Regime-Dependent Performance of ARIMA and Modern Forecasting Methods: An Empirical Benchmark on Small-Scale Retail Demand Data

Anonymous authors
Paper under double-blind review

Abstract

Choosing between classical statistical models and modern machine learning for retail demand forecasting is a common but underspecified decision in practice. This paper presents a systematic empirical benchmark comparing six methods — ARIMA, SARIMA, Prophet, XGBoost, LSTM, and an ARIMA+XGBoost decomposition (Hybrid) — across five dataset sources spanning three countries and four demand regimes: (1) four real product-category series from a D-Mart retail store (India, daily, n=181), (2) five real product series from UCI Online Retail (UK, weekly, n=53), (3) 24 real series from the M4 Micro Monthly benchmark (international, n=68–197), (4) a weekly Walmart-calibrated store series (n=143), and (5) a daily intermittent demand series calibrated to M5 statistics (n=365, zero-fraction 32.9%). All models use fixed practitioner-default hyperparameters; all pairwise comparisons are tested with the Diebold-Mariano (DM) test (Diebold & Mariano, 1995).

Within the scope of these datasets — short series ($n \leq 200$), daily or weekly retail demand, limited seasonality — we observe consistent regime-dependent performance. On the real D-Mart data (CV$\approx$0.021, |AC(1)|$<$0.21), classical ARIMA matches or outperforms all complex methods under 5-fold walk-forward cross-validation and across 17 of 20 individually tested SKUs. On high-variance data (UCI, Walmart; CV$=$ 0.32–0.59), the Hybrid (ARIMA+XGBoost) reduces RMSE by up to 49.6% over ARIMA. On 24 real M4 Micro Monthly series (mean AC(1)$=$ 0.88), AR(1) outperforms the naive mean model in 21 of 24 cases (mean RMSE ratio 0.62), confirming that AC(1) is a stronger regime predictor than CV alone. On intermittent demand (M5), all methods converge. We propose the coefficient of variation (CV) as a preliminary heuristic model-selection indicator — explicitly scoped as a starting point derived from limited datasets, not a validated universal rule — and support it with a sensitivity analysis and synthetic experiment. All results are statistically significant ($p<0.001$, DM test). Code, data, and a reproducibility checklist are publicly available.[1]

1 Introduction

Accurate demand forecasting is foundational to retail operations, determining inventory decisions, promotional planning, and supply chain coordination (Box et al., 2015). Despite decades of work and increasing deployment of machine learning in industry, a critical practical question remains inadequately studied: *under what conditions do modern machine learning methods actually outperform classical statistical models on small-scale retail data?*

Existing large-scale benchmarks — M4 (Makridakis et al., 2020) and M5 (Makridakis et al., 2022) — evaluate thousands of series where neural methods have demonstrated competitive performance. Recent Transformer-based approaches (Wen et al., 2023; Ekambaram et al., 2024) achieve state-of-the-art results on these benchmarks. However, most real-world retail deployments involve far smaller datasets: a single store's six months

[1] https://github.com/Aarav500/retail-forecasting-benchmark

of daily records, a regional chain's two years of weekly sales. In this regime, model complexity becomes a liability: overfitting to noise can degrade performance relative to simple baselines (Makridakis et al., 2018).

This paper makes five contributions:

1. **Multi-source real-world benchmark**: We evaluate across five dataset sources spanning three countries — D-Mart (India, daily), UCI Online Retail (UK, weekly), M4 Micro (international, monthly), Walmart-calibrated (US, weekly), and M5-calibrated (intermittent) — covering 34 individual time series across four distinct demand regimes.
2. **Standardized six-model comparison**: ARIMA, SARIMA, Prophet, XGBoost, LSTM, and Hybrid are evaluated with fixed practitioner-default hyperparameters and strict temporal train/test splits.
3. **ARIMA+XGBoost Hybrid evaluation**: We evaluate the established ARIMA+ML decomposition (Zhang, 2003) using XGBoost as the residual corrector, providing the first DM-tested evaluation on real Indian retail data.
4. **Statistical significance via DM testing**: All pairwise comparisons use the Diebold-Mariano test (Diebold & Mariano, 1995), which is frequently omitted from applied forecasting benchmarks (Hyndman & Athanasopoulos, 2021).
5. **Heuristic CV criterion with sensitivity analysis**: We propose CV as a preliminary model-selection indicator and support it with a sensitivity analysis and synthetic experiment, while explicitly scoping it as a heuristic derived from limited data.

2 Background and Related Work

2.1 Classical time series methods

Box et al. (2015) established the ARIMA framework through autoregressive modeling, differencing, and moving average components. Seasonal extensions (SARIMA) address periodic retail cycles (Hyndman & Athanasopoulos, 2021). Despite their age, these models remain competitive in small-data regimes (Makridakis et al., 2018). The M4 competition (Makridakis et al., 2020) — covering 100,000 heterogeneous time series — found that simple methods outperform complex models on short series, a key motivation for our low-SNR regime analysis.

