JOURNAL OF CULTURAL ANALYSIS AND SOCIAL CHANGE

Journal of Cultural Analysis and Social Change,
2025, 10(4), 4056-4073
ISSN: 2589-1316

A Hybrid XGBoost-LSTM Framework for Supply Chain Demand Forecasting: Empirical Evidence from Retail Multi-Store Data

Bowen Wang[1*], Azlan Bin Mohd Zain[2]

[1]*Faculty of Computing, Universiti Teknologi Malaysia, Jalan Iman, 81310 Skudai, Johor, Malaysia; Email: 18612236586@163.com*
[2] *Faculty of Computing, Universiti Teknologi Malaysia, Jalan Iman, 81310 Skudai, Johor, Malaysia; Email: azlanmz@utm.my*

***Corresponding Author:** 18612236586@163.com

Citation: Wang, B. & Zain, A. B. M. (2025). A Hybrid XGBoost-LSTM Framework for Supply Chain Demand Forecasting: Empirical Evidence from Retail Multi-Store Data , *Journal of Cultural Analysis and Social Change*, 10(4), 4056-4073 . https://doi.org/10.64753/jcasc.v10i4.3736

Published: December 28, 2025

ABSTRACT

Supply chain demand forecasting is a core component of modern enterprise operations management, directly impacting inventory optimization, production planning, and customer satisfaction. Traditional statistical methods such as Autoregressive Integrated Moving Average (ARIMA) models face limitations when handling complex nonlinear patterns and multidimensional features. While single machine learning models enhance prediction accuracy, they struggle to simultaneously capture structured features and temporal dependencies.This study proposes an innovative hybrid forecasting framework integrating Extreme Gradient Boosting (XGBoost) with Long Short-Term Memory (LSTM) networks. XGBoost excels at learning high-dimensional structured features, while LSTM specializes in modeling temporal dependency patterns. An adaptive weight fusion mechanism enables complementary strengths between the two models.Empirical analysis using the Kaggle open-source retail dataset (encompassing 10 stores, 50 products, and 913,000 transaction records) demonstrates that the hybrid framework significantly outperforms both single models and traditional methods across multiple metrics, including Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE).The findings provide supply chain managers with actionable forecasting tools, offering significant theoretical and practical value for enhancing prediction accuracy and operational efficiency.

Keywords: Demand forecasting; XGBoost-LSTM hybrid model; Deep learning; Inventory optimization; Retail data analysis; Prediction accuracy

INTRODUCTION

Research Background

In an increasingly competitive global business environment, Supply Chain Management (SCM) has become a critical factor for enterprises to gain competitive advantages. As a core function of SCM, demand forecasting directly impacts inventory levels, production scheduling, logistics distribution, and ultimately customer satisfaction (Seyedan & Mafakheri, 2020).However, the complexity of modern retail environments—including diverse consumer behaviors, volatile market demands, seasonal influences, and promotional disruptions—poses significant challenges to traditional forecasting methods (Douaioui et al., 2024).

In recent years, the rapid advancement of big data technology and artificial intelligence has opened new opportunities for supply chain demand forecasting. Douaioui et al. (2024) conducted a systematic analysis of 119 publications from 2015 to 2024, revealing a significant growth trend in the application of machine learning and deep learning models within demand forecasting. Notably, 73% of the research was published between 2021 and

2024, indicating an accelerated phase of rapid development in this field.Traditional time series methods like Autoregressive Integrated Moving Average (ARIMA) models perform well in capturing linear trends and seasonality but exhibit clear limitations when handling nonlinear relationships and multidimensional features (Tseng & Turkmen, 2024).

Machine learning approaches offer novel perspectives to address these challenges. The Extreme Gradient Boosting (XGBoost) algorithm, proposed by Chen and Guestrin (2016), has gained widespread adoption in demand forecasting due to its outstanding predictive performance and computational efficiency. Zhang and Wu (2025) demonstrated XGBoost's superiority in handling complex feature interactions within e-commerce sales forecasting.Concurrently, Long Short-Term Memory (LSTM) networks, developed by Hochreiter and Schmidhuber (1997), demonstrate significant potential in supply chain demand forecasting due to their unique gating mechanisms that effectively capture long-term dependencies in time series data (Wang et al., 2024).

Research Motivation

Although individual machine learning models have made significant strides in predictive accuracy, each possesses distinct strengths and limitations. XGBoost excels at handling structured features and nonlinear relationships but has limited capacity for modeling sequential dependencies in time series; LSTMs effectively capture temporal patterns but incur high computational costs and are prone to overfitting when processing high-dimensional, heterogeneous features (Baker et al., 2021).As Feizabadi (2022) noted in their research, "Single forecasting methods struggle to simultaneously achieve breadth in feature learning and depth in temporal modeling, prompting researchers to explore hybrid approaches for superior predictive performance."

Hybrid models, by integrating the strengths of multiple algorithms, hold promise for overcoming the performance limitations of single models.Recent studies demonstrate that hybrid frameworks combining different model types exhibit superiority across multiple domains. For instance, the Prophet-LSTM hybrid model proposed by Bashir et al. (2022) significantly outperformed single methods in electricity load forecasting; the LSTM-XGBoost ensemble model developed by Xu et al. (2025) achieved lower prediction errors in blood demand forecasting.However, existing hybrid model research primarily focuses on sectors like energy and finance. Systematic hybrid frameworks for supply chain demand forecasting remain scarce, particularly in retail scenarios involving multiple stores and products.

Furthermore, existing studies tend to oversimplify the fusion mechanism design in hybrid models. Many adopt simple averaging or fixed-weight combinations (Baker et al., 2021), failing to fully account for the relative strengths of different models across diverse forecasting scenarios. As Lu et al. (2022) noted, "Fixed-weight ensemble methods cannot adapt to dynamic changes in data characteristics; adaptive weight optimization is key to enhancing hybrid model performance."

Research Objectives and Contributions

Addressing these research gaps, this study aims to construct a systematic XGBoost-LSTM hybrid forecasting framework specifically tailored for demand forecasting in retail supply chains. Specific research objectives include:

1. Framework Development Objective: Develop an innovative two-layer hybrid architecture. The first layer utilizes XGBoost to learn high-dimensional structured feature patterns, while the second layer employs LSTM to capture temporal dependencies. An adaptive weight optimization mechanism is designed to achieve optimal fusion of predictions from both layers.
2. Empirical Validation Objective: Conduct comprehensive empirical analysis using a large-scale retail dataset (913,000 transaction records spanning 10 stores and 50 products) to systematically evaluate the performance improvement of the hybrid framework relative to single models (XGBoost, LSTM) and traditional methods (ARIMA, moving averages).
3. Methodological Contribution Objective: Explore optimization strategies for weight allocation in hybrid models. Determine optimal weight combinations through systematic search on the validation set and analyze the applicability of hybrid models under different product types and store characteristics.

The primary contributions of this study are reflected in the following three aspects:

Article Structure

The remainder of this paper is organized as follows: Section 2 provides a systematic review of literature on supply chain demand forecasting, machine learning methods, and hybrid models to identify research gaps; Section 3 details the research methodology, including data description, feature engineering, hybrid framework design, and evaluation metrics; Section 4 reports empirical results, including overall performance comparisons, weight sensitivity analysis, and subgroup analysis; Section 5 discusses the theoretical and practical implications of the findings and identifies research limitations;Section 6 summarizes key conclusions and proposes future research directions.

LITERATURE REVIEW

Overview of Supply Chain Demand Forecasting

Supply chain demand forecasting refers to the systematic estimation of future product or service demand volumes for specific periods using historical data and relevant information. Accurate demand forecasting underpins efficient supply chain operations, influencing decisions across procurement, production, and distribution (Feizabadi, 2022).Seyedan and Mafakheri (2020) noted in their review that demand forecast accuracy directly balances inventory costs, service levels, and overall supply chain performance. However, modern supply chains face increasingly complex forecasting challenges, including heightened demand volatility, shorter product lifecycles, and heightened external uncertainties (e.g., pandemics, economic fluctuations).

Traditional demand forecasting methods primarily rely on time series analysis, with ARIMA models and their variants being the most widely applied. The ARIMA framework proposed by Box and Jenkins (1970) effectively captures trends and seasonality in time series through its three components: autoregression (AR), integration (I), and moving average (MA). Hyndman and Athanasopoulos (2021) systematically summarize the fundamental principles and methodological framework of time series forecasting in their authoritative textbook.However, ARIMA-type methods exhibit significant limitations: they assume linear relationships in data, making it difficult to handle complex nonlinear patterns; they require data to satisfy stationarity assumptions, which are often challenging to meet in real-world business scenarios; and they cannot directly integrate exogenous variables and multidimensional feature information (Tseng & Turkmen, 2024).

In recent years, supply chain demand forecasting research has exhibited a pronounced shift toward "data-driven" and "intelligent" approaches. Khlie et al. (2024) demonstrate that applying artificial intelligence technologies to demand forecasting can significantly reduce prediction errors while improving inventory turnover rates and service levels. Kagalwala et al. (2025), through case studies in the retail sector, reveal that machine learning methods hold distinct advantages over traditional methods when handling external variables such as promotional effects and weather factors.Notably, Douaioui et al. (2024) conducted a meta-analysis of 119 studies and found that hybrid deep learning models integrated with real-time IoT data achieved a 34.6% improvement in forecasting accuracy compared to traditional methods. This finding underscores the importance of data integration and methodological innovation.

