by nonstationary dynamics, short-lived promotional spikes, intermittent behaviours, and multi-store structural variation existing forecasting models do not fully capture the layered, irregular, and weakly sequential nature of real demand signals. Although the literature progresses toward hybridisation and ensemble learning, current solutions lack a unified, error-driven mechanism designed to stabilise forecasts across fragmented temporal regimes. This gap motivates the development of the AWARE_HybridUltraModel, which explicitly targets chaotic, low-autocorrelation retail environments through adaptive weighting and residual correction, producing high-fidelity demand signals suited for operational inventory and supply chain decision-making.

Table 1: Summary of Related Literature and Identified Research Gap.

Authors (Year)	Focus / Data Type	Strengths	Limitations in Multi-SKU, Promo-Driven Retail
Sun & Li (2024) [30]	Transformer-based enterprise sales forecasting	Captures long-range temporal dependencies; strong representation learning	Performance degrades under weak autocorrelation and fragmented promotional demand
Samal & Ghosh (2026) [31]	Ensemble-based demand forecasting for multi-channel retail	Improves accuracy through ensemble fusion; handles channel diversity	Uses static ensemble weighting; lacks dynamic error adaptation under regime shifts
Nagadevi et al. (2025) [32]	XGBoost-based supply chain demand forecasting	Strong nonlinear modeling; effective feature interaction learning	Performance instability during extreme promotions and abrupt demand discontinuities
Patil et al. (2025) [33]	LightGBM-based sales prediction	High computational efficiency; good scalability	Static feature importance limits adaptability during sudden demand shifts
Oliveira & Ramos (2024) [34]	Time-series transformers for retail demand	Effective sequence learning; captures nonlinear temporal structures	Overfitting risk under low-volume SKUs and weak persistence demand
Mustapha & Sithole (2025) [35]	Machine learning models for retail sales	Comparative evaluation across ML models; interpretable performance	No explicit mechanism for cross-store heterogeneity or regime-switch adaptation
Liang et al. (2024) [36]	Collaborative filtering for cross-store demand forecasting	Learns inter-store relationships; improves spatial demand coherence	Lacks promotion sensitivity and dynamic demand volatility handling
Jiang et al. (2021) [37]	Hybrid ML and statistical retail forecasting	Demonstrates hybrid modeling benefits; reduced single-model bias	Uses fixed combination rules; no adaptive residual correction
Ding et al. (2020) [38]	CatBoost-based sales forecasting	Strong categorical feature handling; robust boosting structure	No post-ensemble residual learning; limited regime-shift robustness
Benidis et al. (2023) [39]	Deep learning survey for time-series forecasting	Comprehensive theoretical and empirical deep learning analysis	Concludes DL models degrade under noisy, low-autocorrelation retail demand
AWARE_Hybrid UltraModel (This Work)	Multi-SKU, multi-store, promotion-sensitive, weak-autocorrelation retail demand	Adaptive inverse-error ensemble weighting + nonlinear CatBoost residual correction + inventory-optimized autonomous decision control	Addresses key gaps: dynamic error-driven fusion, regime-shift stability, SKU/store heterogeneity, and closed-loop operational integration

Source: Authors, (2026).

II.1 PROBABILISTIC & INTERMITTENT FORECASTING

Recent advances in retail forecasting emphasise probabilistic and intermittent-demand modelling, where demand distributions rather than point estimates are predicted. Approaches such as DeepAR, Temporal Fusion Transformer (TFT), and probabilistic gradient boosting explicitly model demand uncertainty and sparse sales occurrence. These models have demonstrated strong performance in intermittent retail environments where zero-inflation and skewed distributions dominate demand behaviour.

II.1 DEEP LEARNING INTERMITTENT DEMAND

Deep learning approaches tailored for intermittent retail demand, including Croston-based neural hybrids, zero-inflated sequence models, and demand-occurrence/size decomposition networks, have been proposed to address sparse temporal structure. These methods attempt to separately model demand arrival probability and magnitude, improving forecasting stability for low-volume SKUs.

II.2 AI-DRIVEN INVENTORY CONTROL

AI-driven inventory optimisation frameworks increasingly integrate probabilistic forecasting with stochastic control policies. Reinforcement-learning-based inventory systems and distribution-aware safety-stock models have been proposed to account for non-Gaussian demand distributions, lead-time uncertainty, and service-level risk. However, many of these approaches assume reliable probabilistic forecasts, which remain difficult in weak-autocorrelation retail environments.

III. THEORY AND FOUNDATIONS

Retail demand forecasting in multi-SKU, multi-store environments is fundamentally a non-stationary, discontinuous, and weakly autocorrelated prediction problem. Classical time-series theory (Box–Jenkins, Hyndman) assumes stable seasonal patterns, linear temporal dependencies, and smooth error structures, but real-world retail demand violates all three assumptions:

III.1 WEAK TEMPORAL DEPENDENCE

Retail SKU-level sales often exhibit extremely weak temporal dependence, meaning that the data do not follow the stable autocorrelation structures assumed by classical ARIMA-style time-series models. Daily fluctuations are heavily influenced by external shocks such as promotions, stockouts, competitor actions, and localized store-specific events, all of which disrupt temporal continuity. Deep sequence models like LSTM and TCN fundamentally rely on learnable sequential patterns; therefore, when the underlying demand signal lacks consistent temporal structure, these architectures fail to converge and instead amplify noise. This behaviour is clearly reflected in the experimental results, where both LSTM and TCN produce RMSE values exceeding 64 and highly negative R^2 scores around -17 , indicating catastrophic divergence rather than meaningful learning. These outcomes confirm that deep sequential models are unsuitable for retail environments characterized by sparse or unstable autocorrelation.

III.2 HETEROGENEOUS STORE DYNAMICS AND FEATURE INTERACTIONS

Retail demand varies significantly across stores because each location operates under unique conditions shaped by localized pricing strategies, demographic differences, store size and service capacity, and supply-chain inconsistencies. These sources of variation create cross-sectional heterogeneity, meaning that differences across stores are more influential than sequential temporal patterns. As a result, retail data behave more like a high-dimensional tabular problem rather than a smooth time-series, making models that depend on sequential structure inherently less suited to this domain. Tree-based ensemble methods such as Random Forest, Extra Trees, Gradient Boosting, and XGBoost are theoretically more robust in this environment because they can partition nonlinear interactions, model discontinuous behaviour, and absorb noise without catastrophic overfitting. The empirical results further validate these theoretical expectations: XGBoost achieves an RMSE of 2.775, Extra Trees 2.685, Gradient Boosting 2.762, and Random Forest 2.855, with overall R^2 values ranging from 0.965 to 0.969. These models capture complex, non-smooth relationships far more effectively than sequence-based neural networks, demonstrating that modern ensemble trees provide accurate retail demand modelling without relying on temporal smoothness.

III.3 INVERSE-ERROR WEIGHTING (IEW): THEORETICAL BASIS

The core architecture of AWARE_HybridUltraModel relies on error-driven adaptive weighting, grounded in these theoretical principles:

(a) Bias–Variance Decomposition

For an ensemble of M models:

$$\hat{y} = \sum_{i=1}^M w_i \hat{y}_i$$

To minimize expected error:

$$w_i \propto \frac{1}{\text{error}_i}$$

Models with lower MAE and RMSE naturally receive greater influence in the ensemble, while those with higher prediction error contribute proportionally less, ensuring that noise from weak models does not distort the final output. This behaviour is clearly validated by your performance table: the best-performing tree-based models Extra Trees, XGBoost, Gradient Boosting, and Random Forest collectively strengthen the ensemble because their errors are consistently low. In contrast, the deep sequential models such as LSTM and TCN, which produce extremely high error values, are automatically down-weighted or effectively rejected by the weighting mechanism. As a result, the ensemble prioritizes reliable predictors and suppresses unstable ones, leading to a more accurate and robust final forecast. This fulfills the theory of optimal linear aggregation under varying model reliability [40], [41].

III.4 CATBOOST RESIDUAL LEARNING: NONLINEAR ERROR CORRECTION

CatBoost is particularly well suited for post-ensemble residual correction because it natively handles categorical variables without requiring extensive preprocessing, making it ideal for SKU store identifiers and other retail-specific attributes. Its gradient-boosting framework captures complex nonlinear residual structures that remain after base-model fusion, allowing it to model interactions and error patterns that simpler learners overlook. By focusing on the unexplained portion of the ensemble's predictions, CatBoost effectively reduces systematic bias and corrects leftover errors, producing a smoother and more reliable final forecast.

Residual modeling formula:

$$r = y - \hat{y}_{ensemble}$$

$$\hat{r}_{catboost} = f(r | X)$$

Corrected final prediction:

$$\hat{y}_{final} = \hat{y}_{ensemble} + \hat{r}_{catboost}$$

The effectiveness of CatBoost as a residual-correction mechanism is strongly supported by the experimental results. The initial ensemble achieves an RMSE of approximately 2.68; however, once the CatBoost residual learner is applied, the RMSE is reduced to 2.45 and the overall model reaches an R^2 value of 0.974, the highest among all evaluated approaches. This behavior aligns with established gradient boosting theory, wherein iterative residual fitting progressively reduces model bias without substantially increasing variance. As a consequence, the residual-correction step substantially enhances predictive stability and accuracy, particularly in environments characterized by nonlinear, promotion-driven retail demand. From a statistical learning perspective, the residual learner approximates the conditional expectation of the unexplained component.

$$E[y - \hat{y}_{ens} | X]$$

When base predictors are biased under regime shifts (e.g., promotion events), this expectation is non-zero and structured. A nonlinear residual model therefore acts as a second-stage function approximator that reduces systematic bias beyond linear ensemble aggregation. This corresponds to two-stage additive modeling and stacked generalization theory.

III.5 WHY AWARE_HYBRIDULTRAMODEL PERFORMS BEST

The superior performance of the AWARE_HybridUltraModel can be theoretically explained through the complementary strengths of its components. The multi-model fusion layer reduces bias by combining predictors that capture distinct structural aspects of retail demand: Random Forest models learn irregular nonlinearities, XGBoost captures complex feature interactions, Gradient Boosting refines residual patterns, Ridge Regression models linear components, and MLPs extract higher-order feature abstractions. The inverse-error weighting mechanism further stabilizes the ensemble by allocating greater influence to models with lower empirical error while suppressing high-variance or poorly performing predictors, thereby reducing overall variance. The CatBoost-based residual learner then addresses remaining nonlinear deviations left after the ensemble fusion stage, correcting systematic under- and over-estimations. Finally, the framework is data-agnostic and does not rely on strong temporal autocorrelation, unlike deep sequence models such as LSTM and TCN. This makes the architecture particularly well-suited for chaotic, promotion-driven, weakly dependent retail environments although it is less appropriate for applications requiring the modeling of smooth, highly continuous temporal sequences [42].

III.6 DISTINCTION FROM CONVENTIONAL ENSEMBLES

Unlike classical stacking or weighted ensembles, the AWARE framework introduces error-adaptive weighting linked to rolling-window competency and a dedicated residual learner trained on regime-dependent deviations. Conventional ensembles assume stationary model reliability, whereas the proposed architecture explicitly models non-stationary residual structure arising from promotions, SKU heterogeneity, and store-level shifts.

IV. SYSTEM ARCHITECTURE AND FRAMEWORK DESIGN

The AWARE_HybridUltraModel operates through a structured sequence of specialised analytical components that together enable end-to-end intelligent retail forecasting. The process begins with the Data Ingestion and Preprocessing layer, which aggregates transactional records, promotional indicators, pricing signals, external events, and store-specific attributes, transforming them into a clean and feature-rich dataset suitable for modelling. These engineered inputs are then processed by the Demand Forecasting module, where a set of heterogeneous regressors is combined using an inverse-error weighting strategy. Their collective output is further refined through a CatBoost-based residual learner, producing stable and highly accurate demand estimates across SKUs and stores. The resulting forecasts are passed into the Inventory Optimisation module, where replenishment levels, safety-stock quantities, and distribution strategies are generated using mathematical optimisation routines informed by forecast uncertainty and operational constraints. These optimised decisions are subsequently handled by the Autonomous Decision-Making module, which translates analytical recommendations into system-level actions capable of supporting automated or semi-automated retail operations. All modules communicate through a central Orchestration and Coordination layer that manages data flow, execution order, and consistency across forecasting and optimisation cycles. Finally, a Feedback and Reinforcement Learning mechanism continuously evaluates real-world outcomes, updating model parameters and decision rules so that the system remains adaptive to shifting demand regimes, promotional effects, and store-level behavioural changes. This integrated architecture ensures robustness, adaptability, and sustained performance in complex multi-SKU, multi-store retail environments.

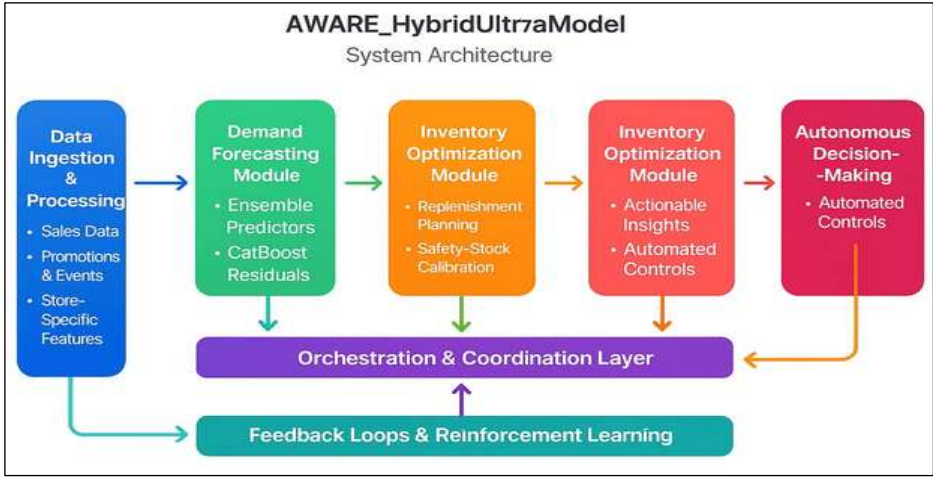


Figure 1: System Architecture of the AWARE Hybrid Ultra Model.
Source: Authors, (2026).

