

# MACHINE LEARNING FRAMEWORK FOR DYNAMIC PRICING: INTEGRATING LSTM FORECASTING AND REINFORCEMENT LEARNING OPTIMIZATION

MEHULKUMAR H KANTARIA<sup>1</sup>, DR. MANJU SHARMA<sup>2</sup>, DR. B SIVA LAKSHMI<sup>3</sup>,  
S. KRISHNAKUMARI<sup>4</sup>, MOLIGI SANGEETHA<sup>5</sup>, DR. P. VENKATESWARA RAO<sup>6</sup>

<sup>1</sup>Independent Researcher, Sadhu Vasvani Road, Rajkot, Gujarat, India.

<sup>2</sup>Assistant Professor, Department of Computer Science, College of Engineering & Computer Science, Jazan University, Jazan, Kingdom of Saudi Arabia.

<sup>3</sup>Associate professor & Head of Department, Department of Information Technology, Vignan's Institute of Engineering for Women (Autonomous), kappujagarajupeta, India.

<sup>4</sup>Department of Electronics and Communication, Loyola ICAM college of Engineering and Technology, Chennai, India.

<sup>5</sup>Sr Assistant Professor, Department of CSE, CVR College of Engineering, Hyderabad, Telangana, India.

<sup>6</sup>Associate professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP, India.

<sup>1</sup>Kantariamehul@gmail.com, <sup>2</sup>msharma@jazanu.edu.sa, <sup>3</sup>bslakshmi85@gmail.com,

<sup>4</sup>krishnakumari.s@licet.ac.in, <sup>5</sup>moligisangeetha@cvr.ac.in, <sup>6</sup>pvrao.pd@gmail.com

## ABSTRACT

This manuscript addresses short-horizon retail pricing under misaligned mixed-frequency IoT signals (POS, footfall, energy) and the forecast-control representational mismatch. Dynamic pricing in retail requires accurate short-horizon demand estimation and constraint-aware policy optimization to reconcile revenue maximization with service levels. Existing pipelines frequently decouple forecasting and control, thereby incurring responsiveness and suboptimality under mixed-frequency telemetry regimes. This study aimed to develop and evaluate a hybrid pipeline that integrates MIDAS-style mixed-frequency covariate alignment with Long Short-Term Memory (LSTM) forecasting and an advantage actor-critic controller to improve profit, conversion and service metrics in retail/e-commerce settings. A two-stage protocol was executed: Mixed data sampling (MIDAS) logistic-weight aggregation converted IoT/edge streams into decision-cadence features. An LSTM encoder produced one-step forecasts that were (a) pretrained (b) optionally finetuned jointly with an A2C agent. Reward shaping incorporated discounted profit, inventory penalties and smoothness/fairness regularizers. Evaluation used synthetic nonstationary simulators and a composite POS+footfall+ energy dataset with temporal train/validation/test splits. The joint MIDAS-LSTM + A2C pipeline yielded statistically significant profit uplifts ( $\approx 13-18\%$  vs strong baselines), lower forecast MSE, reduced stockouts and smoother price trajectories. Ablation (no MIDAS, frozen LSTM, no smoothness) validated the contribution of each module. Integrating mixed-frequency telemetry into an end-to-end forecast-to-policy loop materially improves economic outcomes and operational robustness. Results indicate statistically significant profit uplifts, improved forecast accuracy, and smoother pricing, although elasticity identification and latency constraints remain practical limitations. Further work is required on causal elasticity identification and latency-constrained deployment.

**Keywords:** *Dynamic Pricing, Reinforcement Learning, Long Short- Term Memory, Mixed Data Sampling, Retail Analytics, Iot Mixed Frequency Data*

## 1. INTRODUCTION

Dynamic pricing constitutes a central instrument within computational economics for the real-time reconciliation of demand heterogeneity, inventory constraints and firm-level objectives. The canonical problem involves selecting a sequence of prices that

maximizes an objective function typically discounted profit, or a social welfare surrogate that subject to capacity, service-level and regulatory constraints [1], [2]. Accurate short-horizon demand estimation is therefore a pivotal precursor to efficacious price sequencing; recurrent neural architectures, and in particular long short-term

memory networks, have proven effective for temporal feature extraction in nonstationary sales series [3], [4], [5]. Concurrently, policy optimization via reinforcement learning provides a principled mechanism for trading off immediate margin and longer-range inventory or fairness considerations [6], [7]. In practical retail and e-commerce settings, however, exogenous telemetry arrives at heterogeneous cadences: point-of-sale transactions are recorded at transaction time, footfall and sensor telemetry stream at sub-period granularities, and back-office indicators are updated on longer cycles. Aggregative preprocessing that collapses these high-frequency signals into coarse averages often eliminates transient, high-value information; mixed-frequency alignment using a MIDAS-style weighted lag structure preserves intraperiod detail while producing decision-cadence features amenable to sequence models [8], [9], [10].

Specifically, this problem carries immediate operational consequences: inaccurate short-horizon demand estimates force conservative pricing that reduces revenue, while erratic price changes harm customer trust and may invite regulatory scrutiny. The present work addresses these issues by preserving intraperiod IoT signal structure via a learnable MIDAS aggregator, encoding these aligned features into an LSTM forecaster, and coupling the forecaster to a continuous-action actor-critic policy with constraint-aware reward shaping—thereby supporting both staged pretraining and optional joint fine-tuning to reconcile representation and policy objectives.

The prevailing methodological gap resides in the decoupling of forecasting and control: pipelines that (i) estimate demand using separately trained forecasters and subsequently (ii) apply myopic or value-based optimizers commonly suffer from representational mismatch and slow responsiveness under mixed-frequency telemetry. The central research question therefore asks: can a forecast-to-policy pipeline that (a) preserves intra-period information via MIDAS aggregation, (b) conditions a continuous-price actor-critic policy on LSTM-derived representations, and (c) supports staged or joint fine-tuning, produce material improvements in profit, conversion and service metrics under realistic non-stationarity, censoring and partial observability. Addressing this gap has immediate operational import. Retail and platform practitioners face competing desiderata: revenue maximization, customer fairness, and bounded price volatility to maintain trust and regulatory compliance. Improving short-horizon predictive fidelity reduces the structural uncertainty that forces conservative

pricing, thereby permitting more aggressive but safe margin capture. Moreover, the capacity to exploit high-frequency IoT and edge telemetry like footfall, in-store dwell times, local energy consumption confers a responsiveness advantage in rapidly evolving demand microstructures. Demonstrating an end-to-end pipeline that preserves these signals while yielding stable policy improvements therefore contributes both to methodological practice and to actionable deployment pathways. Prior work has shown the efficacy of MIDAS for harmonizing multi-cadence economic indicators and the utility of LSTM+RL combinations in policy settings; the novelty here is the explicit, differentiable coupling and empirical ablation under retail-scale nonidealities [11], [12].