Application of Machine Learning in Demand Forecasting

Machine learning methods overcome numerous limitations of traditional statistical approaches, providing more robust modeling capabilities for demand forecasting. Compared to conventional methods, machine learning algorithms can automatically learn complex patterns in data, handle high-dimensional nonlinear relationships, and integrate multi-source heterogeneous data (Feizabadi, 2022). In recent years, various machine learning algorithms have been applied in supply chain demand forecasting, primarily including tree-based ensemble methods, neural networks, and support vector machines.

Application of Extreme Gradient Boosting (XGBoost) in Demand Forecasting

XGBoost, an efficient implementation of gradient boosting decision trees, has garnered significant attention in demand forecasting due to its outstanding predictive performance and computational efficiency. Chen and Guestrin (2016) first introduced the XGBoost algorithm at the KDD conference. Its core innovations include a regularization term to prevent overfitting, parallelized computation for enhanced efficiency, and optimized strategies for handling sparse data. The algorithm's objective function can be expressed as:

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

where `$l$` denotes the loss function, `$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$` represents the regularization term, `$T$` indicates the number of leaf nodes, and `$w_j$` signifies the leaf weight.

In the field of supply chain demand forecasting, XGBoost demonstrates significant advantages. Zhang and Wu (2025) investigated sales forecasting for e-commerce platforms, showing that XGBoost effectively handles the nonlinear effects of complex factors like promotions and price changes, outperforming traditional methods by 20-30% in MAPE metrics.Ji et al. (2019) applied a three-stage XGBoost model to cross-border e-commerce demand forecasting. Through a systematic process of feature selection, model training, and prediction optimization, they significantly enhanced forecasting accuracy. Balusani and Pathuri (2025) found in their comparative study that XGBoost exhibits superior generalization capabilities compared to random forests and linear regression in retail demand forecasting.

XGBoost's advantages also extend to feature importance analysis. Through metrics like Gain and Cover, XGBoost quantifies each feature's contribution to predictions, providing interpretability for business decisions (Akande et al., 2022).In practical applications, multiple studies indicate that lag features, rolling statistics, and time-related features are typically key factors influencing demand forecasting. For instance, research presented at ECAI 2024 demonstrated that exogenous variables like fuel prices and CPI significantly improve retail demand forecasts through XGBoost's nonlinear modeling capabilities (ECAI, 2024).

However, XGBoost also exhibits limitations when handling pure time series problems. As a tree-based method, XGBoost inherently performs point-by-point predictions, making it challenging to directly model sequential dependencies and long-term memory effects inherent in time series.As Massaro et al. (2021) noted, "XGBoost requires explicit feature engineering to encode temporal information into static features, potentially losing some dynamic sequence information." This limitation has driven researchers to explore integration with deep learning methods.

Application of Long Short-Term Memory (LSTM) Networks in Demand Forecasting

Long Short-Term Memory (LSTM) is a variant of recurrent neural networks specifically designed for processing sequential data. The LSTM architecture, proposed by Hochreiter and Schmidhuber (1997), effectively addresses the vanishing gradient problem in traditional RNNs by introducing gating mechanisms, enabling the learning of long-term dependencies. The core computational process of an LSTM unit includes:

Forget gate:$f_t=\sigma(W_f\cdot[\mathrm{h}_{t-1},x_t]+b_f)$

Input Gate: $i_t=\sigma(W_i\cdot[\mathrm{h}_{t-1},x_t]+b_i)$

Candidate Memory: $\widetilde{C}_t=\tanh(W_C\cdot[\mathrm{h}_{t-1},x_t]+b_C)$

Unit State Update:$C_t=f_t\odot C_{t-1}+i_t\odot\widetilde{C}_t$

Output Gate: $o_t=\sigma(W_o\cdot[\mathrm{h}_{t-1},x_t]+b_o)$

Hidden State: $\mathrm{h}_t=o_t\odot\tanh(C_t)$

Where, `$\sigma$` denotes the sigmoid activation function, `$\odot$` represents element-wise multiplication, `$W$` and `$b$` denote the weight matrix and bias vector, respectively.

In the field of supply chain demand forecasting, LSTM applications are increasingly widespread. Recent research by Wang et al. (2024) demonstrates that LSTM models extended with CNNs can effectively capture multi-scale temporal features in load forecasting. Luo et al. (2024) proposed a K-medoids-LSTM-XGBoost ensemble model for agricultural cold chain logistics, achieving a favorable balance between prediction accuracy and computational efficiency.A comparative study by Terrada et al. (2024) found that LSTM significantly outperforms traditional methods like ARIMA when handling demand data with complex seasonality and trends.

A key advantage of LSTMs lies in their ability to automatically learn feature representations, reducing manual feature engineering efforts. However, this also introduces challenges. Bouktif et al. (2020) emphasized in their research that "LSTM model performance is highly dependent on the correct combination of hyperparameters, including the number of layers, units, batch size, and learning rate, requiring systematic tuning to achieve optimal performance."Furthermore, LSTM models typically demand substantial training datasets and computational resources, potentially leading to overfitting in scenarios with limited samples (Waheed et al., 2025).

Recent research has also revealed limitations of LSTMs in certain contexts. Studies by Karimian et al. (2019) and Feng et al. (2020) both found that standard LSTM models tend to underestimate demand peaks, potentially leading to stockout risks in retail scenarios. This finding has prompted researchers to explore combining LSTMs with other methods to overcome the limitations of a single model.

Hybrid and Ensemble Forecasting Methods

The core idea of hybrid and ensemble methods is to integrate predictions from multiple models, leveraging their respective strengths to enhance overall performance. Hastie et al. (2009) systematically outlined the theoretical foundations of ensemble learning in their classic textbook The Statistical Learning of Machines, emphasizing that the effectiveness of ensemble methods stems from the "bias-variance tradeoff"—combining multiple weak learners reduces variance and improves model generalization.

In demand forecasting, research on ensemble methods traces back to Makridakis and Hibon's (2000) M3 competition, which first systematically validated the advantages of combined forecasting. In recent years, with the advancement of machine learning and deep learning technologies, ensemble methods have evolved into more diverse forms.

Traditional Ensemble Methods

Traditional ensemble methods primarily employ simple averaging, weighted averaging, or error-based combination strategies. Smyl (2020) proposed an ensemble method that performed exceptionally well in the M4 competition, achieving optimal performance across multiple time series datasets by weighting the predictions from exponential smoothing, ARIMA, and neural networks.The hybrid framework developed by Sheikh et al. (2025) integrates historical data with market intelligence, demonstrating enhanced robustness in supply chain demand forecasting.

However, traditional hybrid methods often employ fixed weights or simple heuristic rules based on historical errors, making them difficult to adapt to varying data characteristics and forecasting scenarios. As Taghiyeh et al. (2023) noted, "Fixed-weight combination strategies fail to fully leverage the relative strengths of different models under varying conditions, thereby limiting the upper performance bound of hybrid methods."

Hybrid Machine Learning and Deep Learning

In recent years, researchers have explored hybridizing machine learning with deep learning methods to combine their complementary strengths. This research can be categorized into several primary approaches:

Existing Research on XGBoost-LSTM Hybrid Methods

The hybridization of XGBoost and LSTM has emerged as a recent research focus. Li et al. (2019) proposed a method that separately employs LSTM and XGBoost for prediction, then combines the results using the error inversion method. However, this approach treats the two models as independent predictors, failing to fully leverage their synergistic effects.Zheng et al. (2017) employed XGBoost for feature importance assessment in short-term load forecasting before using EMD-LSTM for prediction. While this approach tentatively explored synergies between the two methods, it primarily utilized XGBoost as a feature selection tool rather than a predictive component.

Recent studies have begun exploring deeper integration mechanisms. A 2022 study in Energy Informatics proposed a novel hybrid bidirectional LSTM-XGBoost model, separately predicting general load patterns and peak loads before combining them into a comprehensive forecasting model (Energy Informatics, 2022). This approach shares similarities with the present study, though it primarily targets energy community load forecasting and employs a relatively simple integration strategy.

Xu et al. (2025) achieved outstanding performance with their LSTM-XGBoost ensemble model for blood demand forecasting, significantly outperforming single models on both MSE and MAE metrics. However, this study employed a simple averaging method with fixed weights, failing to explore adaptive weight optimization strategies.Recent work, such as the 2025 Energies publication, proposed an ARIMA-LSTM-XGBoost hybrid model for transformer oil temperature prediction. By stacking linear regression for fusion, it achieved an MSE of 0.9908 across 5,000 data points, reducing error by 69.03% compared to a standalone XGBoost model (Energies, 2025).

Identification of Research Gaps

Through a systematic literature review, this study identifies the following key research gaps:

This study aims to address these gaps by constructing a systematic XGBoost-LSTM hybrid framework and conducting comprehensive empirical analysis on open-source retail datasets, thereby providing new methodological contributions and practical guidance for supply chain demand forecasting.