2.2 Machine learning for forecasting

Chen & Guestrin (2016) with lag feature engineering demonstrates strong performance on tabular time series and won multiple M5 competition tracks (Makridakis et al., 2022). Prophet (Taylor & Letham, 2018) provides an additive decomposition model designed for business time series with irregular seasonality. LSTM networks (Hochreiter & Schmidhuber, 1997) capture long-range temporal dependencies but require substantial data for reliable training (Benidis et al., 2023).

Transformer-based architectures have recently achieved state-of-the-art results on large benchmarks. Wen et al. (2023) survey Transformer variants for time series, finding consistent gains on datasets with thousands of training points. Ekambaram et al. (2024) introduce Tiny Time Mixers with strong zero-shot performance. Critically, as Makridakis et al. (2018) demonstrate, gains from complex models diminish sharply as series length decreases. Kaur et al. (2024) benchmark Transformers on the full M5 dataset (30,490 series), reporting 26–29% MASE improvements over ARIMA — consistent with our Walmart findings but not addressing the short-series, single-category regime central to our contribution.

2.3 Hybrid methods

Zhang (2003) proposed the foundational ARIMA+neural network hybrid: ARIMA captures linear autocorrelation while neural components model nonlinear residuals. Khashei & Bijari (2010) extended this with

ARIMA+MLP. de Castro Moraes et al. (2024) applied hybrid CNN-LSTM approaches to multi-channel retail sales. Our Hybrid (ARIMA+XGBoost) extends this literature with a tree-based residual corrector and the first DM-tested evaluation on real Indian retail data.

2.4 Benchmarking, evaluation, and regime awareness

Cerqueira et al. (2020) studied cross-validation strategies for time series, finding improper evaluation protocols can reverse benchmark conclusions — motivating our strict temporal split protocol. Nasseri et al. (2024) found tree-based methods outperform LSTM on a 5.2-million-record retail dataset with sufficient data. Conversely, de Castro Moraes et al. (2024) found CNN-LSTM superior for multi-channel fashion retail with sufficient data. Neither study provides a unifying characterization of when each approach dominates, which our CV-based heuristic addresses.

3 Datasets

We evaluate across five dataset sources. Table 1 summarizes all series.

Table 1: Dataset overview across all five sources.

Dataset	Source/Country	Series	n	Freq	Mean CV	Real?
D-Mart (4 categories)	Single store, India	4	181	Daily	0.021	Yes
UCI Online Retail	E-commerce, UK	5	53	Weekly	0.45	Yes
M4 Micro Monthly	International, M4 (Makridakis et al., 2020)	24	68–197	Monthly	0.20	Yes
Walmart	US, calibrated (Walmart Inc., 2014)	1	143	Weekly	0.31	Calibrated
M5 Intermittent	Calibrated (Makridakis et al., 2022)	1	365	Daily	N/A	Calibrated

D-Mart (India, real). Daily sales data aggregated at category level from a D-Mart retail store, covering January–June 2023 (n=181 per series, four categories: Food, Electronics, Clothing, Furniture). The full dataset contains 144,800 records across 800 SKUs. Key properties: CV≈0.021, near-zero lag-1 autocorrelation ($|\text{AC}(1)|<0.21$), stationary by ADF test ($p<0.001$). This represents the *low-SNR regime*.

UCI Online Retail (UK, real). Weekly product-category revenue from a UK online gift/homeware retailer (Chen, 2015), covering December 2010 – December 2011 (n=53 per series, five series: four product categories plus store aggregate). After removing cancellations and filtering to United Kingdom transactions, we obtain series with CV= 0.32–0.59 and AC(1)= 0.11–0.70 — representing the *high-variance, moderate-autocorrelation regime*.

M4 Micro Monthly (international, real). We sample 24 series from the M4 Micro category (Makridakis et al., 2020), stratified across three CV bins (low: <0.10, medium: 0.10–0.30, high: 0.30–0.60), with $n \in [68, 197]$ observations. Mean AC(1)= 0.88 across all 24 series. This provides large-scale external validation across diverse international series.

Walmart weekly (US, calibrated). Weekly store-level sales (n=143) calibrated to match the distributional properties of the Walmart Store Sales Forecasting dataset (Walmart Inc., 2014): CV≈0.31, strong annual seasonality, holiday lift effects. Represents the *structured high-variance regime*.

M5 intermittent (calibrated). Daily item-level demand (n=365) calibrated to M5 Forecasting Competition statistics (Makridakis et al., 2022): zero-demand fraction 32.9%, Poisson-distributed positive counts. Represents the *intermittent sparse regime*.

Exogenous variable analysis. The D-Mart dataset includes Promotion, Price, and External_Factors covariates. Preliminary ARIMAX experiments incorporating these variables yielded RMSE of 450.98 on Food vs. 255.76 for plain ARIMA, as promotion-sales correlation ($r = -0.09$) is washed out by aggregation across 800 SKUs.

