applied sciences

Article

STL-DCSInformer-ETS: A Hybrid Model for Medium- and Long-Term Sales Forecasting of Fast-Moving Consumer Goods

Yecheng Ma, Lili He and Junhong Zheng *

College of Computer Science and Technology (College of Artificial Intelligence), Zhejiang Sci-Tech University, Hangzhou 310018, China; 202230603092@mails.zstu.edu.cn (Y.M.); llhe@zju.edu.cn (L.H.)
* Correspondence: zjhist@zstu.edu.cn

Abstract: Accurately forecasting sales for fast-moving consumer goods (FMCG) remains a significant challenge due to the volatile and multi-faceted nature of sales data. Existing methods often struggle to capture intricate patterns driven by seasonal trends, external factors, and consumer behavior, hindering effective inventory management and strategic decision-making. To overcome these challenges, we propose STL-DCSInformer-ETS, a hybrid model that integrates three complementary components: STL decomposition, an enhanced DCSInformer model, and the ETS model. The model uses monthly sales data from a FMCG company, with key features including sales volume, product prices, promotional activities, and regulatory factors such as holidays, geographical information, consumer behavior, product factors, etc. STL decomposition partitions time-series data into trend, seasonal, and residual components, reducing data complexity and enabling more targeted forecasting. The enhanced DCSInformer employs dilated causal convolution and a multi-scale feature extraction mechanism to capture long-term dependencies and short-term variations effectively. Meanwhile, the ETS model specializes in modeling seasonal patterns, further refining forecasting precision. To further improve predictive performance, the Random Forest-based Recursive Feature Elimination (RF-RFE) method is applied to optimize feature selection. RF-RFE identifies key predictive factors from multiple dimensions, such as time, geography, and economy, which significantly influence forecasting accuracy. Through numerical experiments, the method demonstrates excellent performance by achieving a 35.9% reduction in Mean Squared Error and a 21.4% decrease in Mean Absolute Percentage Error, significantly outperforming traditional methods. Furthermore, the model effectively captures both medium- and long-term sales trends while addressing short-term fluctuations, leading to more accurate forecasting and improved decision-making for fast-moving consumer goods. This research provides new theoretical insights into hybrid forecasting models and practical solutions for optimizing inventory management and strategic planning in the FMCG industry.

Keywords: STL-DCSInformer-ETS; medium- and long-term sales forecasting; fast-moving consumer goods; random forest recursive feature elimination; multi-scale feature extraction; dilated causal convolution

Academic Editor: Stefan Fischer

Received: 6 January 2025
Revised: 29 January 2025
Accepted: 29 January 2025
Published: 2 February 2025

Citation: Ma, Y.; He, L.; Zheng, J. STL-DCSInformer-ETS: A Hybrid Model for Medium- and Long-Term Sales Forecasting of Fast-Moving Consumer Goods. *Appl. Sci.* **2025**, *15*, 1516. https://doi.org/10.3390/app15031516

1. Introduction

The rapid development and intensifying competition in the FMCG sector have underscored the critical importance of medium- and long-term sales forecasting in optimizing supply chains, managing inventories, and informing strategic market decisions [1]. In recent years, the FMCG industry has witnessed significant changes driven by evolving consumer preferences, the rise of e-commerce, and increasing supply-chain management demands.

These factors have not only altered the competitive landscape but also heightened the complexity of supply-chain management, making accurate forecasting more essential than ever for businesses to stay competitive and responsive to market dynamics. However, the volatile and complex nature of FMCG markets presents significant challenges for existing forecasting methodologies. Conventional models often struggle to effectively address multidimensional dynamic factors, such as price fluctuations, promotional campaigns, and macroeconomic conditions, particularly when it comes to capturing complex temporal dependencies and long-term trends. For example, the author Satyanarayana KV used the k-Nearest Neighbors (k-NN) algorithm for sales prediction in the FMCG sector [2]. While the algorithm performed quite well on certain evaluation metrics, the model was overly simplistic and failed to effectively capture the complex nonlinear relationships and multidimensional factors within the sales data, limiting its predictive capability in practical applications. The author Sun used the XGBoost model, a LightGBM tree-based model, and a Long Short-Term Memory (LSTM) network for sales forecasting [3]. While these individual models showed reasonable performance on certain evaluation metrics, they failed to fully leverage the strengths of each model and struggled to effectively capture the complex nonlinear relationships and multidimensional factors within the sales data, limiting their predictive capability in real-world applications. C Sun et al. proposed a multi-step time-series forecasting method based on Informer, XGBoost, and Genetic Algorithm (GA) [4]. This method combines Informer with XGBoost and optimizes the model weights through GA to improve the accuracy and stability of multi-step time-series forecasting. However, it still faces significant challenges in handling long-term and mid-term data, especially regarding its sensitivity to long-term trend changes. These limitations affect the model's accuracy in handling complex multi-step time-series forecasts, thereby restricting its practical application in long-term prediction tasks. Existing forecasting models for FMCG sales often fail to effectively capture long-term dependencies and complex seasonal trends due to their reliance on simplistic feature extraction methods or limited temporal modeling capacity. While models like LSTM and Transformer show promise, they struggle to address multi-dimensional data and long-term trend variations, especially in dynamic markets. This research aims to fill this gap by proposing a novel hybrid approach, STL-DCSInformer-ETS, which integrates STL decomposition, DCSInformer, and ETS models to better capture both short-term fluctuations and long-term trends.

This study first employs the RF-RFE method to quantitatively identify key features, enhancing feature selection precision and improving predictive performance and model reliability. Building on this, the study proposes an innovative hybrid model, STL-DCSInformer-ETS, combining STL decomposition, an enhanced DCSInformer model, and the ETS model. Compared to previous models, DCSInformer introduces novel capabilities, such as its ability to simultaneously capture both short-term fluctuations and long-term trends through dilated causal convolution and multi-scale feature extraction. This effectively addresses the shortcomings of traditional models, which struggle with long-term dependencies. The STL decomposition reduces data complexity by partitioning time-series data into trend, seasonal, and residual components, streamlining the forecasting process. Meanwhile, the ETS model specializes in capturing seasonal variations, improving the model's ability to model periodic patterns and enhance forecasting precision. In summary, this integrated approach provides a comprehensive solution for FMCG sales forecasting, significantly improving predictive accuracy and offering practical insights for inventory optimization and supply-chain strategic decision-making.

1.1. Principal Contributions

The principal contributions of this paper are as follows:

- Hybrid Model Development : Proposes the STL-DCSInformer-ETS hybrid model, which combines STL decomposition, enhanced DCSInformer, and ETS models to address the unique challenges of FMCG sales forecasting.
- Feature Selection: Introduces the RF-RFE method to enhance feature selection precision, improving model reliability and predictive accuracy.
- Temporal Dependency Modeling: Incorporates dilated causal convolution and multi-scale feature extraction in the DCSInformer component to simultaneously capture short-term fluctuations and long-term trends.
- Seasonal Pattern Modeling: Integrates ETS to effectively model seasonal variations, enhancing the ability to handle periodic patterns in time-series data.

1.2. Paper Structure

The structure of the paper is as follows:

- Section 2: Reviews related work in time-series forecasting, highlighting existing challenges and gaps in FMCG sales forecasting.
- Section 3: Presents the methodology, describing the STL-DCSInformer-ETS model and the RF-RFE feature selection process.
- Section 4: Details the experimental setup and provides a comprehensive evaluation of the model's performance compared to baseline methods.
- Section 5: Concludes the paper, summarizing the findings, contributions, and future research directions.

2. Related Work

2.1. Definition of FMCG Supply Chain and Sales Forecasting

The FMCG supply chain encompasses the end-to-end processes involved in delivering fast-moving consumer goods from manufacturers to consumers [5,6]. This includes procurement of raw materials, manufacturing, inventory management, warehousing, transportation, and retail distribution. An efficient FMCG supply chain ensures that products are available at the right time, in the right quantity, and at the right place to meet consumer demand.

Sales forecasting for FMCG refers to the process of predicting future demand for FMCG products based on historical sales data, market trends, seasonal patterns, promotional activities, and external factors such as economic conditions or weather fluctuations. Accurate sales forecasting plays a critical role in ensuring demand–supply alignment, optimizing inventory levels, and minimizing waste [7]. The key elements of FMCG forecasting are:

- Dynamic Demand Patterns: FMCG products are characterized by high demand volatility driven by seasonality, promotions, and shifts in consumer preferences.
- Short Product Life Cycles: Many FMCG products, such as food and beverages, have limited shelf lives, necessitating precise and timely forecasting.
- Big Data Integration: Modern forecasting incorporates large-scale data from diverse sources, including sales history, social media trends, and real-time consumer feedback.
- Advanced Forecasting Techniques: To improve prediction accuracy, methods such as time-series analysis (e.g., ARIMA, LSTM) and machine learning models (e.g., Informer) are increasingly adopted.

Sales forecasting is a critical foundation of FMCG supply-chain operations, enabling efficient planning and execution by aligning production schedules with market demand, optimizing inventory levels to minimize costs and waste, improving logistics and distribution efficiency through timely deliveries, and enhancing customer satisfaction with consistent product availability [8]. The integration of accurate sales forecasting with a responsive FMCG supply chain is essential for achieving operational excellence and maintaining a competitive edge in the fast-moving consumer goods industry.