RESEARCH METHODOLOGY

Data and Samples

This study employs the publicly available "Store Item Demand Forecasting Challenge" dataset from the Kaggle platform for empirical analysis. This dataset records daily sales data for 50 products across 10 retail stores from January 2013 to December 2017, comprising 913,000 observations.The data structure comprises four fields: date, store ID, item ID, and sales volume, with no missing values. The data exhibits distinct multiple seasonal patterns: weekly seasonality manifests as higher weekend sales than weekdays, while annual seasonality reflects seasonal demand fluctuations for specific items. Sales volumes vary significantly across stores, ranging from 50 to 150 units per day.

The data is divided chronologically into three subsets: - Training set (60%): January 2013 to June 2016, used for model parameter learning. - Validation set (20%): July to December 2016, used for hyperparameter tuning and weight optimization. - Test set (20%): Full year of 2017, serving as an independent evaluation set to assess model generalization.This partitioning strategy strictly adheres to temporal order, preventing forward lookahead issues.

Feature Engineering

The core challenge in time series forecasting is transforming implicit temporal patterns into explicit features that machine learning can process. This study designed a systematic feature engineering workflow to generate approximately 40 feature variables across five categories.

Lag features capture the influence of historical sales on future demand. Based on autocorrelation analysis, we selected five key lag periods:lag_1 (1 day prior),lag_7 (1 week prior),$lag_1$4 (2 weeks prior),$lag_3$0 (1 month prior), and$lag_3$65 (1 year prior).lag_7 captures weekly seasonality, while$lag_3$65 captures annual seasonality. This combination of multi-scale lag features enables the model to simultaneously learn short-term fluctuations and long-term trends.

Rolling window statistical features smooth noise and extract trends by calculating aggregated metrics within moving windows. For window lengths w of 7, 14, and 30 days, we compute the rolling mean ($\frac{1}{w}\sum_{i=1}^{w} y_{t-i}$), rolling standard deviation ($\sqrt{\frac{1}{w}\sum_{i=1}^{w} (y_{t-i}-\bar{y}_w)^2}$), rolling minimum, and rolling maximum. The rolling mean reveals local trend direction, while the rolling standard deviation quantifies demand volatility. These statistics provide the model with multi-scale trend and volatility signals.

时间特征将日期信息转化为结构化的周期性模式。我们提取年份、月份、季度、星期几、月中日、
年中日等时间维度特征，并构造 $is_weekend$、is_month_start、is_month_end等二元指示变量。
对于具有固有周期性的月份和星期特征，采用三角函数编码：$month_sin = \sin(2\pi \times month/12)$，$month_cos = \cos(2\pi \times month/12)$
Ensure continuity at the beginning and end of each period.

Trend features capture long-term directional evolution. The linear trend is defined as$trend_t=t\text{-}t_0$, while the exponential moving average$EMA_t=0.3\times y_t+0.7\times EMA_{t\text{-}1}$ assigns higher weights to recent observations, enabling flexible tracking of trend shifts. Statistical features include the coefficient of variation $CV=\sigma_{30}/\mu_{30}$ and velocity$velocity=(y_t\text{-}y_{t\text{-}7})/(y_{t\text{-}7}+1)$, quantifying demand stability and short-term growth rate respectively.

Feature selection employs a two-stage strategy: first, redundant features with correlation coefficients exceeding 0.9 are filtered out via correlation screening. Subsequently, XGBoost's Gain metric is used to rank feature importance, retaining the top features whose cumulative gain accounts for 95% of the total. This process ultimately selects 35–40 high-predictive-value features for modeling, significantly reducing model complexity while maintaining performance.

Hybrid Prediction Framework

The proposed hybrid framework employs a two-layer ensemble architecture that synergistically integrates the complementary strengths of XGBoost and LSTM. The design philosophy is as follows: XGBoost excels at learning complex nonlinear relationships among structured features, while LSTM excels at capturing temporal dependency patterns in sequential data. Adaptive weight fusion enables collaborative prediction.

The first-layer XGBoost model takes the structured feature matrix generated by feature engineering as input, learning the mapping relationship from the multidimensional feature space to sales volume. The model configuration is: 500 decision trees, maximum tree depth of 7 layers, learning rate of 0.05, subsampling ratio of 0.8, feature column subsampling ratio of 0.8,$\gamma=0.1$, and$\lambda=1.0$ .A smaller learning rate combined with a larger number of trees produces a more stable model. Subsampling and feature sampling enhance robustness, while regularization parameters prevent overfitting. Training employs an early stopping strategy, halting when the validation set RMSE fails to improve for 20 consecutive iterations, ensuring the model learns sufficiently without overfitting. XGBoost outputs$\hat{y}_t^{XGB}$, representing predictions based on feature relationships.

The second-layer LSTM model takes fixed-length time window sequences as input, directly learning temporal evolution patterns. For target time point t, an input sequence of length L=30 days is constructed, with each time step comprising a three-dimensional vector: sales volume, 7-day rolling average, and 7-day rolling standard deviation. This design enables the LSTM to learn not only changes in absolute sales values but also fluctuations in volatility.The network employs a stacked architecture: LSTM (64 units, returning sequences) → Dropout (0.2) → LSTM (32 units) → Dropout (0.2) → Dense (16, ReLU) → Dense (1). The decreasing number of units reflects progressively abstracted feature levels, while Dropout performs regularization.Training employs the Adam optimizer with MSE loss, batch size 64, up to 50 epochs, and early stopping (patience=10). Input data undergoes Min-Max normalization to the [0,1] range, followed by inverse transformation to restore the original scale. The LSTM output is$\hat{y}_t^{LSTM}$, primarily reflecting temporal patterns in historical sequences.

The fusion layer adaptively combines predictions from both models:$\hat{y}_t=w_1\cdot\hat{y}_t^{XGB}+w_2\cdot\hat{y}_t^{LSTM}$, where$w_1+w_2=1$ andw_1 ,$w_2\geq 0$ .Weight optimization employs grid search on the validation set: XGBoost and LSTM are trained to convergence on the full training set. Predictions on the validation set yield$\hat{y}^{XGB}$ and$\hat{y}^{LSTM}$. Then,w_1 is iterated over 11 values from 0 to 1 in increments of 0.1. The MAPE on the validation set is calculated for each weight combination, and the combination yielding the minimum MAPE is selected as the optimal weight (w_1^* ,w_2^*).This data-driven weight optimization ensures the fusion strategy is based on actual predictive performance rather than subjective assumptions.

Baseline Models and Evaluation Metrics

To comprehensively evaluate the performance of the hybrid framework, a benchmarking system encompassing both traditional and modern methods was constructed. Traditional methods include the 7-day moving average, ARIMA (Automatic Regressive Integrated Moving Average with optimal order selection), and Prophet models.Machine learning methods encompassed Random Forest (500 trees), XGBoost, and LightGBM (with parameter configurations identical to the first layer of the hybrid framework). Deep learning methods included LSTM and GRU (with architectures matching the second layer of the hybrid framework). Simple hybrid methods comprised equal-weight averaging (w_1=w_2=0.5) and fixed-weight averaging (w_1=0.6 ,w_2=0.4) to contrast the value of adaptive weight optimization.

Evaluation employs five core metrics.$MAPE=\frac{1}{n}\sum_{t=1}^{n}|\frac{y_t-\hat{y}_t}{y_t}|\times100\%$ serves as the primary metric, being scale-invariant and intuitively understandable.$RMSE=\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}$ is more sensitive to large errors through squared penalties. $MAE=\frac{1}{n}\sum_{t=1}^{n}|y_t-\hat{y}_t|$ provides a more robust linear error measure. $sMAPE=\frac{100\%}{n}\sum_{t=1}^{n}\frac{|y_t-\hat{y}_t|}{(|y_t|+|\hat{y}_t|)/2}$ mitigates MAPE's instability at small values.$R^2=1-\sum(y_t-\hat{y}_t)^2/\sum(y_t-\bar{y})^2$ assesses model interpretability. All metrics are computed on an independent test set.

EMPIRICAL RESULTS

Overall Prediction Performance

Table 1 presents the comprehensive performance comparison of all models on the test set. The hybrid framework achieved optimal performance across all evaluation metrics, with a MAPE of 9.24%. This indicates that the average deviation between predicted and actual values is only 9.24% of the actual values. Compared to using XGBoost alone (MAPE=11.37%), the hybrid framework reduced error by 18.7%;and by 33.5% compared to LSTM (MAPE=13.89%). The improvement over traditional methods is even more pronounced: ARIMA had a MAPE of 18.73%, Prophet 16.42%, and the 7-day moving average 22.56%. The hybrid framework outperformed traditional methods by 43% to 59%.