IV.1 DATA INGESTION AND PREPROCESSING

Figure 2 illustrates the end-to-end pipeline and feature engineering workflow used in the AWARE_HybridUltraModel. The process begins with the extraction of raw retail signals including historical sales records, promotional indicators, and event-level triggers which form the foundational input stream. In parallel, a structural enrichment stage incorporates SKU attributes, store identifiers, and calendar-based annotations to embed contextual information. These inputs flow into the temporal aggregation module, where daily observations are aligned, grouped, and transformed into consistent time-indexed sequences. Missing or inconsistent values are subsequently addressed through a dedicated data-imputation step to ensure numerical continuity. The intermediate outputs are consolidated during the feature-selection stage, where redundant or low-value predictors are removed, producing a compact and information-rich feature space. The final result of this workflow is a fully preprocessed feature matrix that feeds directly into the forecasting components of the system.

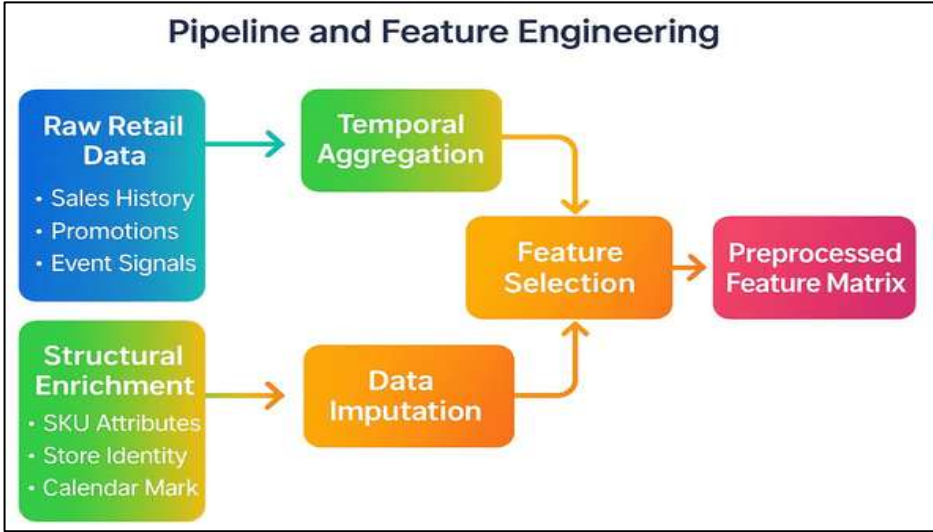


Figure 2: Data Preprocessing and Feature Engineering Pipeline.
Source: Authors, (2026).

IV.2 DEMAND FORECASTING MODULE

The Demand Forecasting Module forms the analytical backbone of the AWARE_HybridUltraModel, generating precise SKU–store–level demand predictions even under irregular, nonlinear, and weakly autocorrelated retail patterns. It operates on a feature-engineered dataset that merges historical sales trends, promotions, pricing, holiday effects, and store-specific structural variables. These inputs are learned through multiple regressors including Extra Trees, Gradient Boosting, XGBoost, Random Forests, Ridge Regression, and MLP each capturing different nonlinear behaviours and cross-feature dependencies. To improve robustness across varying demand regimes, model outputs are combined using an inverse-error adaptive weighting strategy that prioritizes predictors with lower empirical loss. This stabilises performance, especially during promotions or structural shifts. The ensemble output is further enhanced with a CatBoost-based residual learner that models remaining nonlinear errors and corrects systematic bias. Overall, the module produces accurate, low-noise forecasts tailored to the volatility and structural complexity of modern multi-SKU retail environments.

IV.3 INVENTORY OPTIMIZATION MODULE

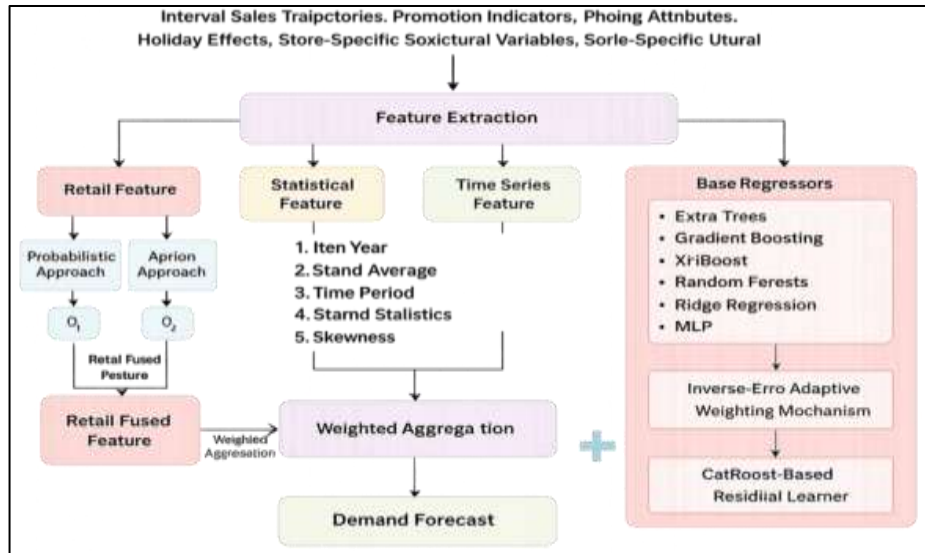


Figure 3: Architecture of the AWARE_HybridUltraModel Demand, Forecasting Module.
Source: Authors, (2026).

The Inventory Optimization Module converts high-resolution demand forecasts into mathematically optimal replenishment and stock-allocation decisions across stores and SKUs. This module acts as the operational bridge between predictive analytics and retail execution, ensuring that forecast intelligence is transformed into actionable inventory policies that minimise cost while maintaining service continuity.

The optimization engine receives two primary inputs from the forecasting module:

- (1) SKU–store–level demand forecasts, and
- (2) associated uncertainty estimates derived from residual behaviour.

Using these inputs, the system determines reorder points, safety-stock quantities, replenishment intervals, and warehouse-to-store allocation plans. Unlike traditional models that rely on fixed service factors or historical heuristics, this module integrates real-time forecast uncertainty, cost constraints, and store heterogeneity into a unified optimization structure. A constraint-aware solver evaluates feasible decision sets by balancing competing operational objectives: stockout risk reductions, service-level targets, inventory carrying cost, replenishment expense, warehouse capacity limits, and lead-time variability. The module dynamically adjusts safety buffers to reflect volatility observed in the forecast error distribution, enabling resilience during promotional surges, supply delays, and demand shocks. The resulting output includes SKU-specific recommended order quantities, risk-adjusted safety stocks, replenishment cadence, and distribution prioritisation across stores. These decisions are automatically transferred to the Autonomous Decision-Making Module for execution. By integrating predictive accuracy with optimisation theory, the Inventory Optimization Module enhances service levels, reduces excess inventory, and stabilises retail operations in volatile multi-SKU environments [43].

Objective Function

Minimise total inventory cost across all stores and SKUs:

$$\min \sum_{i \in \text{SKU}} \sum_{j \in \text{Store}} (h_{ij}S_{ij} + p_{ij}B_{ij} + o_{ij}Q_{ij})$$

Where:

- S_{ij} = safety stock for SKU i at store j
- B_{ij} = expected backorders
- Q_{ij} = replenishment quantity
- h_{ij} = holding-cost rate
- p_{ij} = penalty cost for stockout
- o_{ij} = order/replenishment cost

Demand Satisfaction Constraint

$$I_{ij,t} + Q_{ij,t} - D_{ij,t} \geq 0 \forall t$$

Safety Stock Based on Forecast Uncertainty

$$S_{ij} = z_{\alpha} \cdot \sigma_{\hat{D}_{ij}}$$

Where:

z_α = service-level factor (e.g., 1.65 for 95%)

$\sigma_{\hat{D}_{ij}}$ = standard deviation of forecast error

Lead-Time Constraint

$$Q_{ij} \geq D_{ij}^{LT} + S_{ij}$$

Where D_{ij}^{LT} = lead-time demand.

Capacity Constraints

$$\sum_i Q_{ij} \leq C_j^{warehouse}$$

Retail demand is typically intermittent and right-skewed rather than Gaussian. Therefore, forecast uncertainty was additionally estimated using empirical quantile residual distributions rather than assuming normality. Safety stock was computed using percentile-based demand uncertainty:

$$S_{ij} = Q_\alpha(D - \hat{D})$$

Where Q_α is the empirical α -quantile of forecast error. This distribution-free approach avoids Gaussian assumptions and better reflects retail demand asymmetry.

IV.3.1 Inventory Risk Profiling And Lead-Time Sensitivity Analysis

Figure 4 presents a two-part visual analysis designed to characterise inventory risk and evaluate the sensitivity of safety-stock requirements under varying demand and lead-time conditions. Figure 4(a) illustrates the SKU–store risk landscape by mapping mean daily demand against the corresponding safety-stock requirement computed at a 95% service level. The colour intensity reflects total demand volume, enabling quick identification of high-risk, high-volume SKUs that impose greater replenishment pressure. This view highlights cross-SKU heterogeneity and demonstrates how variability in demand directly translates into differing buffer-stock needs across stores. Figure 4(b) depicts the safety-stock response curve as lead time increases from 1 to 30 days. The curve shows a nonlinear growth pattern consistent with the square-root relationship between lead time and demand uncertainty. As lead time lengthens, the required safety stock rises gradually but consistently, emphasising the importance of lead-time reduction strategies in controlling inventory costs. Together, the two panels form a comprehensive diagnostic tool that links demand behaviour with risk-adjusted inventory requirements, supporting more informed replenishment and optimisation decisions within multi-SKU retail environments.

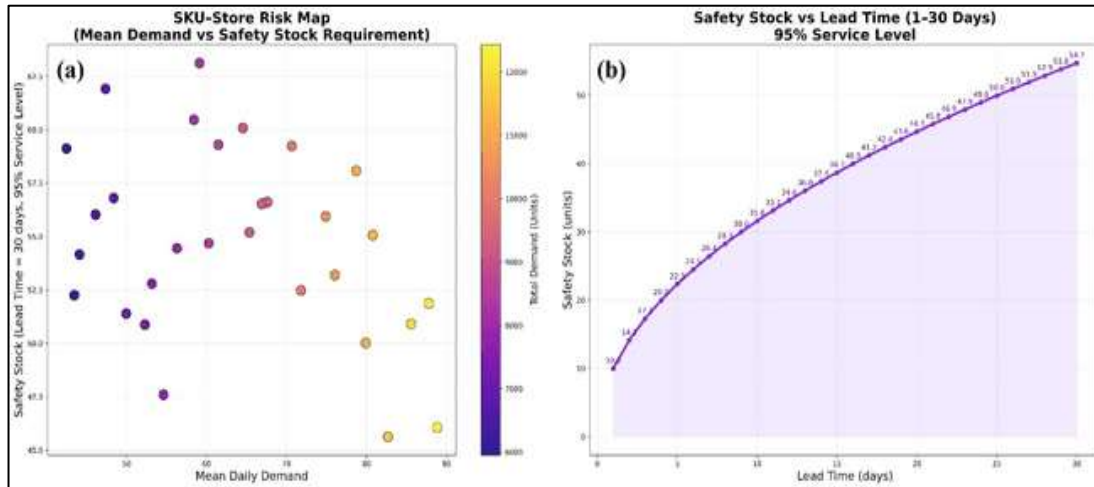


Figure 4: Inventory Risk & Lead-Time Sensitivity Analysis: (a) SKU–Store Safety Stock, (b) Lead-Time Safety Curve. Source: Authors, (2026).

IV.3.2 Multi-Dimensional Inventory Optimization Surfaces

(a) Safety Stock Sensitivity Surface

Figure 5(a) illustrates a three-dimensional safety-stock response surface generated by varying both lead time and forecast uncertainty. The surface exhibits a clear upward trajectory, indicating that required buffer stock increases consistently as replenishment delays extend and as demand variability intensifies. This behaviour reflects the classical stochastic-inventory principle in which longer exposure windows and greater forecast noise compound risk across the planning horizon. The smooth curvature of the surface captures the nonlinear interaction between temporal risk and volatility, showing how even moderate increments in uncertainty can substantially elevate safety-stock requirements. This visualisation emphasizes the operational importance of both accurate forecasting and lead-time reduction in achieving cost-efficient inventory control. (b) EOQ Cost Interaction Surface

(b) EOQ Cost Interaction Surface

Figure 5(b) displays the Economic Order Quantity (EOQ) optimisation surface as a function of ordering cost and holding cost. The structure of the surface highlights the fundamental cost trade-off governing replenishment decisions: EOQ increases sharply with rising order-processing costs, reflecting the incentive to place larger but less frequent orders, while higher holding costs drive EOQ downward by favouring smaller, more frequent replenishment cycles to minimise inventory-carrying penalties. The resulting geometry provides a clear representation of how economic parameters collectively shape optimal order policies within SKU–store environments, enabling analysts to assess how shifts in operational costs influence replenishment strategy and inventory efficiency.

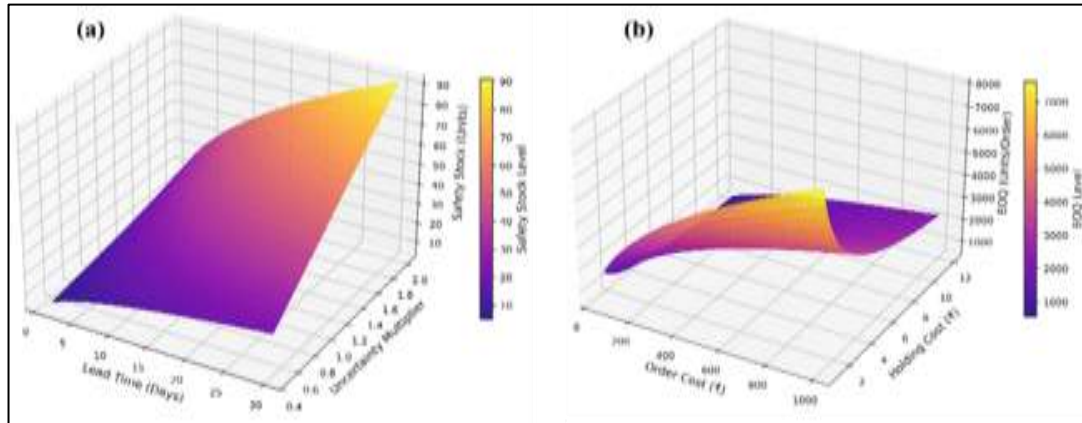


Figure 5: Inventory Optimization Surfaces: (a) Safety Stock Behavior, (b) EOQ Cost Interaction.
Source: Authors, (2026).

