KMANPUB | AI and Tech in Behavioral and Social Sciences Vol. 5 No. 3 (2027): 1-20

Article history:
Received 30 May 2026
Revised 07 October 2026
Accepted 15 October 2026
Initial Publication 26 October 2026
Final Publication 01 July 2027

Transformer-Based Multi-Horizon Traffic Forecasting in Bulk Port Environments

Sanaz. Jafarinia[1], Mohammad Ali. Afshar Kazemi[1*], Seyed Abdollah. Amin Mousavi[1], Hassan. Mehrmanesh[1], Mohammad. Ali Keramati[1]

[1] Department of Industrial Management, CT.C., Islamic Azad University, Tehran, Iran

*** Corresponding author email address**: Moh.afsharkazemi@iauctb.ac.ir

A r t i c l e I n f o

Article type:
Original Research

How to cite this article:
Jafarinia, S., Afshar Kazemi, M. A., Amin Mousavi, S. A., Mehrmanesh, H., & Keramati, M. (2027). Transformer-Based Multi-Horizon Traffic Forecasting in Bulk Port Environments. *AI and Tech in Behavioral and Social Sciences, 5*(3), 1-20.
https://doi.org/10.61838/kman.aitech.6488

© 2027 the authors. Published by KMAN Publication Inc. (KMANPUB), Ontario, Canada. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License.

A B S T R A C T

Reliable multi-horizon traffic forecasts can improve operational planning at bulk ports, but the performance of data-intensive forecasting models depends on training-set size and forecast horizon. This study compares seven architectures—ARIMA, LSTM, GRU, a standard Transformer, PatchTST, iTransformer, and the Temporal Fusion Transformer—using aggregate vehicle counts and mean speeds generated by a statistically checked agent-based discrete-event simulation. Forecasts cover short (15–30-minute), medium (seven-day), and long (30-day) horizons. A preliminary 569-observation short-horizon experiment favored ARIMA; after expanding the simulated short-horizon series to 51,840 observations, all six deep models outperformed that baseline. For 15-minute vehicle-count prediction, PatchTST achieved $R^2 = 0.4921$ compared with ARIMA's $R^2 = 0.3632$. Mean R^2 varied non-monotonically across short, medium, and long horizon tiers (0.290, 0.517, and 0.325, respectively). The greatest transformer advantages emerged in longer-horizon speed forecasts, where PatchTST attained R^2 values of 0.9821 and 0.9741 at the medium and long tiers. Feature-ablation results suggested that the value of additional predictors depended on sample availability. A supplementary evaluation using independent DT-CARGO truck telemetry supported the advantage of deep approaches over ARIMA, although recurrent models slightly surpassed transformer variants on that external series. These findings support horizon- and data-aware model selection rather than a universal architectural ranking. Results based on simulation require validation in operational bulk-port settings.

Keywords: *Transformer-Based Traffic Forecasting, Port Traffic Forecasting, Multi-Horizon Time Series, Discrete-Event Simulation*

1. Introduction

Accurate traffic forecasting is critical for optimizing operations and enhancing efficiency within complex logistical environments such as bulk ports (Lamdjad, 2025) (Wang et al., 2024). However, comparative studies reporting such advantages do not always verify that the training data volume available is sufficient for these architectures to realize their claimed benefits over classical statistical baselines (Zhu, 2023) (Kambale, 2023) (Miño Calero et al., 2026) a methodological gap this paper addresses directly rather than assumes away. This research addresses traffic forecasting for bulk cargo and its associated trailers in port environments across three complementary horizons: short-term (15–30 minutes), medium-term (one week ahead), and long-term (one month ahead). This paper reports both outcomes explicitly, since the comparison itself not only the corrected result carries a methodological lesson for comparative forecasting studies in this domain. Beyond architecture comparison, this study also examines which engineered input features contribute meaningfully to forecasting accuracy at each horizon, and the extent to which findings obtained from simulated data generalize to independent real-world vehicle telemetry. Related transportation research has also explored adaptive route selection under changing traffic conditions using reinforcement learning, emphasizing the connection between intelligent transportation decisions and traffic efficiency.

2. Literature Review

2.1. Time Series Forecasting with Transformers

The field of time series forecasting has witnessed significant advancements with the advent of deep learning, particularly the Transformer architecture (Okamoto, 2026; Zhou, 2020). Unlike traditional recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), Transformers can process sequences in parallel, leading to faster training times and improved performance on long sequences (Dos Santos et al., 2025).

A further architectural development relevant to this study is the Temporal Fusion Transformer (TFT), which combines gated residual processing, a learned variable-selection mechanism over input features, and recurrent local processing with self-attention, and has been proposed specifically for interpretable multi-horizon forecasting (Lim et al., 2021). Unlike the Transformer variants above, TFT is designed to report which input variables the model relies on for a given prediction, a property directly relevant to feature-level analysis such as the ablation study conducted in this paper.

These four Transformer-family variants are not redundant alternatives but represent distinct design philosophies, and this distinction motivates their joint inclusion in this study rather than selecting a single "best" Transformer a priori. Because this study's forecasting task varies enormously in available sample size across the three horizon tiers examined (Section 3.1), and in the number and informativeness of engineered features across tiers (Section 3.3), these four biases are not equally well matched to every tier a consideration returned to when interpreting cross-tier and ablation results (Sections 4.4–4.5, 5.2–5.3).

2.2. Traffic Forecasting in Port Environments

Studies have explored various deep learning architectures for port traffic forecasting, though typically evaluating a single architecture at a single operational configuration rather than a controlled multi-architecture, multi-horizon comparison. Shah (2025) specifically highlights the application of Transformer architecture for real-time traffic flow prediction in port environments, focusing on optimizing the movement of specific vehicle types like fish trucks, and demonstrates that Transformer-based models can effectively manage dynamic port traffic by integrating multimodal sensor data and utilizing attention mechanisms to capture spatial and temporal dependencies, outperforming traditional machine learning models (Shah et al., 2025); this comparison, however, is not extended to classical statistical baselines such as ARIMA under matched data conditions, leaving open whether the reported advantage would persist or reverse under a constrained-data regime of the kind examined in this paper. The integration of explainable AI (XAI) with deep learning models in maritime logistics also addresses the need for transparency and interpretability in forecasting container throughput, thereby supporting data-driven decision-making in port management systems (Lamdjad, 2025); this interpretability-focused contribution is complementary to, rather than in competition with, the systematic architecture-and-data-volume comparison undertaken here, and the TFT architecture's variable-selection mechanism (discussed above) offers a natural

AITBSS AI and Tech in Behavioral and Social Sciences E-ISSN: 3041-9433

point of connection between the two research directions that future work could pursue jointly.

2.3. Deep Learning in Bulk Cargo and Shipping Operations

Deep reinforcement learning frameworks have been developed to optimize bulk cargo terminal unloading and outbound operations, integrating transportation planning, yard selection, and equipment scheduling to improve operational efficiency (Li et al., 2025); this targets terminal-operations scheduling as a control/optimization problem rather than traffic forecasting per se, and is included here for domain relevance rather than as a directly comparable forecasting benchmark. The comprehensive review by Wang (2024) further underscores the growing application of deep learning in maritime time series analysis, covering various aspects of ship and port operations (Wang et al., 2024), providing broader context for the specific architecture comparison undertaken in this paper.