2.2. Sales Time-Series Forecasting

Sales forecasting is a crucial aspect of business operations, providing valuable insights for decision-making and strategic planning. To improve forecasting accuracy and efficiency, various methods and techniques have been explored across different domains, including e-commerce, livestream sales, and B2B environments. In the e-commerce industry, ref. [9] conducted research on sales forecasting for imbalanced classification data, offering practical solutions to enhance market competitiveness and economic benefits for enterprises. With the advent of the big-data era, e-commerce sales forecasting has emerged as a key factor in maintaining market competitiveness. For instance, ref. [10] leveraged machine learning models to predict sales-pattern changes and surges triggered by local events in specific geographical locations. By capturing the complex relationships between events and their impact on sales, this approach significantly improved inventory management and sales surge prediction. In the context of livestream sales prediction, ref. [11] introduced SaleNet, an interpretable deep learning model designed to account for various factors influencing sales volumes. This model provided a powerful tool for businesses to achieve accurate sales forecasts in livestream environments, where dynamic and rapidly changing factors play a significant role. In multi-channel retail settings, ref. [12] proposed OCCPH-MHA, a two-stage deep learning approach for sales forecasting. The model first identifies consumer group preferences based on individual purchasing behavior and then integrates these preferences with time-series demand data using a global–local attention mechanism. This method significantly improved multi-step forecasting accuracy by effectively combining behavioral insights with temporal data. For B2B sales forecasting, ref. [13] explored time-series forecasting techniques, utilizing Recurrent Neural Networks (RNNs) and Echo State Networks (ESNs) enhanced with a modified metaheuristic optimizer for parameter selection. This approach incorporated decomposition techniques to improve prediction accuracy, resulting in a notable performance increase from an initial score of 0.082790 to 0.646967. These results demonstrate the potential of advanced neural network architectures in addressing complex forecasting challenges in the B2B domain. While these studies have significantly advanced sales forecasting in various domains, existing methods still struggle to simultaneously capture short-term fluctuations and long-term trends in dynamic environments like the FMCG sector. To address these challenges, and building on refs. [14–16], this research proposes the STL-DCSInformer-ETS hybrid model, which can enhance forecasting accuracy and scalability for FMCG sales. To more clearly illustrate the distinctive aspects of this work and its contributions to the existing knowledge base. Table 1 presents a comparison between previous related work and the significant contributions of this study.

Table 1. Comparison of related work and significant contributions.

Related Work	Methods/Techniques	Limitations	Contribution	Contribution Improvement
Satyanarayana KV (k-NN for FMCG sales prediction) [2]	k-Nearest Neighbors (k-NN)	Too simple, unable to capture complex non-linear relationships and multidimensional factors	Proposed STL-DCSInformer-ETS model, combining STL decomposition, DCSInformer, and ETS, to capture short-term fluctuations and long-term trends	Can handle more complex multidimensional data, accurately capturing seasonal fluctuations and long-term trends
C Sun et al. (XG-Boost + GA) [4]	XGBoost with Genetic Algorithm optimization	Sensitive to long-term trend changes, struggles with long-term prediction tasks	Introduced the DC-SInformer model, overcoming long-term trend prediction issues with dilated causal convolution and multi-scale feature extraction	Improved long-term sales prediction accuracy, reduced sensitivity to long-term trend changes
zhou et al. (Informer) [17]	Informer	Faces challenges in long-term trend variation when handling multi-step forecasting	Added RF-RFE method for feature selection, optimizing feature selection precision	Improved feature selection accuracy, enhancing model predictive performance and stability
Wu et al. (OCCPH-MHA) [12]	Multi-stage deep learning (OCCPH-MHA)	Limited in capturing short-term fluctuations and long-term trends effectively	Combined STL decomposition and ETS models to capture seasonal fluctuations and long-term trends effectively	Increased accuracy in capturing both short-term fluctuations and long-term trends

3. Materials and Methods

3.1. Data Preprocessing

3.1.1. Feature Engineering

Feature engineering plays a pivotal role in medium- and long-term forecasts of FMCG sales. It not only enhances model performance but also enables the model to identify and capture complex temporal patterns within the data [18]. Given the high volatility and seasonality of the FMCG market, sales volume is impacted by multiple dimensions, including price, promotional activities, macroeconomic conditions, holidays, and weather. Extracting meaningful information through feature engineering can greatly enhance the model's ability to predict medium- and long-term trend variations. To reduce data complexity, this study employs the mutual information method to screen features across several critical dimensions.

3.1.2. Feature Construction

To ensure the model's performance in medium- and long-term predictions, this study constructs features that exert significant influence over extended time spans [19]:

(1) Time Dimension: Seasonal and holiday features were extracted to effectively capture the cyclical changes in consumption patterns and sales fluctuations. These features are particularly significant in explaining sales variability over longer time spans [20].

(2) Geographic Dimension: By utilizing the latitude and longitude data of different regions, the spatial distribution of consumption potential and market demand over the medium to long term can be revealed. These geographical features help the model better

capture regional differences in consumer behavior, enhancing the spatial adaptability of the predictions.

(3) Economic Dimension: Macro-economic indicators, such as consumer confidence indices, national policies, and government spending, were analyzed. In medium to long-term forecasting, macro-economic features are crucial in reflecting overall consumption demand, effectively capturing long-cycle market fluctuations [21].

(4) Product Dimension: Features such as packaging format, advertising investment, store promotion frequency, and supply-chain delays were extracted. These help the model identify sales trends throughout a product's lifecycle, particularly clarifying key stages of the lifecycle in medium to long-term forecasting.

(5) Weather Dimension: Meteorological features such as temperature and rainfall were integrated to analyze the short-term and long-term effects of weather conditions on sales. Especially in markets with significant seasonality, weather features have an undeniable impact on medium- to long-term sales fluctuations [22].

Table 2 shows the features and their meanings. By incorporating these features, the model can accurately capture short-term sales fluctuations while effectively reflecting medium- and long-term trend variations. This comprehensive feature set enhances the model's performance in sales forecasting over extended time spans.

Table 2. Feature description by dimension.

Feature Dimension	**Feature Variable Name**	**Feature Meaning**
Time Dimension	Seasonal feature	Each season is assigned a fixed representative value: 0 for winter, 1 for spring, 2 for summer, and 3 for autumn.
	Holiday feature	Number of holidays in the month.
Geographical Dimension	Latitude and longitude	Regional latitude and longitude.
	Population density	Regional population density.
Economic Dimension	Consumer confidence index	The current consumer confidence is proportional to the confidence in the benchmark year.
	Consumer expenditure	The per capita consumer expenditure.
	Government spending	Proportion of government expenditure on consumer goods in GDP.
Product Dimension	Packaging type	Product packaging type number.
	Advertising frequency	Frequency of advertisements.
	Store promotions	Number of in-store promotional activities.
	Supply-chain delay	Time delay from supplier to retail point.
Weather Dimension	Rainfall days	Days of rainfall in a month.
	Temperature average	Monthly average temperature.

3.1.3. Feature Selection

To further enhance the effectiveness of feature selection, this study employs the RF-RFE method to identify features that exhibit a strong correlation with medium- and long-term sales forecasts.The RF-RFE method was chosen for feature selection due to its ability to efficiently identify important features while handling high-dimensional data. Unlike traditional methods such as Lasso or decision trees, RF-RFE allows for the elimination of irrelevant features while retaining those that contribute the most to predictive performance.This makes it particularly well-suited for the model, where multiple features from diverse dimensions are used [23]. Feature selection was performed using RF-RFE, with features ranked in descending order of importance. Detailed results are shown in the Table 3.

Table 3. "Key Features and Their Importance Selected by RF-RFE".

Feature Name	Symbol	Feature Importance
Seasonal feature	S_f	0.42
Holiday feature	H_f	0.38
Latitude and longitude	L_f	0.12
Population density	P_f	0.20
Consumer confidence index	C_f	0.24
Consumer expenditure	E_f	0.17
Government spending	G_f	0.21
Packaging type	K_f	0.02
Advertising frequency	A_f	0.31
Store promotions	M_f	0.41
Supply-chain delay	D_f	0.16
Rainfall days	R_f	0.08
Temperature average	T_f	0.22

Based on the feature importance values derived from RF-RFE, features were categorized into two groups based on their relevance to the model's predictive performance. Features with importance values below 0.05 were assigned a fixed importance value of 0.01, as they were considered weakly relevant. Features with importance values above 0.05 were retained with their original values, as they were deemed relevant to the prediction task. In the feature selection process, this categorization helped streamline the identification of key variables that significantly contribute to sales predictions, while effectively minimizing the influence of irrelevant or weakly correlated features.