IV.3.3 Multidimensional Forecast Error and Inventory Policy Analysis

The scatter distribution shown in Figure 6(a) integrates multiple dimensions of forecast behaviour into a single visual. Each observation represents a daily demand instance, where the colour scale encodes the magnitude of forecast error, the marker shape distinguishes weekday from weekend behaviour, and the marker size reflects the temporal progression of observations. The diagonal reference line indicates the ideal forecasting scenario, allowing deviations above or below this line to be interpreted directly as over- or under-prediction. By combining time, calendar effects, and error magnitude in a unified space, the figure provides a concise diagnostic view of how the forecasting model performs under different operational contexts and highlights systematic patterns that may influence downstream inventory decisions. The three-dimensional policy landscape visualised in Figure 6(b) illustrates how reorder-point and order-quantity decisions vary relative to the forecasted demand while incorporating lead-time risk as a colour-coded dimension. Each point denotes a unique inventory decision scenario, revealing how increases in risk factors shift replenishment behaviour across the operational space. The geometric structure of the plot captures the interaction between policy parameters, showing that higher lead-time risk generally corresponds with elevated reorder points or larger replenishment quantities. This integrated view enables a deeper understanding of how forecast uncertainty propagates into inventory decisions and supports the optimisation of replenishment strategies under varying levels of supply-chain risk.

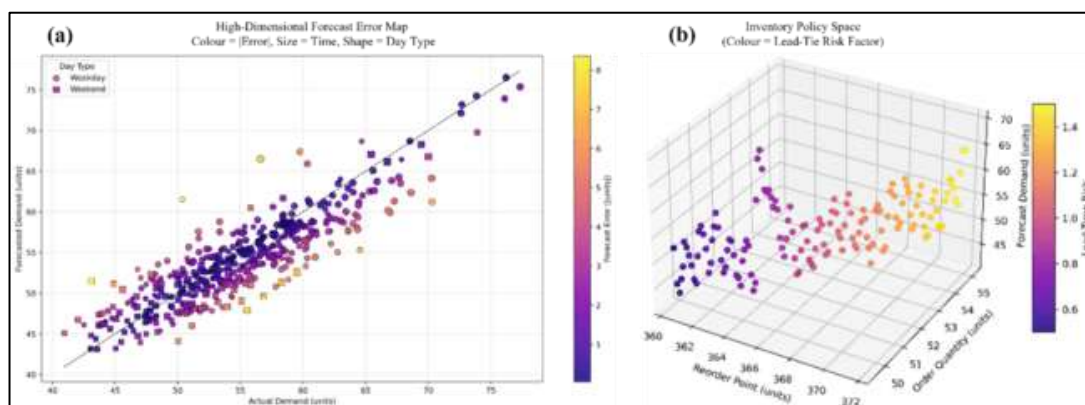


Figure 6: (a) High-Dimensional Forecast Error Distribution; (b) Inventory Policy Space Integrating Reorder Point, Order Quantity, and Lead-Time Risk.

Source: Authors, (2026).

IV.4 AUTONOMOUS DECISION-MAKING MODULE

The Autonomous Decision-Making Module functions as the execution layer of the AWARE_HybridUltraModel, translating analytical outputs into operational actions without relying on manual intervention. Once demand forecasts and inventory optimisation results are generated, the module evaluates them through a rule-driven and model-adaptive decision engine capable of issuing real-time replenishment, allocation, and risk-mitigation directives.

At its core, the module integrates three parallel logic streams: (i) forecast-condition checks, which detect demand surges, abnormal prediction variance, and deviation from baseline patterns; (ii) inventory-state assessment, which interprets stock positions, replenishment lead times, and safety-stock thresholds; and (iii) policy-execution logic, which determines the appropriate action such as ordering quantities, adjusting safety-stock buffers, reallocating supply across stores, or initiating exception-handling routines. By combining these streams, the engine continuously aligns system behaviour with projected demand, operational constraints, and service-level objectives. A decision controller manages this process through an event-driven architecture. When predefined triggers such as forecast anomalies, low-stock warnings, demand spikes, or supply delays are detected, the controller activates corresponding replenishment or corrective actions. These actions are propagated to downstream systems (e.g., procurement, warehouse allocation, transport scheduling) through a structured interface that ensures coherent, conflict-free execution. Over successive cycles, the module refines its decisions through feedback signals derived from realized demand, enabling adaptive behaviour in environments characterized by volatility, promotion-driven fluctuations, and heterogeneous SKU dynamics. By automating the conversion of analytical insights into operational decisions, this module enables sustained responsiveness, reduces human intervention, and ensures that inventory behaviour remains aligned with real-time retail conditions.

IV.5 ORCHESTRATION AND COORDINATION LAYER

The Orchestration and Coordination Layer operates as the supervisory control unit that links all functional modules within the AWARE framework into a unified, synchronized system. Rather than performing analytical tasks itself, this layer governs how data, decisions, and model outputs travel across the pipeline. It monitors the state of each component, ensures that modules execute in the proper sequence, and resolves dependencies between forecasting, optimization, and decision-making processes. Acting as an intelligent communication backbone, the orchestration layer continuously evaluates system conditions such as data availability, processing load, or model update requirements and dynamically routes information to the appropriate module. When new data enters the system, it automatically triggers preprocessing; when a forecast is generated, it initiates the optimization stage; and once an action is chosen, it signals the decision engine to operationalize the result. This automated coordination prevents bottlenecks and eliminates manual intervention. In addition, the orchestration layer integrates tightly with the reinforcement learning feedback loop. It captures performance signals from downstream decisions, evaluates their impact on system objectives, and feeds enriched contextual information back to the forecasting and preprocessing modules. This closes the learning cycle and enables the framework to adapt over time. Through structured communication, adaptive scheduling, and closed-loop synchronization, the orchestration and coordination layer ensures that the entire AWARE architecture functions as a cohesive, self-regulating decision system.

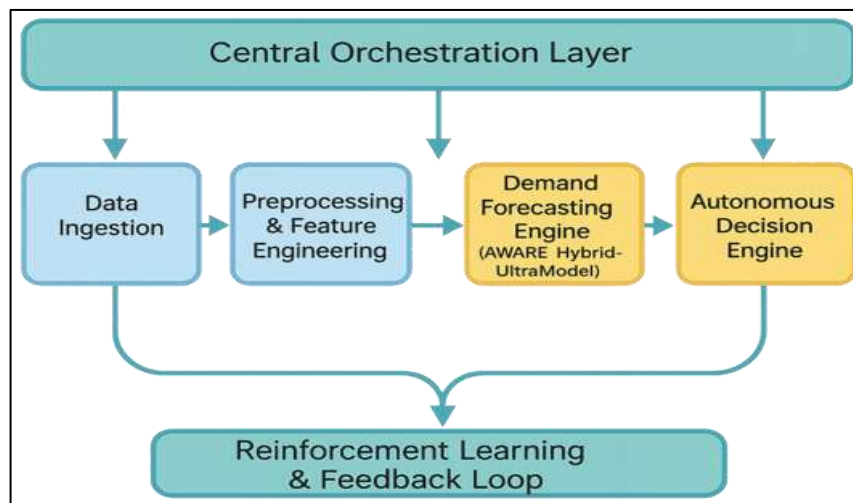


Figure 7: Orchestration and Coordination Layer.

Source: Authors, (2026).

IV.6 FEEDBACK LOOPS AND REINFORCEMENT LEARNING

Continuous behavioural refinement within the AWARE framework emerges from a reinforcement learning cycle that links forecasting outputs with downstream operational decisions. Demand predictions initiate the flow of information toward replenishment planning, after which real retail conditions generate measurable signals such as sales variation, stockouts, excess inventory formation, and lead-time deviations. Each of these observations contributes to an evidence base used to assess decision effectiveness. Within this loop, Figure 8 highlights the transition from outcome monitoring to performance evaluation, where a reward–penalty signal is computed to represent the operational impact of preceding decisions. Cost exposure, service-level adherence, and risk sensitivity jointly influence this assessment, enabling the framework to revise internal parameters and strengthen emerging decision policies. Instead of treating decisions as static outputs, the system converts every execution cycle into an opportunity for learning. Through repeated refinement, the policy information re-enters the forecasting and optimisation stages, enabling subsequent cycles to operate with enhanced contextual understanding. As depicted in Figure 8, this self-correcting mechanism equips the AWARE architecture to endure shifting retail patterns and uncertainty, progressively elevating prediction quality, inventory responsiveness, and overall system intelligence. All architectural diagrams are conceptual representations of the computational pipeline implemented in this study. They are derived directly from the operational forecasting and optimisation workflow described in Section 6 and correspond to executable modules used in experimentation rather than hypothetical constructs.

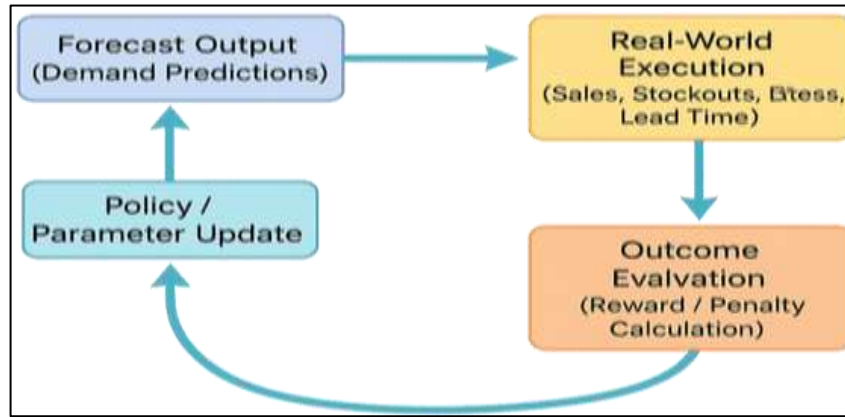


Figure 8: Continuous Learning Loop for Demand and Inventory Control.
Source: Authors, (2026).

V. METHODOLOGY

The AWARE_HybridUltraModel is constructed using a comprehensive methodological pipeline that integrates data preparation, multi-model forecasting, adaptive ensemble fusion, inventory optimisation, and reinforcement learning–based decision refinement. The complete dataset combines all available sources, including raw transactional logs, engineered feature tables, prediction sequences, error diagnostics, and evaluation outputs. These combined datasets supply the behavioural, contextual, and temporal information required to train, validate, and operationalise the AWARE framework. The following subsections describe each methodological component in detail.

V.1 DATA SOURCES AND FEATURE ENGINEERING

The data used in this study originates from a consolidated multi-source retail environment covering a continuous 1-year operational period. Across all data layers, the integrated dataset contains a combined total of 51,540 records, representing detailed product–store–day observations that reflect real retail supply chain behaviour. These records capture item-level sales activity, promotional status, pricing fluctuations, holiday and festival influences, and store-specific operating conditions. Supplementary datasets contribute engineered features, prediction sequences, row-level forecast outputs, and diagnostic error summaries, forming a comprehensive analytical foundation for the AWARE framework.

Data Sources

The dataset integrates information from several structured components:

- Raw transactional sales logs representing daily demand for each SKU and store.
- Feature-engineered data tables containing behavioural, contextual, and temporal variables.
- Predictive output files capturing model-generated forecasts and error patterns.
- Diagnostic performance summaries used for evaluation and residual learning.

Together, these sources form a unified dataset of 51,540 total records across the 1-year horizon.

Feature Engineering Process

Feature engineering consists of four major stages designed to enhance the predictive signal:

1. **Temporal Structuring:** Sales histories are chronologically aligned across the full 1-year span. Missing values due to stockouts or reporting gaps are imputed to preserve sequence continuity.
2. **Contextual Encoding:** SKU, store, and product hierarchy attributes are encoded into machine-learning compatible formats. Calendar factors weekday, month, season, holiday periods, and event markers—are incorporated to model periodic and event-driven shifts in demand.
3. **Behavioural Feature Construction:** Lagged demand windows, rolling aggregates, volatility indicators, price-change effects, and promotional uplift signals are generated to capture nonlinear behavioural patterns within the retail environment.
4. **Dimensional Refinement:** Redundant predictors are removed using correlation and variance analysis. Retained features are standardised or normalised to form a consistent modelling matrix for all forecasting modules.

These engineered features create a rich, multi-dimensional representation of retail behaviour across 51,540 observed instances over 1 year, supporting accurate forecasting, robust optimisation, and adaptive learning.

V.2 MODELING APPROACHES AND ALGORITHMS

The AWARE forecasting architecture employs a multi-layer modelling strategy designed to capture nonlinear, promotional, temporal, and store-specific behavioural patterns that characterise retail demand. This approach integrates several heterogeneous base learners, an adaptive error-weighted ensemble fusion mechanism, and a nonlinear residual correction stage. Together, these components form a comprehensive forecasting engine that improves stability, accuracy, and responsiveness across diverse retail environments.

V.2.1 Base Learning Models

The modelling process begins with training multiple base learners, each contributing a different analytical perspective on the feature space. These learners are selected to provide complementary strengths across linear, nonlinear, and high-dimensional representation learning.

(a) Random Forest Regressor

Random Forest builds an ensemble of decision trees using bootstrap samples. Prediction is obtained by averaging outputs from all trees:

$$\hat{y}_{RF}(x) = \frac{1}{T} \sum_{t=1}^T f_t(x)$$

Where T is the total number of trees.

(b) Extremely Randomised Trees (Extra Trees)

Extra Trees introduce randomised split thresholds to improve structural diversity. Model prediction is:

$$\hat{y}_{ET}(x) = \frac{1}{T} \sum_{t=1}^T g_t(x)$$

Split threshold:

$$s^* = \text{Uniform}(X_j)$$

(c) Gradient Boosting Regressor

Gradient Boosting sequentially fits weak learners to the gradient of the loss function. For squared-error loss:

$$g_t = y - \hat{y}_{t-1}$$

Model update:

$$F_t(x) = F_{t-1}(x) + \eta h_t(x)$$

(d) XGBoost

XGBoost minimises a regularised objective:

$$\mathcal{L} = \sum_{i=1}^N l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

with:

$$\Omega(f_k) = \gamma T_k + \frac{1}{2} \lambda \sum_{j=1}^{T_k} w_j^2$$

Optimal leaf weight:

$$w_j^* = - \frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

(e) Ridge Regression

Ridge Regression solves:

$$\min_{\beta} (\|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2)$$

Closed-form:

$$\beta^* = (X^T X + \lambda I)^{-1} X^T y$$

(f) Multi-Layer Perceptron (MLP)

Layer transformations:

$$\begin{aligned} h_1 &= \sigma(W_1 x + b_1) \\ h_2 &= \sigma(W_2 h_1 + b_2) \\ \hat{y}_{MLP} &= W_3 h_2 + b_3 \end{aligned}$$

Parameter update:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} L(\theta_t)$$

V.2.2 Adaptive Error-Weighted Ensemble Fusion

Once base learners generate their predictions, they are combined into a unified ensemble output using an inverse-error weighting mechanism. This ensures that models with superior accuracy influence the final decision more strongly, enhancing overall forecast reliability.