2.4. Simulation-Based Data Generation for Forecasting Research

Public benchmark datasets from adjacent domains offer an alternative, but as this paper's own experience with the DT-CARGO heavy-vehicle telemetry dataset demonstrates (Section 3.10), an adjacent-domain dataset is not automatically fit for purpose: its speed variable required principled cleaning using the dataset's own documented GPS-signal-loss quality field before it could be used reliably, and even after cleaning, it captures general heavy-vehicle operation rather than bulk-port-specific dynamics, limiting it to an external validation role rather than a primary training data source in this study. Discrete-event simulation offers a third route, generating large synthetic datasets for machine-learning training when production data is inaccessible (Chan, 2022), while complex simulation systems may require validation

2.5. Research Gap and Contribution of This Study

The current research builds upon these advancements by evaluating short-, medium-, and long-horizon traffic forecasting for bulk port environments using both classical and Transformer-based models under conditions that explicitly control for training data volume a factor the literature reviewed above treats as architecturally settled more often than it is empirically tested within a single study, and one that this paper's own initial (pre-correction) results show can reverse an architecture comparison's conclusions entirely if left uncontrolled. Finally, this study benchmarks the same architectures on independent real-world heavy-vehicle telemetry data, providing external validation evidence for whether conclusions drawn from simulated port data generalize beyond the simulation that produced them addressing a gap in the reviewed literature, which largely evaluates architectures on a single data source without a corresponding real-world cross-check, and which, in this study's case, surfaced a genuine discrepancy (a reversal in the relative ranking of recurrent versus Transformer architectures) that a single-data-source evaluation could not have revealed.

3. Methods and Materials

This section describes the complete methodological framework used to generate the results in Section 4: the three-tier forecasting design, the discrete-event simulation used as the primary data source, feature engineering, model architectures, training procedure, evaluation metrics, the ablation study design, the simulation validation procedure, and the external real-world validation dataset (DT-CARGO). All design decisions described here reflect the final, corrected pipeline; two methodological issues identified and resolved during development an insufficient short-horizon training set, and a berth capacity/demand imbalance in the simulation are described in place, since they materially shaped the final design rather than being incidental implementation details.

3.1. Research Design and Three-Tier Forecasting Framework

This study addresses port traffic forecasting at three complementary horizon tiers, motivated by the hypothesis tested directly in Section 4 that the relative performance of transformer-based architectures against classical and recurrent baselines depends on both training data volume and forecast horizon. The Short tier forecasts 15- and 30-minute-ahead traffic from a 5-minute-resolution time series. The Mid tier forecasts 7 days ahead from a 28-day lookback window on daily-aggregated data. The Long tier forecasts 30 days ahead from a 90-day lookback window, also on daily-aggregated data. All three tiers share the same underlying data source (Section 3.2), the same seven candidate architectures (Section 3.5), and the same training and evaluation procedure (Sections 3.6–3.7), differing only

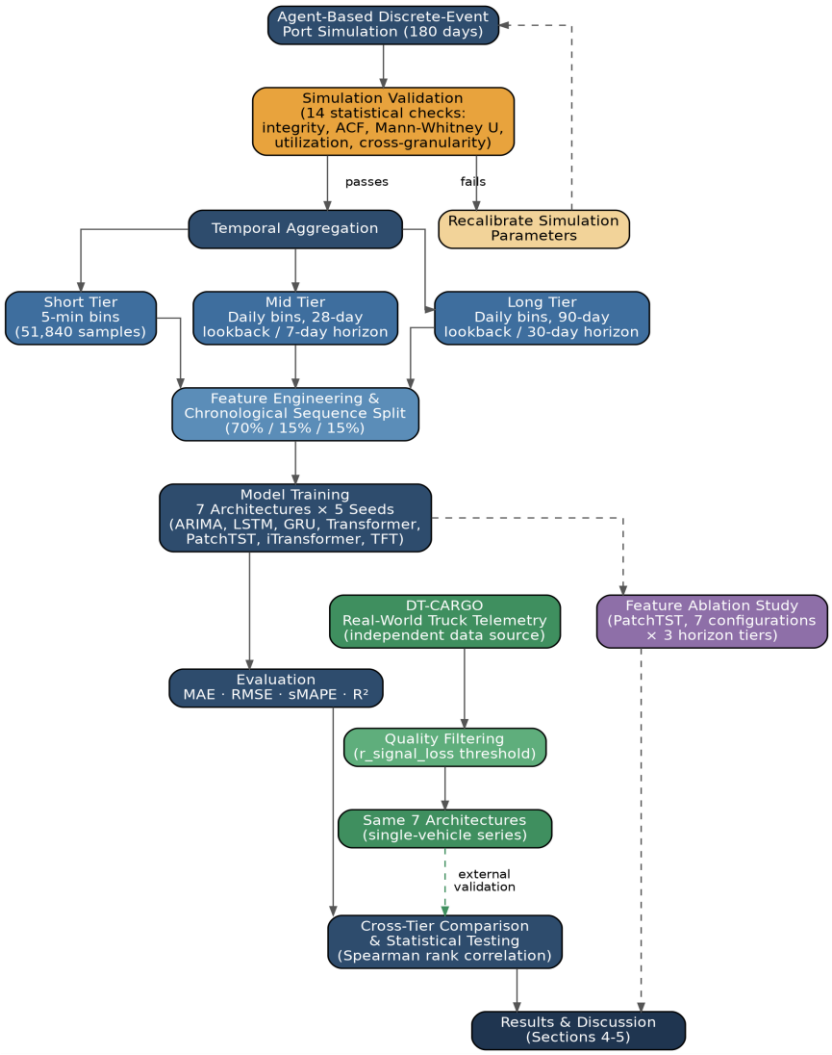
AITBSS AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

in temporal aggregation granularity and lookback/horizon length.

The overall research workflow is summarized in Figure 1.

Figure 1

Overview of the research pipeline, from simulation through cross-tier comparison, including the feature ablation and external (DT-CARGO) validation branches



3.2. Data Source: Discrete-Event Port Simulation

Training data was generated by an original agent-based discrete-event simulation (implemented in Python with SimPy) of a generic bulk port comprising five berths, three stockyards, a blending plant, a weighbridge, a parking area, and entry/exit gates. Fifty trailers are pre-positioned at simulation start; additional trailers are activated by stochastic vessel arrivals, governed by a Poisson process with a mean arrival rate that is approximately 2.5 times higher on weekdays than on weekends, reflecting realistic port operating patterns.

Each trailer agent operates under one of four primary modes, selected by a properly normalized probability distribution summing to 100%: Load and Exit (40%), Load and Internal Unload (35%), Partial Top-Up (15%), and Empty Repositioning (10%). A separate, independent generator produces road-arriving, pre-loaded trailers per vessel arrival, contributing empirically around 15% of total trailer traffic. GPS coordinates are generated via haversine-based interpolation between waypoints with added Gaussian noise, but stated explicitly, since it materially scopes this paper spatial trajectory data is used only for illustrative visualization in this study and is not used as a

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

feature or forecasting target in any reported model comparison; this paper addresses aggregate operational traffic forecasting (vehicle counts and speeds), not spatial trajectory forecasting.

An illustrative view of the simulated port is shown in Figure 2.