The analysis results reveal that S_f demonstrate the highest importance, with values is 0.42, indicating a substantial impact on the sales of this fast-moving consumer good. This is primarily due to consumer demand fluctuations during different seasons and holidays, which strongly highlight the seasonality of the product. The importance of G_f is 0.21; government spending affect sales indirectly by shaping market conditions and influencing consumer behavior. These changes impact variables like pricing and supply-chain dynamics, which in turn affect purchasing decisions. However, the influence of policies is typically less direct compared to more immediate factors like seasonality or promotions, resulting in moderate relevance to short-term sales forecasts. In contrast, features related to less relevant variables, such as K_f, exhibit low importance values, often below 0.05, and are therefore excluded from further analysis.

3.1.4. Feature Contribution Description

To better understand the factors influencing sales predictions, this study decomposes the contributions of different dimensions into distinct models. Each dimension reflects specific aspects of the data and their interactions with the target variable. The contribution calculations are categorized as follows:

- Time Dimension Contribution

  The contribution of the time dimension is modeled as:

$$F_{\text{time}}(S_f, H_f) = w_1 S_f^2 + w_2 H_f \cdot \sin\left(\frac{2\pi}{12} t\right) + w_3 S_f \cdot H_f \tag{1}$$

The weights w_1, w_2, w_3 control the contributions of nonlinear seasonal effects, periodic influences, and their interactions, respectively. This formulation captures seasonal variations, holiday influences, and their interactions. The term S_f^2 accounts for nonlinear seasonal effects, while $H_f \cdot \sin\left(\frac{2\pi}{12} t\right)$ introduces periodicity reflecting the calendar cycle.

- Geographical Dimension Contribution

  The geographical dimension contribution is defined as:

$$F_{\text{geo}}(L_f, P_f) = w_4 \log(1 + P_f) + w_5 L_f \cdot \cos\left(\frac{2\pi}{365}t\right) \tag{2}$$

The weights w_4 and w_5 control the contributions of population density effects and periodic regional geographical influences, respectively. Here, $\log(1 + P_f)$ models the diminishing marginal impact of population density, while the term $L_f \cdot \cos(\frac{2\pi}{365}t)$ reflects the periodic influence of regional geographical factors over time [24].

- Economic Dimension Contribution

  The contribution of the economic dimension is expressed as:

$$F_{\text{econ}}(C_f, E_f, G_f) = w_6 C_f^2 + w_7 E_f + w_8 G_f \cdot C_f \tag{3}$$

The weights w_6, w_7, and w_8 control the contributions of consumer confidence, per capita consumption expenditure, and the interaction between government spending and consumer confidence, respectively. This equation captures key economic drivers, including consumer confidence (C_f), per capita consumption expenditure (E_f), and government spending in GDP (G_f). The interaction term $G_f \cdot C_f$ reflects the combined effect of government policies and consumer sentiment.

- Product and Promotion Contribution

  The contribution of product dimension, advertising, and promotional activities is formulated as:

$$F_{\text{product}}(K_f, A_f, M_f, D_f) = w_9 \frac{A_f}{1 + A_f} + w_{10} M_f \cdot \log(1 + D_f) + w_{11} K_f \tag{4}$$

The weights w_9, w_{10}, and w_{11} control the contributions of diminishing advertising returns, the delayed impact of promotions, and the influence of packaging type, respectively. This model accounts for the diminishing returns of advertising ($\frac{A_f}{1+A_f}$), the delayed impact of promotions through $M_f \cdot \log(1 + D_f)$, and the influence of packaging type (K_f) on the target variable [21].

- Weather Dimension Contribution

  The weather dimension contribution is given by:

$$F_{\text{weather}}(R_f, T_f) = w_{12} R_f \cdot T_f \tag{5}$$

The weight w_{12} regulates the combined impact of rainfall (R_f) and temperature (T_f) on the target variable. This formulation highlights the interaction between rainfall and temperature, capturing their joint influence on the target variable [25].

3.1.5. Adaptive Regulation of Fluctuations

The sales of fast-moving consumer goods (FMCG) exhibit significant fluctuations over time. Exploring the fluctuation patterns of FMCG sales can contribute to more accurate sales forecasting [26]. The actual fluctuation rate of FMCG sales can be calculated using Equations (6) and (7).

$$r = \log\left(\frac{x_1}{x_2}\right) \tag{6}$$

$$RV = \sqrt{\sum_{i=1}^{T} r_t^2} \tag{7}$$

where x_1 represents the sales volume of the current month, x_2 represents the sales volume of the next month, r denotes the logarithmic return, T is the time period, and RV is the realized volatility.

The fluctuation module applies weighted processing to months with significant sales fluctuations, allowing for a deeper understanding of complex sales data and making it more suitable for medium- and long-term sales forecasting [27,28]. The specific process is as follows:

(1) Input Data: X_{data}: Time data, where each time point is denoted as X_{date}. $X_{\text{fluctuations}}$: Key fluctuation months.
weight: The weight factor for fluctuation adjustments.

(2) Normalize Monthly Data: Normalize the monthly data to map it to the range $[-0.5, 0.5]$, making it suitable for model training. The normalization formula is as follows:

$$\text{values}_{\text{date}} = \frac{\text{x}_{\text{date}}}{12} - 0.5 \tag{8}$$

(3) Define Time Window: Specify a *time_window* to define the range of months around each key fluctuation month.

(4) Compute Fluctuation Month Range: Calculate the fluctuation month intervals by expanding around each key fluctuation month ($X_{\text{fluctuations}}$) using the defined time window:

$$\text{month\_range} = X_{\text{fluctuations}} \pm \textit{time\_window} \tag{9}$$

(5) Adjust Values within the Fluctuation Range: Add the specified weight to the normalized time values corresponding to the fluctuation month range:

$$\text{values}_{\text{month\_range}} = \text{values}_{\text{month\_range}} + \text{weight} \tag{10}$$

(6) Return Final Values: Output the adjusted values, which now reflect the weighted impact of fluctuations over the specified months.

This study first employs the Random Forest-based Recursive Feature Elimination (RF-RFE) method to quantitatively identify key features, enhancing feature selection precision and improving predictive performance and model reliability. Once the relevant features have been identified, they form the foundation for the subsequent model construction. Building on this, the study proposes an innovative hybrid model, STL-DCSInformer-ETS, which integrates the complementary strengths of STL decomposition, an enhanced DCSInformer model, and the ETS model.

3.2. Model Construction

3.2.1. STL-DCSInformer-ETS Hybrid Model

To enhance the accuracy of time-series forecasting [29,30], this paper proposes a combined prediction model based on STL decomposition, the DCSInformer model, and the ETS model, as illustrated in Figure 1. STL (seasonal and trend decomposition using Loess) is a robust non-parametric time-series decomposition method that divides the series into three components: trend (trend), seasonality (seasonal), and remainder (remainder) [31,32]. The decomposition formula is expressed as:

$$Y_t = T_t + R_t + S_t \tag{11}$$

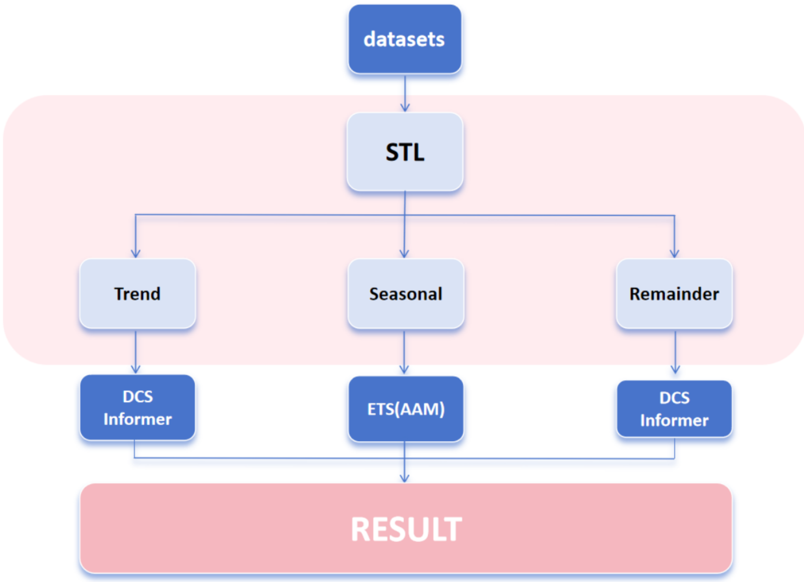

Figure 1. STL-DCSInformer-ETS hybrid model.

By decomposing the series into these components, STL effectively reduces the complexity of model fitting, allowing subsequent models to concentrate on the distinct characteristics of each component, thereby improving predictive accuracy.

For medium- and long-term time-series forecasting [33,34], the Informer model excels, particularly in uncovering latent time-series associations within high-dimensional data. Its sparse self-attention mechanism significantly reduces computational complexity while enhancing its ability to capture both long-term and short-term dependencies. To better address the requirements of medium- and long-term forecasts, this paper employs an enhanced version of the Informer model, DCSInformer. By incorporating dilated causal convolution and multi-scale feature extraction mechanisms, DCSInformer captures features across different time scales more effectively, achieving a balanced approach to processing long-term trends and short-term fluctuations. The ETS (AAM) model is specifically designed to handle seasonal data with clear cyclical patterns, making it highly effective for capturing and forecasting seasonal fluctuations. For macroeconomic indicators like GDP growth rates or CPI, the ETS model accurately identifies seasonality trends and integrates them into sales forecasts, enhancing overall prediction precision.