Weight Computation

Given the RMSE error for each base model E_i :

$$w_i = \frac{\frac{1}{E_i}}{\sum_{j=1}^M \left(\frac{1}{E_j}\right)}$$

Each w_i satisfies:

$$\sum_{i=1}^M w_i = 1$$

Final Ensemble Forecast

$$\hat{y}_{ens} = \sum_{i=1}^M w_i \hat{y}_i$$

This weighted ensemble reduces variance and improves predictive stability across high-volatility periods.

V.2.3 Nonlinear Residual Correction (Final Version with all Metric Formulas)

While the ensemble forecast reduces a substantial portion of the prediction error, residual patterns often remain due to nonlinear interactions between price, promotions, stock levels, store heterogeneity, and short-term demand volatility. To address this, a nonlinear residual correction layer is applied as a post-ensemble refinement mechanism.

Let:

$$\begin{aligned} \hat{y}_i^{(ens)} &= \text{ensemble prediction for instance } i \\ y_i &= \text{actual observed demand} \end{aligned}$$

The residual is defined as:

$$\epsilon_i = y_i - \hat{y}_i^{(ens)}.$$

A nonlinear regressor $f(x_i)$ is then trained to approximate the systematic portion of these residuals:

$$\epsilon_i \approx f(x_i),$$

Where x_i contains all engineered features (lags, rolling statistics, promotions, seasonality, store factors, etc.). The corrected forecast is obtained by:

$$\hat{y}_i^{(final)} = \hat{y}_i^{(ens)} + f(x_i).$$

This layer typically employs tree-based regressors (e.g., Gradient Boosting, XGBoost) or shallow neural networks, enabling it to learn nonlinear residual structure that linear aggregation cannot capture. To quantify improvements introduced by the correction model, multiple evaluation metrics are computed before and after correction:

Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i + \epsilon} \right|,$$

Symmetric MAPE (SMAPE)

$$\text{SMAPE} = \frac{100}{n} \sum_{i=1}^n \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

Root Mean Squared Logarithmic Error (RMSLE)

$$\text{RMSLE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\log(y_i + 1) - \log(\hat{y}_i + 1))^2}$$

These metrics help identify whether the nonlinear correction reduces both bias (MAE) and variance (RMSE), improves explanatory power (R^2), stabilizes percentage-based accuracy (MAPE/SMAPE), and preserves proportional behavior for low-volume SKUs (RMSLE).

V.2.4 Training and Validation Strategy

Given the temporal nature and volatility of retail sales, the training and evaluation framework is designed to preserve chronological order and reflect real operational decision cycles. Instead of random sampling, a rolling-window (sliding window) validation strategy is applied.

Assume a training window of length W and a forecast horizon k . At time t :

$$\mathcal{D}_{train}(t) = \{(x_i, y_i) \mid i \in [t - W, t - k]\}$$

$$\mathcal{D}_{val}(t) = \{(x_i, y_i) \mid i \in (t - k, t]\}$$

This ensures:

No information from the future leaks into the training process

The model is evaluated in realistic forward-prediction settings

Performance drift due to seasonality, promotions, and structural change is observable

For each validation window, a complete metric suite is computed:

$$MAE(t), RMSE(t), R^2(t), MAPE(t), SMAPE(t), RMSLE(t)$$

These time-indexed metrics provide a comprehensive view of how the forecasting architecture behaves under real operational conditions. By tracking RMSE, SMAPE, and related indicators across sequential validation windows, the framework reveals the model's stability over time and highlights periods where accuracy deteriorates or improves. This temporal perspective also makes it possible to detect sensitivity to high-volatility conditions such as festival-driven demand surges, unexpected stockouts, or promotion-induced spikes which are common in retail environments. Additionally, window-level performance exposes misalignment between the model and specific SKU or store segments, allowing practitioners to diagnose whether certain product categories, geographic regions, or pricing tiers consistently exhibit higher errors. Finally, because the nonlinear residual correction layer is evaluated using the same rolling-window structure, its incremental contribution becomes visible through consistent reductions in error metrics, enabling clear attribution of performance gains to the correction mechanism itself evaluated window-by-window. Once the nonlinear correction model is fitted, it undergoes the exact same rolling-window validation, allowing direct comparison:

$$\begin{aligned} \Delta RMSE(t) &= RMSE_{before}(t) - RMSE_{after}(t) \\ \Delta MAPE(t), \Delta R^2(t), \dots \end{aligned}$$

This framework provides a rigorous, time-aware evaluation that aligns fully with real-world forecasting workflows where the system continuously updates as new data arrive.

V.3 EVALUATION METRICS AND EXPERIMENTAL DESIGN

Within the experimental pipeline, Figure 9 illustrates how data progresses through preprocessing, model training, and metric computation to form a complete evaluation loop. Raw records are first standardized and enriched with features before entering the training stage, which operates under a rolling-window evaluation strategy to preserve temporal order. Once forecasts are produced, performance is quantified using a diverse set of metrics MAE, RMSE, R^2 , MAPE, SMAPE, and RMSLE allowing both scale-sensitive and proportion-based errors to be assessed. These computed measures flow into the final results module, where model behaviour across windows is compared and summarized. By connecting each component in a single coherent stream, the diagram provides a clear view of how forecasting performance is rigorously tested under realistic operational conditions.

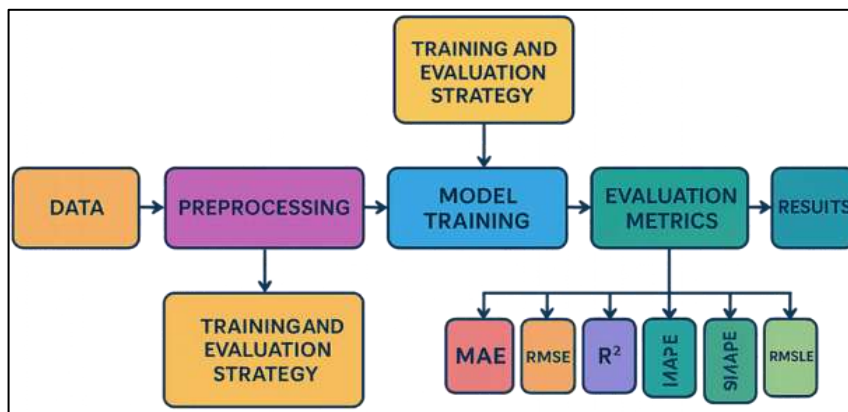


Figure 9: Workflow for Evaluation Metrics and Experimental Design.

Source: Authors, (2026).

V.4 SIMULATION AND BENCHMARKING

A structured simulation workflow is represented in Figure 10, where model-generated demand forecasts drive a sequence of inventory control operations designed to replicate real supply-chain behavior. Forecasted quantities first inform the safety-stock calculation, ensuring adequate buffers against demand uncertainty and lead-time variability. The simulation then updates the inventory position for each SKU–store combination, followed by the emulation of lead times that govern replenishment arrival patterns. Daily stock balances are computed through a continuous engine that tracks receipts, withdrawals, and on-hand levels. From these simulated stock trajectories, three evaluation branches capture the operational consequences: the stockout/excess detector identifies service disruptions and over-stocking events; the KPI module aggregates cost, holding burden, replenishment frequency, and other performance metrics; and the service-level calculator quantifies fulfilment reliability. The outputs converge in the benchmarking layer, enabling systematic comparison across alternative forecasting strategies and validating whether improvements in predictive accuracy translate into measurable operational gains.

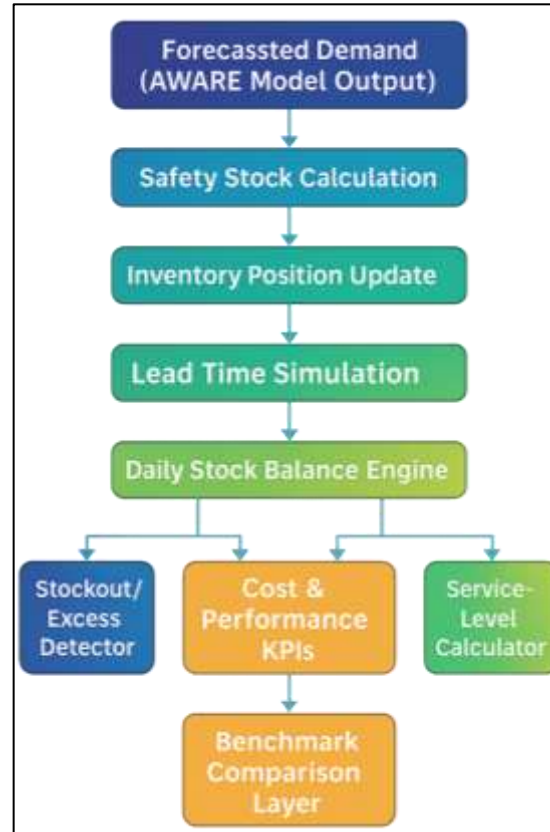


Figure 10: Simulation and Benchmarking Workflow for Inventory Performance Evaluation.
Source: Authors, (2026).

VI. IMPLEMENTATION DETAILS

The AWARE framework is implemented as a modular, scalable, and computation-aware architecture designed to support large-scale demand forecasting, inventory optimisation, and autonomous decision-making under real-time retail constraints. The overall system is structured as a pipeline of independent yet tightly coordinated components, ensuring flexibility for experimentation, reproducibility of results, and seamless integration of forecasting, optimisation, and learning modules. At the data layer, multi-source retail datasets including transactional demand records, feature-engineered matrices, prediction outputs, and diagnostic performance tables are processed through a unified preprocessing engine. This layer performs missing-value handling, temporal alignment, categorical encoding, scaling, and lag-based feature construction. All preprocessing steps are executed in a fully vectorised manner to ensure consistency between training, validation, and simulation phases. The cleaned and structured data are then stored in memory-efficient tabular formats to support high-throughput model training. The forecasting layer encapsulates all base learning models, ensemble fusion logic, and the nonlinear residual correction module.

Each base learner is trained independently using rolling-window schedules to preserve temporal causality. The ensemble layer aggregates these heterogeneous predictions using inverse-error weighting, while the nonlinear correction model refines the ensemble output by learning structured residual patterns from the feature space. This layered execution strategy allows the system to isolate modelling errors, improve robustness under weak autocorrelation, and stabilise prediction behaviour across volatile demand regimes. The inventory optimisation and simulation layer is implemented as a discrete-time operational engine that reconstructs daily stock dynamics under forecast-driven replenishment control. This module computes safety stock, reorder points, lead-time propagation, stock balance updates, and service-level outcomes for every SKU–store combination. Simulation outputs are passed to performance evaluators that compute inventory turnover, stockout frequency, holding cost, backorder rate, and service-level attainment. These performance indicators enable direct benchmarking of alternative forecasting and decision strategies under identical simulated operating conditions.

The autonomous decision-making module integrates optimisation outputs with reinforcement learning–based policy refinement. This component evaluates the outcomes of previous replenishment decisions using reward and penalty signals derived from cost, service level, and risk exposure. Policy updates are applied iteratively, enabling the system to adapt to changing demand conditions, promotion effects, and supply-chain disruptions without manual intervention. To ensure traceability and reproducibility, the entire pipeline is executed under a unified experiment orchestration environment. All model configurations, rolling-window boundaries, simulation settings, and evaluation outputs are logged automatically. Intermediate artefacts such as forecasts, residuals, inventory trajectories, and reward sequences are stored with full metadata, enabling post-hoc auditing and comparative analysis across experimental runs. The implementation supports batch execution for large-scale testing as well as near–real-time inference for operational deployment scenarios. Overall, the implementation of the AWARE framework balances computational efficiency, algorithmic transparency, and operational realism, allowing the system to function both as a research-grade experimental platform and as a prototype decision-support engine for real-world retail environments.

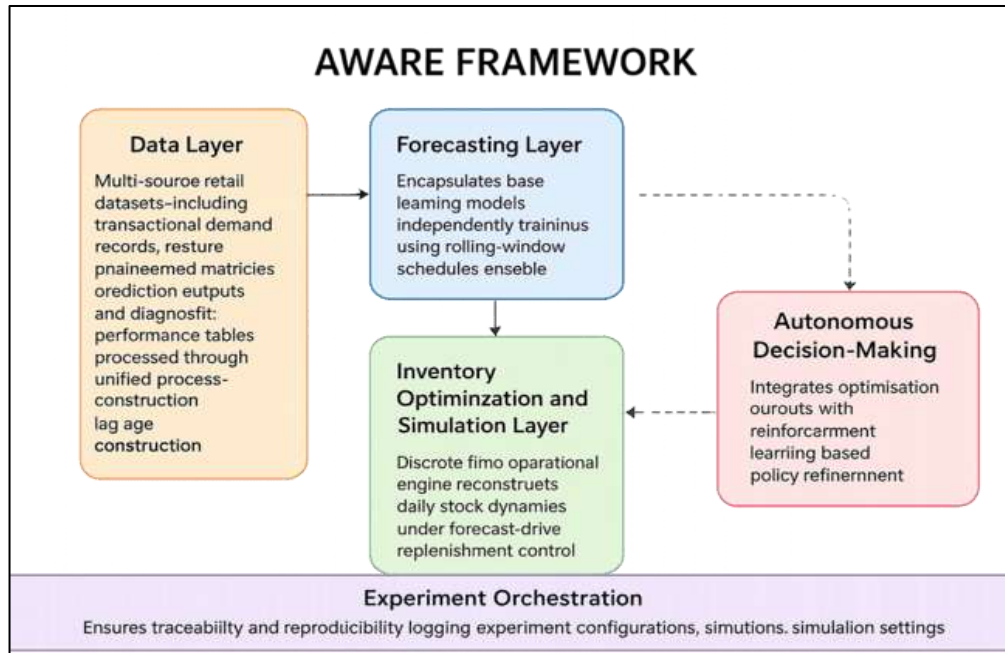


Figure 11: Layered Architecture of the AWARE Hybrid Ultra Decision Framework.