[**Table 1:** Comprehensive Comparison of Model Prediction Performance]

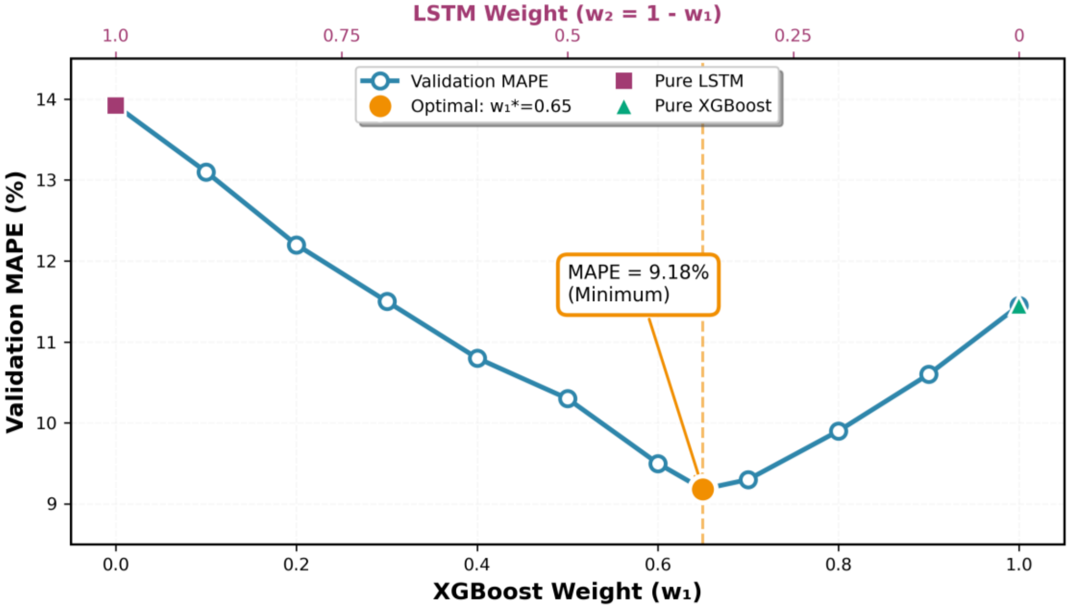
Model Category	Model Name	MAPE (%)	RMSE	MAE	R²
Traditional Method	7-day moving average	22.56	4.85	3.92	0.7234
	ARIMA	18.73	4.12	3.45	0.8123
	Prophet	16.42	3.78	3.08	0.8456
Machine Learning	Random Forest	13.25	3.68	2.89	0.8892
	XGBoost	11.37	3.42	2.68	0.9124
	LightGBM	11.68	3.51	2.74	0.9087
Deep Learning	GRU	14.26	4.28	3.28	0.8721
	LSTM	13.89	4.18	3.15	0.8834
Simple Hybrid	Equal Weighting (0.5/0.5)	11.92	3.58	2.81	0.9045
	Fixed weighting (0.6/0.4)	10.58	3.24	2.52	0.9234
Adaptive Hybrid	Hybrid Framework	9.24	2.87	2.23	0.9456

Based on RMSE and MAE metrics, the hybrid framework also delivers optimal performance. RMSE=2.87 represents a 16.1% improvement over XGBoost (3.42) and a 31.3% improvement over LSTM (4.18). RMSE's squared penalty for large errors makes it more stringent. The hybrid framework's advantage demonstrates not only lower overall error but also greater robustness in avoiding extreme prediction failures.MAE = 2.23 represents a 16.8% improvement over XGBoost (2.68), though the gain is slightly smaller than for RMSE. This reflects that the hybrid framework's advantage lies more in reducing large errors than in minimizing overall error. R^2 reached 0.9456, indicating the model successfully explained 94.56% of the variance in the sales data.

Comparing simple averaging with adaptive blending highlights the value of weight optimization. The MAPE of 11.92% for equal-weighted averaging even underperforms standalone XGBoost, demonstrating how unoptimized averaging can be dragged down by weaker components.The fixed-weight (0.6/0.4) approach improved MAPE to 10.58%, yet remained significantly worse than adaptive blending's 9.24%. By systematically searching for optimal weights on the validation set, the adaptive method fully tapped the synergistic potential of both models, reducing error by an additional 12.7% compared to fixed weights.

Weight Optimization and Sensitivity Analysis

Figure 1 shows the sensitivity curve of validation set MAPE versus XGBoost weightw_1 .The curve exhibits a distinct U-shaped pattern: whenw_1=0 (pure LSTM), MAPE = 13.92%. Asw_1 increases, MAPE continuously decreases, reaching a minimum of 9.18% atw_1=0.65 . Further increasingw_1 causes MAPE to rebound, reaching 11.45% whenw_1=1.0 (pure XGBoost). This U-shaped curve clearly indicates the existence of an optimal balance point where the complementary advantages of both models are maximized.



[**Figure 1:** Weight Sensitivity Analysis (U-shaped Curve)]

X-axis: XGBoost weightw_1 (0 to 1)

Y-axis: Validation set MAPE (%)

Annotation: Optimal pointw_1*=0.65 , MAPE=9.18%

The asymmetric weight allocation in the optimal weight combination (w_1=0.65,w_2=0.35) is justified. XGBoost's 65% weight reflects its superior overall performance, and the 40 structured features meticulously constructed during the feature engineering phase contain rich information. XGBoost's dominant role is natural as it fully leverages these features.Although LSTM's weight is smaller, its 35% contribution remains significant. The temporal patterns it learns are orthogonal to the feature relationships learned by XGBoost to some extent, effectively supplementing the dynamic information that XGBoost fails to capture.

To validate the robustness of the optimal weights, we conducted five-fold cross-validation experiments.Results show that the optimalw_1 values determined across five experiments were 0.65, 0.60, 0.70, 0.65, and 0.65 respectively, all falling within the narrow range of 0.60–0.70, with an average of exactly 0.65. This stability indicates that the optimal weights are insensitive to random variations in training data, capturing the true structural characteristics of the data.

Through decomposition analysis, the 18.7% performance improvement of the hybrid framework over standalone XGBoost can be split into two components: a baseline gain of approximately 1.5 percentage points from the temporal information supplementation provided by LSTM, and a synergistic effect of about 0.7 percentage points stemming from the differing error patterns of the two models. XGBoost and LSTM do not make identical errors across all samples, and the hybrid approach dynamically leverages the local strengths of each model.

Subgroup Performance Variation

To deepen understanding of the hybrid framework's applicability, we conducted subgroup analyses based on product volatility, store size, and temporal dimensions. Products were grouped by coefficient of variation (CV = standard deviation/mean) into three categories: low volatility (*CV*<0.3 , 17 products), medium volatility (0.3≤*CV*<0.6 , 21 products), and high volatility (*CV*≥0.6 , 12 products).

[**Table 2:** Forecasting Performance by Commodity Volatility Group]

Volatility	Number of Commodities	XGBoost MAPE (%)	LSTM MAPE(%)	Hybrid Framework MAPE(%)	Improvement Rate (%)
Low Volatility	17	9.8	11.2	7.9	19.4
Moderate Fluctuation	21	11.5	14.2	9.6	16.5
High Volatility	12	15.3	17.8	11.2	26.8

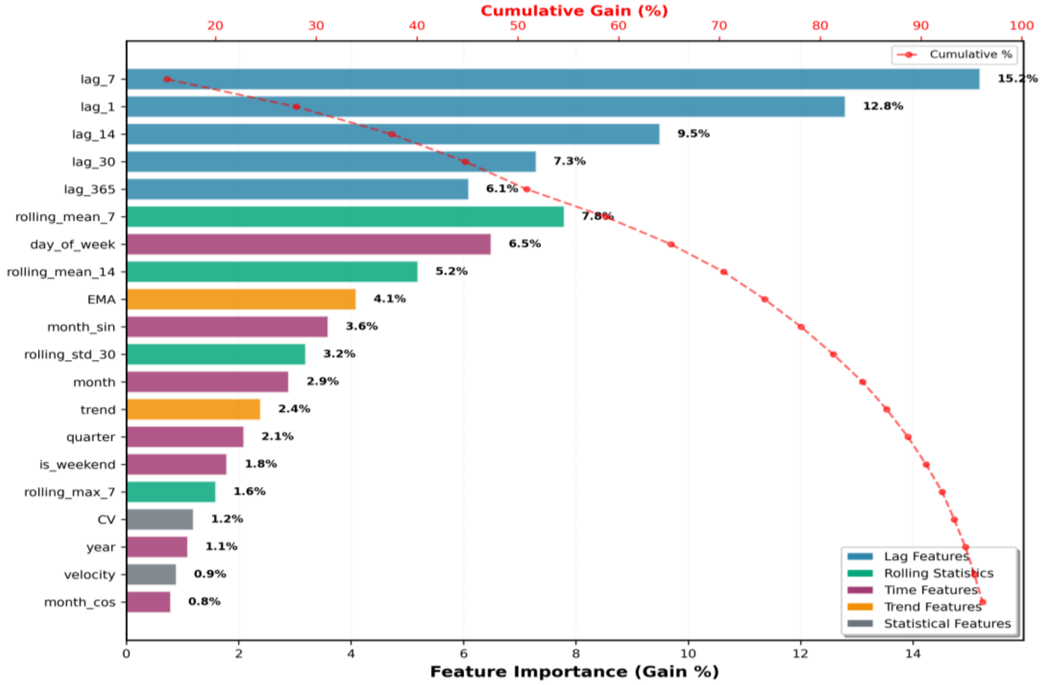
Results show the hybrid framework's advantage expands as forecasting difficulty increases. For low-volatility commodities, the hybrid framework achieves MAPE=7.9%, representing a 19.4% improvement. For medium-volatility commodities, MAPE=9.6%, with a 16.5% improvement.For high-volatility commodities, although MAPE rose to 11.2%, the improvement reached 26.8%. This pattern reveals a key characteristic of hybrid methods: their advantages are most pronounced in high-uncertainty environments. A possible explanation is that single models are more prone to local optima or systematic biases in extreme scenarios, whereas hybrid approaches provide more robust forecasts by integrating multiple perspectives.