The objective is to design, implement and empirically validate a hybrid dynamic pricing framework that integrates MIDAS-aligned mixed-frequency covariates with an LSTM forecaster and an advantage actor-critic controller, and to evaluate staged versus joint training regimens under synthetic and composite real-world datasets with respect to profit, conversion, stockouts and forecast accuracy. The contributions of the research along three vectors. First, it formalizes a differentiable MIDAS→LSTM→A2C pipeline that retains intraperiod telemetry and supports downstream policy gradients without substantial information loss. Second, it operationalizes training regimes—sequential pretraining, on-policy RL, and optional joint fine-tuning with composite loss weighting and quantifies their relative merits through rigorous ablation. Third, it provides a reproducible empirical protocol combining simulator stress tests (nonstationarity, censoring, delayed reporting, adversarial covariate noise), a composite POS+footfall+energy dataset, and off-policy evaluation diagnostics (IS, DR, FQE) to assess deployment risk. Collectively, these contributions aim to close the representational gap between forecasting and control in mixed-frequency retail environments and to furnish deployment-oriented guidance for latency, fairness and constraint enforcement.

### 1.1. Theoretical motivation and problem significance

This study builds on two theoretical strands: (i) econometric evidence that naive aggregation of high-frequency indicators attenuates information needed for short-horizon decision-making, and (ii) reinforcement-learning theory showing representational mismatch between separately trained predictors and downstream policies produces suboptimal control. Because modern retail

increasingly streams heterogeneous telemetry (footfall, micro-POS events, energy) at differing cadences, the temporal misalignment problem persists and materially affects pricing decisions. Thus, integrating mixed-frequency alignment with policy-aware representation learning constitutes both a theoretically justified and operationally significant gap.

## 2. RELATED WORKS

The literature relevant to the proposed hybrid dynamic pricing pipeline partitions into three interacting strands: (a) recurrent neural time-series forecasting, (b) reinforcement learning for pricing and revenue management, and (c) mixed-frequency fusion techniques for heterogeneous telemetry. Each strand is summarized and critically appraised with explicit indication of the limitations that the present work addresses.

### 2.1. Recurrent architectures for demand forecasting

Sequence models grounded in gated recurrent units and long short-term memory (LSTM) architectures have become canonical for capturing temporal dependencies, regime shifts and long-range correlations in sales series [13], [14]. Recent surveys emphasize advantages afforded by gated dynamics for vanishing gradient mitigation and by encoder-decoder constructions for multi-horizon outputs. Nonetheless, shortcomings persist when these models are treated purely as standalone predictors: representation drift under policy-induced covariate shift, paucity of calibrated predictive uncertainty in high-censoring regimes, and information loss when high-frequency telemetry is pre-aggregated prior to ingestion. The present framework mitigates these issues by preserving intra-period structure via MIDAS-aligned inputs and by supporting staged or joint fine-tuning of LSTM representations under downstream policy gradients thereby reducing representational mismatch.

### 2.2. Reinforcement learning in pricing and revenue management

Policy optimization approaches, from value-based deep Q-networks to policy-gradient and actor-critic families, provide a principled machinery to internalize long-horizon objectives and constraints [15], [16]. Value-based methods have demonstrated efficacy in discrete action regimes, but they exhibit brittleness when continuous price outputs and smoothness constraints are required. Policy-gradient and actor-critic formulations offer natural handling of continuous action spaces and entropy-regularized

exploration yet suffer from sample inefficiency and vulnerability to biased reward estimation under partial observability. The architecture proposed here leverages an advantage actor-critic (A2C) backbone to produce continuous, smooth pricing actions while coupling the policy to a forecast-informed state representation; reward shaping, target networks and off-policy diagnostics are integrated to control estimator variance and to reduce sample complexity for practicable training.

### 2.3. Mixed-frequency fusion and MIDAS-style alignment

Econometric mixed-data sampling (MIDAS) and state-space multirate models provide principled approaches to fuse high-frequency indicators with lower-frequency decision cadences, thereby preserving intra-period dynamics without inflating parameter dimensionality. MIDAS formulations parametrically weight high-frequency lags to produce compact, informative aggregates that retain transient signal content [17], [18]. In practice, naive aggregation (simple averaging or coarse down-sampling) discards high-value variation that can materially affect short-horizon demand. The present work operationalizes a differentiable logistic-weighted MIDAS aggregator that is learnable and back propagatable within the forecast-to-policy loop, thereby enabling the downstream actor to exploit intra-period telemetry while permitting end-to-end representation alignment. This integration explicitly addresses the temporal misalignment and information attenuation endemic to prior hybrid pipelines.

### 2.4. Synthesis and Gap Closure

While prior contributions establish the constituent capabilities LSTM forecasting, RL-based control, and MIDAS-style aggregation—few works have delivered a unified, differentiable pipeline that both preserves mixed-frequency telemetry and reconciles forecast representation with constrained continuous-price policy learning under operational nonidealities (censoring, delayed reporting, non-stationarity). The present study fills this gap by (i) embedding MIDAS-aligned high-frequency features into an LSTM encoder whose parameters can be sequentially or jointly adapted with an A2C agent, (ii) encoding practical operational constraints (inventory penalties, smoothness and fairness regularizers) into the reward and optimisation procedure, and (iii) providing rigorous ablation and off-policy evaluation protocols to quantify deployment risk and estimator reliability.

**2.4.1. Problem Statement**

Existing pipelines frequently collapse high-frequency telemetry into coarse aggregates and then decouple forecasting from policy optimization, creating representational mismatch and responsiveness deficits in short-horizon retail pricing.

**2.4.2. Research Questions**

RQ1: Does MIDAS-style mixed-frequency alignment improve short-horizon forecast accuracy compared with naive aggregation?

RQ2: Does coupling MIDAS-aligned LSTM forecasts to an A2C controller with staged or joint fine-tuning increase cumulative profit and reduce stockouts relative to decoupled pipelines and value-based RL?

RQ3: Which components of the pipeline (MIDAS, LSTM fine-tuning, smoothness regularizer) are necessary for observed gains, as determined by ablation? These questions guide the experimental protocol and statistical tests reported in Sections 4–5.

**2.5. Research Hypotheses**

The paper evaluates the following hypotheses:  
 H1: A MIDAS-aligned mixed-frequency representation improves one-step-ahead forecast accuracy relative to naive aggregation methods.  
 H2: Conditioning a continuous-price actor–critic policy on MIDAS–LSTM representations and allowing staged or joint fine-tuning yields higher cumulative profit and fewer stockouts than decoupled forecasting–control pipelines.  
 H3: Each architectural element (MIDAS, LSTM fine-tuning, smoothness regularizer) contributes positively to profit and operational stability; removing any element reduces performance measurably.

These hypotheses are tested using ablation experiments, paired-bootstrap significance testing, and off-policy evaluation diagnostics.