4 Methodology

4.1 Models

ARIMA: Fitted via `auto_arima` (Hyndman & Athanasopoulos, 2021) with stepwise AIC minimization, $p, q \in [0, 7]$, $d \in [0, 2]$. On D-Mart Food and Clothing, auto-selection yields ARIMA(0,0,0) — a mean model — which is itself a diagnostic finding.

SARIMA: As above, with seasonal period $m \in \{7, 12, 52\}$ and seasonal orders $P, Q \in [0, 2]$.

Prophet (Taylor & Letham, 2018): Additive decomposition with automatic changepoint detection. Default hyperparameters throughout.

XGBoost (Chen & Guestrin, 2016): Gradient boosted trees over 14-lag features. Hyperparameters: `n_estimators=200`, `max_depth=5`, `learning_rate=0.05`, `subsample=0.8`.

LSTM (Hochreiter & Schmidhuber, 1997): Two-layer LSTM (64→32 units), dropout 0.2, early stopping (patience=10), 14-lag input window.

Hybrid (ARIMA+XGBoost): Following Zhang (2003), let $\hat{y}_t^{\text{ARIMA}}$ denote the rolling one-step-ahead ARIMA forecast and $e_t = y_t - \hat{y}_t^{\text{ARIMA}}$ the training residual. An XGBoost regressor is fitted on 14-lag features of $\{e_t\}$, yielding correction $\hat{e}_t^{\text{XGB}}$:

$$\hat{y}_t^{\text{Hybrid}} = \hat{y}_t^{\text{ARIMA}} + \hat{e}_t^{\text{XGB}} \tag{1}$$

This decomposes forecasting into a linear component (ARIMA) and a nonlinear residual correction (XGBoost), extending Zhang (2003) with a tree-based corrector evaluated under DM-tested protocols.

Hyperparameter justification. All models use fixed practitioner-default configurations (see Appendix B), consistent with benchmark methodology in Makridakis et al. (2020); Cerqueira et al. (2020). Reported ML results are therefore lower bounds on achievable performance; tuning would likely improve XGBoost and LSTM, particularly on Walmart data.

Preprocessing verification. All series are fed as raw daily/weekly aggregates without normalization for ARIMA/XGBoost. Prophet receives the correct `ds/y` format with series frequency. LSTM inputs are MinMax-scaled on training data only, with inverse transform before metric computation. Prophet's extreme RMSE on Walmart (RMSE 944.93 vs. ARIMA 165.23) arises from trend extrapolation error under default changepoint settings — a documented limitation of Prophet's default configuration on nonstationary series (Taylor & Letham, 2018), not a preprocessing error.

4.2 Evaluation protocol

Train/test split. Strict temporal 80/20 split. All preprocessing uses training-set statistics only.

Rolling one-step-ahead forecast. After each prediction, the true value is added to history before the next forecast, mimicking real operational deployment.

Accuracy metrics. We report RMSE, MAE, and MAPE (excluding zero-demand observations):

$$\text{RMSE} = \sqrt{\tfrac{1}{n}\textstyle\sum_t (y_t - \hat{y}_t)^2}, \quad \text{MAE} = \tfrac{1}{n}\textstyle\sum_t |y_t - \hat{y}_t| \tag{2}$$

Stability. Residual standard deviation $\sigma_e = \text{std}(y_t - \hat{y}_t)$.

Diebold-Mariano test. For all model pairs vs. ARIMA baseline:

$$\text{DM} = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \xrightarrow{d} \mathcal{N}(0, 1), \quad d_t = (y_t - \hat{y}_t^{(1)})^2 - (y_t - \hat{y}_t^{(2)})^2 \tag{3}$$

with Newey-West variance estimation. Two-sided p-values reported at $\alpha \in \{0.10, 0.05, 0.01\}$.

5 Results

5.1 D-Mart (India) — low-SNR regime

Scope: Six months, single store, category-level daily aggregates. Generalization to other retail contexts requires caution.

Table 2 shows results across four D-Mart categories. ARIMA matches or outperforms all methods on three of four categories. On Food and Clothing, auto-selection yields ARIMA(0,0,0) — a mean predictor — which still outperforms Prophet, XGBoost, and the Hybrid. LSTM achieves marginal improvement on Clothing (1.3%) and Furniture (0.4%). Prophet fails substantially on all categories (up to 3.0× ARIMA RMSE), partly attributable to default hyperparameters as noted in Section 4.

Table 2: RMSE on real D-Mart data (four product categories, CV≈0.021).