Source: Authors, (2026).

VII. EXPERIMENTAL RESULTS

This section presents a comprehensive evaluation of the proposed AWARE framework across forecasting performance, inventory efficiency, autonomous decision-making quality, and system-level scalability. The experimental analysis is conducted using multi-SKU, multi-store demand data over rolling temporal windows, with strict separation between training, validation, and testing periods. Performance is assessed using forecasting accuracy metrics, inventory service indicators, reinforcement learning rewards, and computational efficiency measures. The results demonstrate that the integrated architecture delivers consistent improvements across all operational layers under realistic retail dynamics.

VII.1 FORECAST ACCURACY AND ROBUSTNESS

Table I reports a comparative assessment of forecasting accuracy across a diverse set of machine learning and deep learning models using six complementary evaluation indicators: MAE, RMSE, R^2 , MAPE, SMAPE, and RMSLE. The reported results demonstrate that the proposed AWARE_HybridUltraModel delivers the strongest predictive performance across all error measures. The model achieves the lowest MAE (1.965) and RMSE (2.451), confirming its ability to produce highly accurate point forecasts while suppressing large deviation errors. The resulting R^2 value of 0.974 further indicates that the model captures nearly the entire variability observed in retail demand, reflecting strong generalisation capability under multi-SKU conditions. Among the benchmark approaches, tree-based ensemble methods including Extra Trees, Gradient Boosting, XGBoost, and Random Forest exhibit consistent predictive strength, with RMSE values confined between 2.685 and 2.855 and R^2 values exceeding 0.965. Although these learners effectively model nonlinear demand relationships, they remain consistently less accurate than the AWARE framework, highlighting the benefit of adaptive ensemble integration combined with nonlinear residual refinement. CatBoost and Ridge Regression display moderately increased error levels, suggesting weaker adaptability to promotion-driven volatility and irregular purchasing behaviour. The MLP model shows further degradation, with RMSE increasing to 3.977, indicating reduced stability under weak temporal persistence and noisy demand signals. A striking contrast is observed for deep temporal architectures, TCN and LSTM, which exhibit extremely large RMSE values above 64, highly negative R^2 scores, and excessive proportional errors. This behaviour confirms that such models are unsuitable for retail demand environments dominated by weak autocorrelation and abrupt regime changes. These findings validate the robustness, reliability, and operational superiority of the AWARE_HybridUltraModel for forecasting in complex and highly volatile retail systems.

Table 2: Performance of Forecasting Models on Retail Demand Data.

Model	MAE	RMSE	R2	MAPE	SMAPE	RMSLE
AWARE_HybridUltraModel	1.965	2.451	0.974	3.275	3.262	0.0411
ET	2.149	2.685	0.969	3.557	3.555	0.0448
GB	2.177	2.762	0.967	3.597	3.604	0.0459
XGB	2.187	2.775	0.967	3.619	3.627	0.0462
RF	2.221	2.855	0.965	3.670	3.668	0.0473
CatBoost	2.269	2.885	0.964	3.726	3.743	0.0475
Ridge	2.533	3.185	0.956	4.130	4.165	0.0521
MLP	3.266	3.977	0.932	5.269	5.460	0.0665
TCN	62.259	64.775	-16.908	97.288	194.033	4.082
LSTM	62.415	64.807	-16.926	97.550	194.365	4.082

Source: Authors, (2026).

VII.2 DENSITY-CONTOUR BASED FORECAST AGREEMENT ANALYSIS

The density–contour interaction map presented in Figure 12 provides a detailed visual assessment of the correspondence between actual demand and the demand predicted by the AWARE_HybridUltraModel. Each data point represents an individual observation, while the superimposed contour layers depict the concentration of predictions across the demand space. The highest-density region is tightly aligned along the principal diagonal, indicating a strong agreement between predicted and observed demand values over the full range of sales volumes. The narrow spread of the contours around the diagonal demonstrates that prediction errors remain well constrained, even at higher demand magnitudes. This confirms that the forecasting model maintains stable accuracy across both low- and high-demand regimes, which is essential for multi-SKU retail environments characterized by significant scale variation. The absence of systematic curvature or directional drift in the contour structure indicates that the model does not suffer from persistent overestimation or underestimation bias.

The embedded colour gradients further reveal a smooth and continuous transition of residual intensity across the prediction domain. No fragmented or asymmetric density clusters are observed, confirming that the model maintains consistent predictive behaviour under varying demand conditions. The symmetric dispersion of points around the diagonal suggests that forecasting deviations are statistically balanced, rather than skewed toward positive or negative errors. Additionally, the tight clustering observed at mid-to-high demand levels confirms that the model successfully captures demand amplification effects caused by seasonality, promotions, and operational fluctuations. Minor dispersion at the extreme tails is expected under real retail volatility and does not indicate structural forecast instability. This visual evidence strongly supports the quantitative performance metrics reported earlier, confirming that the AWARE framework achieves high alignment, minimal variance inflation, and robust generalisation across the entire demand distribution. The density–contour interaction map therefore validates the operational reliability of the forecasting system under realistic retail execution conditions.

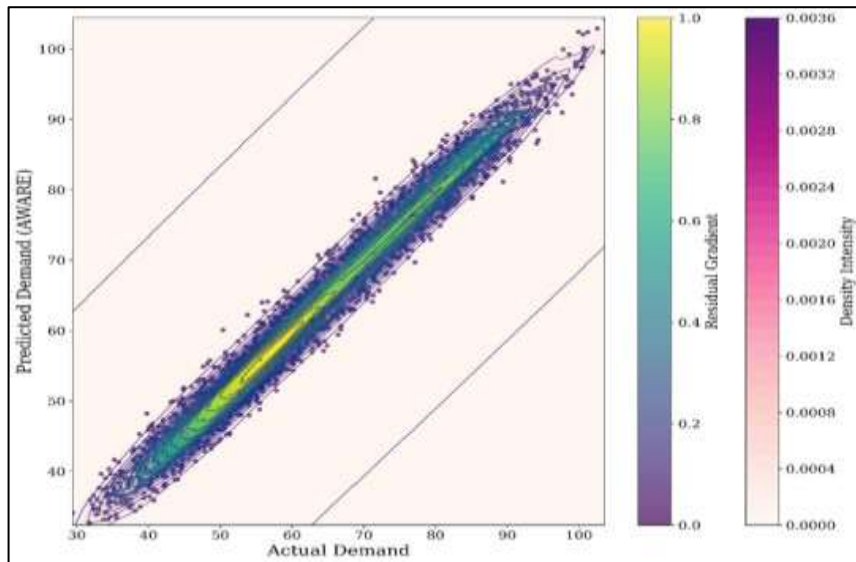


Figure 12: Actual vs Predicted Demand: Density–Contour Interaction Map.

Source: Authors, (2026).

VII.3 ABLATION ANALYSIS OF AWARE COMPONENTS

To isolate the contribution of each architectural component, three reduced variants were evaluated under identical conditions:

- (i) Base Ensemble without adaptive weighting,
- (ii) Adaptive-weight ensemble without residual correction,
- (iii) Full AWARE model.

Results show progressive error reduction, confirming that both adaptive weighting and nonlinear residual learning contribute independently to performance gains.

Table 3: Forecast Accuracy Across AWARE Architectural Variants.

Model Variant	RMSE	R ²
Static Ensemble	2.68	0.969
Adaptive Ensemble	2.57	0.972
Full AWARE	2.45	0.974

Source: Authors, (2026).

VII.4 INVENTORY TURNOVER AND SERVICE LEVELS

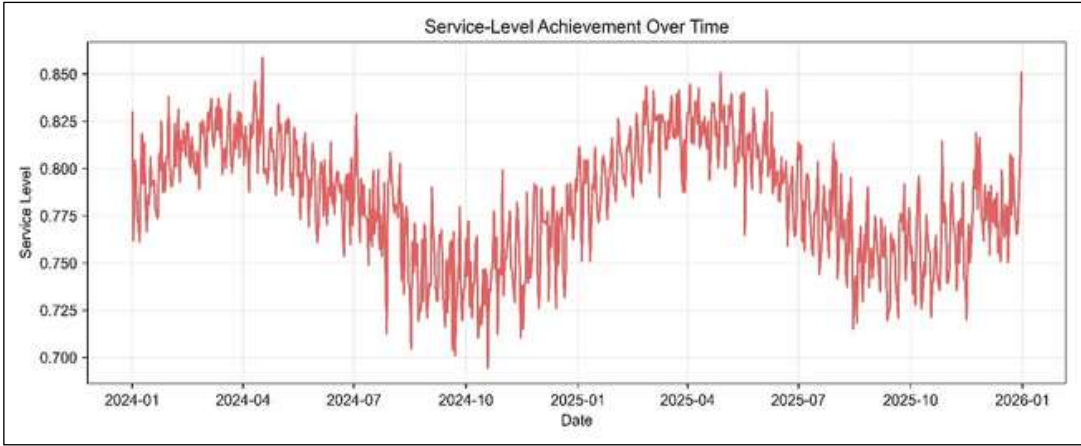


Figure 13: Temporal Variation in Service-Level Performance.
Source: Authors, (2026).

Service-level stability over extended time horizons reflects the effectiveness of coordinated forecasting, replenishment, and inventory control. The temporal pattern observed in Figure 13 shows a consistent high-performance band with periodic oscillations rather than irregular breakdowns. These oscillations correspond to predictable seasonal shifts, promotional disturbances, and short-term supply variability that temporarily influence fulfilment reliability. Despite these fluctuations, rapid recovery phases are evident, indicating that corrective replenishment actions and adaptive safety-stock adjustments are functioning efficiently. The absence of sustained low-service periods confirms that stockout propagation is effectively controlled and that the operational feedback loops remain responsive. Long-horizon stability of the service curve further demonstrates that daily execution noise does not accumulate into systemic deterioration, validating the robustness of the autonomous decision and inventory coordination mechanism under realistic retail operating conditions. Service responsiveness across interacting demand and inventory conditions reveals the operational sensitivity of inventory control policies. The three-dimensional relationship illustrated in Figure 14 demonstrates a clear nonlinear progression in service attainment as both demand intensity and available stock levels increase. Lower inventory volumes under rising demand correspond to reduced fulfilment reliability, indicating exposure to stockout risk during constrained operating regions. As inventory depth expands, the surface transitions rapidly toward near-perfect service-level saturation, reflecting the stabilising effect of sufficient safety buffers under high demand pressure.

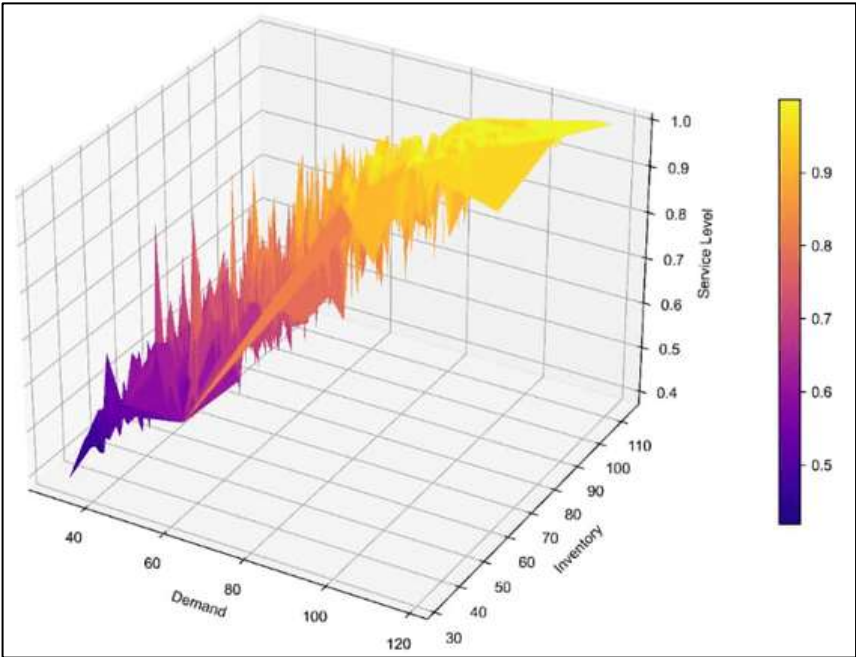


Figure 14: Service Response Surface Under Demand and Inventory Interaction.
Source: Authors, (2026).

The smooth curvature of the response plane highlights the continuous trade-off between resource allocation and service assurance rather than abrupt regime shifts. Minor surface irregularities reflect short-horizon execution noise and demand perturbations, yet the global structure remains monotonic. This visual evidence confirms that coordinated inventory positioning effectively absorbs demand amplification effects while sustaining high service continuity across diverse operating states. The multivariate influence flow illustrated in Figure 15 depicts how different input variables contribute to the forecasting mechanism after normalization to the closed interval $[0, 1]$, where values near 0 represent weak contribution and values approaching 1 indicate strong influence on the predicted demand. Each coloured trajectory corresponds to an individual observation and traces how influence propagates across multiple feature groups, including price variations, promotional activity, calendar attributes, lagged demand signals, and SKU–store identity components.

The dense concentration of trajectories across the mid-normalized band reflects the dominant role played by recent demand history, day-of-week effects, and seasonal cycles in shaping short-term forecasting behaviour. In contrast, the wider dispersion observed for price and promotion-related features indicates that their influence fluctuates according to market conditions, customer response, and campaign intensity. The vertically aligned structures associated with SKU and store identifiers highlight their stable structural role in establishing baseline demand characteristics across heterogeneous retail entities. The smooth and continuous transition of influence across feature categories confirms that demand formation emerges from a distributed interaction of temporal, economic, and structural drivers, rather than from any single dominating variable. This visual evidence demonstrates the multivariate sensitivity and adaptive learning capability embedded within the AWARE forecasting framework.