Figure 2

Simulation View

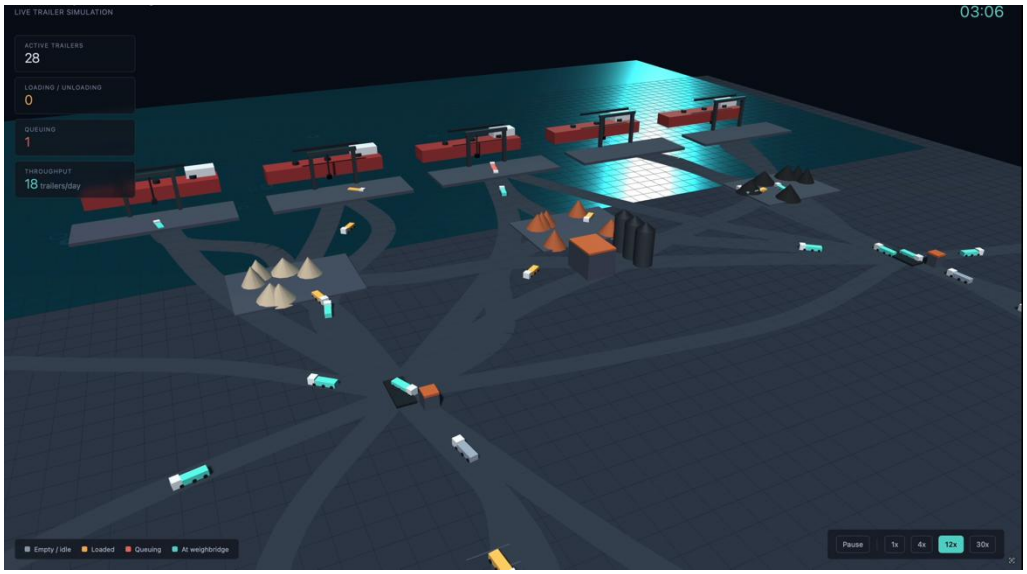


For the Mid and Long tiers, an interval-logging approach records one entry per operational segment (e.g., loading, queuing, weighbridge service) rather than dense fixed-interval GPS pings, keeping data volume tractable across multi-month simulated durations while preserving the same underlying agent dynamics. Trailer travel speed is tied to same-day congestion itself a function of the weekday/weekend vessel arrival-rate difference so that average speed carries genuine weekly structure rather than being sampled independently of operational conditions.

Figure 3 provides an additional visualization of the simulation environment.

Figure 3

Simulation view



AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

During development, a berth capacity/demand imbalance was identified via the validation procedure described in Section 3.9: queuing accounted for over 65% of total simulated time in an early configuration, indicating that trailer arrival rate substantially exceeded berth service capacity. This was resolved by recalibrating fleet trip cadence, berth count, vessel-to-trailer arrival stagger window, and per-vessel trailer count distribution, reducing queuing to 24.4% of total simulated time (Section 4.1) without disrupting the underlying weekly statistical structure.

3.3. *Feature Engineering and Temporal Aggregation*

For the Short tier, raw simulation output is aggregated into uniform 5-minute bins. Seven features are extracted per bin: active vehicle count, average speed (km/h), average cargo level, queue ratio (fraction of records in a queuing state), and cyclical hour-of-day encoding (sine and cosine components) plus a binary peak-hour indicator. Cyclical encoding avoids the discontinuity a linear hour representation would introduce between hour 23 and hour 0.

For the Mid and Long tiers, the same underlying interval-logged records are aggregated to daily resolution. In addition to vehicle count, average speed, average cargo level, and queue ratio, cyclical day-of-week encoding (sine and cosine) and a binary weekend indicator are added, making weekly structure explicitly available to the model at these coarser horizons a design choice directly relevant to the horizon-dependent results reported in Section 4.3–4.4.

A methodological correction made during development is reported explicitly: the original short-horizon dataset consisted of only 569 five-minute bins (48 simulated hours), which proved insufficient for training the data-hungry deep architectures evaluated in this study (Section 5.1). The final pipeline instead derives the Short-tier dataset from the same 180-day simulation used for the Mid/Long tiers, aggregated to 5-minute bins, yielding 51,840 bins a roughly 90-fold increase while keeping the feature schema identical.

3.4. *Sequence Construction and Data Partitioning*

Both temporal resolutions are converted to supervised learning windows via a sliding-window procedure: an input window of length L produces a forecasting target of length H. For the Short tier, L = 3 time steps (a 15-minute lookback) with H ∈ 3, 6 steps (15- and 30-minute horizons). For the Mid tier, L = 28 days and H = 7 days; for the Long tier, L = 90 days and H = 30 days.

Data is partitioned chronologically (70% train, 15% validation, 15% test) rather than randomly, preserving temporal order and preventing information leakage from future observations into training. Feature and target scalers (Min-Max) are fit exclusively on the training portion and then applied to the full series correcting an earlier implementation in which scalers were fit on the complete series including validation/test data, artificially inflating apparent performance.

3.5. *Model Architectures*

Seven architectures are compared at every horizon tier. ARIMA (2,1,2) represents classical statistical forecasting, operating on a single univariate series; for computational tractability at the 51,840-bin Short-tier scale, full parameter re-estimation occurs periodically rather than at every window, with incremental state updates (via statsmodels' append mechanism) between full refits. LSTM and GRU (two layers, 64 hidden units, linear multi-step projection head) represent gated recurrent baselines.

Four transformer-family architectures are evaluated, all sharing scaled dot-product self-attention as their core operation (Vaswani, 2017). The plain Transformer is a standard encoder-only architecture with sinusoidal positional encoding. PatchTST (Nie et al., 2023) splits the input sequence into non-overlapping patches, treating each as a token; patch length and stride were corrected during development after an initial configuration (patch length equal to the full lookback window) collapsed the architecture to effectively a single token with no local structure. iTransformer (Liu et al., 2025) inverts the conventional attention axis, attending across variates rather than across time steps. The Temporal Fusion Transformer (TFT) implementation includes the architecture's defining components as described in the original paper (Lim et al., 2021): a Variable Selection Network assigning learned softmax weights to each input feature, Gated Residual Network blocks as the primary processing unit throughout, an LSTM encoder for local temporal processing, and gated self-attention a materially more faithful implementation than an earlier version of this pipeline, which used a generic transformer with sigmoid gates appended before and after a standard encoder.

AITBSS AI and Tech in Behavioral and Social Sciences E-ISSN: 3041-9433

3.6. Training Procedure

All deep models are trained with the Adam optimizer (learning rate = 0.001, weight decay = 0.0001) and cosine-annealing learning-rate scheduling for up to 40 epochs, with gradient clipping at norm 1.0. Patience-based early stopping (patience = 10 epochs, monitoring validation loss) halts training when no improvement is observed, with the best validation checkpoint restored added specifically to mitigate overfitting risk observed in higher-parameter-count architectures (notably TFT) on smaller datasets. Because deep models can produce high-variance results on modest-sized datasets, each architecture is trained with five random seeds per (target, horizon, tier) combination for the primary results (Sections 4.2–4.4), with mean and standard deviation reported; the feature ablation study (Section 3.8) and the DT-CARGO external validation (Section 3.10) use single-seed training, explicitly flagged as a lower rigor level appropriate to a supplementary rather than primary analysis.

3.7. Evaluation Metrics

Four complementary metrics are reported, computed on inverse-transformed (original-scale) predictions: mean absolute error (MAE), root mean squared error (RMSE), symmetric mean absolute percentage error (sMAPE), and the coefficient of determination (R^2). sMAPE is used in place of conventional MAPE because MAPE diverges toward infinity for near-zero target values (e.g., speed during a full stop or queuing state), whereas sMAPE remains bounded within [0%, 200%] by dividing by the mean of the absolute actual and predicted values.

3.8. Ablation Study Design

A feature ablation study using PatchTST (vehicle_count target) was conducted at all three horizon tiers to quantify the marginal contribution of each feature group. Seven configurations were evaluated at each tier: the full feature set; the full set minus cargo level; the full set minus temporal encoding (hour-of-day for Short, day-of-week for Mid/Long); the full set minus queue ratio; speed only; vehicle count plus speed; and a univariate configuration (vehicle count only). This design isolates the marginal contribution of each engineered feature group rather than only comparing a fixed feature set across architectures.

3.9. Simulation Validation Methodology

Before using simulated data for model training, a four-category validation procedure is applied. Statistical evidence for weekly structure is obtained via the autocorrelation function of daily vehicle count (expected to peak at lag 7, 14, 21, and 28 days) and a non-parametric Mann-Whitney U test for the weekday/weekend difference in both vehicle count and average speed a formal significance test rather than a visual comparison of means. Finally, cross-granularity consistency is verified by comparing segment counts between the daily and 5-minute aggregations, which are derived independently from the same underlying segment log and should match exactly.