The hybrid model proposed in this paper further refines this process by precisely decomposing time-series data into trend, seasonal, and remainder components. This approach significantly improves the accuracy of long-term forecasts, particularly within the Chinese market.

3.2.2. DCSInformer Model

The Informer model is an efficient deep learning model specifically designed to address long-series time-series prediction problems [17]. As an improved variant of the traditional Transformer model, the Informer excels in both natural language processing and time-series data processing [14]. The overall architecture of the Informer model, as shown in Figure 2, comprises two main components: an encoder and a decoder.

However, it still faces several limitations. The lack of hierarchical feature processing across different time scales makes it difficult to balance long-term and short-term dependencies effectively [35]. This limitation is particularly evident in time series with high-frequency fluctuations or abrupt trends, where the model underperforms in predicting short-term features, thereby affecting trend accuracy. Further optimization is required to fully address the computational challenges posed by long-term time-series forecasting.

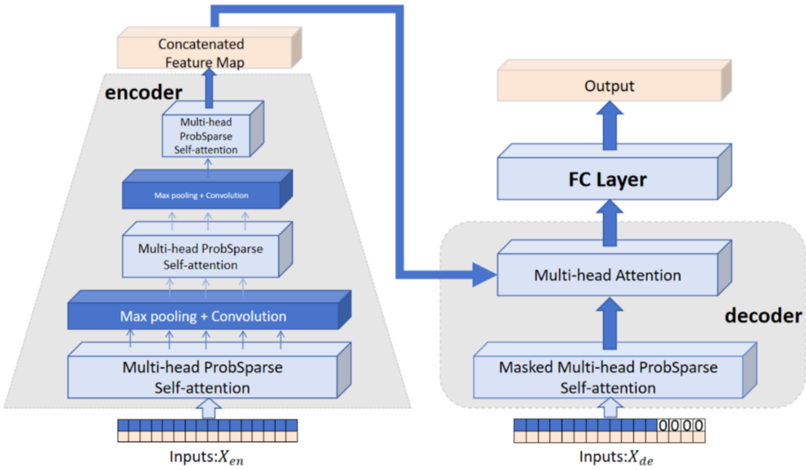

Figure 2. Informer model architecture diagram.

Therefore, this article proposes a DCSInformer model that combines dilated causal convolution and multi-scale feature extraction, significantly enhancing its ability to capture long time series and features across different time scales, making it particularly suitable for medium- and long-term forecasts, as shown in Figure 3.

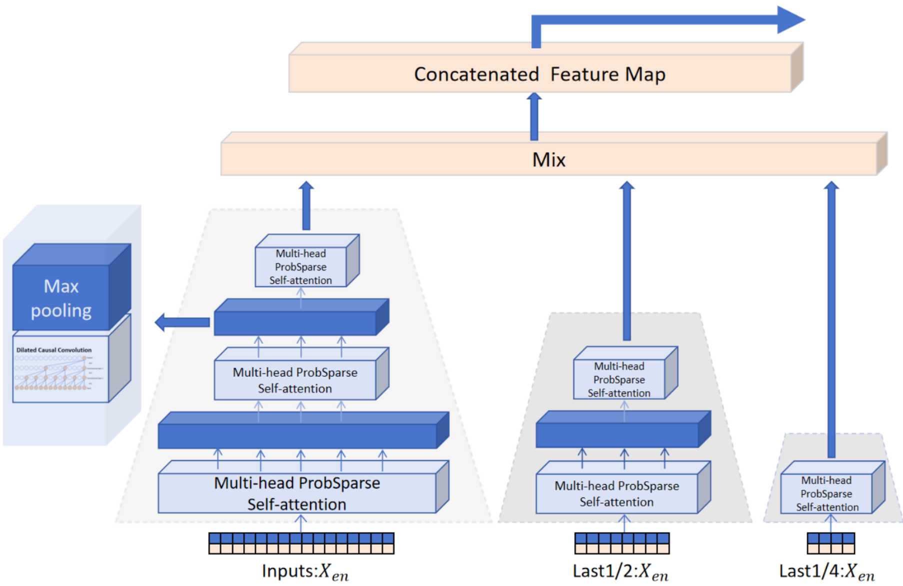

Figure 3. The DCSInformer model architecture diagram, illustrating the various components of the model, including the Max Pooling, Multi-Head Attention, and Self-Attention mechanisms, which work together to process and capture features from the input time series (inputs X_t).

Therefore, this article proposes a DCSInformer model. The DCSInformer component introduces innovative techniques that differentiates it from previous hybrid models. Unlike traditional models that primarily focus on capturing short-term fluctuations, the DCSInformer component effectively handles both short-term variations and long-term dependencies simultaneously. This is achieved through the use of dilated causal convolution, which allows the model to capture broader temporal relationships, and a multi-scale feature extraction mechanism that enhances its ability to process varying time scales. These advancements significantly improve the model's ability to predict complex FMCG sales patterns, addressing the limitations of conventional forecasting methods, which often struggle to capture long-term trends. By incorporating these novel features, the DCSInformer model provides a more accurate and robust solution for medium- and long-term sales forecasting in dynamic and volatile markets.

3.2.3. ProbSparse Self Attention

Self-Attention [17,36], a fundamental operation in Transformer models, computes relationships among query, key, and value pairs. However, its computational complexity scales quadratically as $O(L_Q L_K)$, where L_Q and L_K denote the lengths of the query and key sequences, respectively. This quadratic scaling poses a significant challenge for processing long sequences. ProbSparse Self-Attention addresses this issue by introducing sparsity into the attention mechanism, drastically reducing computational and memory requirements. The standard Self-Attention mechanism computes the output as:

$$A(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{12}$$

where $Q \in \mathbb{R}^{L_Q \times d}$, $K \in \mathbb{R}^{L_K \times d}$, $V \in \mathbb{R}^{L_V \times d}$, d is the feature dimension, and q_i, k_i, v_i represent the i-th rows of Q, K, V respectively.

As shown in Figure 4, the Self-Attention mechanism computes the attention scores ($\alpha'_{1,i}$) for each query by comparing it with all keys in the sequence. These scores are then normalized using the *Softmax* function and used to compute a weighted sum of the value vectors (v^i). The resulting output vector (b^1) is given by:

$$b^1 = \sum_i \alpha'_{1,i} v^i \tag{13}$$

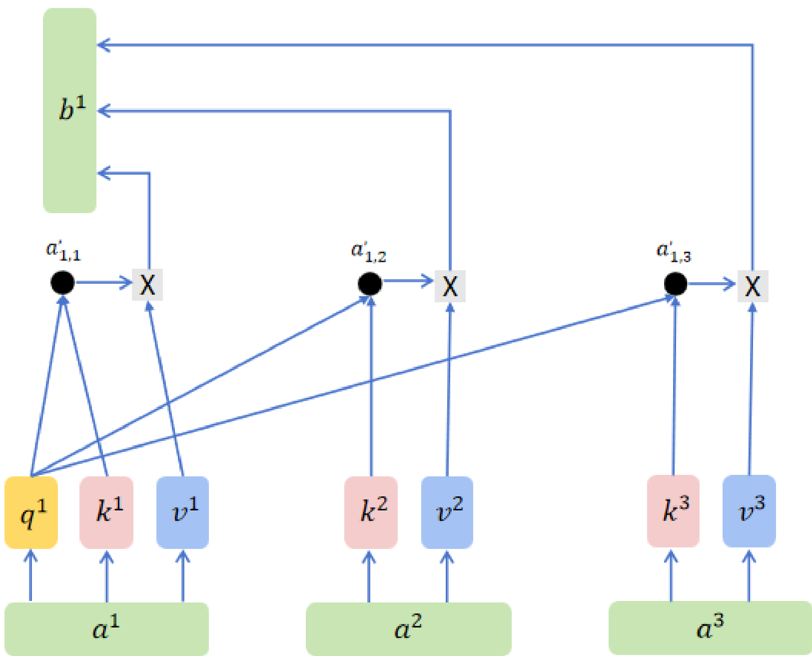

Figure 4. An illustration of the Self-Attention mechanism.

This equation illustrates how the attention mechanism aggregates information from all value vectors, weighted by their respective attention scores.