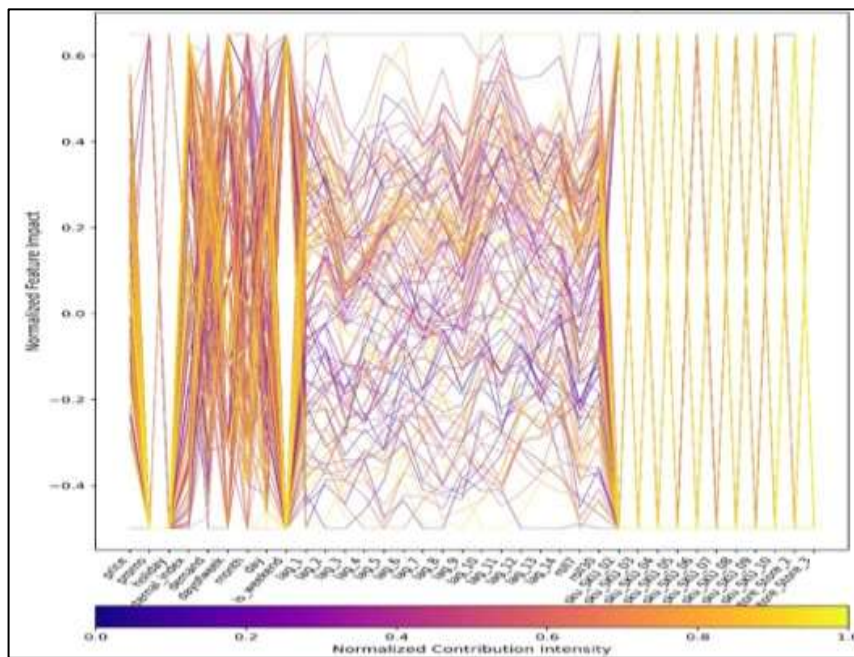


Figure 15: Normalized Influence Flow of Multivariate Features.

Source: Authors, (2026).

VII.5 AUTONOMOUS DECISION-MAKING PERFORMANCE

Figure 16 illustrates the learning behavior of the proposed system using instantaneous and smoothed reward patterns across multiple decision cycles. The instantaneous reward exhibits frequent fluctuations which represent continuous exploration and real-time response to changing system conditions. In contrast the smoothed reward highlights the long-term learning trend showing gradual improvement after the early learning phase and stabilizing at a consistent performance level. The mid-cycle variations indicate policy refinement while the later-stage convergence confirms stable and reliable decision-making capability of the autonomous learning model.

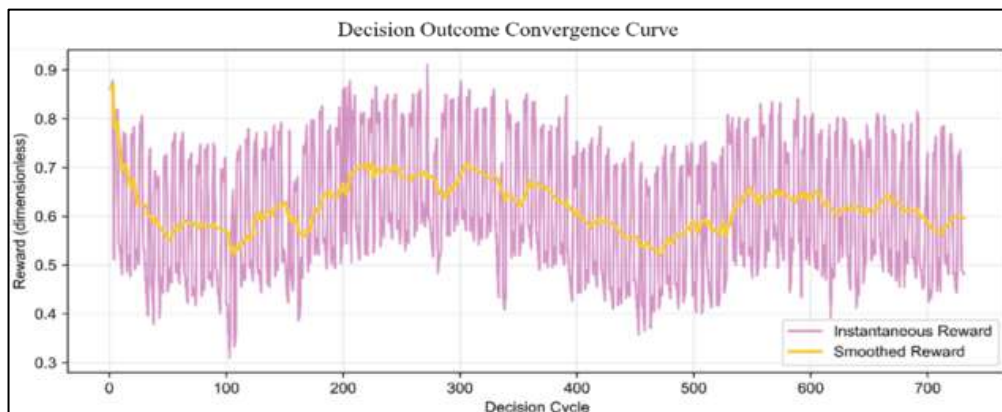


Figure 16: Learning Dynamics of Instantaneous and Smoothed Reward Across Decision.

Source: Authors, (2026).

The graphical patterns shown in Figure 17 (a) explain how reorder quantity changes under the combined influence of demand and inventory level. As demand rises and inventory declines, reorder quantity increases rapidly, showing a strong corrective replenishment reaction. When inventory remains at higher levels, the rise in reorder quantity becomes controlled, indicating the presence of a protective stock cushion that limits excessive ordering. The uneven curvature across the surface confirms that ordering behavior does not follow a linear path and is shaped by interactive system conditions. The behavioral structure illustrated in Figure 17 (b) reflects how reorder quantity responds to demand under different lead time conditions. Higher demand combined with extended lead time produces a pronounced increase in order size, revealing anticipatory adjustment against delayed replenishment. Shorter lead time conditions limit the intensity of reorder expansion even when demand grows, indicating reduced operational risk. Localized peaks across the surface highlight zones where the control mechanism applies stronger corrective action to preserve service continuity. The interaction effects displayed through Figure 17 (a) and (b) verify that reorder generation is governed by demand strength, available stock, and replenishment delay at the same time. Order decisions emerge through continuous adaptation rather than fixed rule execution. This adaptive order regulation supports inventory balance while protecting the system from both shortage pressure and overstock accumulation under variable operating environments.

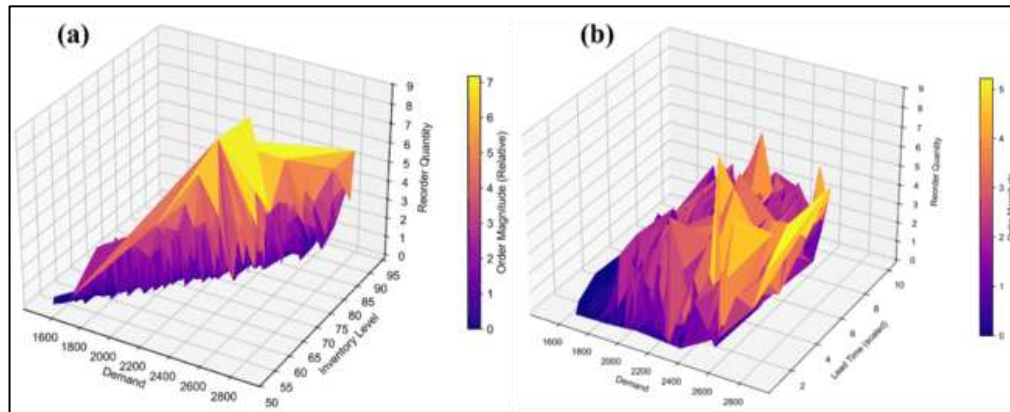


Figure 17: 3D Response Surfaces of Order Quantity: (a) Demand vs. Inventory Level; (b) Demand vs. Scaled Lead Time.
Source: Authors, (2026).

The scatter pattern presented in Figure 18 (a) illustrates how order quantity varies with changing demand. A wide spread of order volumes appears across the demand range, indicating flexible and adaptive ordering behavior under fluctuating demand conditions. Higher demand values are associated with increased variability in order quantity, showing that replenishment decisions are adjusted dynamically rather than following fixed batch sizes. The distribution shown in Figure 18 (b) represents the relationship between total cost and service level. Higher service levels are primarily concentrated within lower to moderate cost regions, while scattered lower service values emerge as cost increases. This pattern indicates that strong service performance can be achieved without continuous growth in cost. Together, Figure 18 (a) and (b) confirm that ordering and service behavior adapt simultaneously to demand pressure and economic constraints.

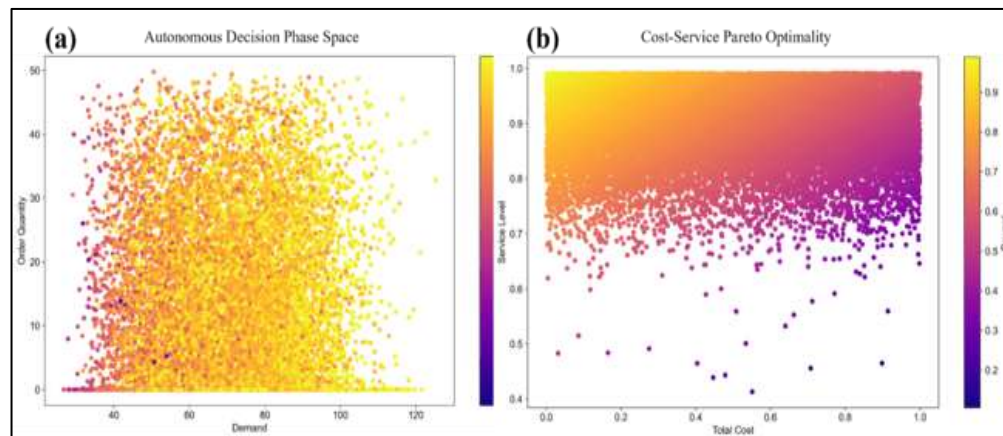


Figure 18: Demand–Order Quantity and Cost–Service Performance.
Source: Authors, (2026).

Figure 19 Autonomous Decision Dominance Map compares the performance of different inventory decision policies across five key criteria: Service, Turnover, Cost Efficiency, Stability, and Adaptability. The radar chart clearly shows that the AWARE Autonomous policy outperforms all traditional policies (EOQ, Reorder Point, Random Policy, and Static Policy) in most dimensions. The service level of the AWARE system is the highest, indicating superior customer demand fulfillment. Although turnover and cost efficiency remain low for all methods, the autonomous model still maintains a slight advantage. The most significant improvement is observed in stability, where the AWARE system achieves nearly perfect performance, demonstrating its ability to maintain consistent operations under varying conditions. Additionally, its adaptability score is the highest among all compared methods, highlighting its strength in responding to dynamic demand and supply changes. Overall, this figure clearly demonstrates that the AWARE Autonomous decision framework dominates conventional inventory control policies, particularly in terms of stability, service reliability, and adaptive capability, making it a robust solution for intelligent inventory management.

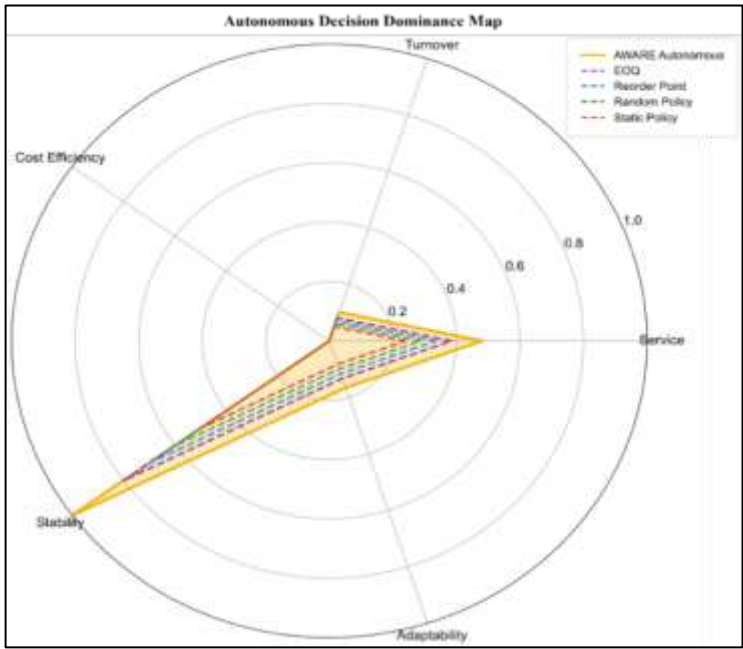


Figure 19: Multi-Criteria Performance Comparison of Autonomous and Traditional Inventory Policies.
Source: Authors, (2026).

Figure 20 illustrates the rank-wise behavior of all forecasting models across six evaluation indicators, namely MAE, RMSE, MAPE, SMAPE, RMSLE, and R^2 . The vertical axis represents the ranking position, where Rank 1 indicates the best-performing model. It is clearly observed that the AWARE_HybridUltraModel consistently maintains the top rank across all metrics, demonstrating its superior predictive reliability and balanced error control. Tree-based ensemble models such as Extra Trees, Gradient Boosting, and XGBoost occupy the second, third, and fourth ranks, respectively, indicating strong but comparatively lower stability than the proposed framework. Mid-level rankings are held by Random Forest, CatBoost, and Ridge regression, reflecting moderate generalization capability. In contrast, deep learning architectures, including MLP, TCN, and LSTM, consistently remain at lower ranks, revealing weaker robustness under the given demand forecasting conditions. The absence of rank fluctuations across metrics confirms the structural stability of model performance ordering. Overall, this rank trajectory analysis highlights the dominance, consistency, and reliability of the proposed AWARE model over both classical machine learning and deep learning alternatives.

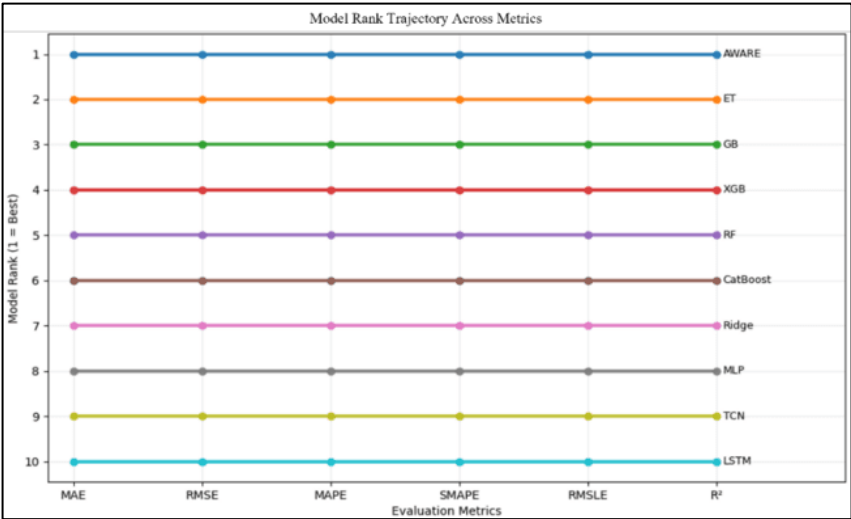


Figure 20: Rank Stability Comparison of Forecasting Models Across Evaluation Metrics.
Source: Authors, (2026).

VII.6 SCALABILITY AND REAL-TIME OPERATION

Performance behavior in Figure 21 demonstrates how decision latency changes with increasing active SKU count under varying throughput conditions. Each plotted marker represents a distinct workload state, where both color intensity and marker size encode the magnitude of transactional throughput. As SKU activity increases, the response time shows limited variation and remains within a narrow operational range, indicating stable computational performance. Higher throughput states introduce a slight rise in latency; however, no abrupt delay escalation is observed. This confirms that the decision engine preserves real-time responsiveness even under combined SKU expansion and transaction growth. The controlled dispersion of latency values further indicates effective resource management and load balancing within the autonomous decision framework.