3.10. External Validation Data Source: DT-CARGO

To assess whether findings generalize beyond simulation, DT-CARGO (Balke & Adenaw, 2023; DOI: 10.5281/zenodo.7599687) a public dataset of high-resolution GPS telemetry from 54 heavy commercial trucks is used as an independent real-world benchmark (Section 4.6). Only the dataset's track-level summary file (one row per completed trip, including average speed, maximum speed, distance, and a documented GPS-signal-loss quality field, r_signal_loss) is used, rather than the full 5.7 GB raw 10 Hz speed archive, extracted via HTTP range requests against the dataset's public Zenodo archive rather than a full download.

Trips are filtered using the dataset's own documented quality field (excluding trips with r_signal_loss above 5%, and average speeds outside a physically plausible 0.5–40 m/s range for a loaded commercial truck) a principled cleaning procedure using DT-CARGO's own designed-in quality indicator, correcting an earlier attempt that applied an undocumented threshold directly to the (extremely outlier-contaminated) raw speed variable. The vehicle with the most recorded trips after cleaning (3,169 trips) is used as a single real-world comparison series, analogous to how the simulated port is treated as a single series in the primary results. The forecasting task predicting a truck's next few trips' average speed from its preceding trips, using average speed, maximum speed, and distance as features, with the same L = 3, H = 3 window structure as the Short tier is a coarser-grained analog to the port's fixed-interval task, given DT-CARGO records one observation per completed trip rather than per fixed time interval; this distinction is maintained explicitly in the reported results (Section 4.6) rather than treated as a like-for-like replication.

AITBSS AI and Tech in Behavioral and Social Sciences E-ISSN: 3041-9433

3.11. Reproducibility

All experiments use fixed random seeds for NumPy and PyTorch. The complete simulation, preprocessing, model implementation, and experiment execution code is available as supplementary material.

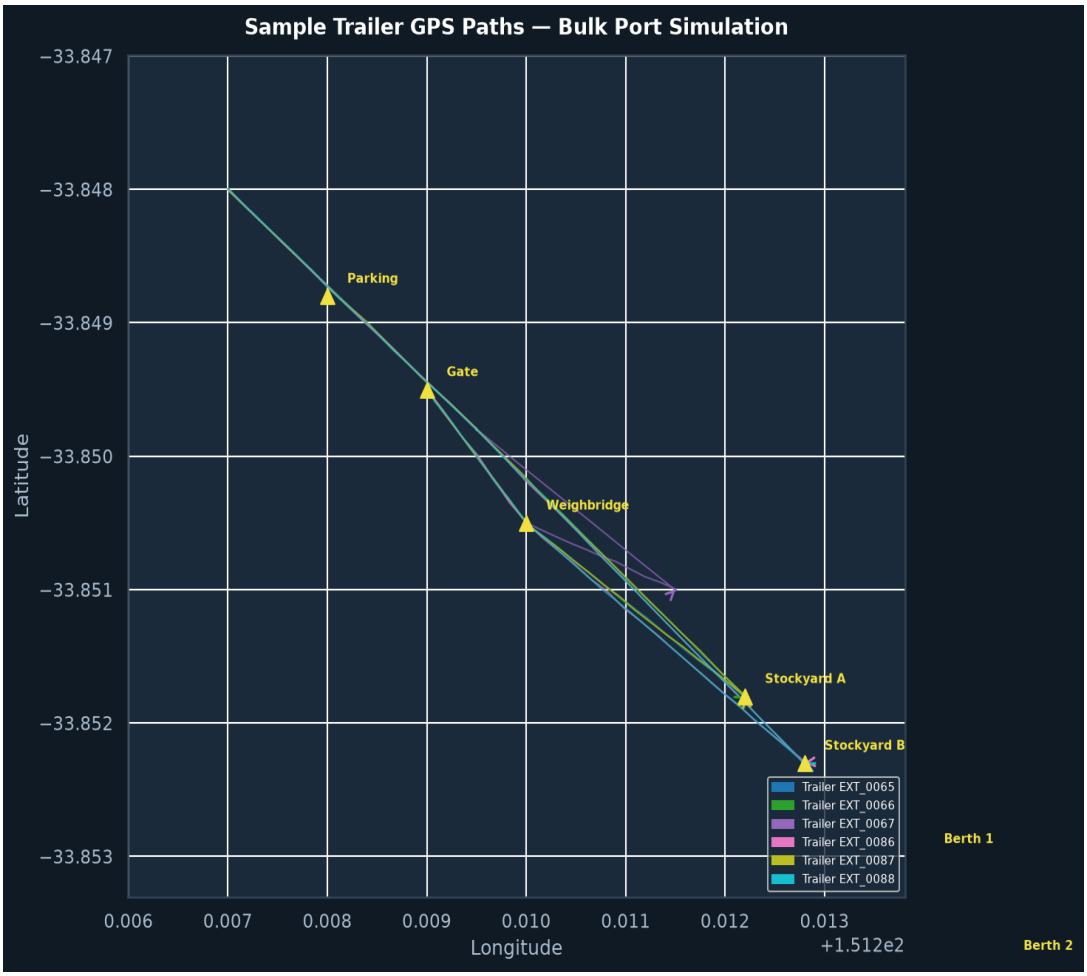
4. Findings and Results

This section presents the empirical findings from the corrected experimental pipeline described in Section 3. All results reported here reflect the final, fingerprint-verified run: the short-horizon experiments were trained on the full 51,840-bin 5-minute aggregate (not the original 569-bin 48-hour subset), and the mid/long-horizon experiments were trained on the recalibrated 180-day simulation in which the berth capacity/demand imbalance identified during validation (Section 3.9) had been resolved.

Illustrative trailer trajectories generated by the simulation are shown in Figure 4.

Figure 4

Sample Trailer Paths - Bulk Port Simulation



4.1. Simulation Validation

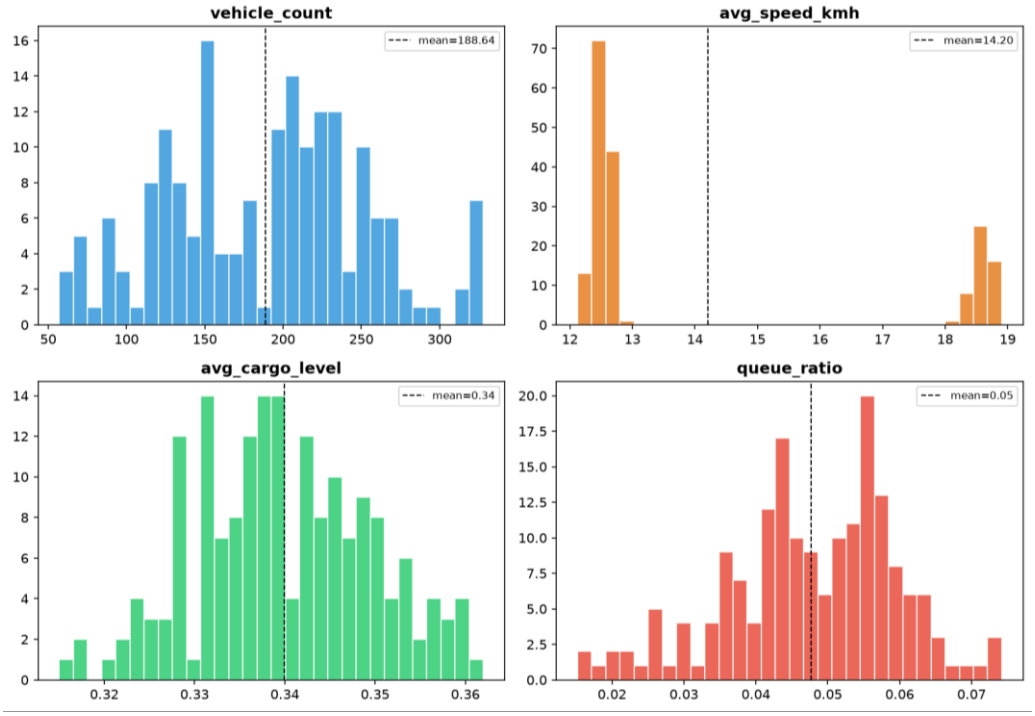
Before using the simulated dataset for model training, it was subjected to the four-category validation procedure described in Section 3.9. All 14 checks passed on the final 180-day run (432,214 segment records): no missing values, no negative durations, cargo levels and speeds within physically plausible ranges, and all expected operational states present.

