Methodological Framework

1. Overview of Forecasting Paradigms

Short-term retail demand forecasting sits at the intersection of statistical signal processing and machine learning. On one end of the spectrum, classical time-series models such as SARIMA and SARIMAX impose explicit parametric structure on the data, leveraging stationarity and autocorrelation to produce interpretable forecasts. On the other end, deep learning architectures such as LSTM and CNN–LSTM learn nonlinear representations directly from data, relaxing distributional assumptions at the cost of higher complexity and reduced transparency. Additive decomposition models like Facebook Prophet occupy a middle ground, encoding trend and seasonality with a predefined structure that is flexible yet remains interpretable. This study conducts a rigorous empirical comparison of these five prominent paradigms. Crucially, the evaluation extends beyond traditional point-forecast accuracy to assess the *distributional fidelity* of each model's predictions—a dimension critical for inventory and risk management but largely overlooked in comparative forecasting studies. This section formalizes the theoretical foundations of each model class and the comprehensive evaluation framework.

2. Statistical Time-Series Models

Statistical time-series models explicitly encode temporal dependence through autoregressive and moving-average operators applied to a (possibly differenced) series. Their main advantages are interpretability, relatively low data requirements, and the availability of well-established diagnostic tools. However, their linear structure can be restrictive when demand is driven by complex, nonlinear interactions.

2.1 Seasonal ARIMA (SARIMA)

Let $\{y_t\}$ denote a univariate time series observed at equally spaced intervals. The Seasonal Autoregressive Integrated Moving Average model, denoted as SARIMA$(p, d, q)(P, D, Q)_s$, combines non-seasonal autoregressive and moving-average terms with seasonal counterparts. Define the backshift operator B such that $By_t = y_{t-1}$. The non-seasonal differencing operator is $(1 - B)^d$ and the seasonal differencing operator is $(1 - B^s)^D$, where s is the seasonal period. The fully differenced series is:

$$w_t = (1 - B)^d (1 - B^s)^D y_t.$$

The general SARIMA model is then written as:

$$\Phi(B^s)\phi(B)w_t = \Theta(B^s)\theta(B)\varepsilon_t,$$

where ε_t is a white-noise process with zero mean and variance σ^2; $\phi(B)$ and $\theta(B)$ are the non-seasonal autoregressive (AR) and moving-average (MA) polynomials

of orders p and q; $\Phi(B^s)$ and $\Theta(B^s)$ are the seasonal AR and MA polynomials of orders P and Q. A key strength of SARIMA is its transparent mapping between past observations and future forecasts. However, its reliance on linear operators implies that sudden regime changes or nonlinear responses may not be fully captured.

2.2 SARIMA with Exogenous Regressors (SARIMAX)

SARIMAX extends SARIMA by allowing the conditional mean to depend on exogenous variables. Let X_t be a vector of exogenous regressors at time t, such as promotion indicators or calendar dummies. The SARIMAX model can be written as:

$$\Phi(B^s)\phi(B)w_t = \Theta(B^s)\theta(B)\varepsilon_t + \beta^\top X_t,$$

where β is a vector of regression coefficients. This formulation is advantageous in retail settings where demand is strongly affected by known external drivers. However, SARIMAX inherits the limitations of linear regression: effects are assumed additive and constant over time, and unobserved or misspecified factors can lead to biased estimates.

3. Deep Learning Forecasting Models

Deep learning models approach time-series forecasting as a sequence-learning problem, learning complex mappings from past trajectories to future outcomes through multiple nonlinear transformations.

3.1 Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory networks are a class of recurrent neural networks (RNNs) designed to mitigate vanishing-gradient problems. An LSTM cell maintains an internal memory state c_t and a hidden state h_t. Given the input vector x_t, previous hidden state h_{t-1}, and previous cell state c_{t-1}, the cell updates are:

$$\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f),\\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i),\\
\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c),\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t,\\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o),\\
h_t &= o_t \odot \tanh(c_t),
\end{aligned}$$

where $\sigma(\cdot)$ is the logistic sigmoid function, $\tanh(\cdot)$ is the hyperbolic tangent, $\odot$ denotes elementwise multiplication, and $W_\cdot, U_\cdot, b_\cdot$ are trainable parameters. The gated structure allows LSTM networks to adaptively learn which temporal patterns are relevant. Its main disadvantage is a large parameter space, increasing the risk of overfitting on small datasets.

3.2 CNN–LSTM Hybrid Architecture

The CNN–LSTM hybrid architecture combines convolutional neural networks (CNNs) with LSTMs to exploit both local and long-range temporal structure. A one-dimensional convolutional layer first applies learnable filters across sliding windows of the input sequence. For a filter j of length K, the convolutional operation at time t can be expressed as:

$$z_{t,j} = \sigma\left(\sum_{k=1}^{K} w_{j,k} x_{t-k+1} + b_j\right),$$

where $z_{t,j}$ is the j-th feature at time t, $w_{j,k}$ are learnable filter weights, b_j is a bias term, and $\sigma(\cdot)$ is a nonlinear activation function such as ReLU. The resulting feature map sequence $\{Z_t\}$ is then provided as input to an LSTM layer: $h_t = \mathrm{LSTM}(Z_t, h_{t-1}, c_{t-1})$. Theoretically, this hybrid can approximate both local and global temporal structures more efficiently than a pure LSTM.