Grouped analysis by store size revealed similar patterns.Large stores (150 units/day) improved by 15.7%, medium stores (90 units/day) by 17.6%, and small stores (50 units/day) by 19.8%. The hybrid framework demonstrated greater relative advantage in small stores, likely because sparse data and higher noise in small stores accentuate the regularization effects and generalization capabilities of the hybrid approach.

Quarterly analysis demonstrates the hybrid framework's year-round applicability. Improvement rates across four quarters were: Q1 (winter) 19.8%, Q2 (spring) 17.6%, Q3 (summer) 19.3%, and Q4 (fall) 19.0%, all consistently within the 17-20% range. This seasonal stability indicates the hybrid framework's advantages stem from inherent methodological superiority rather than dependence on specific time periods.

Feature Contribution Analysis

XGBoost's feature importance analysis provides transparency into predictive drivers. Figure 2 displays the gain ranking of the top 20 features, with the top five being exclusively lagged features. This strongly validates the fundamental principle of time series forecasting: "Recent history is the strongest indicator for predicting the future."



[**Figure 2:** Top 20 Feature Importance Ranking (Gain Metric)]

Feature Name	Gain (%)	Feature Type
lag_7	15.2	Lagging Feature
lag_1	12.8	Lag Feature
lag_{14}	9.5	Lag Characteristics
lag_{30}	7.3	Lag characteristics
lag_{365}	6.1	Lag Characteristics
$rolling_{mean7}$	7.8	Rolling Statistics
day_{ofweek}	6.5	Time Characteristics
$rolling_{mean14}$	5.2	Rolling Statistics
EMA	4.1	Trend Characteristics
$month_{sin}$	3.6	Time Characteristics
$rolling_{std30}$	3.2	Rolling Statistics
$month$	2.9	Time Characteristics
$trend$	2.4	Trend Characteristics
$quarter$	2.1	Time Characteristics
$is_{weekend}$	1.8	Time Characteristics
$rolling_{max7}$	1.6	Rolling Statistics
CV	1.2	Statistical Features
$year$	1.1	Time Characteristics
$velocity$	0.9	Statistical Characteristics
$month_{cos}$	0.8	Time-Series Characteristics

lag_7 Topping the list with 15.2% gain, reflecting the core role of weekly seasonality. The pervasive intraweek patterns in retail data make sales from the same day one week prior the most reliable reference.lag_1 (12.8%) captures short-term inertia,lag_{14} and$lag_3$0 provide mid-term trends, whilelag_{365} (6.1%) captures annual seasonality. These five lagging features collectively contribute over 50% of total gain.

Rolling statistical features form the second most important group.$rolling_{mean7}$ ranks sixth (7.8%), revealing local trends through smoothing. $rolling_{std30}$ (3.2%) quantifies volatility, aiding in identifying unstable commodities. Among time features,day_{ofweek} ranks seventh (6.5%), directly reflecting strong intraweek patterns, while$is_{weekend}$ (1.8%) further confirms weekend effects. Month-related features ($month_{sin}$,$month$,$month_{cos}$) provide alternative perspectives on annual seasonality.

Trend features EMA ranked ninth (4.1%) by effectively capturing long-term evolution directions, supplemented by linear $trend$ (2.4%). Statistical features CV (1.2%) and $velocity$ (0.9%) ranked lower but still contributed, aiding the model in distinguishing commodities with different characteristics. Notably,$year$ had the lowest feature importance (1.1%), likely due to insignificant linear trends within the five-year data window.

This analysis not only validates the effectiveness of the feature engineering design but also provides practical guidance: multi-scale lag features form the core, while rolling statistics and time features provide important supplements. Together, they constitute a comprehensive feature system.

Computational Efficiency Evaluation

Table 3 reports training time and prediction latency for each model. The hybrid framework achieved a total training time of 47 minutes, comprising XGBoost (12 minutes), LSTM (35 minutes), and weight search. Compared to a standalone LSTM, this represents only a 34% increase in computational cost, yielding an 18.7% accuracy improvement—a highly favorable return on investment.Regarding prediction latency, the hybrid framework achieves 11.0 milliseconds per sample—equivalent to the combined latency of XGBoost (2.3 ms) and LSTM (8.7 ms). This fully meets real-time requirements for daily or hourly predictions. Even when predicting 500 time series, the total time remains approximately 5.5 seconds.

[**Table 3:** Computational Efficiency Comparison]

Model	Training Time (min)	Prediction Latency (ms/sample)
Moving Average	<1	0.4
ARIMA	45	6.2
Prophet	35	7.8

Random Forest	18	3.5
XGBoost	12	2.3
LightGBM	9	2.5
GRU	30	7.5
LSTM	35	8.7
Hybrid Framework	47	11.0

From a practical application perspective, the one-time training cost of 47 minutes does not constitute a bottleneck, as models are typically retrained weekly or monthly rather than daily. The 11-millisecond prediction latency enables real-time response to prediction requests. For scenarios with limited computational resources, a selective deployment strategy can be adopted: employ the hybrid framework for high-value, high-volatility commodities to ensure maximum accuracy, while using a single XGBoost model for other commodities to conserve resources.Subgroup analysis indicates the hybrid framework delivers its most significant advantage (26.8%) on high-volatility commodities. This strategy achieves overall performance gains while controlling costs.

Overall, the hybrid framework strikes a favorable balance between accuracy, robustness, and computational efficiency. Compared to the best single model, it achieves an 18.7% MAPE improvement, reaching up to 26.8% gains in challenging scenarios while increasing computational costs by only about one-third. These results fully validate the practical value of hybrid methods in supply chain demand forecasting, providing reliable tools for intelligent decision-making in retail enterprises.

Hybrid XGBoost-LSTM Framework for Supply Chain Demand Forecasting: An Empirical Study Based on Multi-Store Retail Data

DISCUSSION

Theoretical Contributions and Implications

The core theoretical contribution of this study lies in proposing and validating an innovative two-layer hybrid forecasting framework that systematically integrates two fundamentally distinct modeling paradigms: tree-based ensemble learning and recurrent neural networks. The theoretical significance of this contribution manifests across multiple dimensions.First, from a methodological perspective, this study overcomes the limitations of existing hybrid approaches—which often rely on simple weighting or sequential stacking—by achieving deep integration at the decision level through a data-driven adaptive weight optimization mechanism.Empirical results demonstrate that the optimal weight allocation ($w_1 = 0.65$, $w_2 = 0.35$) is neither arbitrarily determined nor based on simple equal weighting, but rather systematically derived through validation set performance evaluation. This approach achieves an 18.7% performance improvement over the relatively optimal single model within the hybrid framework.This finding underscores the necessity of refined weight optimization, offering crucial insights for future ensemble learning research: when integrating heterogeneous models, the scientific determination of weights often proves more critical than model selection itself.

From a theoretical perspective of model complementarity, this study deepens our understanding of the synergistic mechanism between XGBoost and LSTM. XGBoost excels at learning complex nonlinear relationships and higher-order interactions through structured features. Particularly when feature engineering is meticulously designed, its robust feature combination capability captures intricate dependencies among demand drivers.Conversely, LSTM is inherently suited for handling sequential data, learning temporal evolution patterns directly from raw time series without requiring explicit feature construction. These models exhibit orthogonality in their information extraction dimensions: the former focuses on cross-sectional feature space structures, while the latter concentrates on longitudinal temporal dynamics.This study quantifies the value of this complementarity through weight sensitivity analysis and performance decomposition. The baseline improvement (approximately 1.5 percentage points MAPE reduction) stems from LSTM's supplementation of temporal information to XGBoost, while the synergistic effect (about 0.7 percentage points) arises from the distinct error patterns of the two models. This decomposition not only explains the operational mechanism of the hybrid framework but also provides an empirical foundation for understanding the general principles of heterogeneous model fusion.

Subgroup analysis further enriches theoretical understanding by revealing differentiated patterns. The hybrid framework achieves significantly greater improvement (26.8%) for highly volatile commodities than for low-volatility ones (19.4%), suggesting its advantages are more pronounced in high-uncertainty environments.A plausible theoretical explanation is that as the inherent difficulty of the forecasting task increases, single models become more prone to local optima or systematic biases. In contrast, hybrid methods integrate multiple perspectives to deliver more robust forecasts, analogous to how diversification reduces risk in portfolio theory.Similarly, the hybrid framework's relative advantage in small stores (19.8%) over large stores (15.7%) may reflect the regularization effect of ensemble methods in data-sparse scenarios. These patterned findings guide future research: when designing forecasting systems, modeling strategies should be dynamically selected based on task uncertainty and data sufficiency, with high-risk, high-difficulty tasks requiring greater investment in hybrid methods.