The simulated daily-variable distributions are presented in Figure 5.

AITBSS
E-ISSN: 3041-9433

Figure 5

Distribution Sanity Check - Daily Aggregate

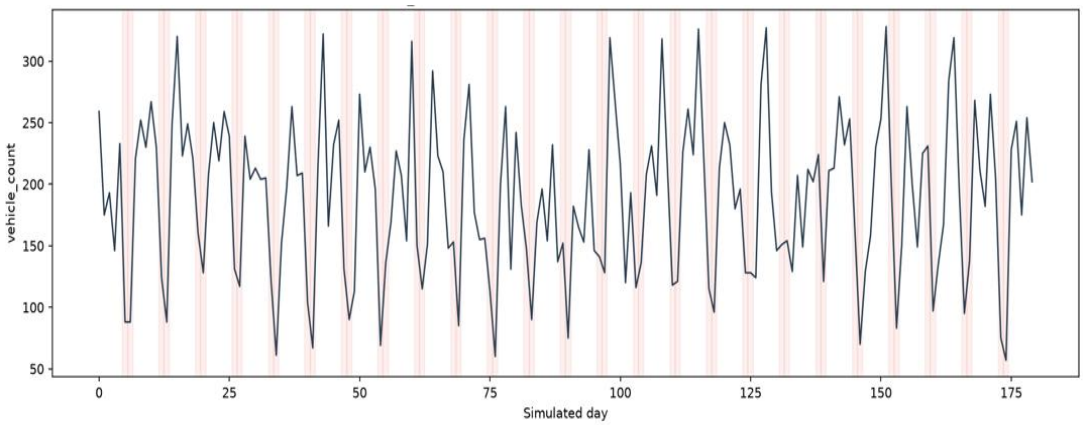


The weekly structure hypothesis was confirmed statistically rather than only visually: the autocorrelation function of daily vehicle_count peaked at lag 7 (ACF = 0.529, exceeding the 95% confidence band of ±0.146), and a Mann-Whitney U test found the weekday/weekend difference highly significant for both vehicle_count ($p = 1.25\times10^{-19}$) and avg_speed_kmh ($p = 3.12\times10^{-25}$).

Figure 6 illustrates the simulated vehicle-count series over time.

Figure 6

Vehicle_count over full simulation (weekends shaded)



The daily vehicle-count autocorrelation is presented in Figure 7.

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Figure 7

ACF of daily vehicle_count

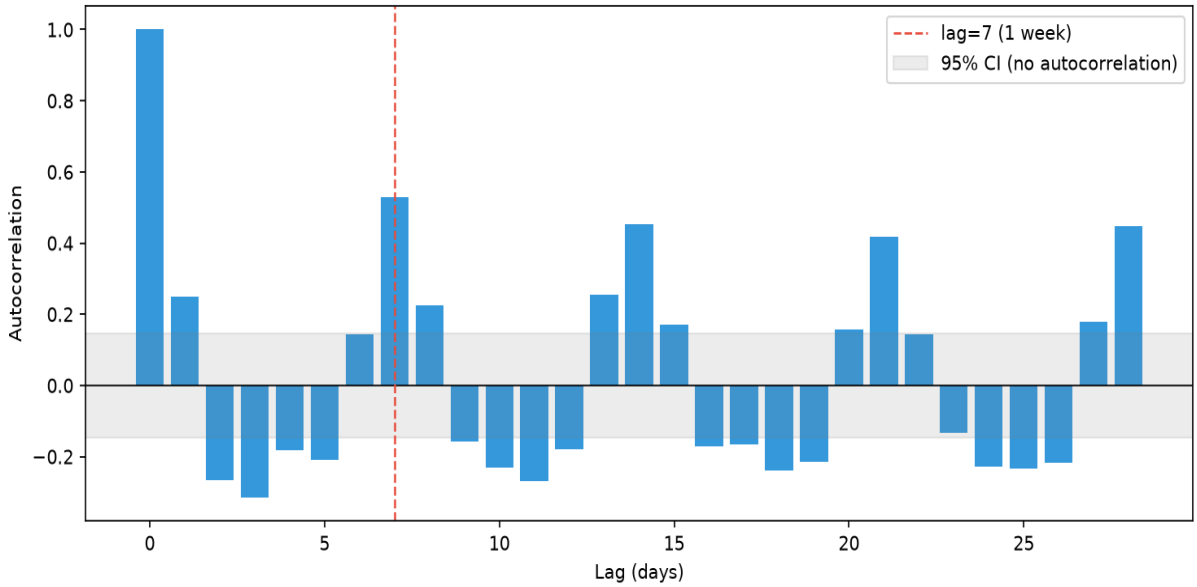


Figure 8 compares weekday and weekend distributions for the selected traffic variables.

Figure 8

Weekday vs Weekend - statistical significance, not just visual

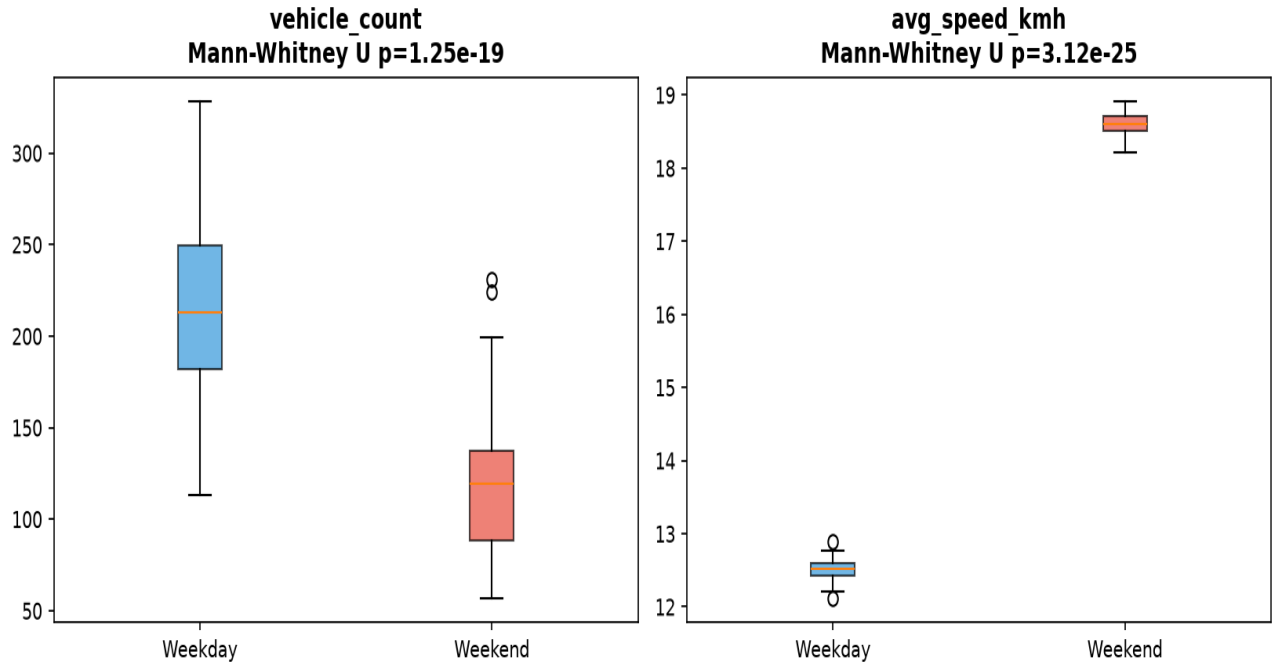
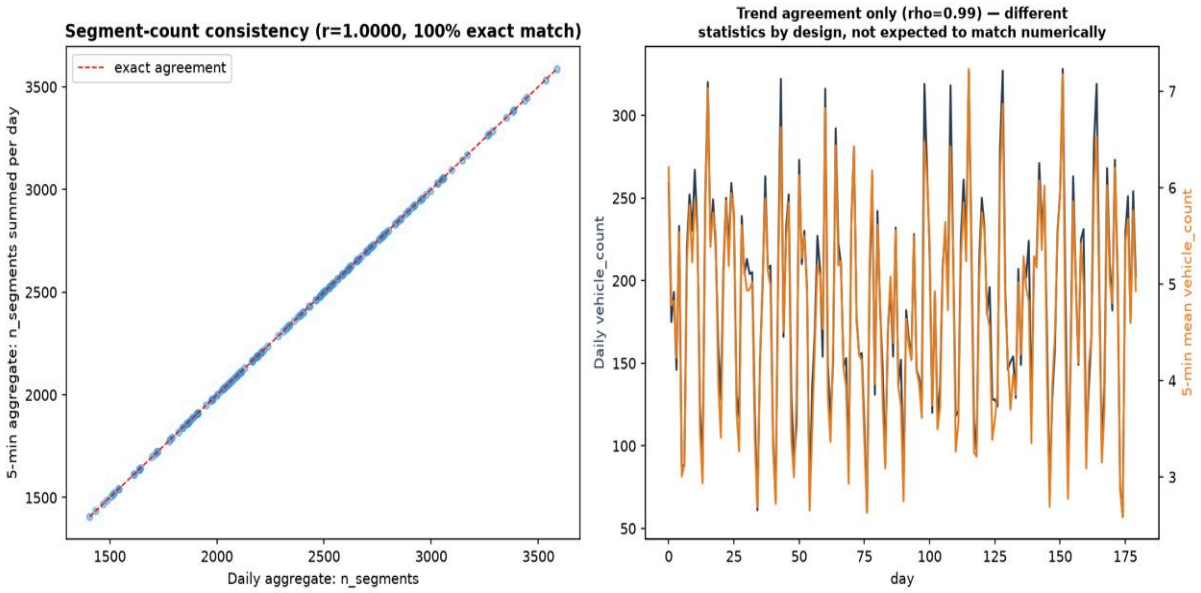