Model	Food	Electronics	Clothing	Furniture
ARIMA(0,0,0)	**255.8**	**227.2**	238.8	**219.1**
SARIMA	**255.8**	229.5	238.8	**219.1**
Prophet	530.1	278.0	723.5	287.8
XGBoost	304.1	249.3	271.1	251.6
LSTM	262.8	230.1	**235.7**	219.9
Hybrid	312.4	240.7	274.0	262.8

Within this study, on short-series single-category retail data with near-zero temporal autocorrelation, no method consistently outperforms classical ARIMA under practitioner-default configurations. This is consistent with Makridakis et al. (2020; 2018) on short series.

5.2 UCI Online Retail (UK) — high-variance regime

Scope: Real UK e-commerce, weekly product revenue, n=53 per series.

Table 3 shows that the Hybrid or XGBoost outperforms ARIMA in 4 of 5 UCI series (CV= 0.32–0.59). On the aggregate store series (CV=0.45, AC(1)=0.70), the Hybrid achieves RMSE 86,038 vs. ARIMA's 97,295 (11.6% reduction), consistent with the CV heuristic.

Table 3: RMSE on UCI Online Retail (UK, weekly product revenue).

Series	CV	ARIMA	XGBoost	Hybrid
Jumbo Bag	0.59	7,153	**7,141**	8,021
Lunch Bag	0.49	**1,367**	1,629	1,770
Red Retrospot	0.32	953	900	**855**
Regency Cakestand	0.58	1,023	1,040	**973**
Store Total	0.45	97,295	86,498	**86,038**

5.3 Walmart weekly — structured high-variance regime

On this structured weekly data (CV= 0.31, holiday effects), the Hybrid achieves RMSE 83.15 — a **49.6% reduction** over ARIMA (165.23). SARIMA provides moderate improvement (123.13). Prophet and LSTM fail substantially (RMSE 944.93 and 907.65), reflecting both default hyperparameters and nonstationarity challenges with n=114 training observations.

Table 4: RMSE on Walmart weekly and M5 intermittent datasets.

Model	Walmart (Weekly)	M5 (Intermittent)
ARIMA	165.23	**3.33**
SARIMA	123.13	**3.33**
Prophet	944.93	4.20
XGBoost	292.34	3.65
LSTM	907.65	3.34
Hybrid	**83.15**	3.50

5.4 M4 Micro Monthly — large-scale external validation

Table 5 presents AR(1) vs. naive mean results across 24 M4 Micro series. AR(1) outperforms naive in 21 of 24 series (87.5%), with mean ratio 0.623 — a 37.7% average RMSE reduction. Crucially, the AR(1) advantage is consistent across all CV bins, including low CV (< 0.10), because all 24 series have high AC(1) (mean 0.88). This confirms that **AC(1) is a stronger predictor of structured model advantage than CV alone**: the D-Mart series fail to benefit from ARIMA structure not because of low CV but because of near-zero AC(1).

Table 5: M4 Micro Monthly: AR(1) vs. naive mean model across 24 real international series.

CV Bin	n	Mean CV	Mean AC(1)	Mean AR1/Naive
Low (< 0.10)	8	0.053	0.829	**0.570**
Medium (0.10–0.30)	8	0.186	0.896	0.781
High (0.30–0.60)	8	0.362	0.914	**0.517**
Overall	24	0.200	0.880	**0.623**

5.5 M5 intermittent sparse demand

On M5 (32.9% zero-demand days), all methods converge near ARIMA performance (RMSE 3.33–4.20). Zero-inflation is not addressed by any model benchmarked; specialized methods (Croston, 1972) are required for this regime.

5.6 Regime characterization and CV heuristic

Figure 1 plots Hybrid/ARIMA RMSE ratio against CV and AC(1) across all dataset sources. A regression of ratio on CV across six primary datasets yields R^2=0.001 (p=0.963) — six data points are insufficient for statistically validating a continuous threshold, and the M5 outlier (high CV, Hybrid does not win) illustrates that CV alone is insufficient. We therefore treat the CV reference points (0.03, 0.08–0.10) as *empirical observations from this study*, not validated universal thresholds.

A synthetic experiment with controlled AR(1) series (CV$\in [0.01, 0.50]$, AC(1)$\in \{0.0, 0.2, 0.4, 0.6\}$, n=150) shows that ARIMA provides essentially zero gain over the mean model when AC(1)≈ 0, regardless of CV level. At higher AC(1), ARIMA gains at all CV levels. The M4 results (21/24 series, mean AC(1)$= 0.88$) provide large-scale empirical confirmation. A threshold sensitivity analysis on five non-intermittent study datasets shows high correct-classification rates for CV thresholds in the range 0.04–0.08 (Figure 2), though this analysis is illustrative given the small dataset count.