4. Additive Decomposition Model: Facebook Prophet

Facebook Prophet is an additive time-series model designed to capture trend, seasonality, and holiday effects within a unified framework. It assumes that the observed series y_t can be decomposed as:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t,$$

where $g(t)$ represents the long-term trend component, $s(t)$ is a sum of seasonal components, $h(t)$ captures the effect of known events or holidays, and ε_t is an error term. Seasonality is represented using a Fourier series expansion:

$$s(t) = \sum_{n=1}^{N} \left[a_n \cos(2\pi n t/P) + b_n \sin(2\pi n t/P)\right],$$

where P is the seasonal period and a_n, b_n are parameters to be estimated. Prophet's decomposable structure offers interpretability and robustness to missing data, but its additive assumption may be restrictive in the presence of strong nonlinear interactions.

5. Point Forecast Evaluation Metrics

To compare models on their ability to predict the central tendency of demand, we employ three widely used point-forecast error metrics.

5.1 Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum_{t=1}^{n} |y_t - \hat{y}_t|,$$

where y_t denotes the actual observations, $\hat{y}_t$ denotes the corresponding forecasts, and n is the number of evaluation points. MAE measures the average absolute deviation and is robust to outliers.

5.2 Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}.$$

RMSE penalizes larger errors more heavily due to the squaring of deviations, which can be useful when occasional large errors are particularly costly.

5.3 Mean Absolute Percentage Error (MAPE)

$$\text{MAPE} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right|.$$

MAPE expresses forecast error as a percentage of the actual demand, facilitating comparison across products with different scales. However, it becomes unstable when y_t is close to zero.

6. Distribution-Based Evaluation for Predictive Reliability

Point-error metrics assess how close the forecasts are to the realized values on average but do not evaluate how well a model captures the overall *distribution* of demand. For operational decisions involving safety stock or service-level guarantees, the shape and calibration of the predictive distribution are as important as the point forecast. To address this critical gap, we complement standard error metrics with two rigorous distribution-based measures: the **Wasserstein distance** and the **Kolmogorov–Smirnov (KS) statistic**. Their application in a comparative forecasting study represents a novel contribution to the methodological evaluation framework.

6.1 First-Order Wasserstein Distance

The first-order Wasserstein distance (Earth Mover's Distance) quantifies the minimal "work" required to transform the predicted probability distribution into the observed empirical distribution. Let F and G denote the cumulative distribution functions (CDFs) of the predictive and empirical distributions, respectively. The continuous form is:

$$W_1(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)|\, dx.$$

For empirical samples $\{y_t\}$ and $\{\hat{y}_t\}$, sorted in ascending order as $y_{(i)}$ and $\hat{y}_{(i)}$, an empirical approximation is:

$$W_1 \approx \frac{1}{n} \sum_{i=1}^{n} |y_{(i)} - \hat{y}_{(i)}|.$$

Unlike MAE, which compares observations and forecasts pointwise in time, the Wasserstein distance compares the entire *empirical distributions*, making it sensitive to differences in location, spread, and shape. A smaller W_1 indicates that the predictive distribution more closely matches the observed distribution, which is critical for decisions depending on quantiles or tail probabilities.

6.2 Kolmogorov–Smirnov (KS) Statistic

The Kolmogorov–Smirnov statistic quantifies the maximum discrepancy between two cumulative distribution functions. Let F_n and G_n be the empirical CDFs of the predictive and actual samples. The KS statistic is defined as:

$$D = \sup_x |F_n(x) - G_n(x)|.$$

This statistic underlies the Kolmogorov–Smirnov test for the null hypothesis that F_n and G_n are generated from the same underlying distribution. A large value of D suggests that the predictive distribution is systematically misaligned with the observed distribution. In the context of forecast evaluation, the KS statistic provides a formal measure of distributional calibration.

7. Critical Theoretical Comparison and Novel Evaluation Synthesis

The models considered in this study represent complementary philosophical approaches to forecasting. SARIMA and SARIMAX are grounded in linear stochastic processes, offering transparency and diagnostic clarity but limited flexibility for complex nonlinearities. LSTM and CNN–LSTM architectures relax structural assumptions, learning flexible representations at the cost of interpretability and data hunger. Prophet offers a structured, interpretable middle ground but may be constrained by its additive form.

The choice of evaluation metrics fundamentally shapes conclusions about model performance. Traditional point-error metrics (MAE, RMSE, MAPE) answer the question of which model, on average, predicts the central tendency more accurately. They do not, however, indicate whether a model reliably captures the *variability and tail behavior* of demand that are crucial for risk-sensitive operational decisions.

Therefore, this study introduces a **dual-layer evaluation framework**. The first layer employs standard point-forecast metrics to assess central tendency accuracy. The second, more rigorous layer applies the Wasserstein distance and the Kolmogorov–Smirnov statistic to evaluate distributional fidelity. This dual approach provides a stricter, more operationally relevant standard: models are judged not only on how close their point forecasts are to realized values but on how well their implied *predictive distributions* match the empirical demand

distribution. From a methodological standpoint, this comprehensive framework is essential for aligning forecast assessment with the true objectives of retail supply chain management, where both average accuracy and distributional reliability are paramount for performance.