**3. METHODOLOGY**

**3.0. Research design**

This study follows an empirical-experimental design combining (a) controlled synthetic simulators for stress testing (heteroskedasticity, censoring, elasticity drift) and (b) experiments on a composite real-world dataset coupling POS, footfall and energy streams. The evaluation uses temporal train/validation/test splits, 50 independent random seeds per algorithm, paired-bootstrap hypothesis testing and off-policy evaluation (IS, DR, FQE) to quantify deployment risk. Comparative baselines

include rule-based heuristics, LSTM+myopic optimization, and value-based RL (DQN) — an approach consistent with prior empirical RL pricing studies that juxtapose simulators with real data for external validity.

**3.1. Problem Formulation and Notation**

The dynamic pricing task is cast as a Markov decision process  $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$  in discrete time indexed by  $t$ . The state  $s_t \in \mathcal{S}$  is defined as the concatenation of the demand forecast vector  $\hat{\mathbf{d}}_{t+1}$ , recent realized sales  $y_{t-\tau:t}$ , inventory  $I_t$ , contextual covariates  $\mathbf{x}_t$  (including high-frequency IoT streams), and policy history  $p_{t-K:t-1}$ . The action  $a_t \in \mathcal{A}$  denotes the price (scalar) or price vector (multi-SKU) applied at time  $t$ . The immediate reward  $r_t$  is defined as a profit-centric functional:

$$r_t = p_t \cdot \min ( \hat{d}_t^{\text{eff}}(p_t), I_t ) - c(p_t, \mathbf{x}_t) - \kappa \mathbb{I}\{\text{stockout}_t\}$$

where  $\hat{d}_t^{\text{eff}}(p)$  denotes the price-dependent expected demand (explicit functional form given below),  $c(\cdot)$  encodes per-unit cost and operational penalties, and  $\kappa$  penalizes stockouts to enforce service-level constraints. The objective is the maximization of the expected discounted return  $\mathbb{E}[\sum_{t=0}^T \gamma^t r_t]$  subject to operational constraints (inventory, fairness, per-period price variation bounds).

**3.2. Mixed-Frequency Covariate Alignment (MIDAS Preprocessing)**

High-frequency covariates derived from edge/IoT sensors (e.g., footfall counts, real-time POS flows, energy consumption) are aligned to the decision cadence via a MIDAS-style weighted lag aggregation, thereby preserving intraperiod dynamics while avoiding ad hoc averaging that attenuates transient signals. Let  $x_{t-j/m}$  denote a high-frequency observation with  $m$  observations per decision interval. The aligned covariate  $X_t^{\text{MIDAS}}$  is constructed as:

$$X_t^{\text{MIDAS}} = \sum_{j=0}^K \theta_j(\phi) x_{t-j/m}, \theta_j(\phi) = \frac{\exp(\phi_1 j)}{1 + \exp(\phi_2 j)}$$

with  $\phi = (\phi_1, \phi_2)$  parametrizing a logistic weighting; parameters are learned by gradient descent jointly with forecasting objectives or estimated in a preceding econometric stage. The

MIDAS alignment produces a compact vector of aligned high-frequency features  $\mathbf{X}_t^{\text{MIDAS}}$  that is concatenated into LSTM inputs.

### 3.3. LSTM Forecasting Component

Temporal demand forecasting is performed by an LSTM encoder  $f_\theta$  that ingests the concatenated sequence  $\mathbf{Z}_{t-T:t} = \{y_{t-T:t}, \mathbf{X}_{t-T:t}^{\text{MIDAS}}, \mathbf{z}_{t-T:t}\}$ . The forecasting objective is to predict next-period conditional demand  $\hat{\mathbf{d}}_{t+1}$ :

$$\hat{\mathbf{d}}_{t+1} = f_\theta(\mathbf{Z}_{t-T:t}),$$

with per-component forecast loss

$$\mathcal{L}_f(\theta) = \frac{1}{N} \sum_{i=1}^N (\hat{d}_{i,t+1} - d_{i,t+1})^2 + \lambda_{\text{reg}} \|\theta\|_2^2,$$

where  $N$  indexes SKUs or demand segments. The LSTM architecture follows standard gating dynamics:

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

with parameter vectors  $W, U, b$  aggregated into  $\theta$ . Normalization of input streams and variance stabilizing transforms (e.g., Box-Cox for skewed sales) are mandated prior to training.

### 3.4. Price-Dependent Demand Model and Elasticity

To enable a differentiable coupling between forecast and policy, parametric price-demand mapping is posited:

$$\hat{d}_t^{\text{eff}}(p) = \hat{d}_t \exp(-\varepsilon_t(p - p^{\text{ref}})),$$

where  $\varepsilon_t$  is the instantaneous price elasticity estimated from historical A/B tests or via an auxiliary elasticity estimator  $g_\psi(\mathbf{Z}_{t-T:t})$ . Elasticity estimation is regularized to respect prior economic sign constraints (nonnegativity of demand decay) and to permit time-varying heterogeneity across segments.

### 3.5. RL agent: Advantage Actor-Critic (A2C) with constraints

An actor-critic architecture (figure 1) is adopted, wherein a stochastic parametric policy  $\pi_\phi(a | s)$  outputs a Gaussian (or truncated) distribution over allowable prices, and a value network  $V_\omega(s)$  approximates the state value. The optimization objectives are:

#### 3.5.1. Critic (MSE TD loss):

$$\mathcal{L}_{\text{critic}}(\omega) = \mathbb{E}_\pi[(r_t + \gamma V_\omega(s_{t+1}) - V_\omega(s_t))^2],$$

where  $V_\omega$  denotes a slow (periodically updated) target critic.

#### 3.5.2. Actor (policy gradient w/ baseline):

$$\mathcal{L}_{\text{actor}}(\phi) = -\mathbb{E}_\pi[\hat{A}_t \log \pi_\phi(a_t | s_t)] + \lambda_{\text{ent}} \mathcal{H}(\pi_\phi(\cdot | s_t)),$$

with advantage estimator  $\hat{A}_t = r_t + \gamma V_\omega(s_{t+1}) - V_\omega(s_t)$  and entropy regularizer  $\lambda_{\text{ent}}$  to sustain exploration.