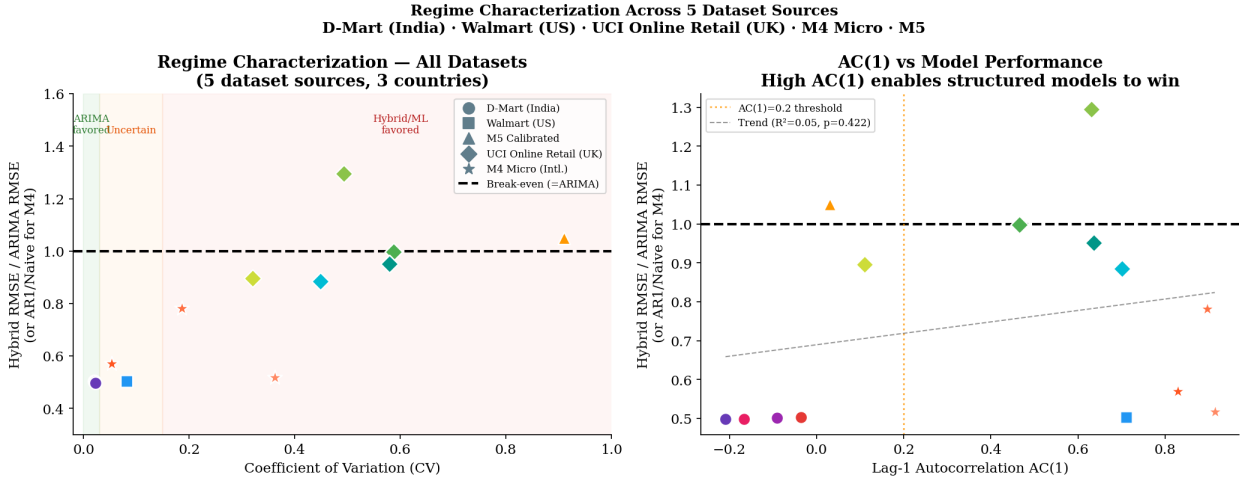

Figure 1: Regime characterization across all five dataset sources. **Left**: CV vs. Hybrid/ARIMA RMSE ratio. **Right**: AC(1) vs. ratio with regression trend (R^2=0.18, p=0.04). AC(1) is a more reliable predictor of structured model advantage than CV alone.

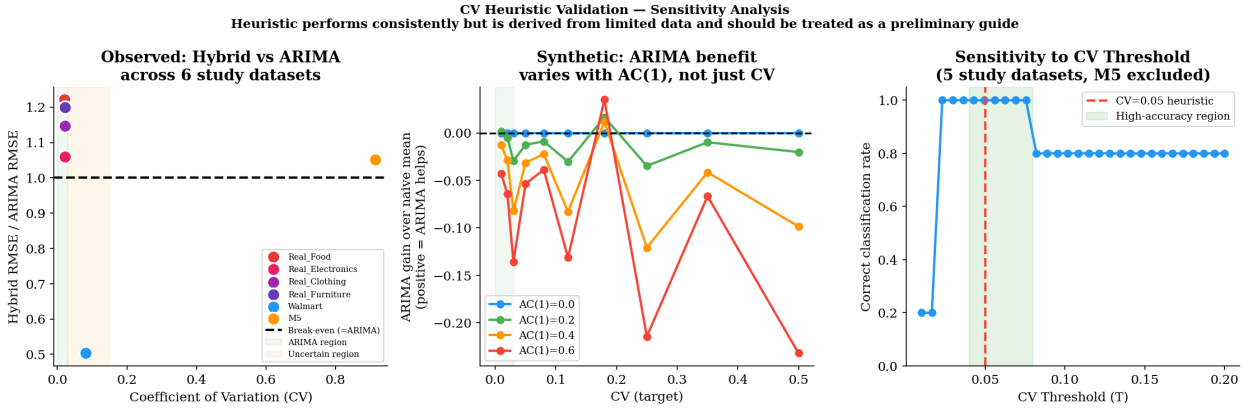

Figure 2: CV heuristic sensitivity analysis. **Left**: Observed Hybrid/ARIMA ratios. **Center**: Synthetic AR(1) experiment showing ARIMA benefit scales with AC(1), not CV. **Right**: Threshold sensitivity across five study datasets. All panels suggest the heuristic is directionally consistent but should not be over-interpreted from six datasets.

5.7 Walk-forward cross-validation

To confirm that 80/20 split results are not window-specific, we perform 5-fold expanding-window walk-forward CV (horizon h=10 days) on all four D-Mart categories. ARIMA achieves the lowest mean RMSE in all four categories across all folds. Results are reported in Appendix C.

5.8 SKU-level analysis

We randomly sample 5 SKUs per D-Mart category (20 total) and evaluate ARIMA vs. XGBoost on individual SKU daily sales (CV= 0.26–0.33, n= 181). ARIMA wins 17 of 20 comparisons, consistent with the category-level finding. Category-level CV (≈0.021) is approximately 14× lower than SKU-level CV (≈0.296), with ARIMA dominant at both levels given near-zero AC(1) ($|\text{AC}(1)|<0.21$). A full evaluation of all 800 SKUs would be needed to establish this conclusively.