ProbSparse Self-Attention modifies this formulation by recognizing that attention scores are inherently sparse: not all queries strongly relate to all keys. By focusing only on the most relevant key–value pairs for each query, the mechanism reduces complexity while maintaining performance. The attention for the i-th query is reformulated as a kernel-smoothing operation:

$$A(q_i, K, V) = \sum_j \frac{k(q_i, k_j)}{\sum_j k(q_i, k_j)} v_j = \mathbb{E}_{p(k_j|q_i)}[v_j] \tag{14}$$

Here, the kernel function $k(q_i, k_j)$ is defined as an asymmetric exponential kernel:

$$k(q_i, k_j) = \exp\left(\frac{q_i^\top k_j}{\sqrt{d}}\right) \tag{15}$$

and $p(k_j|q_i)$ is the normalized probability distribution. By focusing on these probabilistically significant pairs, ProbSparse reduces the computational complexity. To identify the sparsity in attention weights, ProbSparse uses the Kullback–Leibler (KL) divergence to measure the deviation between the distribution $p(k_j|q_i)$ and a uniform distribution. The KL divergence is defined as:

$$KL(q||p) = \ln L_K + \sum_{l=1}^{L_K} \exp\left(\frac{q_i^\top k_l}{\sqrt{d}}\right) - 1 - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{q_i^\top k_j}{\sqrt{d}} \tag{16}$$

Removing the constant terms, the sparsity measure for the i-th query is defined as:

$$M(q_i, K) = \ln \sum_{l=1}^{L_K} \exp\left(\frac{q_i^\top k_l}{\sqrt{d}}\right) - \frac{1}{L_K} \sum_{j=1}^{L_K} \frac{q_i^\top k_j}{\sqrt{d}} \tag{17}$$

The first term is the Log-Sum-Exp (LSE) operation, highlighting the strongest key–query relationships, while the second term reflects the average similarity across all keys. A larger $M(q_i, K)$ value indicates greater sparsity in attention weights. Using the sparsity measure $M(q_i, K)$, ProbSparse Self-Attention selects the top $u = c \ln L_Q$ queries, where c is a constant factor. These queries form the subset $\hat{Q}$, which is used to compute the attention with reduced complexity:

$$A(Q, K, V) = \text{softmax}\left(\frac{\hat{Q}K^\top}{\sqrt{d}}\right)V \tag{18}$$

This formulation ensures that only the most important query–key interactions are retained. As a result, ProbSparse Self-Attention achieves a computational complexity of $O(L_Q \ln L_Q)$ and a memory usage of $O(L_K \ln L_Q)$, making it suitable for long-sequence modeling.

3.2.4. Dilated Causal Convolution

Dilated Causal Convolution is a technique that combines causal convolution and dilated convolution. In causal convolution, it is ensure that the model's predictions depend only on the current or past input data, without using future information. This is crucial for time-series forecasting tasks, as future data cannot be 'seen' when predicting current values. Dilated convolution, on the other hand, introduces spacing between the convolutional kernel elements, allowing the model to handle longer time spans of data without increasing computational cost. The combination of these two techniques enables the model to effectively capture long-term temporal dependencies, thereby improving forecasting accuracy for time series.

Dilated convolutions address this limitation by introducing a dilation factor d, which determines the spacing between input values used in the convolution [37]. The formula for a dilated convolution becomes:

$$(y_t) = \sum_{i=0}^{k-1} w_i \cdot x_{t-d\cdot i} \tag{19}$$

where d is the dilation factor. For example, with $d = 2$, the convolution skips every other input value, effectively doubling the receptive field. This allows the model to cover a larger context without increasing the number of parameters.

The receptive field of a dilated convolution is given by:

$$\text{Receptive Field} = (k-1) \cdot d + 1 \tag{20}$$

This formula shows how the dilation factor d exponentially increases the receptive field, enabling the model to capture long-range dependencies efficiently. In time-series modeling, causality is a critical requirement. A causal convolution ensures that the output y_t at time t only depends on the inputs $x_t, x_{t-1}, \ldots$, and not on any future inputs $x_{t+1}, x_{t+2}, \ldots$. This is achieved by masking future inputs during the convolution operation. Combining causality with dilation, the formula for a dilated causal convolution becomes:

$$(y_t) = \sum_{i=0}^{k-1} w_i \cdot x_{t-d\cdot i}, \quad \text{where } t - d \cdot i \geq 0 \tag{21}$$

This masking ensures that the model adheres to the causality constraint, making it suitable for autoregressive tasks. To capture increasingly larger contexts, multiple layers of dilated causal convolutions can be stacked. At each layer, the dilation factor is typically increased exponentially (e.g., $d = 1, 2, 4, \ldots$), which further expands the receptive field without increasing the computational cost significantly. The effective receptive field for L layers is:

$$\text{Effective Receptive Field} = \sum_{l=0}^{L-1} (k-1) \cdot d_l + 1 \tag{22}$$

where d_l is the dilation factor at layer l.

Dilated causal convolution combines the strengths of dilated convolution and causal convolution [25], ensuring that the model effectively captures long-term dependencies while maintaining temporal order. By expanding the receptive field of the convolution kernel, dilated convolution excels at extracting key information over extended time spans, while mitigating the impact of short-term fluctuations on long-term trends.

Figure 5 illustrates the operation of dilated causal convolution, where the kernel size is set to $K = 2$ and the dilation factor d is configured as $(1, 2, 4)$. When $d = 1$, the kernel slides sequentially, processing adjacent input points. The causality constraint ensures that the convolution output at each time step depends solely on the current and previous inputs. When $d = 2$, the convolution operation skips one input point, leaving a gap of one point between the inputs corresponding to each convolution output. This expands the receptive field of the neuron, allowing the convolution to capture information over a longer time span without increasing computational cost. When $d = 4$, the convolution operation skips three input points, further enlarging the receptive field and enabling the model to capture a broader range of input data. By introducing the dilation factor, dilated causal convolution effectively extracts critical features from long time series while maintaining temporal causality. This ensures that predictions at each time step depend only on past information, avoiding any leakage of future data.

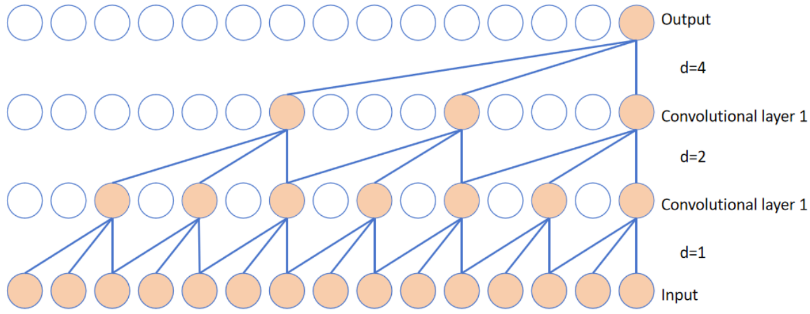

Figure 5. Dilated causal convolution.

3.2.5. Multi-Scale Feature Extraction Mechanism

To enhance the feature extraction capability of DCSInformer across different time scales, the model adopts a multi-scale feature extraction mechanism [38]. Specifically, the input time series X is divided into three distinct parts: the full series, the half series, and the quarter series. These sub-series correspond to features extracted at different time granularities. Assuming the total time steps of the input series is T, the three series are represented as follows:

$$\text{Full sequence: } X_{\text{full}} = X \tag{23}$$

$$\text{Half sequence: } X_{\text{half}} = \{x_1, x_3, \ldots, x_{T/2}\} \tag{24}$$

$$\text{Quarter sequence: } X_{\text{quarter}} = \{x_1, x_5, \ldots, x_{T/4}\} \tag{25}$$

These sub-series are processed independently through Multi-Head ProbSparse Self-Attention, which prioritizes computations by selecting the most relevant query–key pairs for each sub-series. While the model reduces the computational burden, it retains key contextual information from the data at each granularity, ensuring that important temporal dependencies are effectively captured. By combining features from all three series, the model achieves a comprehensive representation of the input time series, capable of addressing variations across multiple time scales.

For each sub-series X_{sub}, the attention mechanism computes the weighted sum of the values V, given the queries Q and keys K:

$$A(Q, K, V) = \text{softmax}\left(\frac{\hat{Q}K^{\top}}{\sqrt{d}}\right)V \tag{26}$$

Here, d is the feature dimension. The computation is independently performed for X_{full}, X_{half}, and X_{quarter}.

After extracting features from each sub-series, the outputs are concatenated and fused to form the final representation:

$$H = [H_{\text{full}}, H_{\text{half}}, H_{\text{quarter}}] \tag{27}$$

where $H_{\text{full}}, H_{\text{half}}, H_{\text{quarter}}$ are the attention-weighted representations for the respective sub-series.

Compared to the traditional Informer model, DCSInformer demonstrates significant improvements in capturing temporal dependencies across different time scales. By introducing the dilated causal convolution and multi-scale extraction mechanisms, the model enhances its ability to focus on key features over longer time spans while maintaining sensitivity to short-term variations [39]. This capability is especially valuable for medium-and long-term predictions, where balancing short-term dynamics and long-term trends is critical for improving forecast accuracy.

3.2.6. ETS Model

The ETS model is a statistical model widely used in time-series forecasting, especially for processing data with obvious seasonality and trend [40]. The ETS model can capture the patterns in the data more accurately by decomposing the time series into three parts: error, trend, and seasonality. In ETS (AAM), AAM means:

- Error: The error is additive (A), which is suitable for situations where the error remains relatively stable at different levels.
- Trend: The trend component is additive (A), which is suitable for time series with relatively linear trend changes.
- Seasonality: The seasonal component is multiplicative (M), which is applicable when seasonal fluctuations change with changes in the level.

The ETS (AAM) model has good adaptability and flexibility. It uses the maximum likelihood method to estimate parameters. Especially in the sales forecast of FMCG, it can effectively capture seasonal fluctuations and provide accurate sales forecasts.