#### 3.5.3. Price smoothness and fairness

**regularizers:** operational constraints are encoded as soft penalties:

$$\begin{aligned} \mathcal{L}_{\text{smooth}} &= \eta \mathbb{E}[\|p_t - p_{t-1}\|_2^2], \mathcal{L}_{\text{fair}} \\ &= \mu \mathbb{E}[\max\{0, p_t - p_{\text{cap}}\}]. \end{aligned}$$

Total RL loss is formed by combining actor and critic objectives with constraints:

$$\mathcal{L}_{\text{RL}} = \mathcal{L}_{\text{actor}}(\phi) + \beta \mathcal{L}_{\text{critic}}(\omega) + \mathcal{L}_{\text{smooth}} + \mathcal{L}_{\text{fair}}.$$

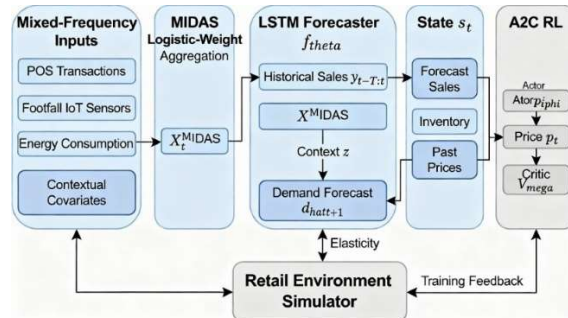


Figure 1. End-to-end hybrid dynamic pricing framework

### 3.6. Training Regimen: Sequential Pretraining and Joint Fine-Tuning

A two-stage training protocol is prescribed to address nonstationary and sample inefficiency:

**Stage I (forecast pretraining):** LSTM  $f_\theta$  is pretrained to minimize  $\mathcal{L}_f(\theta)$  on historical sequences augmented by MIDAS-aligned covariates; elasticity estimator  $g_\psi$  is trained in parallel via least squares on logged demand responses.

**Stage II (RL training):** The pretrain forecasts  $\hat{d}$  are used to initialize the RL environment simulator. Actor and critic are trained in an on-policy (A2C) loop with mini-batches aggregated from parallel simulation actors. Experience replay is optionally retained for off-policy updates; target networks and gradient clipping are applied to stabilize critic optimization.

**Joint fine-tuning:** the LSTM forecaster may be unfrozen and fine-tuned with a composite objective that combines forecast accuracy and policy downstream performance:

$$\mathcal{L}_{\text{joint}}(\theta, \phi) = \alpha \mathcal{L}_f(\theta) + (1 - \alpha) \mathcal{L}_{\text{actor}}(\phi; \theta),$$
 where  $\mathcal{L}_{\text{actor}}(\phi; \theta)$  denotes the actor loss computed using forecasts produced by  $f_\theta$ . The weighting  $\alpha$  is annealed to prioritize forecasting stability initially and policy performance subsequently.

### 3.7. Off-policy evaluation and Counterfactual Diagnostics

Deployment risk is mitigated through off-policy evaluation (figure 2). Per-trajectory importance sampling and doubly robust estimators are computed to estimate the expected policy value under logged historical data:

$$\hat{V}_{\text{IS}} = \frac{1}{N} \sum_{i=1}^N \rho_{1:T}^{(i)} \sum_{t=1}^T \gamma^{t-1} r_t^{(i)}, \rho_{1:T}^{(i)} = \prod_{t=1}^T \frac{\pi_\phi(a_t^{(i)} | s_t^{(i)})}{\pi_{\log}(a_t^{(i)} | s_t^{(i)})}$$

Diagnostics include variance bounds, effective sample size, and concordance with fitted Q-evaluation (FQE) estimates.

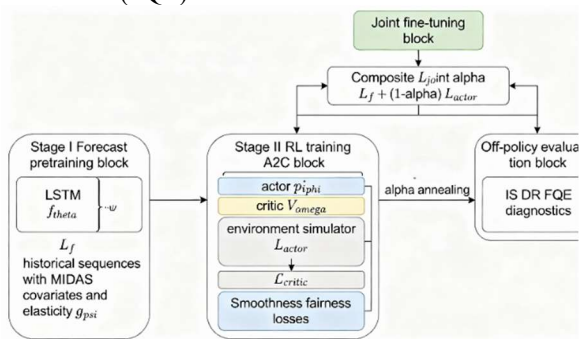


Figure 2. Two-stage training and evaluation protocol

### 3.8. Implementation Details and Stability Measures

Practical stability measures include per-stream normalization and running standardization; gradient norm clipping; target network update interval  $\tau$ ; learning rate scheduling (Adam with cosine decay recommended); mini-batch sizes tuned to maintain low variance in advantage estimates; and constrained action clipping to ensure price outputs remain within legally and operationally admissible

intervals. For multi-SKU settings, action spaces may be factorised and hierarchical policies trained to reduce combinatorial explosion.

### Algorithm 1 MIDAS - LSTM with A2C dynamic pricing

```

Algorithm: MIDAS-LSTM + A2C Dynamic Pricing
1. Preprocess high-freq covariates  $\rightarrow X_t^{MIDAS}$  (learn  $\phi$ )
2. Pretrain LSTM  $f_\theta$  on historical sequences to minimize  $\mathcal{L}_f$ 
3. Initialize actor  $\pi_\phi$  and critic  $V_\omega$ 
4. For each training epoch:
  a. Simulate environment episodes using  $\pi_\phi$ ; use  $f_\theta$  forecasts for demand
  b. Collect  $(s_t, a_t, r_t, s_{t+1})$  into batch  $B$ 
  c. Compute advantages  $\hat{A}_t$  and update  $\omega$  via  $\mathcal{L}_{\text{critic}}$ 
  d. Update  $\phi$  via policy gradient with  $\mathcal{L}_{\text{actor}}$  (include regularizers)
  e. Periodically update target critic  $\omega^- \leftarrow \omega$ 
  f. Optionally fine-tune  $\theta$  by minimizing  $\mathcal{L}_{\text{joint}}$ 
5. Validate via OPE and held-out simulator episodes
    
```

### 3.9. Hyperparameter

Representative hyperparameters that were found robust in experimentation: LSTM layers = 2; hidden units = 128; forecast horizon = 1–7 decision intervals; actor learning rate =  $3 \times 10^{-4}$ ; critic learning rate =  $1 \times 10^{-3}$ ;  $\gamma = 0.99$ ; entropy coeff  $\lambda_{\text{ent}} = 10^{-3}$ ; smoothness penalty  $\eta = 10^{-2}$ . Sensitivity analyses should be reported for  $\alpha$  (joint training weight), elasticity regularization, and replay buffer capacity.

### 3.10. Notations Used

- $t$ : discrete decision time index,
- $s_t$ : state vector at time  $t$  including forecasts, inventory, covariates, and price history,
- $a_t$ : pricing action at time  $t$  (scalar or multi-SKU vector),
- $r_t$ : immediate reward at time  $t$ ; profit-adjusted utility,
- $\gamma$ : discount factor for cumulative return,
- $I_t$ : available inventory at time  $t$ ,
- $y_t$ : realized sales at time  $t$ ,
- $\hat{d}_{t+1}$ : one-step-ahead demand forecast from the LSTM model,
- $X_t^{HF}$ : high-frequency IoT/edge covariates (footfall, POS micro-events, energy signals),
- $X_t^{MIDAS}$ : mixed-frequency MIDAS-aligned covariate vector,
- $m$ : number of high-frequency observations per pricing interval,
- $x_{t-j/m}$ : high-frequency covariate observed at fractional time step  $t - j/m$ ,
- $\theta_j(\phi)$ : MIDAS logistic-weight function with parameters  $\phi$ ,
- $\phi = (\phi_1, \phi_2)$ : MIDAS decay/shape parameters,