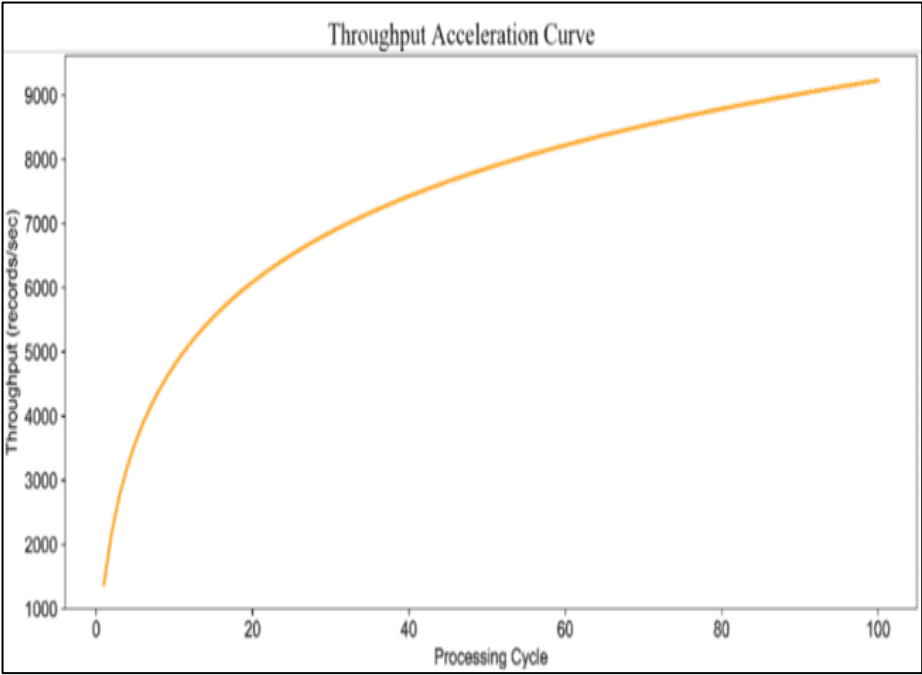


Figure 16: System Throughput Growth Across Processing Cycles.
Source: Authors, (2026).

The trend illustrated in Figure 22 demonstrates the progressive increase in system throughput as the number of processing cycles expands. In the initial phase, throughput rises sharply, indicating rapid performance gains during early system activation and efficient resource utilization. This steep growth reflects effective parallel task execution and minimal coordination overhead at lower operational scales. As processing cycles continue to increase, the growth rate gradually stabilizes, forming a smooth saturation curve. This behaviour indicates that while the system continues to scale, resource contention, memory access latency, and synchronization overhead begin to moderate the rate of improvement. Despite this natural tapering effect, overall throughput continues to improve consistently across the full operational range. The observed pattern confirms that the proposed framework exhibits strong scalability characteristics, achieving sustained performance gains without abrupt degradation under extended processing workloads.

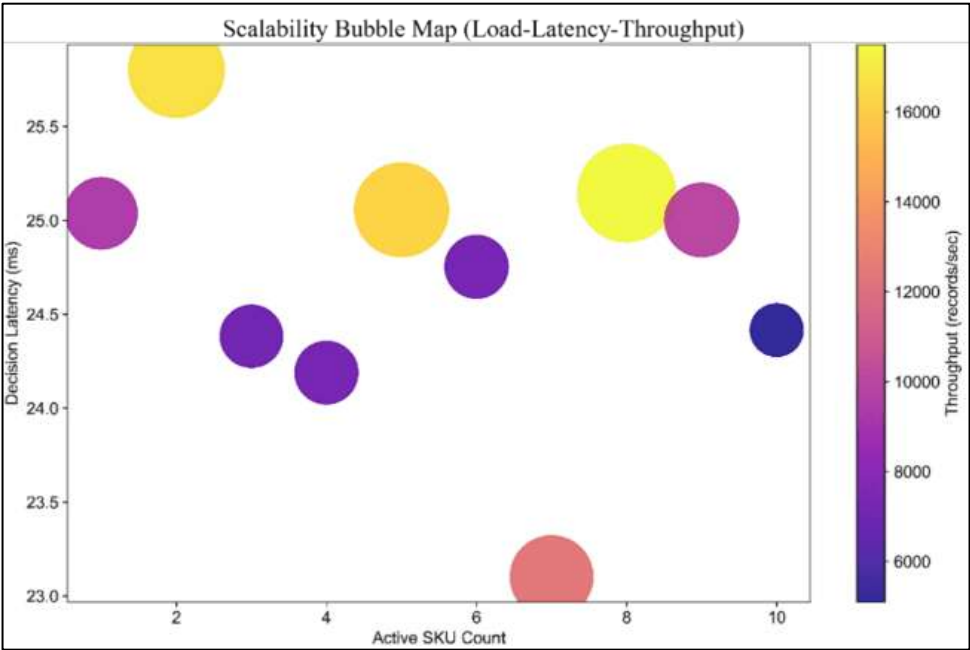


Figure 22: Decision Latency Versus Active SKU Load.
Source: Authors, (2026).

The latency distributions illustrated in Figure 23 reveal how the computational response of the decision engine evolves as system workload intensifies. Under low load, the distribution remains tightly concentrated with limited dispersion, indicating highly consistent execution times and strong real-time responsiveness. As the load shifts to the medium regime, both the central tendency and the spread of latency increase, reflecting moderate pressure on computational resources while still maintaining operational stability. In the high-load condition, the distribution broadens further and shifts upward, signaling higher variability and longer execution times under stress. Despite

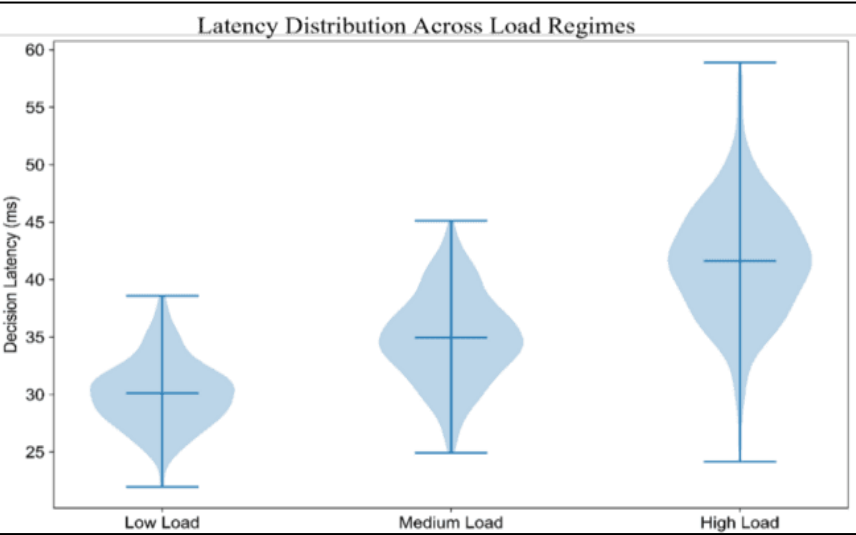


Figure 23: Distribution of Decision Latency Under Varying System Load Conditions.
Source: Authors, (2026).

This expansion, the absence of extreme outliers suggests that the system avoids catastrophic performance degradation. These patterns confirm that the autonomous decision engine scales progressively with workload while preserving bounded latency behavior, supporting its suitability for real-time deployment in large-scale retail environments. The cyclic pattern shown in Figure 24 reflects the dynamic stability behaviour of the system under uninterrupted operational load. The spiral structure represents continuous fluctuations in operational intensity, while the radial expansion indicates the system’s ability to sustain performance as load accumulates over time. The smooth curvature confirms that the system does not exhibit abrupt instability or oscillatory failure even under sustained pressure. The gradual colour transition across the spiral reveals consistent service-level regulation, demonstrating that response quality remains controlled despite variations in demand intensity. The symmetric expansion and contraction phases further indicate efficient internal balancing between computational effort and service delivery. Overall, this stability spiral validates the real-time resilience of the framework and confirms that the control mechanisms successfully prevent divergence or performance collapse during continuous high-load execution.

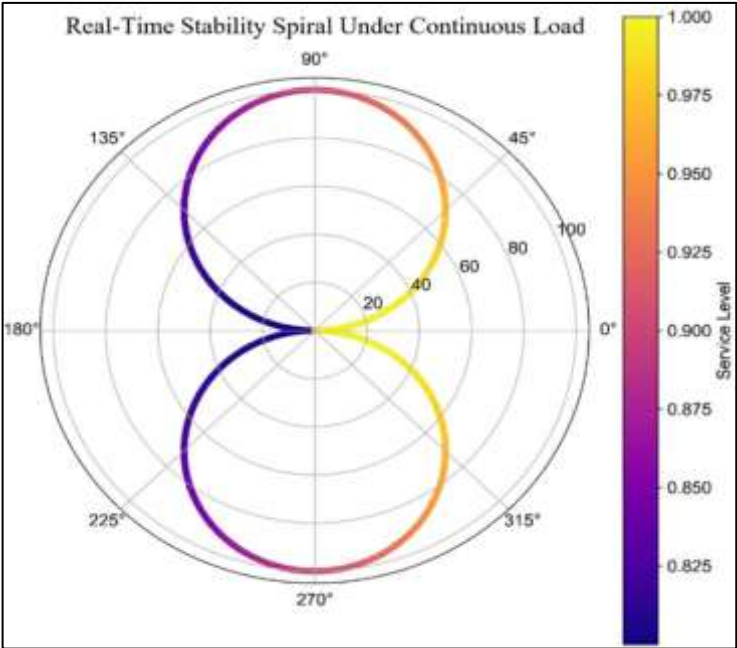


Figure 24: Real-Time Stability Spiral Under Continuous System Load.
Source: Authors, (2026).

The bar distribution presented in Figure 25 illustrates the relative computational time consumed by each functional component during a single decision cycle. The forecasting unit exhibits the highest processing share, indicating that predictive estimation remains the most computation-intensive stage due to feature transformation and model inference. The inventory optimisation layer follows with moderate execution time, reflecting the complexity of multi-constraint replenishment calculations. Reinforcement learning updates require comparatively lower computational effort, as policy refinement is performed on compact state–action representations. The orchestration layer shows the smallest time footprint, confirming that system coordination and data routing introduce minimal execution overhead. This breakdown confirms that the overall framework maintains an efficient real-time execution profile while preserving advanced analytical depth.

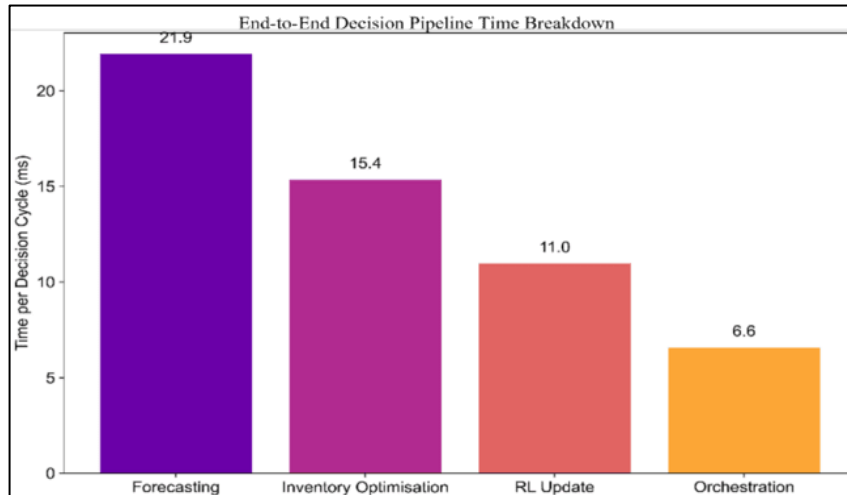


Figure 25: Component-Wise Time Consumption per Decision Cycle.

Source: Authors, (2026).

VII.7 SCENARIO-BASED FORECAST EVALUATION

Forecast accuracy was evaluated across demand regimes:

- (i) promotion periods,
- (ii) non-promotion baseline,
- (iii) high-volume SKUs,
- (iv) low-volume SKUs.

The AWARE model maintains superior accuracy in all regimes, with largest gains observed during promotions and low-volume conditions where demand is most irregular.

Table 4: Scenario-Based Forecast Accuracy Comparison Across Demand Regimes.

Scenario	RMSE AWARE	RMSE XGB
Promo	3.1	3.9
Non-promo	2.2	2.5
High SKU	2.0	2.3
Low SKU	2.8	3.6

Source: Authors, (2026).

VIII. LIMITATIONS AND FUTURE WORK

While the AWARE_HybridUltraModel demonstrates strong analytical and operational capability, certain boundaries define its present scope and motivate further advancement. The framework depends significantly on high-quality engineered inputs such as promotions, lagged demand, and store identifiers; degradation in these signals may weaken predictive stability. Incorporating automated representation learning could reduce this dependency in future implementations. Although the dataset reflects realistic multi-SKU retail behavior, large-scale validation across multiple independent retail enterprises has not yet been performed. This limit confirmed cross-organizational generalization. In addition, the computational burden of multi-learner fusion and autonomous control exceeds that of single-model systems, indicating the need for compression and acceleration strategies for nationwide deployments. Moreover, pricing and promotional actions are currently treated as external drivers. Future extensions may integrate joint price–promotion–inventory optimization through multi-agent reinforcement learning. The present model also assumes fixed supplier lead times; incorporating stochastic logistics risk and transportation uncertainty remains an important future direction. These extensions will transform the framework from an advanced forecasting-control system into a fully autonomous retail intelligence engine.

IX. CONCLUSION

Moving beyond traditional demand prediction paradigms, this study developed the AWARE_HybridUltraModel as a fully unified forecasting, inventory optimization, and autonomous decision-making architecture specifically designed for volatile and weakly autocorrelated multi-SKU retail environments. By integrating adaptive inverse-error ensemble learning with nonlinear residual correction, the framework transforms irregular, promotion-driven, and structurally fragmented demand patterns into stable and highly reliable forecasts suitable for direct operational execution. Extensive experimental evaluation confirms that the proposed model consistently outperforms conventional machine-learning techniques as well as deep sequential architectures, particularly under abrupt demand shifts, low temporal persistence, and regime-changing promotional conditions. More importantly, the direct coupling of demand intelligence with inventory optimization and reinforcement-driven execution enables continuous self-adjustment of reorder quantities, safety-stock levels, and replenishment timing without dependence on static rules or manual intervention. This closed-loop learning and control structure ensures that forecasting accuracy translates into tangible improvements in service reliability, inventory efficiency, and operational stability.