4. Experimental Results and Analysis

4.1. Dataset Description

The dataset used in this study originates from a real-world FMCG company, encompassing monthly sales data from January 2018 to December 2023, covering a total of 60 months and over 120,000 records. It provides a detailed perspective on FMCG market dynamics.

The core features of the dataset include sales volume (target variable), price, promotion, seasonal factors, regional attributes, and macroeconomic indicators. Specifically, sales volume represents the number of units sold for each product in a given time period, while price refers to the average selling price of each product category. Promotion captures the number of promotional activities conducted during the period. Seasonal factors include time-related features such as months and holidays, regional attributes cover demographic and economic information such as population density and average income, and macroeconomic indicators include metrics such as the Consumer Confidence Index. These features collectively constitute the multi-dimensional information in the dataset, supporting the forecasting tasks. Figure 6 highlights the performance of various product categories, while Figure 7 illustrates sales performance across different regions.

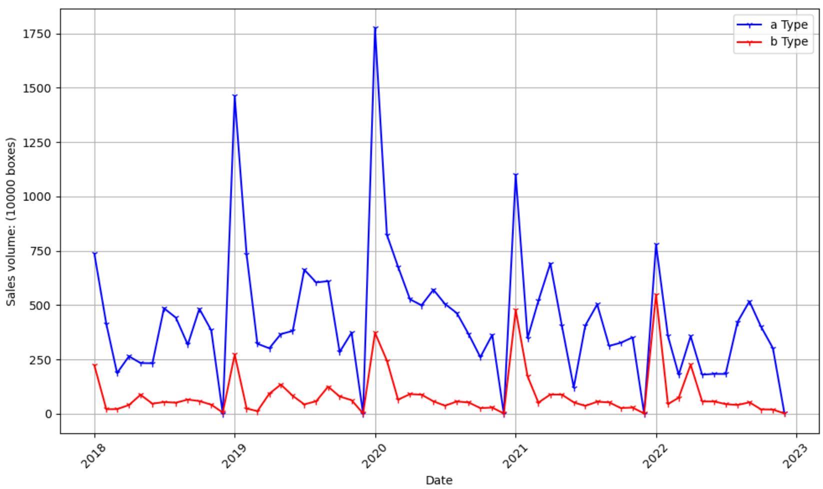

Figure 6. Monthly sales trend chart of different types.

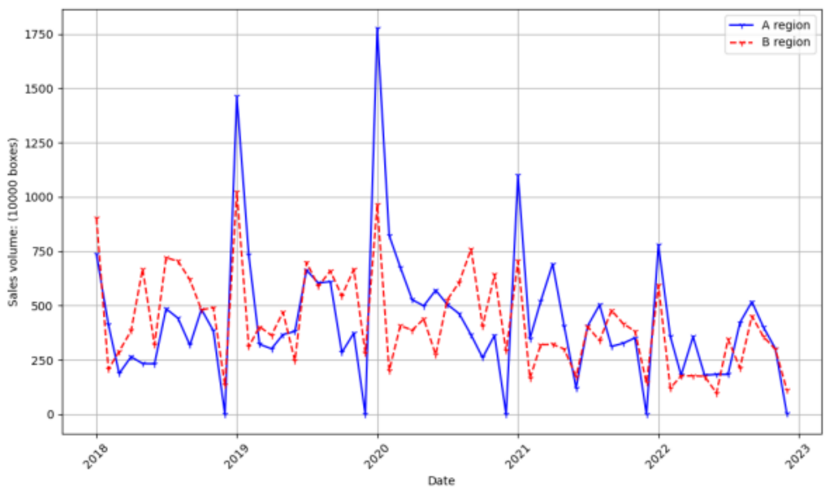

Figure 7. Monthly sales trend in different regions.

The dataset exhibits significant seasonality and high variability across regions and product categories, making it an ideal benchmark for evaluating forecasting models.

4.2. Normalization of Data

In the process of data collection, incomplete data are often encountered. To ensure data integrity and improve the prediction accuracy of the model, this study employs interpolation methods to address missing values. Simultaneously, to simplify model computation and enhance efficiency, the min–max normalization method is applied to scale the data to the range of [0, 1]. The normalization formula is as follows:

$$X' = \frac{X - \min}{\max - \min} \tag{28}$$

Here, the maximum and minimum values in the sample data are denoted as max and min, respectively.

4.3. Evaluation Indicators

This study evaluates model performance using four metrics after logarithmic transformation: MSE, MAE, RMSE, MAPE, and IoA. The transformation reduces the impact of large magnitude differences, improving evaluation robustness.

The formula for IoA is:

$$IoA = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)^2} \tag{29}$$

The formula for MSE is:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (\log(\hat{y}_i + 1) - \log(y_i + 1))^2 \tag{30}$$

The formula for MAE is:

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\log(\hat{y}_i + 1) - \log(y_i + 1)| \tag{31}$$

The formula for RMSE is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\log(\hat{y}_i + 1) - \log(y_i + 1))^2} \tag{32}$$

The formula for MAPE is:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\log(\hat{y}_i+1)-\log(y_i+1)}{\log(y_i+1)}\right| \times 100\% \tag{33}$$

Here, $\hat{y}_i$ represents the predicted value of the model, y_i is the actual value, and n is the total number of samples. By applying a logarithmic transformation ($\log(x+1)$), the metrics adjust for differences in scale and magnitude, making the evaluation more robust to large deviations and extreme values. Logarithmic transformation offers several advantages for FMCG sales data. It helps reduce scale differences between data points, especially when the data spans a wide range of values. This makes the loss functions, such as MSE and MAE, more stable by minimizing the impact of extreme values. Sales data often have a skewed distribution due to promotional spikes, and the logarithmic transformation brings the data closer to a normal distribution, which benefits models assuming normality. The transformation also improves model stability, reducing the influence of outliers and ensuring robustness. Additionally, it is effective for handling seasonal variations, allowing the model to capture cyclical trends more accurately.

4.4. Result Analysis

To enhance the performance of the STL-DCSInformer-ETS model, hyperparameter optimization was carried out using Grid Search and Bayesian Optimization methods. Initially, Grid Search was employed to exhaustively explore a predefined hyperparameter space, adjusting parameters such as the learning rate, number of epochs, and batch size. The learning rate was tested across values of [0.0001, 0.001, 0.01], batch size was varied between [16, 32, 64, 128], and the number of epochs was set to [50, 100, 150]. Each combination was evaluated based on the Mean Squared Error (MSE) on the validation set, ensuring a thorough initial search that captured a wide range of possibilities. Following this, Bayesian Optimization was applied to refine the results obtained from Grid Search. This method allowed for a more efficient exploration of the hyperparameter space by leveraging a probabilistic model (Gaussian Process) to prioritize the most promising parameter combinations. The learning rate was narrowed to [0.0005, 0.001] based on prior performance, while the batch size was fine-tuned between 32 and 64 to strike a balance between training stability and efficiency. The objective of this optimization was to minimize MSE on the validation set, ultimately converging on the optimal parameter settings, as presented in Table 4. These optimizations led to a significant improvement in the model's predictive accuracy, enhancing both its robustness and performance across various data scenarios.

Table 4. Parameter settings.

Hyperparameters	**Numeric**
Learning Rate	0.001
Batch size	32
Number of training rounds	100
Model size	512
Encoder input step size	72
Decoder input step size	36
Early Stopping Strategy	3
Number of attention heads	8
Optimizer	Adam
Activation Function	Leaky ReLU

For the forecasting task, the window size was set to 12 months. The forecasting method used in this study is a direct multi-step prediction approach, where the model predicts multiple time steps at once, without relying on recursive predictions. Each prediction step is independent, and the model generates all predictions for the forecast horizon simultaneously. To further prevent overfitting, early stopping is used during training. The process is halted once the validation error stops improving.

4.4.1. Sensitivity Analysis of Feature Importance

In this experiment, a sensitivity analysis was conducted to validate the effectiveness of the previous feature selection results. By systematically perturbing the features—removing certain features or adding noise to simulate real-world variability—the impact of each feature on the model's forecasting accuracy was evaluated, and model performance changes were measured using evaluation metrics.

Table 5 presents the results of the feature analysis, validating the effectiveness of the previous feature selection. The seasonal feature and holiday feature have a significant impact on model accuracy, with all evaluation metrics increasing when these features are modified. In contrast, the packaging type has a relatively small effect on model performance, further validating its insignificance in prediction and supporting the decision to discard it during the feature selection phase.

Table 5. Sensitivity analysis of feature importance.