Figure 9 illustrates the agreement between aggregation levels and the simulated temporal patterns.

AITBSS
E-ISSN: 3041-9433

Figure 9

Trend and segment-count

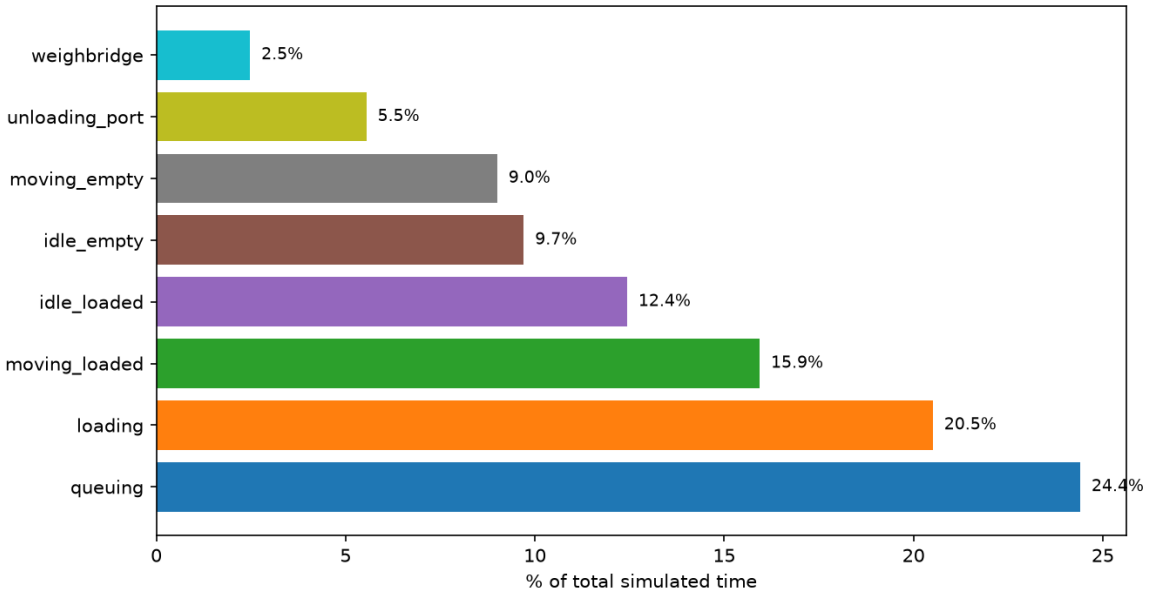


System-wide resource utilization showed queuing accounting for 24.4% of total simulated time down from over 65% in an earlier, uncalibrated version of the simulator with loading (20.5%) as the next-largest state, indicating a functioning rather than pathologically congested port. Finally, the daily and 5-minute aggregates, despite being built independently from the same underlying segment log, agreed exactly on segment counts for 100% of days (r = 1.0000), confirming the aggregation pipeline itself introduces no inconsistency.

The allocation of simulated operational time among resource states is illustrated in Figure 10.

Figure 10

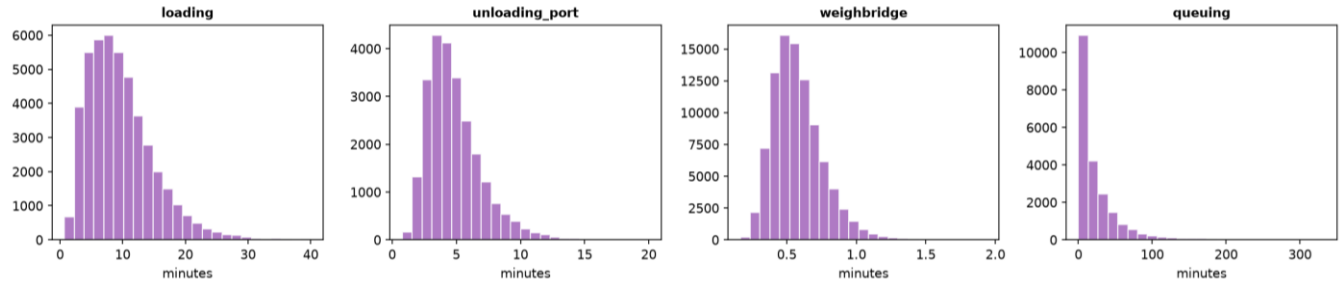
System-wide time in state (resource utilization)



AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Figure 11

Segment duration distributions by state



4.2. Short-Horizon Forecasting Results

Table 1 and Table 2 report short-horizon results (15- and 30-minute forecasts) for vehicle_count and avg_speed_kmh, respectively, trained on the corrected 51,840-bin dataset. Unlike an earlier run on the original 569-bin subset in which ARIMA outperformed every deep learning architecture all six deep models now outperform the ARIMA baseline once adequate training data is available, for both targets and both horizons.

Table 1

Short-horizon vehicle_count forecasting results (post data-volume correction).

Model	MAE (15-min)	R^2 (15-min)	MAE (30-min)	R^2 (30-min)
ARIMA	1.8550	0.3632	1.9646	0.2901
LSTM	1.6559	0.4810	1.7816	0.3992
GRU	1.6538	0.4826	1.7782	0.4007
Transformer	1.6429	0.4876	1.7615	0.4105
PatchTST	1.6340	0.4921	1.7578	0.4129
iTransformer	1.6420	0.4892	1.7613	0.4109
TFT	1.6446	0.4874	1.7671	0.4084

Table 2

Short-horizon avg_speed_kmh forecasting results (post data-volume correction).