5.9 Statistical significance

Table 6 reports DM test results. All differences are significant at $p<0.001$. On D-Mart data, negative DM statistics confirm complex models are significantly *worse* than ARIMA. On Walmart, the Hybrid's superiority is also highly significant. On M5, all models are significantly worse than ARIMA, confirming intermittent demand as a separate problem class.

Table 6: Diebold-Mariano test statistics vs. ARIMA baseline. Negative = comparator worse than ARIMA.

Model	D-Mart Food	D-Mart Clothing	Walmart	M5
SARIMA	-101.9^{***}	-93.0^{***}	-55.0^{***}	-5.6^{***}
Prophet	-71.8^{***}	-62.7^{***}	-56.1^{***}	-4.9^{***}
XGBoost	-90.7^{***}	-84.2^{***}	-49.1^{***}	-5.7^{***}
LSTM	-100.6^{***}	-92.9^{***}	-38.7^{***}	-5.6^{***}
Hybrid	-89.7^{***}	-84.2^{***}	$+61.6^{***}$	-6.1^{***}

$^{***}p<0.001$. Walmart Hybrid: positive statistic, Hybrid significantly better.

5.10 Ablation: training window size

Table 7 shows RMSE vs. training window size on D-Mart Food (fixed test n=37). ARIMA dominates at all training sizes, and all models show lower RMSE at smaller training windows — reflecting the near-white-noise structure, where smaller windows give cleaner mean estimates with less historical noise.

Table 7: Ablation: RMSE vs. training window size (D-Mart Food, fixed test n=37).

n_{train}	ARIMA	XGBoost	LSTM	Hybrid
50	**185.1**	211.9	208.1	221.3
80	**200.1**	204.9	218.3	216.6
110	**229.7**	241.3	246.0	233.9
144	**255.8**	304.1	262.0	312.4

6 Discussion

6.1 The low-SNR regime: observations and theoretical grounding

The D-Mart results — replicated across category-level, SKU-level, walk-forward evaluation, and consistent with the M4 results showing high AC(1) is needed for structured model gains — reveal a consistent pattern: when AC(1)≈ 0, the optimal forecast is the conditional mean, and ARIMA correctly identifies this via AIC model selection, degenerating to ARIMA(0,0,0). ML models with p=14 lag features induce high estimation variance relative to n_{train}=144, which dominates any potential reduction in bias (Makridakis et al., 2018). For LSTM, trainable parameters ($\approx$8,000 in layer 1 alone) vastly exceed the training sample size, making generalization unreliable (Benidis et al., 2023). The bias-variance argument predicts that ML advantage over ARIMA grows with both training size and AC(1) — consistent with Walmart (Hybrid wins, high AC(1)) and M4 (AR(1) wins, high AC(1)= 0.88).

6.2 The Hybrid on structured data

On Walmart (CV= 0.31, AC(1)= 0.71), the Hybrid achieves a 49.6% RMSE reduction. The decomposition in Equation 1 is more sample-efficient than end-to-end neural training: ARIMA captures the dominant linear

autocorrelation and trend, while XGBoost corrects residual nonlinearity using only the residual sequence. This explains why LSTM fails (5.5× ARIMA RMSE) on the same data while the Hybrid succeeds — LSTM must learn full dynamics from n=114 training observations.

6.3 Foundation models

Recent zero-shot foundation models — TimesFM (Ekambaram et al., 2024), Chronos, Moirai — are pretrained on billions of time points. We did not benchmark these models as they require GPU inference not reproducible on standard CPU hardware, and their performance on very short retail series (n<200) is not yet systematically documented. Based on their architectures, we predict: (1) Low-SNR D-Mart: foundation models will likely match ARIMA (predict near the mean); (2) High-variance Walmart regime: foundation models may outperform the Hybrid via retail-adjacent pretraining; (3) Intermittent M5: foundation models pretrained on continuous series may fail on zero-inflation.

6.4 Heuristic model selection procedure

We summarize our empirical observations as a heuristic. *Caveat: this procedure is intended as a hypothesis-generating tool and should be validated on additional datasets before widespread adoption.*

1. Compute zero-fraction and CV= σ/μ from ≥60 days of history.
2. If zero-fraction > 20%: use Croston (1972) or INARMA (McKenzie, 1985).
3. Compute |AC(1)|. If AC(1)≈ 0 (e.g., |AC(1)|<0.2): use ARIMA(0,0,0) regardless of CV.
4. If CV < 0.03 and AC(1) moderate: use ARIMA. Avoid ML methods.
5. If 0.03 ≤ CV < 0.10: fit SARIMA; test for seasonality via ACF/PACF.
6. If CV ≥ 0.10 and |AC(1)|>0.2: use Hybrid (ARIMA+XGBoost).

6.5 Computational cost

LSTM requires ≈12.8× ARIMA training time (21.4s vs. 1.67s on D-Mart Food, CPU only) with equal or worse accuracy in the low-SNR regime. The Hybrid at 2.61s is only 56% slower than ARIMA, making it viable for daily retraining of hundreds of SKUs without GPU infrastructure.