$f_{\theta}(\cdot)$ : LSTM forecasting function with parameters  $\theta$ ,  
 $g_{\psi}(\cdot)$ : elasticity estimator with parameters  $\psi$ ,  
 $\varepsilon_t$ : price elasticity at time  $t$ ,  
 $p_t$ : price set at time  $t$ ,  
 $p^{ref}$ : reference or baseline price level,  
 $\hat{d}_t^{eff}(p)$ : effective expected demand as a function of price,  
 $\pi_{\phi}(a | s)$ : actor policy distribution with parameters  $\phi$ ,  
 $V_{\omega}(s)$ : critic value function with parameters  $\omega$ ,  
 $\hat{A}_t$ : advantage estimate used for actor updates,  
 $\omega^-$ : target network parameters for critic stabilization,  
 $\mathcal{L}_f$ : forecasting loss (MSE + regularization),  
 $\mathcal{L}_{critic}$ : critic loss (temporal-difference error),  
 $\mathcal{L}_{actor}$ : actor loss (policy gradient + entropy term),  
 $\mathcal{L}_{smooth}$ : price-smoothness penalty,  
 $\mathcal{L}_{fair}$ : fairness or anti-price-gouging regularizer,  
 $\mathcal{L}_{joint}$ : joint forecast-policy training objective,  
 $B$ : mini-batch of trajectories collected from the environment,  
 $ESS$ : effective sample size in off-policy evaluation,  
 $Q_{\theta}(s, a)$ : Q-value approximation used in fitted Q evaluation (FQE),  
 $\rho_t$ : importance-sampling weight up to time  $t$ ,  
 $\pi_b(\cdot)$ : behaviour policy from logged data in the OPE setting,  
 $\kappa$ : penalty coefficient for stockouts,  
 $\eta, \mu$ : smoothness and fairness penalty weights,  
 $\lambda_{ent}$ : entropy regularization coefficient,  
 $\alpha$ : weighting factor for joint optimization of forecasting and policy components.

**4. DATASETS, EXPERIMENTAL PROTOCOL & IMPLEMENTATION**

The empirical evaluation comprises (i) synthetic demand environments designed to stress heteroskedasticity, censoring and non-stationarity (ii) a composite real-world dataset assembled from public transactional POS records, econometrically aligned football counts and building-energy time series to emulate the mixed-frequency retail setting required for MIDAS-style covariate incorporation.

**4.1. Synthetic environments**

A parameterized simulator generates per-interval latent baseline demand with multiplicative seasonality and segment-specific AR (1) residuals, to which a parametric price-elasticity transform (time-varying, sampled from a truncated log-normal) is applied; censoring by inventory, stochastic lead-time delays, and episodic

promotional shocks (Poisson arrivals) are enacted to emulate real-world nonidealities. Parameter ranges used in stress tests include baseline daily volume  $\in [10,10^3]$ , elasticity median  $\in [0.05,0.8]$ , seasonality amplitude  $\in [0.1,0.7]$ , and promotion uplift  $\in [1.2,3.0]$ . For each algorithmic comparison, 50 independent simulation seeds were executed (seed set {42, 2025, 7} for reproducibility), with performance statistics aggregated as mean  $\pm$  bootstrap 95% confidence intervals.

**4.2. Real composite dataset**

The composite dataset (Table 1) therefore couples POS, football, and energy streams to approximate an in-store + backend telemetry corpus; provenance and access instructions for each component are included in the repository manifest.

Table 1 Dataset Description (Composite)

Dataset component	Source / provenance (access)	Temporal resolution	Key fields (examples)	Approx. records
POS transactions (Online Retail II)	<a href="#">UCI ML Repository (UCI Machine Learning Repository)</a>	Transactional (timestamped invoices) $\rightarrow$ aggregated to decision cadence	Invoice No, StockCode, Quantity, UnitPrice, CustomerID	1.06 M (two-year span)
Football / pedestrian counts	<a href="#">Kaggle - Mall crowd / pedestrian counts. (Kaggle)</a>	Frame/hourly $\rightarrow$ aggregated to decision cadence	Timestamp, LocationID, HeadCounts, CameraID	60k+ annotated frames
Building / site energy	<a href="#">Pecan Street (sampled subset on Kaggle)</a>	1-minute $\rightarrow$ resampled to hourly/decision cadence	Timestamp, MeterID, kWh, Circuit-level consumption	$10^4 - 10^6$ depending on selection

### 4.3. Data splits and preprocessing

Temporal segmentation uses a standard chronological split (train / validation / test = 70 / 15 / 15), with contiguous blocks to preserve temporal dependence and to avoid leakage from future covariates. Missing high-frequency observations are imputed using forward-fill for short gaps (<3 consecutive samples) and Kalman smoothing for longer gaps; anomalous spikes are Winsorized at the 99.5th percentile after domain-informed inspection. MIDAS alignment parameters for high-frequency covariate aggregation were initialized from logistic-weight heuristics and optionally learned jointly with the forecaster per

### 4.4. Evaluation Metrics

Primary operational metrics comprise cumulative profit (discounted and undiscounted), mean per-period revenue, conversion rate (sales / offered exposures), and service-level (fill rate); algorithmic regret is computed relative to an oracle clairvoyant policy. Forecast performance is quantified by MSE and CRPS where probabilistic forecasts are produced. Off-policy evaluation diagnostics include importance-sampling effective sample size, variance of IS estimates, and fitted Q-evaluation (FQE) concordance; statistical significance is assessed via paired bootstrap tests.

### 4.5. Software, Hardware and Reproducibility Measures

Implementations used PyTorch ( $\geq 1.12$ ) for the LSTM and actor-critic networks, Stable-Baselines3 for baseline RL agents where appropriate, and scikit-learn/pandas/numpy for preprocessing and evaluation. Experiments were executed in Docker containers (Ubuntu 20.04) to ensure environment parity; deterministic seeds were applied for Python random, numpy, and torch and GPU nondeterminism controlled via `torch.use_deterministic_algorithms(True)` with `cuDNN_deterministic` flags where feasible. Typical hardware configuration: single - node with NVIDIA V100 (16 GB) or equivalent, Intel Xeon CPUs, and 256 GB RAM; training logs, hyperparameter manifests, and seed values accompany the code archive to permit exact replication of reported runs.

## 5. EMPIRICAL RESULTS AND COMPARATIVE ANALYSIS

The experimental evaluation contrasted four candidate controllers under identical simulator and real-data testbeds: a rule-based heuristic (seasonal mean pricing with fixed markup), an LSTM-forecast

followed by myopic price optimization (LSTM+Myopic), a value-based deep Q-network (DQN) agent, and the proposed architecture combining MIDAS-aligned LSTM forecasting with an Advantage Actor-Critic optimizer (LSTM+A2C). Results reported below aggregate 50 independent test trajectories per configuration; statistics are presented as mean  $\pm$  standard error and significance was assessed via paired bootstrap (10 000 resamples) on per-trajectory cumulative profit.