Model	MAE (15-min)	R^2 (15-min)	MAE (30-min)	R^2 (30-min)
ARIMA	3.3797	0.0514	3.3254	0.0803
LSTM	3.2034	0.1486	3.2093	0.1417
GRU	3.1984	0.1484	3.2078	0.1422
Transformer	3.1262	0.1826	3.1399	0.1714
PatchTST	3.1324	0.1816	3.1464	0.1694
iTransformer	3.1400	0.1766	3.1668	0.1608
TFT	3.1345	0.1770	3.1500	0.1667

For vehicle_count at the 15-minute horizon, PatchTST achieved the best performance (R^2 = 0.4921) against ARIMA's R^2 = 0.3632 a relative improvement of roughly 35%. For avg_speed_kmh, the plain Transformer architecture performed marginally best (R^2 = 0.1826 at 15 minutes), though all deep models clustered closely together and well ahead of ARIMA (R^2 = 0.0514). Absolute R^2 values for avg_speed_kmh at the short horizon remain modest across every model, a point returned to in Section 5.2.

4.3. Mid- and Long-Horizon Forecasting Results

Table 3 reports R^2 for all seven models across the Mid (7-day-ahead, 28-day lookback) and Long (30-day-ahead, 90-day lookback) tiers, both derived from the same 180-day daily aggregate.

AITBSS
E-ISSN: 3041-9433

Table 3

Mid- and Long-horizon forecasting results (R^2), both targets.

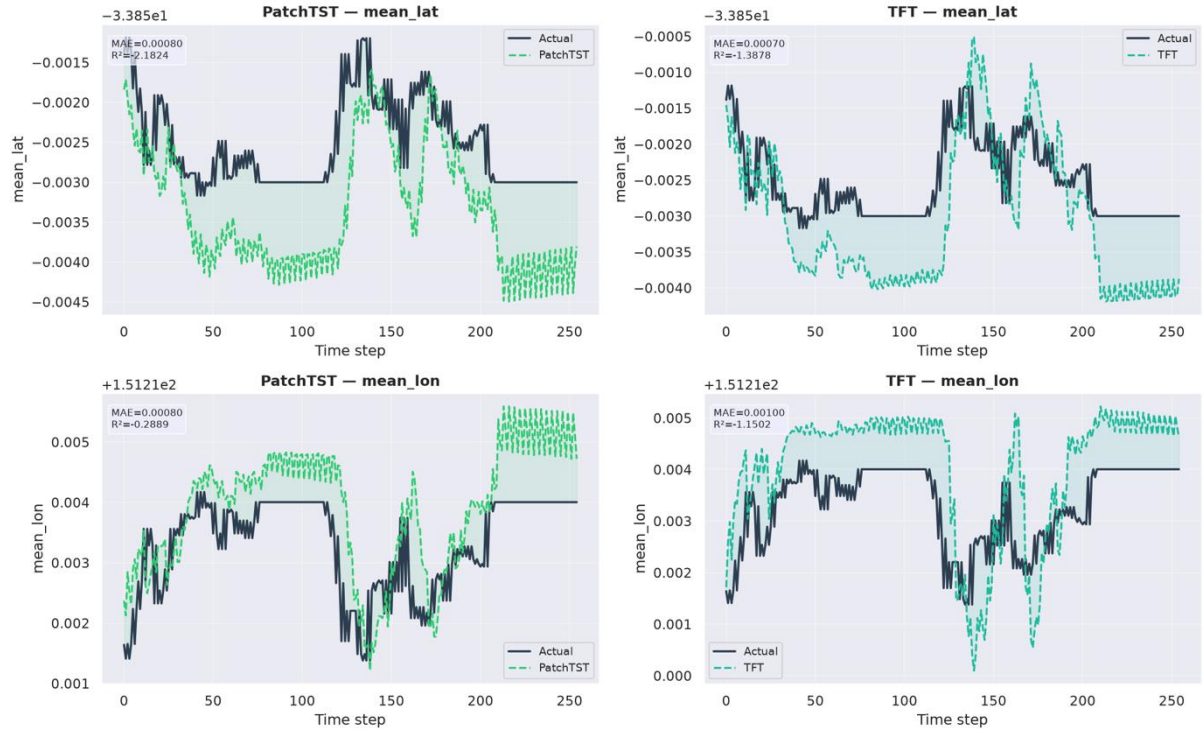
Model	Mid count	Long count	Mid speed	Long speed
ARIMA	-0.0138	0.0180	0.3984	0.1029
LSTM	0.0118	0.0001	0.6073	0.0015
GRU	0.2298	0.0244	0.6932	0.1765
Transformer	0.3694	0.1803	0.9511	0.7396
PatchTST	0.3843	0.2336	0.9821	0.9741
iTransformer	0.3342	0.2077	0.9681	0.9332
TFT	0.4058	0.1976	0.9183	0.7537

The advantage of transformer-family architectures over ARIMA and the recurrent baselines is far more pronounced at these longer horizons than at the short horizon, particularly for avg_speed_kmh: PatchTST reaches R^2 = 0.9821 at the Mid tier and R^2 = 0.9741 at the Long tier, compared to ARIMA's 0.3984 and 0.1029, respectively. One result warrants explicit attention rather than being absorbed into the aggregate picture: LSTM's Long-horizon avg_speed_kmh performance is effectively null (R^2 = 0.0015), a near-total failure discussed further in Section 5.6.

Figure 12 illustrates the spatial forecasting output included for visualization; spatial prediction is not assessed as a primary forecasting target.

Figure 12

Spatial Prediction – Mean Forecasting



4.4. *Cross-Tier Comparison*

Figure 13 and Tables 1–3 allow direct comparison across all three horizon tiers. Mean R^2 across all model/target pairs with complete data at every tier was 0.290 (Short), 0.517 (Mid), and 0.325 (Long) a non-monotonic pattern in which the Mid tier outperforms both the data-rich Short tier and the calendar-feature-equipped Long tier. A Spearman rank-correlation test between horizon-tier order and R^2 found no statistically significant

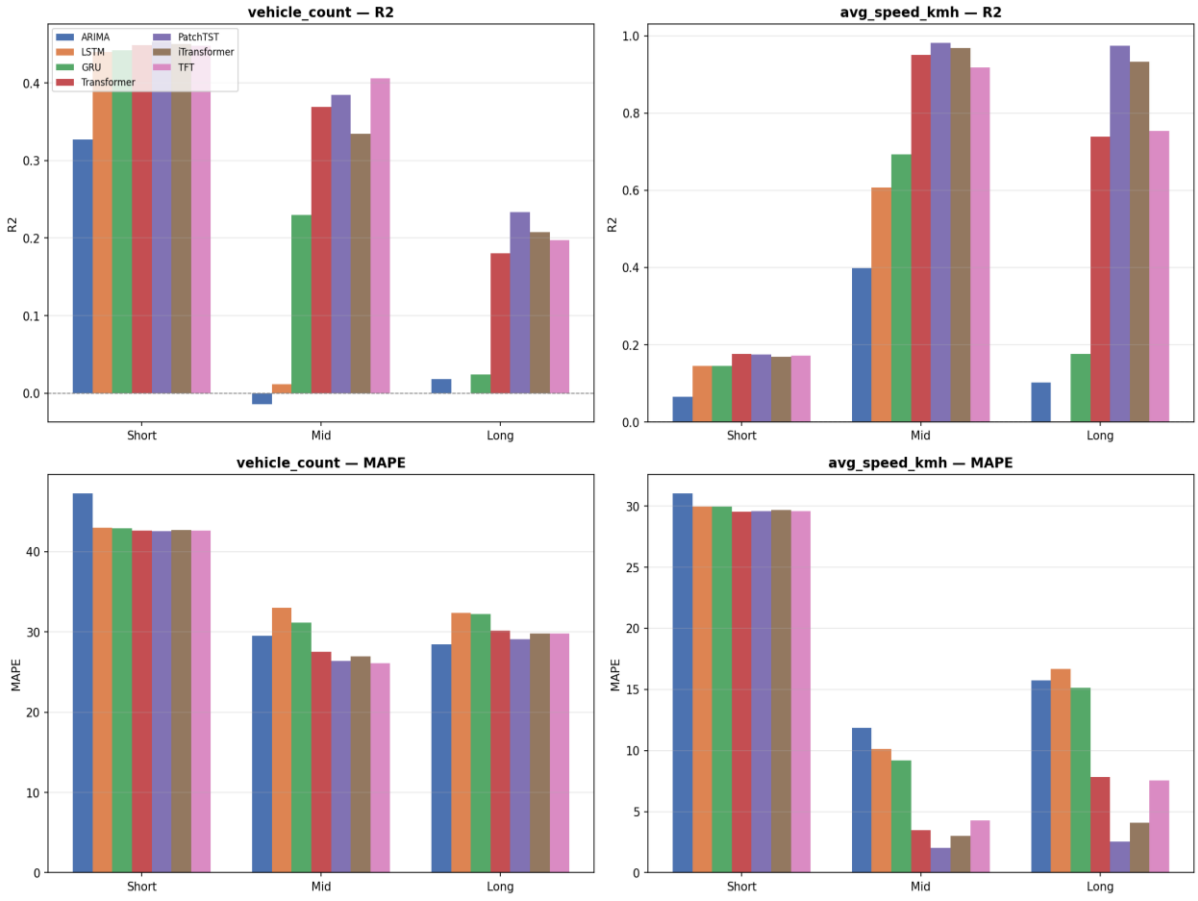
AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