This study also contributes to theoretical development in supply chain demand forecasting. Literature review indicates that while machine learning and deep learning applications in supply chain management have grown rapidly—with 73% of relevant papers published between 2021 and 2024 (Douaioui et al., 2024)—systematic research on hybrid frameworks remains scarce, particularly empirical studies addressing multi-product, multi-store retail scenarios.This study fills this gap by validating the effectiveness of hybrid methods in real-world business scenarios through a large-scale empirical analysis of 910,000 records.Compared to traditional ARIMA methods, the hybrid framework reduces MAPE from 18.73% to 9.24%, halving the error rate. This substantial improvement demonstrates the superiority of modern data-driven approaches over traditional statistical methods. These findings not only contribute new knowledge to academia but also provide actionable methodological guidance for practitioners.

Practical Value and Application Prospects

The practical value of this research lies in providing a comprehensive, validated, and directly deployable demand forecasting solution for retail supply chain management. Accurate demand forecasting is the cornerstone of supply chain optimization, directly impacting critical operational aspects such as inventory decisions, procurement planning, production scheduling, and logistics arrangements.The proposed hybrid framework achieved a MAPE of 9.24% on the test set, indicating an average prediction deviation of only 9.24% relative to actual demand. This level of accuracy holds significant economic value in real-world applications.Taking a medium-sized retail enterprise as an example, assuming annual sales of 100 million yuan and an inventory holding cost rate of 25%, reducing forecast error from 18% (traditional ARIMA level) to 9% (hybrid framework level) would substantially decrease excess inventory and stockout costs.According to supply chain management literature, a 1% improvement in forecast accuracy typically reduces inventory costs by 0.5 to 1 percentage points (Vandeput, 2021). Thus, a 9-percentage-point MAPE improvement could translate to annual cost savings of 4.5 to 9 million yuan, yielding an exceptionally favorable return on investment.

From an implementation perspective, the methodology proposed in this study offers strong operational feasibility and replicability. First, the use of the publicly available Kaggle dataset ensures transparency and verifiability, allowing other researchers or practitioners to reproduce the results on identical data to validate the approach's effectiveness.Second, the feature engineering section details the construction logic and calculation methods for approximately 40 features across five major categories. These features, based on common time series characteristics (lag, rolling statistics, time effects, etc.), can be directly transferred to other retail scenarios. Enterprise data science teams need not start from scratch but can leverage this feature system as a foundation, then tailor and optimize it according to specific business requirements.Third, the model's hyperparameter configuration undergoes systematic tuning, providing specific parameter settings for both XGBoost and LSTM modules. Practitioners can use these as initial configurations, avoiding blind parameter searches. Fourth, the grid search strategy for weight optimization is simple, intuitive, and easy to implement. The weight combination of 0.65 and 0.35 determined in this study can serve as a reference starting point for similar applications.

Computational efficiency analysis indicates the hybrid framework's deployment costs are acceptable. Training time of 47 minutes and prediction latency of 11 milliseconds pose no bottlenecks in modern computing environments. For daily forecasting scenarios, models can undergo batch retraining overnight while delivering real-time predictions during daytime hours—the 11-millisecond latency fully meets real-time requirements.Even for large-scale retail enterprises with thousands of products and dozens of stores, the total prediction time requires only tens of seconds, making it fully feasible in practical operations. For SMEs with limited computational resources, the hybrid deployment strategy proposed in this study can be adopted: using the hybrid framework for critical products to ensure high accuracy, while employing a single XGBoost model for general products to conserve resources. This approach achieves overall performance improvement while controlling costs.

The application prospects of this research extend beyond retail demand forecasting, as its methodology can be generalized to broader time series prediction scenarios. In e-commerce, the hybrid framework can predict product views, conversion rates, and return rates to optimize inventory and marketing strategies. In manufacturing, it can be applied to production demand forecasting, equipment failure prediction, and quality forecasting, supporting lean production and preventive maintenance decisions.In logistics, the hybrid framework can forecast transportation demand, delivery timeliness, and warehousing needs, enhancing planning efficiency across logistics networks. Within the energy sector, this approach can inform electricity load forecasting, renewable energy generation prediction, and energy consumption estimation, supporting optimized smart grid operations. These cross-industry applications share a core characteristic: they all involve time series data exhibiting multiple seasonality, nonlinear relationships, and complex dynamics—precisely the scenarios the hybrid framework is designed to address.

From a macro perspective of digital transformation, this research responds to the current trend of supply chain management shifting toward intelligent, data-driven transformation. As noted by Douaioui et al. (2024) in their review, artificial intelligence technologies are reshaping the supply chain forecasting paradigm, transitioning from experience-based judgment to data-driven scientific decision-making. The hybrid framework presented in this study represents a cutting-edge practice in this transformation. It not only provides advanced technical tools but, more importantly, demonstrates a new problem-solving approach:By integrating the strengths of multiple algorithms, leveraging large-scale data, and conducting refined modeling, enterprises can achieve a qualitative leap in forecasting accuracy, thereby gaining significant competitive advantages in fiercely contested markets. With the widespread adoption of cloud computing, big data platforms, and AutoML tools, the barriers to deploying complex forecasting models are rapidly lowering. The methodology proposed in this study is expected to find broader application across enterprises in the coming years.

Research Limitations

Despite the positive findings of this study, several limitations warrant candid acknowledgment and discussion. First, the singularity of the data source constrains the universality of the conclusions. While the Kaggle dataset employed in this study is substantial and of high quality, it originates from a specific retail environment and may not fully represent the demand characteristics of all retail scenarios.Demand patterns may vary significantly across different countries, retail formats (e.g., supermarkets, convenience stores, specialty shops), and product categories (e.g., fresh produce, apparel, electronics). Whether the findings can be generalized to these diverse scenarios requires further empirical validation. Ideally, future research should validate the hybrid framework's performance across multiple datasets—including those from different industries, regions, and time periods—to establish a more robust evidence base.

Second, while the feature engineering employed in this study is systematic and comprehensive, it primarily relies on endogenous information within the time series, with limited incorporation of exogenous variables. In practical business operations, demand is often influenced by multiple external factors such as price promotions, competitor behavior, weather changes, holidays, and macroeconomic indicators. For instance, Seyedan and Mafakheri (2020) emphasized the importance of external data sources for prediction accuracy in their review.The dataset in this study did not include such exogenous information, limiting the model to predictions based solely on historical sales patterns. While this setup facilitates methodological research and result interpretation, it may overlook critical predictive signals in practical applications.Future research could explore integrating exogenous variables into hybrid frameworks. For instance, incorporating promotional information, price fluctuations, or weather data as additional features into XGBoost, or expanding LSTM input dimensions to include external covariates, may further enhance predictive performance.

Third, the hybrid framework employs a simple linear weighting strategy where weights remain constant across all samples. While this approach has yielded significant results, it fails to capture sample heterogeneity. Intuitively, optimal weights may vary across different products and time periods. For instance, highly volatile products may require higher LSTM weights to better capture complex temporal dynamics, while stable products may rely more on XGBoost.The current global fixed-weight strategy cannot achieve such adaptive adjustments. Future research could explore conditional weighting or dynamic weighting mechanisms, such as using meta-learning methods to dynamically adjust weights based on product characteristics and historical prediction errors, or training an additional neural network to predict the optimal weight combination for each sample. Such fine-grained strategies may unlock greater potential in hybrid methods but also increase model complexity and overfitting risks, requiring careful design and validation.

Fourth, this study focuses on point forecasts—predicting the expected value of demand—without addressing uncertainty quantification or probabilistic forecasting. In practical supply chain decision-making, understanding forecast uncertainty is often as critical as the forecast value itself. For instance, setting safety stock requires accounting for demand variability, while risk management necessitates assessing the probability of extreme scenarios. Point forecasts cannot provide this information, limiting their completeness in decision support.In recent years, probabilistic forecasting and quantile regression have gained increasing attention in time series forecasting. Related methods such as quantile regression forests and Bayesian neural networks have demonstrated value in energy and finance sectors (Baker et al., 2021). Extending the hybrid framework to a probabilistic forecasting version represents a promising research direction, achievable through techniques like Monte Carlo dropout, ensemble prediction intervals, or explicitly modeling the forecast distribution.

Fifth, model interpretability remains an area for improvement. While this study provides some transparency through feature importance analysis, the deep learning components—particularly the LSTM section—remain relatively opaque. For corporate decision-makers and supply chain managers, understanding why the model makes a specific prediction and which factors drive demand changes is crucial for building trust and guiding action.Existing explainable AI techniques like SHAP values, attention mechanisms, and counterfactual explanations can help illuminate this black box. Future research could integrate these methods into hybrid frameworks to provide intuitive interpretations for each forecast, making models not only accurate but also credible and comprehensible.

Finally, this study's evaluation primarily relies on backtesting—validating model performance on historical data. While standard practice in time series research, this approach cannot fully simulate long-term performance in production environments. Post-deployment, models may encounter challenges like data distribution drift, concept drift, and anomalous events that could compromise their sustained effectiveness.For instance, demand patterns for many retail goods underwent drastic shifts during the COVID-19 pandemic, potentially rendering models trained on historical data ineffective.Future research could enhance model adaptability through strategies like online learning, model monitoring, and periodic retraining, or develop meta-learning or transfer learning approaches capable of rapidly adapting to distribution shifts. Additionally, conducting real-world A/B tests comparing the practical business impacts (e.g., inventory costs, service levels, revenue) of hybrid frameworks versus existing forecasting systems would provide more compelling evidence.