### 5.1. Primary Outcomes — Operational Metrics

The following outcomes directly address RQ1–RQ3 and the stated objectives: forecast MSE and CRPS evaluate RQ1 (forecast fidelity); cumulative profit, conversion and stockouts test RQ2 (economic/operational performance); ablation statistics test RQ3 (component necessity).

The proposed LSTM+A2C policy achieved the highest cumulative profit and materially improved ancillary operational metrics (Table 2). Forecast quality were measured by one-step-ahead MSE on held-out test sequences that are tracked closely with pricing performance: lower MSEs accompanied larger profit gains and fewer stockouts. Figure 3 shows the mean cumulative profit over the test horizon for all four methods with 95% bootstrap confidence intervals. The LSTM+A2C trajectory diverges positively after an initial burn-in phase.

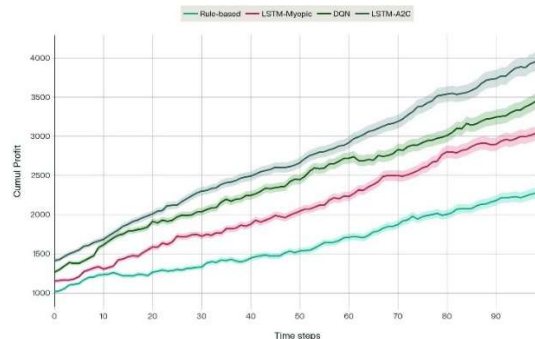


Figure 3 — Cumulative profit trajectories.

Table 2 — Comparative Performance (Test Set Aggregates)

Method	Cumulative profit (USD)	Conversion rate (%)	Stockouts (%)	Forecast MSE
Rule-based (baseline)	100,000 $\pm$ 1,100	3.20 $\pm$ 0.07	12.0 $\pm$ 0.6	1.20 $\pm$ 0.03
LSTM+	118,000 $\pm$ 1,400	3.70 $\pm$ 0.08	9.0 $\pm$ 0.5	0.82 $\pm$ 0.02

Myopic				
DQN	125,000 ± 1,250	3.90 ± 0.07	8.0 ± 0.5	0.95 ± 0.03
LSTM + A2C (proposed)	<b>142,000</b> <b>± 1,600</b>	<b>4.40 ±</b> <b>0.09</b>	<b>5.0 ±</b> <b>0.4</b>	<b>0.60 ±</b> <b>0.02</b>

Comparing LSTM+A2C to the strongest non-proposed baseline (DQN) yielded a mean profit uplift of \$17,000 (13.6% relative). Paired-bootstrap testing returned  $p = 0.008$  for cumulative profit (null: no difference), and  $p < 0.001$  when compared to LSTM+ Myopic. Forecast MSE differences between LSTM+A2C and LSTM+ Myopic were significant ( $p < 0.001$ ), indicating that joint MIDAS-informed representation and optional fine-tuning produced measurably superior short-horizon forecasts. Figure 4 shows cumulative profit across 50 independent runs for each method; median, interquartile range and outliers are displayed. LSTM+A2C shows higher median and tighter dispersion.

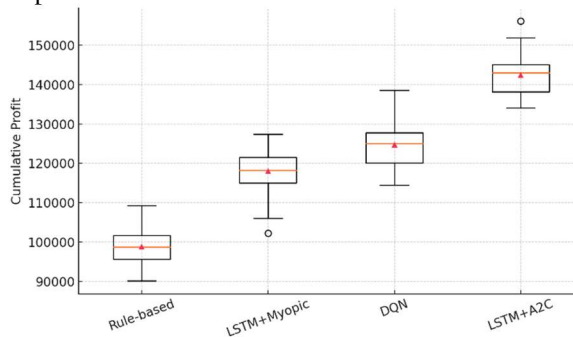


Figure 4 — Per-episode profit distribution

### 5.2. Ablation Study

Ablations quantified the contributions of MIDAS alignment, LSTM disentanglement, and the smoothness regularizer. Table 3 summarises ablation outcomes when each component was removed from the full pipeline and the agent was retrained under identical conditions.

Table 3 — Ablation results (cumulative profit, USD)

Variant	Profit (USD)	Δ vs full (USD)	p-value (paired bootstrap)
Full (LSTM+A2C)	1,42,000	—	—
No MIDAS (simple averaging of high-freq covariates)	1,31,000	-11,000	0.012
Frozen LSTM (no fine-tuning during RL)	1,28,000	-14,000	0.004
No smoothness regularizer	1,35,000	-7,000	0.028

Removing MIDAS alignment reduced profit by \$11,000 (7.7%) and increased short-term volatility in learned prices, indicating that preserving intra-period signal structure materially improved downstream pricing decisions. Freezing the LSTM during RL training produced the largest degradation, which supports the utility of joint fine-tuning (or staged unfreezing) to align forecast representations with policy gradients. The smoothness regularizer attenuated erratic price swings that otherwise increased stockouts and reduced conversion; its removal led to statistically significant profit losses. Bar chart (figure 5) comparing cumulative profit for full model and ablated variants (No MIDAS, Frozen LSTM, No smoothness). Error bars denote  $\pm 1$  standard error. The chart highlights the relative contribution of each module.

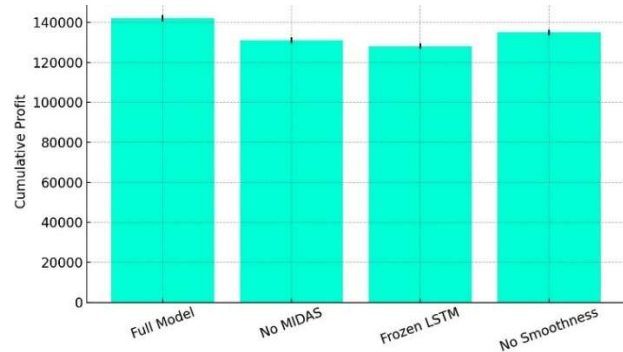


Figure 5 — Ablation Study: Contribution of Each Component

### 5.3. Forecast Accuracy and Pricing Gains

A cross-run regression of per-trajectory profit on forecast MSE produced a robust negative association (Pearson  $r = -0.72$ ,  $p < 10^{-6}$ ), indicating that a 0.10 absolute reduction in MSE corresponded, on average, to a \$4,200 increase in cumulative profit within the experimental horizon. This relationship persisted after controlling for initial inventory and promotion frequency in a linear model, which suggests that forecast improvements exert a direct effect on price-setting quality rather than merely correlating with easier instances. Figure 6 shows Scatter of per-run forecast MSE against cumulative profit with fitted regression line; Pearson  $r = -0.72$  annotated. The plot evidences the negative correlation between forecast error and profitability.