Feature	**Change in MSE**	**Change in MAE**	**Change in RMSE**	**Change in MAPE (%)**	**Change in IoA**
Seasonal feature	+0.011	+0.012	+0.015	+0.23%	−0.006
Holiday feature	+0.011	+0.011	+0.012	+0.22%	−0.006
Packaging type	0.000	0.000	+0.00	0.00%	0.000
Supply-chain delay	+0.002	+0.002	+0.003	0.01%	−0.002
Rainfall days	+0.001	0.000	+0.001	0.00%	−0.001

4.4.2. Ablation Experiment

To validate the effectiveness and reliability of the model components, the Informer model was selected as the baseline for ablation experiments. The experimental results are shown in the last two rows of Table 6. As observed from the table, the introduction of the hybrid models led to a consistent improvement in all prediction metrics. Specifically, the MAE decreased from 0.32 to 0.29, with corresponding decreases in MSE and RMSE. These results demonstrate the effectiveness of the hybrid models in enhancing predictive performance. With the introduction of the DCSInformer, all metrics were further improved, highlighting the significant advantage of DCSInformer in capturing long-term dependencies and reducing errors in time-series forecasting tasks.

Table 6. Model performance comparison.

Model	**MSE**	**MAE**	**RMSE**	**MAPE**	**IoA**
LSTM	0.4449	0.3835	0.6670	19.90%	0.8304
Transformer	0.3265	0.3182	0.5714	16.96%	0.8880
Preformer	0.2130	0.2350	0.4615	13.56%	0.9011
Informer	0.2858	0.3217	0.5346	16.14%	0.9043
STL-Informer-ETS	0.2599	0.2930	0.5098	15.27%	0.9288
STL-DCSInformer-ETS	**0.2092**	**0.2211**	**0.4574**	**13.33%**	**0.9313**

4.4.3. Comparative Experiment

In this experiment, three classic and commonly used time-series forecasting models are selected: LSTM, Transformer, and Preformer. These models have wide applications and show good performance in the field of time-series forecasting. To comprehensively evaluate the performance of the proposed STL-DCSInformer-ETS model, this study compares it with the three classic models. The aim is to conduct an in-depth analysis through experimental comparisons, exploring the accuracy and effectiveness of each model in handling time-series data.

The Figure 8 shows the performance comparison of different models in the medium and long-term sales forecast of fast-moving products. The performance comparison of different models for medium- and long-term sales forecasting of fast-moving products is shown in Table 6. From the table, it is evident that the STL-DCSInformer-ETS model outperforms all other models across all evaluation metrics, demonstrating superior capability in capturing sales trends and fluctuations. Specifically, the MSE of the STL-DCSInformer-ETS model is 0.2092, and the Mean Absolute Error (MAE) is 0.2211, both significantly lower than other models. Additionally, its RMSE and MAPE are 0.4574 and 13.33%, respectively, highlighting its outstanding accuracy and reliability. Furthermore, the IoA score of 0.9313 emphasizes its exceptional consistency with the observed data. The Preformer model, as seen in the third row of Table 6, achieves competitive results, with an MSE of 0.2130, an MAE of 0.2350, and an RMSE of 0.4615. Its MAPE is 13.56%, and it attains an IoA of 0.9011. While the Preformer model demonstrates strong performance, it falls slightly behind STL-DCSInformer-ETS in all metrics, particularly in capturing subtle sales fluctuations. The Transformer model, as shown in the second row of Table 6, achieves moderate results with an MSE of 0.3265 and an MAE of 0.3182, indicating limited effectiveness in medium- and long-term forecasting tasks. As seen in the first row of Table 6, the LSTM model exhibits the weakest performance, with an MSE of 0.4449 and an MAE of 0.3835. This shows that compared with the advanced model, its ability to capture sales trends and fluctuations has great limitations.

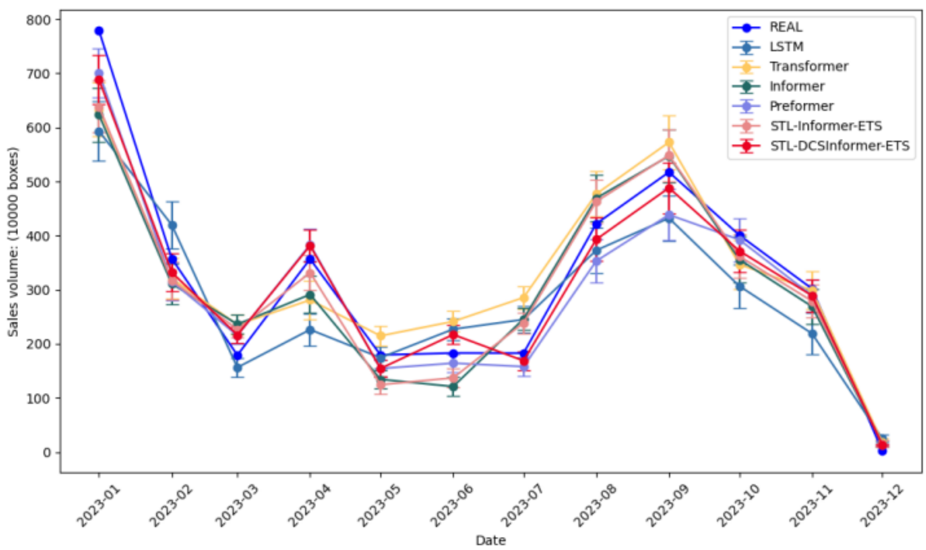

Figure 8. Comparison of ablation and comparative experiment prediction results.

In summary, the STL-DCSInformer-ETS model demonstrates significant advantages in forecasting medium- and long-term sales trends and capturing complex seasonal variations.

5. Generalization and Limitations of the Model

Although the STL-DCSInformer-ETS model has demonstrated strong performance in forecasting sales for FMCG, its applicability to other products and industries requires further validation. The model incorporates the flexible components of STL, DCSInformer, and ETS, which are theoretically applicable to any time-series data with seasonal and trend components. This suggests the model has potential for use in other product categories or industries, especially those with similar temporal patterns and fluctuations. However, different industries or products may present specific market factors and external variables that could affect the model's performance. For instance, industries such as pharmaceuticals and electronics may require additional adjustments, including consideration of regulatory factors in the pharmaceutical sector or price elasticity and technological changes in electronics. Therefore, while the STL-DCSInformer-ETS model performs well in this study, it may require further refinement when applied to these or other sectors. The model does have certain limitations. Its reliance on historical data means that its performance could be affected by the quality and completeness of the input data. Missing or noisy data can reduce predictive accuracy. Furthermore, while the model effectively captures temporal dependencies and seasonal patterns, its adaptability to entirely new industries or domains still needs further exploration. Industries with high volatility or non-seasonal trends, such as technology, may require additional feature engineering or model adjustments to achieve comparable performance. Finally, computational constraints during hyperparameter optimization may limit the model's application in resource-constrained or real-time prediction scenarios. Future work should focus on simplifying the optimization process and evaluating the model's performance in such environments.

6. Conclusions

This study proposes the STL-DCSInformer-ETS hybrid model for medium- and long-term sales forecasting of FMCG. By combining STL decomposition, the enhanced DCSInformer model, and the ETS model, the proposed approach effectively captures complex temporal dependencies, seasonal fluctuations, and long-term trends. Feature engineering, using the RF-RFE method, played a crucial role in identifying and optimizing key predictive features across time, geography, economy, product, and weather dimensions, thus providing a solid foundation for accurate forecasting. The STL method decomposes the data into trend, seasonal, and residual components, simplifying the data structure. The enhanced DCSInformer model incorporates dilated causal convolutions and multi-scale feature extraction to better capture both long-term trends and short-term fluctuations. Meanwhile, the ETS model specializes in modeling seasonal variations, further refining the forecasting accuracy. This integration enables the model to effectively adapt to features across multiple time scales, ensuring a comprehensive understanding of sales patterns. The experimental results show that the STL-DCSInformer-ETS model outperforms traditional methods, including LSTM, Transformer, and Informer models, across key metrics such as MSE, MAE, RMSE, MAPE. The STL-DCSInformer-ETS hybrid model stands out in FMCG sales forecasting by addressing the unique challenges of high volatility, short-term promotional effects, and seasonal trends. This demonstrates the model's superiority in handling the volatility of FMCG sales, making it a valuable tool for supply-chain optimization, inventory management, strategic decision-making, and providing policymakers with actionable insights for policy formulation in economic planning and market regulation Future research could enhance the model by incorporating real-time consumer sentiment data or broader macroeconomic indicators, which could provide more dynamic insights. Additionally, exploring the application of the model in other industries, such as electronics, retail, and pharmaceuticals, could validate its versatility and generalizability. Additionally, future research can integrate real-time data. Integrating real-time data requires efficient algorithms and data-processing techniques, and adapt the model to different market scenarios to ensure its robustness and accuracy in diverse environments.