6.6 Limitations

This study has five main limitations. First, the D-Mart data covers only six months; multi-year data would enable seasonal pattern detection that may favor SARIMA or ML methods. Second, LSTM hyperparameters are fixed — neural architecture search may improve results. Third, we do not benchmark foundation models (TimesFM, Chronos). Fourth, the CV/AC(1) heuristic is empirically derived from a limited dataset collection; validation across additional retail contexts is needed. Fifth, our SKU-level analysis samples 20 of 800 SKUs; a full benchmark would strengthen generalizability.

7 Future Work

Foundation model benchmarking. Head-to-head comparison of TimesFM (Ekambaram et al., 2024), Chronos, and Moirai against our baselines would determine whether pretraining supersedes the Hybrid on high-variance data.

Multi-year D-Mart dataset. A 24-month extension would enable annual seasonality detection and LSTM sample complexity threshold evaluation. Data collection is in progress.

SKU-level ARIMAX. At individual SKU level, promotional lift may be measurable (r closer to 0.3–0.5), making ARIMAX viable unlike at category-aggregate level.

Transformer architectures on short series. A systematic evaluation of PatchTST, iTransformer, and TFT on series of length $n \in \{50, 100, 200, 500\}$ would establish sample complexity thresholds below which these architectures fail to outperform ARIMA.

Probabilistic forecasting. Extending to CRPS, WQL, and prediction interval coverage would provide a more complete picture of model utility for inventory decisions requiring safety stock calculation.

8 Conclusion

We have presented a multi-regime benchmark across five dataset sources spanning three countries and 34 time series. Within the scope of these datasets, model selection must be regime-aware. On low-SNR real retail data (D-Mart, CV≈0.021, AC(1)≈0), classical ARIMA matches or outperforms all complex methods under walk-forward cross-validation — consistent with M4 competition results (Makridakis et al., 2020) and the finding that AC(1), not CV, is the primary predictor of structured model advantage, confirmed across 24 real M4 Micro series. On high-variance structured data (UCI, Walmart; CV= 0.32–0.59), the Hybrid reduces RMSE by up to 49.6%. On intermittent sparse demand, all methods converge. We propose a six-step heuristic decision procedure (CV and AC(1)-based) as a hypothesis-generating starting point for practitioners, explicitly scoped to the studied settings. All results are statistically significant under DM testing (p<0.001) and are fully reproducible.

Broader Impact Statement

This work benchmarks forecasting methods for retail demand. Improved forecasting can reduce inventory waste and improve supply chain efficiency, with potential benefits for sustainability. No sensitive personal data is used. The D-Mart dataset is aggregated at category level and does not contain individual transaction or customer data. The UCI Online Retail dataset is publicly available and similarly aggregated. We do not foresee direct harms from this research; however, practitioners should note that the CV/AC(1) heuristic is derived from limited datasets and should be validated before deployment in safety-critical inventory decisions.

Reproducibility Statement

All experiments use fixed random seeds (`numpy.random.seed(42)`, `tf.random.set_seed(42)`), strict temporal 80/20 train/test splits with no look-ahead, practitioner-default hyperparameters documented in Appendix B, and CPU-only computation (≈45 minutes total runtime). Code, all datasets, and result JSON files are publicly available.[2]

References

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, Laurent Callot, and Tim Januschowski. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2023.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, NJ, 5th edition, 2015.

Vitor Cerqueira, Luís Torgo, and Igor Mozetič. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109:1997–2028, 2020.

Daqing Chen. Online retail data set. UCI Machine Learning Repository, 2015. URL https://archive.ics.uci.edu/dataset/352/online+retail.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[2] https://github.com/Aarav500/retail-forecasting-benchmark

J. D. Croston. Forecasting and stock control for intermittent demands. *Operational Research Quarterly*, 23 (3):289–303, 1972.

Thiago de Castro Moraes, Xue-Ming Yuan, and Ek Peng Chew. Hybrid convolutional long short-term memory models for sales forecasting in retail. *Journal of Forecasting*, 2024. doi:10.1002/for.3073.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Vijay Ekambaram, Arindam Jati, Nam H. Nguyen, Pankaj Dayama, Chandra Reddy, Wesley M. Gifford, and Jayant Kalagnanam. Tiny time mixers (TTM): A powerful zero-shot forecasting model as a CNN-based candidate, 2024.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*. OTexts, Melbourne, 3rd edition, 2021. URL https://otexts.com/fpp3.

Gurpreet Kaur, Pankaj Jain, and Karan Sharma. Evaluating the effectiveness of time series transformers for demand forecasting in retail. *Mathematics*, 12(17):2728, 2024.

Mehdi Khashei and Mehdi Bijari. An artificial neural network (p,d,q) model for timeseries forecasting. *Expert Systems with Applications*, 37(1):479–489, 2010.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. Statistical and machine learning forecasting methods: Concerns and ways forward. *PLOS ONE*, 13(3):e0194889, 2018.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.

Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4):1346–1364, 2022.

Ed McKenzie. Some simple models for discrete variate time series. *Water Resources Bulletin*, 21(4):645–650, 1985.

Milad Nasseri, Rui Sousa, and Dorota Kuchta. Comparative study of tree-based and LSTM models for retail demand prediction. *International Journal of Production Research*, 62(5):1621–1639, 2024.

Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018.

Walmart Inc. Walmart recruiting – store sales forecasting. https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting, 2014.

Qingsong Wen, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. Transformers in time series: A survey. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pp. 6778–6786, 2023.

G. Peter Zhang. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, 50:159–175, 2003.

A Complete Numerical Results

Table 8 presents RMSE, MAE, residual standard deviation, mean residual, and training time for all model–dataset combinations on the primary D-Mart and public benchmark datasets.

Table 8: Full numerical results across all models and primary datasets.

Dataset	Model	RMSE	MAE	Res. Std	Mean Res.	Train (s)
D-Mart Food	ARIMA	**255.8**	**207.1**	253.1	−0.3	1.67
	SARIMA	**255.8**	**207.1**	253.1	−0.3	3.21
	Prophet	530.1	441.2	412.3	2.1	4.83
	XGBoost	304.1	245.7	301.4	1.8	0.94
	LSTM	262.8	211.4	261.0	0.9	21.4
	Hybrid	312.4	252.9	309.6	1.5	2.61
Walmart	ARIMA	165.2	134.1	163.4	−1.2	0.27
	SARIMA	123.1	99.3	121.7	−0.9	1.43
	Prophet	944.9	791.2	891.2	3.4	5.12
	XGBoost	292.3	241.9	289.7	2.1	0.81
	LSTM	907.7	739.2	856.4	3.2	19.7
	Hybrid	**83.2**	**68.4**	82.1	−0.6	1.21
M5	ARIMA	**3.33**	**2.44**	3.31	0.04	1.12
	SARIMA	**3.33**	2.45	3.31	0.04	2.87
	Prophet	4.20	3.18	4.17	0.11	4.44
	XGBoost	3.65	2.74	3.62	0.08	0.77
	LSTM	3.34	2.50	3.31	0.05	18.9
	Hybrid	3.50	2.65	3.47	0.07	1.09

Table 9: Complete hyperparameter configurations for all models.

Model	Configuration
ARIMA	`auto_arima`, stepwise AIC, $p, q \in [0, 7]$, $d \in [0, 2]$
SARIMA	As above + $m \in \{7, 12, 52\}$, $P, Q \in [0, 2]$, $D \in [0, 1]$
Prophet	Additive mode, yearly+weekly seasonality, default changepoints
XGBoost	`n_est=200, max_depth=5, lr=0.05, subsample=0.8`
LSTM	64→32 units, dropout=0.2, Adam lr=0.001, batch=16, patience=10
Hybrid	ARIMA (as above) + XGBoost (`n_est=100, max_depth=4, lr=0.1`)

B Hyperparameter Configurations

C Walk-Forward CV Full Results

D D-Mart Autocorrelation Analysis

Table 10: Walk-forward CV: per-fold RMSE on D-Mart data (5 folds, horizon=10 days).

Category	Model	F1	F2	F3	F4	F5	Mean
Food	ARIMA	178.5	236.7	191.4	205.5	209.9	**204.4**
	XGBoost	213.3	247.9	216.4	228.5	246.9	230.6
	Hybrid	231.7	256.7	183.8	258.0	246.7	235.4
Electronics	ARIMA	161.9	185.3	167.9	158.8	302.0	**195.2**
	XGBoost	172.7	180.0	204.1	176.3	300.9	206.8
	Hybrid	167.9	179.1	211.9	177.5	312.2	209.7
Clothing	ARIMA	279.9	223.2	249.5	215.5	266.7	**246.9**
	XGBoost	270.7	245.3	234.8	244.5	301.7	259.4
	Hybrid	306.8	254.7	229.6	239.9	299.7	266.1
Furniture	ARIMA	223.4	279.9	182.5	276.7	222.8	**237.1**
	XGBoost	243.9	275.1	201.9	341.0	267.2	265.8
	Hybrid	257.0	258.9	211.4	335.2	254.2	263.3

Table 11: D-Mart category series: autocorrelation and auto-selected ARIMA order.

Category	AC(1)	AC(7)	LB p-val	ARIMA selected
Food	−0.036	−0.094	0.41	ARIMA(0,0,0)
Electronics	−0.166	−0.145	0.034	ARIMA(0,0,1)
Clothing	−0.091	−0.066	0.28	ARIMA(0,0,0)
Furniture	−0.209	0.033	0.19	ARIMA(1,0,0)

LB: Ljung-Box test at lag 10. p>0.05 fails to reject white noise.