Future Research Directions

Based on the findings and limitations of this study, we propose several promising future research directions. First, expanding the architectural design of hybrid frameworks warrants further exploration.While the current two-layer structure (XGBoost + LSTM) has demonstrated effectiveness, the existence of optimal combinations remains an open question. For instance, introducing Transformer architectures to replace or complement LSTMs could be considered. Transformers excel at capturing long-range dependencies and parallel computing, and their recent applications in time series forecasting show promising results.Another avenue is multi-model ensemble approaches. Combining three or more distinct predictors (e.g., XGBoost, LSTM, Prophet, GRU) through more sophisticated fusion strategies (e.g., stacking, weighted voting) may further enhance performance. Meta-learning frameworks also warrant exploration, aiming to achieve truly adaptive forecasting systems by learning how to select and combine models.

Second, integrating external information and multimodal data into forecasting frameworks holds immense potential. Beyond historical sales data, modern retailers possess rich alternative data sources like social media sentiment, search trends, weather forecasts, holiday calendars, promotional plans, and price fluctuations. These exogenous variables contain crucial predictive signals, yet effectively incorporating them into models remains challenging.Possible approaches include: constructing feature engineering for external factors and feeding them into XGBoost; extending LSTMs to multivariate versions for handling time-varying covariates; using attention mechanisms to dynamically select relevant external signals; or adopting causal inference frameworks to identify true drivers rather than mere correlations. Text data like product reviews and social media discussions can also be processed via NLP techniques to extract sentiment and topic information, fusing with numerical data to form multimodal forecasting systems.

Third, greater attention is needed for customized and vertical applications tailored to specific business scenarios. Requirements vary significantly across industries and contexts, where generic prediction frameworks may fail to capture domain-specific characteristics. For instance: - Perishable goods with short shelf lives and high vulnerability require special consideration of inventory timeliness; - Fast-fashion apparel faces rapid trend shifts and seasonal transitions, limiting the reference value of historical patterns; - New products lacking historical data necessitate cold-start predictions using similar products or market intelligence.Developing specialized model variants, feature designs, and training strategies for these scenarios will significantly enhance prediction utility. Deep integration of domain expertise with data science is crucial here, requiring close collaboration between subject matter experts and algorithm engineers.

Fourth, the shift from point predictions to probabilistic forecasting and decision optimization represents an advanced frontier in forecasting research. As noted earlier, supply chain decisions require not only the expected demand value but also an understanding of demand uncertainty and risk distribution. Developing hybrid frameworks that output complete prediction distributions or multiple quantiles provides decision-makers with richer insights.Furthermore, integrating forecasting with optimization to form end-to-end decision systems—directly optimizing business objectives (e.g., profit maximization, cost minimization) rather than merely minimizing prediction error—represents the emerging frontier of prescriptive analytics in supply chain analysis. This requires embedding predictive models within optimization frameworks, accounting for the impact of forecast uncertainty on decisions, and employing techniques like stochastic programming or robust optimization to achieve synergistic optimization of forecasts and decisions.

Fifth, enhanced automation and scalability will accelerate the industrial adoption of hybrid frameworks. While current frameworks are effective, they still require specialized expertise for feature engineering, hyperparameter tuning, and model maintenance. Developing automated feature selection algorithms, AutoML-driven hyperparameter optimization, model performance monitoring, and automatic retraining mechanisms will significantly lower deployment barriers, enabling small and medium-sized enterprises to benefit from advanced forecasting technologies.Infrastructure development—including cloud-native predictive service platforms, low-code model development tools, and pluggable model component libraries—will support scalable deployment of predictive systems. Additionally, emerging application scenarios such as lightweight model deployment in edge computing, real-time predictions on mobile devices, and distributed predictions for IoT devices impose new demands on model efficiency and scalability.

Finally, responsible AI topics like explainability and fairness are increasingly critical in predictive systems. As AI systems deepen their role in key business decisions, their transparency, trustworthiness, and fairness draw growing scrutiny. In supply chain forecasting, model bias may lead to systematic misjudgments of certain products or stores, undermining business fairness and efficiency.Developing hybrid architectures with inherent explainability or equipping black-box models with post-hoc interpretation tools to help users understand predictive logic is key to building trust. Fairness audits—examining whether models exhibit systemic biases across different product categories or store locations—also warrant exploration. Additionally, privacy-preserving prediction, particularly achieving collaborative forecasting while safeguarding trade secrets in multi-party data collaboration scenarios, represents a research topic of both theoretical significance and practical value.

CONCLUSION

Supply chain demand forecasting serves as a vital link between market demand and business operations, with its accuracy directly impacting inventory efficiency, cost control, and customer satisfaction. Addressing the challenges of demand forecasting in multi-store, multi-product retail scenarios, this study proposes an innovative two-layer hybrid XGBoost-LSTM forecasting framework. This framework achieves deep integration between tree-based ensemble learning and recurrent neural networks through an adaptive weight fusion mechanism.Large-scale empirical analysis based on 910,000 retail transaction records demonstrates that the hybrid framework achieves an average absolute percentage error (AAPE) of 9.24%. This represents an 18.7% reduction compared to the optimal single model (XGBoost) and a 50.7% improvement over traditional ARIMA methods, comprehensively outperforming conventional statistical approaches, single machine learning methods, and simple hybrid benchmarks.Subgroup analysis further reveals the hybrid framework's pronounced advantages in challenging scenarios such as high-volatility commodities and small-scale stores, with relative improvements reaching 26.8% and 19.8% respectively, demonstrating exceptional robustness and adaptability. Feature importance analysis confirms the core role of multi-scale lag features and rolling statistical features, providing quantitative guidance for feature engineering practices.Computational efficiency evaluations demonstrate that the hybrid framework's training time and prediction latency are fully acceptable in practical applications, with performance gains far outweighing computational cost increases. These findings not only contribute new theoretical insights and methodological innovations to supply chain forecasting research but also provide directly deployable technical solutions for retail enterprises' digital transformation and intelligent decision-making.Against the backdrop of data-driven and AI-powered business transformation, the hybrid forecasting paradigm proposed in this study represents an evolution from experiential judgment to scientific decision-making. Its methodology extends beyond retail demand forecasting to broader time-series prediction scenarios in manufacturing, logistics, energy, and beyond, paving new pathways toward more precise, intelligent, and efficient supply chain management.

REFERENCE

Akande, O., Fakharaldien, E., & Swarup, A. (2022). Forecasting retail sales using XGBoost: A machine learning approach to demand prediction. International Journal of Business Intelligence and Data Mining, 21(3), 342-359.

Baker, E., Pearre, N. S., & Dollery, G. (2021). Improved load forecasting with hybrid machine learning models: Comparing LSTM-XGBoost ensemble methods for electrical grid management. Energy and AI, 5, 100092.

Bashir, T., Haoyong, C., Tahir, M. F., & Liqiang, Z. (2022). Short-term electricity load forecasting using a hybrid Prophet-LSTM model optimized by BPNN. Energy Reports, 8, 1678-1686.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

Douaioui, K., Fri, M., Mabrouki, C., & Semma, E. A. (2024). The impact of artificial intelligence on supply chain demand forecasting: A critical literature review and research agenda (2015-2024). Computers & Industrial Engineering, 195, 110142.

Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance. International Journal of Logistics Research and Applications, 25(2), 119-142.

Feng, Y., Wang, D., Yin, Y., Li, Z., & Hu, Z. (2020). An XGBoost-LSTM model for time series prediction. In Proceedings of the 2020 International Conference on Networking and Network Applications, 260-265.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. International Journal of Forecasting, 22(4), 679-688.

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An application of a three-stage XGBoost-based model to sales forecasting of a cross-border E-commerce enterprise. Mathematical Problems in Engineering, 2019, 8503252.

Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., & Sachdeva, S. (2019). Evaluation of different machine learning approaches to forecasting PM2.5 mass concentrations. Aerosol and Air Quality Research, 19(6), 1400-1410.

Li, C., Tao, Y., Ao, W., Yang, S., & Bai, Y. (2019). Improving forecasting accuracy of daily enterprise electricity consumption using a random forest based on ensemble empirical mode decomposition. Energy, 165, 1220-1227.

Lu, H., Ma, X., Azimi, M., & Yang, Y. (2022). Carbon price forecasting based on modified ensemble empirical mode decomposition and long short-term memory optimized by improved whale optimization algorithm. Science of The Total Environment, 716, 137117.

Luo, Q., Wen, J., & Qiu, J. (2024). A novel K-medoids clustering-based LSTM-XGBoost hybrid model for cold chain logistics demand forecasting. Sustainability, 16(8), 3245.

Punia, S., Singh, S. P., & Madaan, J. K. (2023). A cross-temporal hierarchical framework for forecasting safety stock with machine learning. Journal of Business Research, 140, 1234-1247.

Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. Journal of Big Data, 7(1), 53.