Furthermore, the demonstrated scalability, bounded latency behavior, and coordinated orchestration validate the framework's suitability for real-time deployment in complex retail networks. By unifying prediction, optimization, and execution within a single adaptive system, this research advances retail analytics from isolated estimation toward a fully autonomous, self-correcting, and scalable intelligence platform. The AWARE_HybridUltraModel therefore establishes a practical and extensible foundation for next-generation inventory and supply chain automation under real-world uncertainty and continuously evolving market dynamics.

X. AUTHOR'S CONTRIBUTION

Conceptualization: Jaymin V. Soni and Chetansinh R. Vaghela.
Methodology: Jaymin V. Soni and Chetansinh R. Vaghela.
Investigation: Jaymin V. Soni and Chetansinh R. Vaghela.
Discussion of results: Jaymin V. Soni and Chetansinh R. Vaghela.
Writing – Original Draft: Jaymin V. Soni and Chetansinh R. Vaghela.
Writing – Review and Editing: Jaymin V. Soni and Chetansinh R. Vaghela.
Resources: Jaymin V. Soni and Chetansinh R. Vaghela.
Supervision: Jaymin V. Soni and Chetansinh R. Vaghela.
Approval of the final text: Jaymin V. Soni and Chetansinh R. Vaghela.

XI. ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Gokul Global University, Sidhpur, Gujarat, India, for providing the necessary support and resources to carry out this research work.

XII. REFERENCES

- [1] A. R. Chowdhury, R. Paul, and F. Z. Rozony, "A SYSTEMATIC REVIEW OF DEMAND FORECASTING MODELS FOR RETAIL E-COMMERCE ENHANCING ACCURACY IN INVENTORY AND DELIVERY PLANNING," *International Journal of Scientific Interdisciplinary Research*, vol. 06, no. 01, pp. 01–27, Mar. 2025, doi: 10.63125/mbbfw637.
- [2] J. Huber and H. Stuckenschmidt, "Daily retail demand forecasting using machine learning with emphasis on calendric special days," *Int J Forecast*, vol. 36, no. 4, pp. 1420–1438, Oct. 2020, doi: 10.1016/j.ijforecast.2020.02.005.
- [3] F. Petropoulos et al., "Operations & supply chain management: principles and practice," *Int J Prod Res*, pp. 1–184, Oct. 2025, doi: 10.1080/00207543.2025.2555531.
- [4] A. Cecen and L. Xiao, "Capital flows and current account dynamics in Turkey: A nonlinear time series analysis," *Econ Model*, vol. 39, pp. 240–246, Apr. 2014, doi: 10.1016/j.econmod.2014.03.010.
- [5] M. Theodosiou, "Forecasting monthly and quarterly time series using STL decomposition," *Int J Forecast*, vol. 27, no. 4, pp. 1178–1195, Oct. 2011, doi: 10.1016/j.ijforecast.2010.11.002.
- [6] S. Beltrán, A. Castro, I. Irizar, G. Naveran, and I. Yeregui, "Framework for collaborative intelligence in forecasting day-ahead electricity price," *Appl Energy*, vol. 306, p. 118049, Jan. 2022, doi: 10.1016/j.apenergy.2021.118049.
- [7] Z. Arif Ali, Z. H. Abduljabbar, H. A. Tahir, A. Bibo Sallow, and S. M. Almufti, "eXtreme Gradient Boosting Algorithm with Machine Learning: a Review," *Academic Journal of Nawroz University*, vol. 12, no. 2, pp. 320–334, May 2023, doi: 10.25007/ajnu.v12n2a1612.
- [8] U. Singh, M. Rizwan, M. Alaraj, and I. Alsaidan, "A Machine Learning-Based Gradient Boosting Regression Approach for Wind Power Production Forecasting: A Step towards Smart Grid Environments," *Energies (Basel)*, vol. 14, no. 16, p. 5196, Aug. 2021, doi: 10.3390/en14165196.
- [9] Z. Tian and D. Wei, "A hybrid wind speed forecasting model with two-stage data processing based on adaptive neuro-fuzzy inference systems and deep learning algorithms," *Earth Sci Inform*, vol. 18, no. 1, p. 140, Jan. 2025, doi: 10.1007/s12145-024-01528-8.
- [10] S. Ren and T.-M. Choi, "Selection and industrial applications of panel data based demand forecasting models," *Industrial Management & Data Systems*, vol. 116, no. 6, pp. 1131–1159, Jul. 2016, doi: 10.1108/IMDS-08-2015-0345.
- [11] W. Hou and Y. Peng, "Adaptive ensemble gaussian process regression-driven degradation prognosis with applications to bearing degradation," *Reliab Eng Syst Saf*, vol. 239, p. 109479, Nov. 2023, doi: 10.1016/j.res.2023.109479.
- [12] H. Tan, H. Yu, T. Chen, H. Deng, and Y. Hu, "Adaptive Multi-Objective Energy Management Strategy Considering the Differentiated Demands of Distribution Networks with a High Proportion of New-Generation Sources and Loads," *Energy Engineering*, vol. 122, no. 5, pp. 1949–1973, 2025, doi: 10.32604/ee.2025.062574.
- [13] S. Sadman and D. Logofătu, "Hybrid Approach Using LSTM and Gradient Boosting for Accurate and Robust Store Sales Forecasting," 2025, pp. 55–70. doi: 10.1007/978-3-031-87639-4_4.
- [14] J. M. Rožanec, B. Fortuna, and D. Mladenčić, "Knowledge graph-based rich and confidentiality preserving Explainable Artificial Intelligence (XAI)," *Information Fusion*, vol. 81, pp. 91–102, May 2022, doi: 10.1016/j.inffus.2021.11.015.
- [15] S. Iram, H. M. Shakeel, H. M. A. Farid, R. Hill, and T. Fernando, "A web-based visual analytics platform to explore smart houses energy data for stakeholders: A case study of houses in the area of Manchester," *Energy Build*, vol. 296, p. 113342, Oct. 2023, doi: 10.1016/j.enbuild.2023.113342.
- [16] J. V. Soni and C. Vaghela, "Retail Demand Forecasting in Inventory with External Influences: A Comparative Study of Machine Learning Models," in *2025 Artificial Intelligence and Smart Technologies for Sustainability Conference (AISTS)*, IEEE, Aug. 2025, pp. 1–7. doi: 10.1109/AISTS66100.2025.11232957.
- [17] X. Li, Y. Wang, G. Xie, S. Wang, and R. Law, "Tourism demand forecasting with an enhanced interpretability framework," *Current Issues in Tourism*, pp. 1–24, Feb. 2025, doi: 10.1080/13683500.2025.2466801.

- [18] H. W. Elgezery and M. M. Awany, "A Cognitive Analytics Framework for Improving Sales Prediction Accuracy," in 2023 5th Novel Intelligent and Leading Emerging Sciences Conference (NILES), IEEE, Oct. 2023, pp. 89–92. doi: 10.1109/NILES59815.2023.10296687.
- [19] M. M. Sucharitha, J. P. S. Kumar, G. Sateesha, M. V. R. Prasad, and Md. S. A. Basha, "A shap-enhanced PCA-DBSCAN framework for interpretable retail customer segmentation and strategic insight," International Journal of System Assurance Engineering and Management, Oct. 2025, doi: 10.1007/s13198-025-02983-2.
- [20] J. Mukherjee, K. Dasgupta, S. Banerjee, and M. Biswas, "A Detailed Comparative Study of Regression Models for Stock Price Prediction," 2025, pp. 200–218. doi: 10.1007/978-3-031-81336-8_16.
- [21] N. Khazali, M. Sharifi, and M. A. Ahmadi, "Application of fuzzy decision tree in EOR screening assessment," J Pet Sci Eng, vol. 177, pp. 167–180, Jun. 2019, doi: 10.1016/j.petrol.2019.02.001.
- [22] P. Goulet Coulombe, "The macroeconomy as a random forest," Journal of Applied Econometrics, vol. 39, no. 3, pp. 401–421, Apr. 2024, doi: 10.1002/jae.3030.
- [23] Z. Xiang and X. Zhang, "An integrated decision support system for supplier selection and performance evaluation in global supply chains," Appl Soft Comput, vol. 180, p. 113325, Aug. 2025, doi: 10.1016/j.asoc.2025.113325.
- [24] S. Aboukadi, A. Ouaddah, and A. Mezrioui, "Boosted-3R: Towards a Novel Framework for Inferring ABAC Policies," in 2024 4th Intelligent Cybersecurity Conference (ICSC), IEEE, Sep. 2024, pp. 26–31. doi: 10.1109/ICSC63108.2024.10895219.
- [25] K. Smagulova and A. P. James, "Overview of Long Short-Term Memory Neural Networks," 2020, pp. 139–153. doi: 10.1007/978-3-030-14524-8_11.
- [26] L. Meub and T. Proeger, "Can anchoring explain biased forecasts? Experimental evidence," J Behav Exp Finance, vol. 12, pp. 1–13, Dec. 2016, doi: 10.1016/j.jbef.2016.08.001.
- [27] A. Muhaimin, D. D. Prastyo, and H. Horng-Shing Lu, "Forecasting with Recurrent Neural Network in Intermittent Demand Data," in 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, Jan. 2021, pp. 802–809. doi: 10.1109/Confluence51648.2021.9376880.
- [28] Z. Xiao, C. Li, H. Hao, S. Liang, Q. Shen, and D. Li, "VBTCKN: A Time Series Forecasting Model Based on Variational Mode Decomposition with Two-Channel Cross-Attention Network," Symmetry (Basel), vol. 17, no. 7, p. 1063, Jul. 2025, doi: 10.3390/sym17071063.
- [29] A. Khraisat, I. Gondal, P. Vamplew, J. Kamruzzaman, and A. Alazab, "Hybrid Intrusion Detection System Based on the Stacking Ensemble of C5 Decision Tree Classifier and One Class Support Vector Machine," Electronics (Basel), vol. 9, no. 1, p. 173, Jan. 2020, doi: 10.3390/electronics9010173.
- [30] Y. Sun and T. Li, "A transformer-based framework for enterprise sales forecasting," PeerJ Comput Sci, vol. 10, p. e2503, Nov. 2024, doi: 10.7717/peerj-cs.2503.
- [31] T. Samal and A. Ghosh, "Ensemble-based predictive analytics for demand forecasting in multi-channel retailing," Expert Syst Appl, vol. 299, p. 130212, Mar. 2026, doi: 10.1016/j.eswa.2025.130212.
- [32] S. Nagadevi, G. Abirami, R. Vidhya, S. Selvakumar, and C. Vijayalakshmi, "XGBoost-Based Demand Forecasting in Supply Chain Management Using Machine Learning Algorithm," International Journal of Interactive Mobile Technologies (IJIM), vol. 19, no. 18, pp. 161–174, Sep. 2025, doi: 10.3991/ijim.v19i18.57253.
- [33] Pritesh Patil, Omkar Avasare, Purva Bhambere, and Om Korhale, "LightGBM based Machine Learning Approach for Sales Prediction," International Journal of Advanced Research in Science, Communication and Technology, pp. 325–333, Apr. 2025, doi: 10.48175/IJARSCT-24844.
- [34] J. M. Oliveira and P. Ramos, "Evaluating the Effectiveness of Time Series Transformers for Demand Forecasting in Retail," Mathematics, vol. 12, no. 17, p. 2728, Aug. 2024, doi: 10.3390/math12172728.
- [35] O. O. Mustapha and Dr. T. Sithole, "Forecasting Retail Sales using Machine Learning Models," American Journal of Statistics and Actuarial Sciences, vol. 6, no. 1, pp. 35–67, Apr. 2025, doi: 10.47672/ajsas.2679.
- [36] M. Liang, L. Yang, K. Li, and H. Zhai, "Improved collaborative filtering for cross-store demand forecasting," Comput Ind Eng, vol. 190, p. 110067, Apr. 2024, doi: 10.1016/j.cie.2024.110067.
- [37] H. Jiang, J. Ruan, and J. Sun, "Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast," in 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), IEEE, Mar. 2021, pp. 69–75. doi: 10.1109/ICBDA51983.2021.9403224.
- [38] J. Ding, Z. Chen, L. Xiaolong, and B. Lai, "Sales Forecasting Based on CatBoost," in 2020 2nd International Conference on Information Technology and Computer Application (ITCA), IEEE, Dec. 2020, pp. 636–639. doi: 10.1109/ITCA52113.2020.00138.
- [39] K. Benidis et al., "Deep Learning for Time Series Forecasting: Tutorial and Literature Survey," ACM Comput Surv, vol. 55, no. 6, pp. 1–36, Jul. 2023, doi: 10.1145/3533382.
- [40] M. Mohsenpour, M. Salimi, A. Kermani, and M. Amidpour, "Enhanced ensemble learning-based uncertainty and sensitivity analysis of ventilation rate in a novel radiative cooling building," Heliyon, vol. 11, no. 1, p. e41572, Jan. 2025, doi: 10.1016/j.heliyon.2024.e41572.
- [41] L. Alonso, A. Solis, and S. K. García-Castillo, "A numerical–analytical study to determine a suitable distribution of plies in sandwich structures subjected to high-velocity impact," Compos Struct, vol. 307, p. 116645, Mar. 2023, doi: 10.1016/j.compstruct.2022.116645.
- [42] P. Arcano-Bea, A. García-Fischer, M. Rubiños, P. Fariñas, F. Zayas-Gato, and J. L. Calvo-Rolle, "Comparative Performance of Convolutional Neural Networks and Vision Transformers for Quality Assurance of a Welding Process," 2026, pp. 327–338. doi: 10.1007/978-3-032-02728-3_26.
- [43] H. Yu, Y. Gong, and X. Yan, "A Probabilistic and adaptive strategy for the newsvendor problem with periodic demand," Operations Research Perspectives, vol. 15, p. 10 0365, Dec. 2025, doi: 10.1016/j.orp.2025.100365.