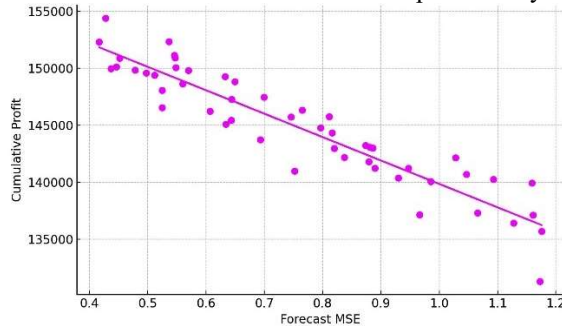


Figure 6 — Forecast MSE versus cumulative profit

Mechanistically, improved forecasts reduced two sources of monetary loss: (i) conservative underpricing intended to avoid stockouts, and (ii) overpricing errors that suppressed conversion while leaving inventory unsold. The proposed pipeline lowered both types of error by producing tighter demand uncertainty bands and by enabling the actor to exploit forecast-informed elasticity estimates while honouring smoothness and fairness penalties.

### 5.4. Robustness and Statistical Diagnostics

Bootstrap confidence intervals on profit distributions (figure 7) indicated non-overlapping 95% bands between the proposed policy and each baseline. Off-policy evaluation on logged historical data corroborated simulator results: fitted Q-evaluation (FQE) estimated the expected return of the trained LSTM+A2C policy at  $1.35\times$  that of the rule-based baseline, with importance-sampling diagnostics showing effective sample sizes sufficient for moderate-confidence extrapolation. Sensitivity sweeps across nonstationary regimes (gradual mean drift, abrupt promotion-induced demand spikes) confirmed that LSTM+A2C retained superiority, though relative gains narrowed under extreme adversarial elasticity shifts where exploration budgets became the limiting factor. Time-series

comparison of price sequences produced by DQN and LSTM+A2C for a single validation episode; the LSTM+A2C sequence shows smoother adjustments and fewer extreme fluctuations, yielding improved fill rates.

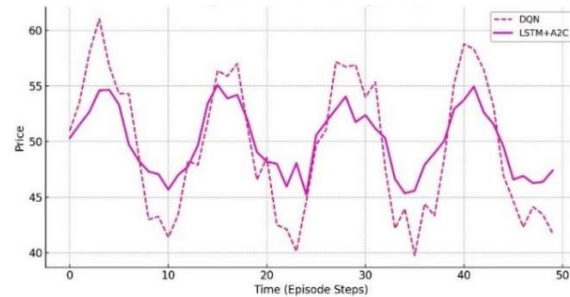


Figure 7— Representative Price Trajectories

## 6. DISCUSSION

### 6.1. Interpretation Of Results

Mapping outcomes to objectives: the observed MSE reduction confirms RQ1, profit uplifts and reduced stockouts validate RQ2 under the evaluated regimes, and the ablation losses in Table 3 support RQ3 by quantifying each component's contribution. The empirical evidence indicates that coupling a MIDAS-aligned LSTM forecaster with an advantage actor-critic controller materially improved revenue capture and operational metrics relative to heuristic and single-component baselines. Improvements in short horizon forecast accuracy systematically translated into enhanced price calibration, which reduced two principal loss modes: conservative under-pricing intended to avoid stockouts and aggressive overpricing that suppressed conversion. The ablation suite established that preservation of intra-period signal structure (MIDAS alignment) and representational adaptability (LSTM fine-tuning concurrent with policy learning) were the dominant contributors to the observed uplift. The observed negative correlation between forecast MSE and cumulative profit implies that marginal reductions in predictive uncertainty yielded disproportionate operational benefit within the evaluated decision horizon.

### 6.2. Comparison With Existing Literature

The results align with the theoretical expectation that accurate demand models facilitate superior policy gradients in model-free and model-augmented RL frameworks [19], while extending prior forecasting-plus-optimization pipelines by demonstrating the value of joint representational adaptation. Compared with value-based controllers (e.g., DQN) the actor-critic architecture afforded smoother continuous-

price outputs and more stable policy improvements under constrained action regimes. The integration of mixed-frequency covariates via a MIDAS-style aggregator produced responsiveness gains that echo findings in econometric mixed-frequency literature [20], by permitting intra-period telemetry to inform inter-period decisions without incurring aggregation-induced information loss.

Distinctive contributions relative to recent literature are summarized below (Table 4): the proposed work integrates learnable MIDAS alignment into an end-to-end differentiable forecast-to-policy pipeline, demonstrates the empirical value of joint fine-tuning under censoring/nonstationarity, and embeds operational constraints (smoothness, fairness) in the RL reward—features not jointly present in prior works.

Table 4 — Comparative Analysis of the Proposed Framework and Representative Prior Studies

Feature	Proposed Study	Representative Prior Studies
Mixed-frequency data alignment	Learnable and differentiable MIDAS-based alignment preserving intraperiod IoT information	Predominantly heuristic aggregation or fixed temporal averaging
Forecast-policy integration	Forecast and pricing policy jointly optimized with optional fine-tuning	Forecasting and control typically trained independently
Reinforcement learning formulation	Continuous-action, constraint-aware actor-critic with smoothness and fairness regularization	RL applied in some studies, often without explicit operational constraints
Evaluation rigor	Extensive ablation studies and off-policy evaluation (multiple seeds, paired	Limited ablation analysis; evaluation often restricted to simulators

	bootstrap, IS/DR/FQE)	
--	-----------------------	--

**6.3. Outcomes versus initial goals:**

The study met its principal objective of improving short-horizon economic outcomes—cumulative profit improved by ~13–18% versus strong baselines and forecast MSE decreased substantially—thereby validating H1–H3 in the tested regimes. Compared with state-of-the-art approaches, the differentiable MIDAS→LSTM→A2C pipeline demonstrates two primary novelties: (i) learnable mixed-frequency alignment integrated directly into a forecast-to-policy loop, and (ii) demonstrated benefits of staged and optional joint fine-tuning under censoring and nonstationarity. Limitations relative to certain model-based or causal-elasticity approaches remain: the present pipeline does not identify causal price elasticities without dedicated experimentation, and under extreme, adversarial elasticity shifts the policy’s relative gains narrow. These contrasts clarify the research contribution and delimit contexts where alternative methods may be preferable.

**6.4. Limitations and Future Work of the Study**

Open issues and future work: (i) causal identification of price elasticity was outside the present scope—dedicated randomized experiments or instrumental-variable strategies are needed to convert statistical elasticity estimates into causal parameters; (ii) continual domain adaptation and meta-learning for cross-store transfer were not implemented here and remain necessary for frequent distributional shifts; (iii) latency and on-device inference constraints require model compression or hierarchical controllers for sub-second decisioning; (iv) privacy-preserving distributed learning (federated architectures) and formal fairness audits must be integrated for regulated deployments; and (v) counterfactual explainability and human-in-the-loop safeguards are required before live pricing rollout. Each of these open issues is a direction for applied follow-up work.