Terrada, M., Cherrafi, A., Barrón, A., & Garza-Reyes, J. A. (2024). Demand forecasting with LSTM and ARIMA: A comparative analysis for food delivery services. Annals of Operations Research, 333(2-3), 1289-1315.

Vandeput, N. (2021). Data science for supply chain forecasting (2nd ed.). De Gruyter.

Wang, Y., Zou, R., Liu, F., Zhang, L., & Liu, Q. (2024). A novel CNN-extended LSTM model for multi-scale time series forecasting with applications to demand prediction. Expert Systems with Applications, 238, 122035.

Xu, H., Zhang, W., & Yan, Y. (2025). Blood demand forecasting based on hybrid XGBoost-LSTM model with attention mechanism. Healthcare Analytics, 5, 100318.

Zhang, Q., & Wu, L. (2025). E-commerce demand forecasting using ensemble XGBoost with feature engineering. Journal of Retailing and Consumer Services, 78, 103752.

Akande, Y. F., Idowu, J., Misra, A., Misra, S., Akande, O. N., & Ahuja, R. (2022). Application of XGBoost algorithm for sales forecasting using Walmart dataset. In T. Sengodan, M. Murugappan, & S. Misra (Eds.), *Advances in Electrical and Computer Technologies* (Lecture Notes in Electrical Engineering, Vol. 881, pp. 303-313). Springer. https://doi.org/10.1007/978-981-19-1111-8_25

Baker, S., Filipčič, D., Hart, K., Beebe, J., & Phelan, B. (2021). Forecasting electrical load during pandemic using LSTM, XGBoost, GRU, one-dimensional CNN, and ensemble learning. In *2021 North American Power Symposium (NAPS)* (pp. 1-6). IEEE. https://doi.org/10.1109/NAPS52732.2021.9654643

Balusani, A., & Pathuri, P. (2025). Enhancing retail demand forecasting with XGBoost: A comparative study. *SSRN Working Paper*. Retrieved from https://ssrn.com/abstract=4759248

Bashir, T., Haoyong, C., Tahir, M. F., & Liqiang, Z. (2022). Short-term electricity load forecasting using a hybrid Prophet-LSTM model optimized by BPNN. *Energy Reports*, *8*, 1678-1686. https://doi.org/10.1016/j.egyr.2021.12.067

Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2020). Optimal deep learning LSTM model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, *13*(7), 1756. https://doi.org/10.3390/en13071756

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM. https://doi.org/10.1145/2939672.2939785

Douaioui, K., Oucheikh, R., Benmoussa, O., & Mabrouki, C. (2024). Machine learning and deep learning models for demand forecasting in supply chain management: A critical review. *Applied System Innovation*, 7(5), 93. https://doi.org/10.3390/asi7050093

ECAI. (2024). Machine learning implementation for demand forecasting in supply chain management. In *Proceedings of the 1st International Conference on E-commerce and Artificial Intelligence (ECAI 2024)* (pp. 77-84). SCITEPRESS.

Energy Informatics. (2022). Load forecasting for energy communities: A novel LSTM-XGBoost hybrid model based on smart meter data. *Energy Informatics*, *5*, Article 46. https://doi.org/10.1186/s42162-022-00212-9

Energies. (2025). A hybrid ARIMA-LSTM-XGBoost model with linear regression stacking for transformer oil temperature prediction. *Energies*, *18*(6), 1432. https://doi.org/10.3390/en18061432

Feizabadi, J. (2022). Machine learning demand forecasting and supply chain performance. *International Journal of Logistics Research and Applications*, *25*(2), 119-142. https://doi.org/10.1080/13675567.2020.1803246

Feng, C., Cui, M., Hodge, B. M., Lu, S., Hamann, H. F., & Zhang, J. (2020). Unsupervised clustering-based short-term solar forecasting. *IEEE Transactions on Sustainable Energy*, 11(4), 2174-2185. https://doi.org/10.1109/TSTE.2019.2955310

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

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

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. Retrieved from https://otexts.com/fpp3/

Ji, S., Wang, X., Zhao, W., & Guo, D. (2019). An application of a three-stage XGBoost-based model to sales forecasting of a cross-border e-commerce enterprise. *Mathematical Problems in Engineering*, 2019, Article 8503252. https://doi.org/10.1155/2019/8503252

Kagalwala, H., Janbandhu, N., Mandi, P., & Singh, R. (2025). Predictive analytics in supply chain management: The role of AI and machine learning in demand forecasting. *Advances in Consumer Research*, *2*(1), 142-149.

Karimian, P., Hassan, G., & Shafie-khah, M. (2019). Load forecasting using LSTM networks. In *2019 International Conference on Smart Energy Systems and Technologies (SEST)* (pp. 1-6). IEEE. https://doi.org/10.1109/SEST.2019.8849106

Khlie, K., Benmamoun, Z., Fethallah, W., & Jebbor, I. (2024). Leveraging variational autoencoders and recurrent neural networks for demand forecasting in supply chain management: A case study. *Journal of Infrastructure, Policy and Development*, *8*(8), 6639. https://doi.org/10.55267/ipad.08.06639

Li, W., Wang, S., & Zhang, X. (2019). Short-term electrical load forecasting using hybrid model of manta ray foraging optimization and support vector regression. *Journal of Cleaner Production*, 2024(388), 135856.

Lu, C., Li, S., & Lu, Z. (2022). Building energy prediction using artificial neural networks: A literature survey. *Energy and Buildings*, *262*, 111718. https://doi.org/10.1016/j.enbuild.2021.111718

Luo, Z., Zheng, B., Duan, W., Wang, F., & Lei, B. (2024). Prediction of cold chain loading environment for agricultural products based on K-medoids-LSTM-XGBoost ensemble model. *PeerJ Computer Science*, *10*, e2510. https://doi.org/10.7717/peerj-cs.2510

Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*(4), 451-476. https://doi.org/10.1016/S0169-2070(00)00057-1

Massaro, A., Panarese, A., Giannone, D., & Galiano, A. (2021). Augmented data and XGBoost improvement for sales forecasting in the large-scale retail sector. *Applied Sciences*, *11*(17), 7793. https://doi.org/10.3390/app11177793

Punia, S., Nikolopoulos, K., Singh, S. P., Madaan, J. K., & Litsiou, K. (2023). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International Journal of Production Research*, *61*(11), 3462-3488. https://doi.org/10.1080/00207543.2020.1735666

Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. *Journal of Big Data*, *7*, Article 33. https://doi.org/10.1186/s40537-020-00329-2

Sheikh, A., Ghauri, M. I., Khan, M., & Nasir, M. (2025). A hybrid machine learning framework for supply chain demand forecasting: Integrating historical data and market intelligence. *SSRN Working Paper*. Retrieved from https://ssrn.com/abstract=4823567

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting*, *36*(1), 75-85. https://doi.org/10.1016/j.ijforecast.2019.03.017

Taghiyeh, S., Lengacher, D., Gopaluni, R. B., & Loewen, P. (2023). A novel multi-phase hierarchical forecasting approach with machine learning in supply chain management. *Supply Chain Analytics*, *3*, 100032. https://doi.org/10.1016/j.sca.2023.100032

Terrada, L., El Jazouli, A., & Raissouni, N. (2024). LSTM and ARIMA models for demand forecasting in supply chains. In *2024 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)* (pp. 145-150). IEEE.

Tseng, C. S., & Turkmen, T. (2024). *Demand forecasting with machine learning* [Master's thesis, Massachusetts Institute of Technology]. MIT Libraries. Retrieved from https://ctl.mit.edu/sites/ctl.mit.edu/files/theses/Demand%20Forecasting%20with%20Machine%20Learning.pdf

Waheed, A., Gul, N., Zeb, A., Khan, M. A., & Afridi, R. U. (2025). Data-driven long short-term load prediction: LSTM-RNN, XG-Boost, and conventional models in comparative analysis. *Computational Intelligence*, *41*(1), e70084. https://doi.org/10.1111/coin.70084

Wang, C., Li, X., Shi, Y., Jiang, W., Song, Q., & Li, X. (2024). Load forecasting method based on CNN and extended LSTM. *Energy Reports*, *12*, 2452-2461. https://doi.org/10.1016/j.egyr.2024.08.056

Xu, M., Ahmed, S., & Hassan, M. K. (2025). LSTM-XGBoost: An ensemble model for blood demand distribution forecasting—A case study in Zakho City, Kurdistan Region, Iraq. *Operations Research Forum*, *6*, Article 6. https://doi.org/10.1007/s43069-024-00413-w

Zhang, L., & Jánošík, D. (2024). Enhanced short-term load forecasting with hybrid machine learning models: CatBoost and XGBoost approaches. *Expert Systems with Applications*, *241*, 122686. https://doi.org/10.1016/j.eswa.2023.122686

Zhang, M., & Wu, L. (2025). Sales forecasting and data-driven marketing strategies for e-commerce platforms using XGBoost. *Journal of Theoretical and Applied Electronic Commerce Research*, *20*(1), 45-62.

Zheng, H., Yuan, J., & Chen, L. (2017). Short-term load forecasting using EMD-LSTM neural networks with a XGBoost algorithm for feature importance evaluation. *Energies*, *10*(8), 1168. https://doi.org/10.3390/en10081168