Author Contributions: Conceptualization, L.H. and J.Z.; methodology, Y.M. and J.Z.; software, J.Z. and Y.M.; validation, J.Z. and Y.M.; writing—original draft preparation, Y.M.; writing—review and editing, L.H. and J.Z.; funding acquisition and resources, L.H. and J.Z.; supervision, L.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "Leading Goose" R&D Program of Zhejiang (Grant No. 2022C01238).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Nozari, H.; Fallah, M.; Szmelter-Jarosz, A.; Krzemiński, M. Analysis of security criteria for IoT-based supply chain: A case study of FMCG industries. *Cent. Eur. Manag. J.* **2021**, *29*, 149–171. [CrossRef]
2. Satyanarayana, K.V.; Ekambaram, V.G.; Kakumani, V.R. Sales Prediction in the Fast-Moving Consumer Goods (FMCG) Sector, Wielding Machine Learning Techniques. Available online: https://ssrn.com/abstract=4608641 (accessed on 5 January 2025).
3. Sun, Y. Sales Prediction Based on Machine Learning Approach. In Proceedings of the 2024 9th International Conference on Social Sciences and Economic Development (ICSSED 2024), Beijing, China, 22–24 March 2024; Atlantis Press: Amsterdam, The Netherlands, 2024; pp. 1016–1023.
4. Sun, C.; Chen, Z.; Qin, Y.; Wang, B. Multi-step Time Series Forecasting Based on Informer-XGBoost-GA. *J. Phys. Conf. Ser.* **2022**, *2333*, 012009. [CrossRef]
5. Paciarotti, C.; Torregiani, F. The logistics of the short food supply chain: A literature review. *Sustain. Prod. Consum.* **2021**, *26*, 428–442. [CrossRef]
6. Sukati, I.; Hamid, A.B.; Baharun, R.; Yusoff, R.M. The study of supply chain management strategy and practices on supply chain performance. *Procedia-Soc. Behav. Sci.* **2012**, *40*, 225–233. [CrossRef]
7. Waller, M.A.; Fawcett, S.E. Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *J. Bus. Logist.* **2013**, *34*, 77–84. [CrossRef]
8. Doshi, M. A Comparison of Data Mining Approaches for Forecasting Sales of FMCG Food Products. In Proceedings of the 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 6–8 July 2023; IEEE: Piscataway, NJ, USA, 2023, pp. 1–7.
9. Lu, T. Research on sales forecasting in e-commerce industry for imbalanced classification data. *Sci. Technol. Eng. Chem. Environ. Prot.* **2024**, *1*. [CrossRef]
10. Seethapathy, K. Unlocking Inventory Efficiency: Harnessing Machine Learning for Sales Surge Prediction. *Int. J. Supply Chain Logist.* **2024**, *8*, 57–66. [CrossRef]
11. Wang, L.; Zhang, X. Livestream sales prediction based on an interpretable deep-learning model. *Sci. Rep.* **2024**, *14*, 20594. [CrossRef] [PubMed]
12. Wu, J.; Liu, H.; Yao, X.; Zhang, L. Unveiling consumer preferences: A two-stage deep learning approach to enhance accuracy in multi-channel retail sales forecasting. *Expert Syst. Appl.* **2024**, *257*, 125066. [CrossRef]
13. Jovanovic, A.; Bacanin, N.; Jovanovic, L.; Damaševičius, R.; Antonijevic, M.; Zivkovic, M.; Kaljevic, J.; Dobrojevic, M. Performance evaluation of metaheuristics-tuned recurrent networks with VMD decomposition for Amazon sales prediction. *Int. J. Data Sci. Anal.* **2024**, *12*, 1–19 . [CrossRef]
14. Zhu, Q.; Han, J.; Chai, K.; Zhao, C. Time series analysis based on informer algorithms: A survey. *Symmetry* **2023**, *15*, 951. [CrossRef]
15. Dong, P.; Zuo, Y.; Chen, L.; Li, S.; Zhang, L. VSNT: Adaptive Network Traffic Forecasting via VMD and Deep Learning with SCI-Block and Attention Mechanism. In Proceedings of the 2023 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Wuhan, China, 21–24 December 2023; IEEE: Piscataway, NJ, USA, 2023, pp. 104–109.

16. Li, Q.; Yao, C.; Yao, X.; Zhou, Z.; Ren, K. Time Series Prediction of Reservoir Bank Slope Deformation Based on Informer and InSAR: A Case Study of Dawanzi Landslide in the Baihetan Reservoir Area, China. *Remote Sens.* **2024**, *16*, 2688. [CrossRef]
17. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on Artificial Intelligence, Virtually, 2–9 February 2021; Volume 35, pp. 11106–11115.
18. Verdonck, T.; Baesens, B.; Óskarsdóttir, M.; vanden Broucke, S. Special issue on feature engineering editorial. *Mach. Learn.* **2024**, *113*, 3917–3928. [CrossRef]
19. Thejovathi, M.; Rao, D.M.C.S. Evaluating the performance of xgboost and gradient boost models with feature extraction in fmcg demand forecasting: A feature-enriched comparative study. *J. Theor. Appl. Inf. Technol.* **2024**, *102*, 4158–4163.
20. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in time series: A survey. *arXiv* **2022**, arXiv:2202.07125.
21. Jackson, I.; Saenz, M.J.; Li, Y.; Moreno, M.S.R. Synchromodal Supply Chains for Fast-Moving Consumer Goods. *Appl. Sci.* **2023**, *13*, 3119. [CrossRef]
22. Rose, N.; Dolega, L. It's the weather: Quantifying the impact of weather on retail sales. *Appl. Spat. Anal. Policy* **2022**, *15*, 189–214. [CrossRef]
23. Jeon, H.; Oh, S. Hybrid-recursive feature elimination for efficient feature selection. *Appl. Sci.* **2020**, *10*, 3211. [CrossRef]
24. Darst, B.F.; Malecki, K.C.; Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* **2018**, *19*, 65. [CrossRef] [PubMed]
25. Tully, J.; Haight, R.; Hutchinson, B.; Huang, S.; Lee, J.Y.; Katipamula, S. Dilated causal convolutional neural networks for forecasting zone airflow to estimate short-term energy consumption. *Energy Build.* **2023**, *286*, 112890. [CrossRef]
26. Mishra, A.K.; Renganathan, J.; Gupta, A. Volatility forecasting and assessing risk of financial markets using multi-transformer neural network based architecture. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108223. [CrossRef]
27. Gao, M.; Yan, X.; Xiao, H. Dynamic window-based volatility adjustment for multivariate time series. *Appl. Energy* **2023**, *331*, 120200.
28. Tashman, L.; Tesfaye, Y. Time-window selection for forecasting models: A novel weighted framework. *Int. J. Forecast.* **2022**, *38*, 815–829.
29. Mahmoud, A.; Mohammed, A. Leveraging hybrid deep learning models for enhanced multivariate time series forecasting. *Neural Process. Lett.* **2024**, *56*, 223. [CrossRef]
30. Abbasimehr, H.; Behboodi, A.; Bahrini, A. A novel hybrid model to forecast seasonal and chaotic time series. *Expert Syst. Appl.* **2024**, *239*, 122461. [CrossRef]
31. Zeng, H.; Zhang, H.; Guo, J.; Ren, B.; Cui, L.; Wu, J. A novel hybrid STL-transformer-ARIMA architecture for aviation failure events prediction. *Reliab. Eng. Syst. Saf.* **2024**, *246*, 110089. [CrossRef]
32. Yuan, H.; Ma, X.; Ma, M.; Ma, J. Hybrid framework combining grey system model with Gaussian process and STL for CO2 emissions forecasting in developed countries. *Appl. Energy* **2024**, *360*, 122824. [CrossRef]
33. Yuan, Z.; Lin, X.; Xu, Y.; Zhao, J.; Du, N.; Cai, X.; Li, M. Zenith Tropospheric Delay Forecasting in the European Region Using the Informer–Long Short-Term Memory Networks Hybrid Prediction Model. *Atmosphere* **2024**, *16*, 31. [CrossRef]
34. Li, J.; Sun, L.P.; Wu, X.; Tao, C. Enhancing Financial Time Series Forecasting with Hybrid Deep Learning: Ceemdan-Informer-Lstm Model. Available online: https://www.ssrn.com/abstract=4923415 (accessed on 10 October 2024).
35. Cui, Y.; Li, Z.; Wang, Y.; Dong, D.; Gu, C.; Lou, X.; Zhang, P. Informer model with season-aware block for efficient long-term power time series forecasting. *Comput. Electr. Eng.* **2024**, *119*, 109492. [CrossRef]
36. Liu, W.; Gu, Y.; Ge, Y. Multi-factor stock trading strategy based on DQN with multi-BiGRU and multi-head ProbSparse self-attention. *Appl. Intell.* **2024**, *54*, 5417–5440. [CrossRef]
37. Moreno-Pino, F.; Zohren, S. Deepvol: Volatility forecasting from high-frequency data with dilated causal convolutions. *Quant. Financ.* **2024**, *24*, 1105–1127. [CrossRef] [PubMed]
38. Zhang, Y.; Lu, Y.; Zhu, W.; Wei, X.; Wei, Z. Traffic sign detection based on multi-scale feature extraction and cascade feature fusion. *J. Supercomput.* **2023**, *79*, 2137–2152. [CrossRef]
39. Wang, W.; Li, Y.; Yan, X.; Xiao, M.; Gao, M. Enhancing liver segmentation: A deep learning approach with eas feature extraction and multi-scale fusion. *Int. J. Innov. Res. Comput. Sci. Technol.* **2024**, *12*, 26–34. [CrossRef]
40. Qi, L.; Li, X.; Wang, Q.; Jia, S. fETSmcs: Feature-based ETS model component selection. *Int. J. Forecast.* **2023**, *39*, 1303–1317. [CrossRef]