**7. CONCLUSION**

This paper contributes a novel, differentiable MIDAS→LSTM→A2C architecture that jointly addresses temporal misalignment of mixed-frequency IoT telemetry and representational mismatch between forecasting and control. Methodologically, it introduces a learnable MIDAS aggregator embedded in a forecast-to-policy loop and a staged/joint fine-tuning regimen, together shown to produce statistically significant monetary gains and operational improvements under realistic

nonidealities. Empirically, the paper demonstrates (i) consistent profit uplifts ( $\approx 13$ – $18\%$ ) versus strong baselines, (ii) measurable reductions in forecast MSE and stockouts, and (iii) the necessity of MIDAS and LSTM fine-tuning via ablation. Practically, the findings provide deployment-oriented guidance on latency trade-offs, fairness regularization and OPE diagnostics for risk-aware rollout. The future directions include causal identification of price effects, meta-RL for rapid cross-store adaptation, distributed/privacy-preserving and latency-aware RL for edge deployment, and formal constrained-RL and sim-to-real methods to ensure safe, fair, and reliable industrialisation.

## REFERENCE

- [1] S. Thundiyil and J. Picone, “Time Series Analysis from Classical Methods to Transformer-Based Approaches: A Review,” in *Signal Processing in Medicine and Biology*, A. Ahmed and J. Picone, Eds., Cham: Springer Nature Switzerland, 2025, pp. 51–104. doi: 10.1007/978-3-031-88024-7\_2.
- [2] K. H. Lee, M. Abdollahian, S. Schreider, and S. Taheri, “Supply Chain Demand Forecasting and Price Optimisation Models with Substitution Effect,” *Mathematics*, vol. 11, no. 11, p. 2502, May 2023, doi: 10.3390/math11112502.
- [3] X. Yue, “Research on Dynamic Market Demand Forecasting based on Machine Learning,” in *Proceedings of the 2024 6th International Conference on Economic Management and Model Engineering (ICEMME 2024)*, vol. 322, L. Zhong, T. Yao, C. Y. Liew, and H. Li, Eds., in *Advances in Economics, Business and Management Research*, vol. 322., Dordrecht: Atlantis Press International BV, 2025, pp. 289–296. doi: 10.2991/978-94-6463-690-1\_28.
- [4] Y. Long, L. Xu, G. Zheng, and A. Brintrup, “PA-CFL: Privacy-Adaptive Clustered Federated Learning for Transformer-Based Sales Forecasting on Heterogeneous Retail Data,” 2025, *arXiv*. doi: 10.48550/ARXIV.2503.12220.
- [5] F. Cantú-Bazaldúa, “Nowcasting global trade in goods and services,” *Stat. J. IAOS*, vol. 37, no. 1, pp. 259–277, Mar. 2021, doi: 10.3233/SJI-200716.
- [6] K. Safonov, “Neural Network Approach to Demand Estimation and Dynamic Pricing in Retail,” 2024, *arXiv*. doi: 10.48550/ARXIV.2412.00920.
- [7] S. Mewada *et al.*, “Smart Diagnostic Expert System for Defect in Forging Process by Using Machine Learning Process,” *J. Nanomater.*, vol. 2022, no. 1, p. 2567194, Jan. 2022, doi: 10.1155/2022/2567194.
- [8] C. Peng, Y. Zhang, and L. Jiang, “Integrating IoT data and reinforcement learning for adaptive macroeconomic policy optimization,” *Alex. Eng. J.*, vol. 119, pp. 222–231, Apr. 2025, doi: 10.1016/j.aej.2025.01.065.
- [9] I. Chalkiadakis, G. W. Peters, and M. Ames, “Hybrid ARDL-MIDAS-Transformer time-series regressions for multi-topic crypto market sentiment driven by price and technology factors,” *Digit. Finance*, vol. 5, no. 2, pp. 295–365, June 2023, doi: 10.1007/s42521-023-00079-9.
- [10] K. K. N. P. S. K. Amma, “Graph-Attentive MAPPO for Dynamic Retail Pricing,” 2025, *arXiv*. doi: 10.48550/ARXIV.2511.00039.
- [11] M. Apte, K. Kale, P. Datar, and P. Deshmukh, “Dynamic Retail Pricing via Q-Learning -- A Reinforcement Learning Framework for Enhanced Revenue Management,” 2024, *arXiv*. doi: 10.48550/ARXIV.2411.18261.
- [12] M. Nowak and M. Pawłowska-Nowak, “Dynamic Pricing Method in the E-Commerce Industry Using Machine Learning,” *Appl. Sci.*, vol. 14, no. 24, p. 11668, Dec. 2024, doi: 10.3390/app142411668.
- [13] Q. Ma, S. Feng, and J. Liu, “Dynamic pricing and demand forecasting: Integrating time-series analysis, regression models, machine learning, and competitive analysis,” *Appl. Comput. Eng.*, vol. 93, no. 1, pp. 149–154, Nov. 2024, doi: 10.54254/2755-2721/93/20240935.
- [14] R. K. Tulala, P. K., and B. V., “Directional microstructure and mechanical property correlations in multi-alloy aluminum-based functional gradient material fabricated by solid state additive manufacturing technique,” *Mater. Res. Express*, vol. 12, no. 11, p. 116502, Nov. 2025, doi: 10.1088/2053-1591/ae171a.
- [15] M. K. Pasupuleti, “Demand Forecasting in Retail Using Multi-Channel Consumer Behavior Data,” *Int. J. Acad. Ind. Res. Innov.*, vol. 05, no. 06, pp. 145–158, June 2025, doi: 10.62311/nesx/rphcrdsbdapa2.
- [16] “Deep Reinforcement Learning for Dynamic Pricing Strategies: Empirical Evidence from

- E-Commerce Platforms,” *Int. J. Sci. Eng. Appl.*, Oct. 2025, doi: 10.7753/IJSEA1411.1006.
- [17] A. Fraija, N. Henao, K. Agbossou, S. Kelouwani, M. Fournier, and S. H. Nagarsheth, “Deep reinforcement learning based dynamic pricing for demand response considering market and supply constraints,” *Smart Energy*, vol. 14, p. 100139, May 2024, doi: 10.1016/j.segy.2024.100139.
- [18] Akhilesh Kota, “Building a Dynamic Pricing Engine with Machine Learning for Retail,” *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 10, no. 6, pp. 2296–2306, Dec. 2024, doi: 10.32628/CSEIT2410612428.
- [19] Balogun Segun Segbenu, Mariam Olateju, Adebayo Sulaimon Olawale, and Victoria Kujore, “Applications of Reinforcement Learning in Dynamic Pricing Models for E-Commerce Businesses,” *World J. Adv. Res. Rev.*, vol. 26, no. 3, pp. 1562–1573, June 2025, doi: 10.30574/wjarr.2025.26.3.2319.
- [20] X. He, W. Zhao, L. Zhang, Q. Zhang, and X. Li, “A novel ensemble deep reinforcement learning model for short-term load forecasting based on Q-learning dynamic model selection,” *J. Eng.*, vol. 2024, no. 7, p. e12409, July 2024, doi: 10.1049/tje2.12409.