monotonic trend (ρ = -0.038, p = 0.809, n = 42 points across 14 model/target pairs). Given the small number of model/target pairs with complete three-tier coverage, this pattern should be treated as descriptive rather than confirmatory; we discuss a plausible explanation in Section 5.2.

The model performances across forecast horizons and targets are compared graphically in Figure 13.

Figure 13

Cross-tier comparison of R^2 and sMAPE across all seven models and both prediction targets



4.5. *Feature Ablation Study*

A feature ablation study using PatchTST (vehicle_count target) was conducted at all three horizon tiers to quantify the marginal contribution of each engineered feature group.

Table 4

Short-horizon PatchTST feature ablation.

Configuration	MAE (Short)	R^2 (Short)
Full (7 features)	1.6303	0.4939
No cargo level	1.6495	0.4831
No time encoding	1.6365	0.4925
No queue ratio	1.6411	0.4888
Speed only	2.0670	0.1979
Count + speed	1.6635	0.4752
Univariate (count only)	1.6853	0.4695

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Table 5

Mid- and Long-horizon PatchTST feature ablation.

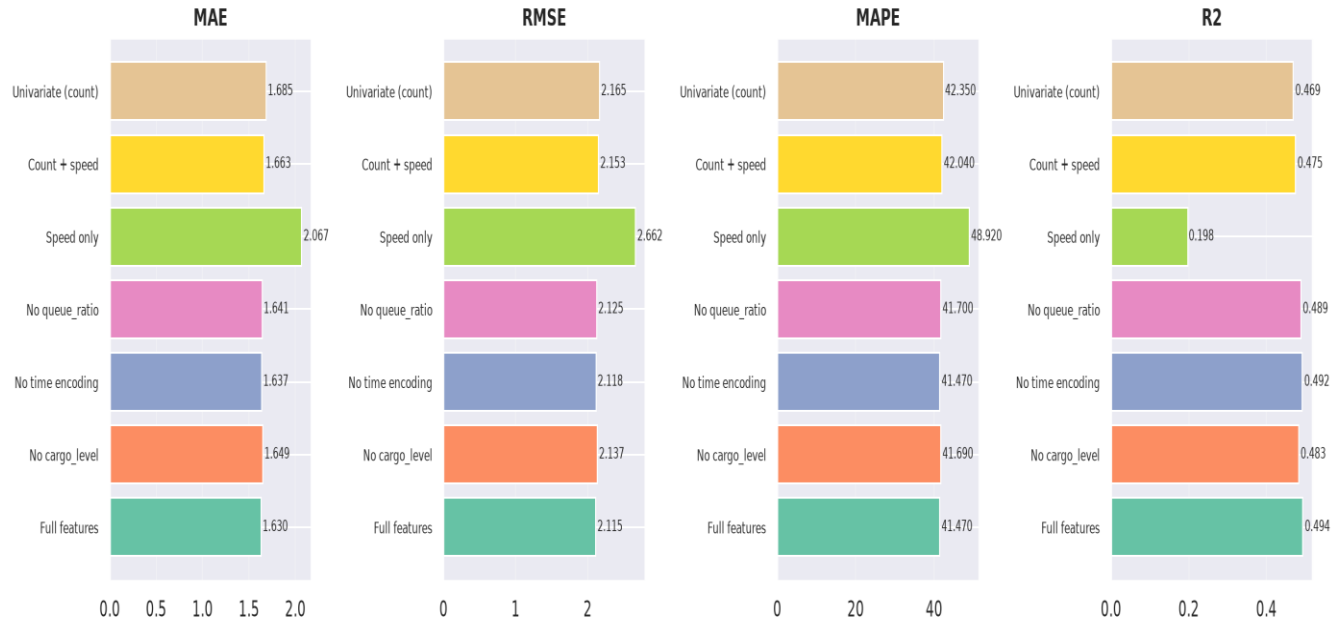
Configuration	MAE (Mid)	R^2 (Mid)	MAE (Long)	R^2 (Long)
Full (7 features)	45.49	0.4134	52.74	0.2260
No cargo level	47.63	0.3388	53.89	0.1568
No day-of-week	47.06	0.3775	52.38	0.2112
No queue ratio	45.43	0.4063	51.69	0.2475
Speed only	47.02	0.3419	52.15	0.2301
Count + speed	47.12	0.3523	52.83	0.2254
Univariate (count only)	49.78	0.2640	55.56	0.1320

At the Short and Mid tiers, the full feature set achieves the best or near-best R^2, consistent with conventional feature-engineering expectations. At the Long tier, however, two reduced configurations "no queue ratio" (R^2 = 0.2475) outperforms the full configuration (R^2 = 0.2260). We interpret this as an overfitting effect specific to the most severe data scarcity condition in the study (only 61 raw windows before splitting), rather than evidence that these features are uninformative in general: the univariate (count-only) configuration is the weakest performer at every single tier without exception, indicating that auxiliary features carry real value, and that the Long-tier exception is a sample-size effect rather than a feature-relevance one.

The short-horizon PatchTST feature-ablation results are visualized in Figure 14.

Figure 14

Ablation PatchTST vehicle_count @ 15min



4.6. *External Validation Against Real-World Data (DT-CARGO)*

All results reported above are derived from the simulated dataset. To assess whether the central finding of this study that deep learning architectures outperform ARIMA once given adequate training data generalizes beyond simulation, the same seven architectures and training code were additionally evaluated on independent real-world data from DT-CARGO (Balke & Adenaw, 2023), a public dataset of high-resolution GPS telemetry from 54 heavy commercial trucks. Rather than the cross-dataset fusion originally envisioned (Section 5.4), this is a direct, independent benchmark: the same seven models, trained and evaluated on DT-CARGO's own operational

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

speed data, using the vehicle with the most recorded trips after data-quality filtering (vehicle 17, 3,169 trips, filtered using DT-CARGO's own documented GPS-signal-loss quality field rather than an ad hoc threshold). The forecasting task here is necessarily coarser-grained than the port simulation's 5-minute bins predicting a truck's next few trips' average speed from its preceding trips and results are based on a single training run per model rather than the five-seed averaging used elsewhere in this paper, since this is a supplementary validation rather than a primary result.

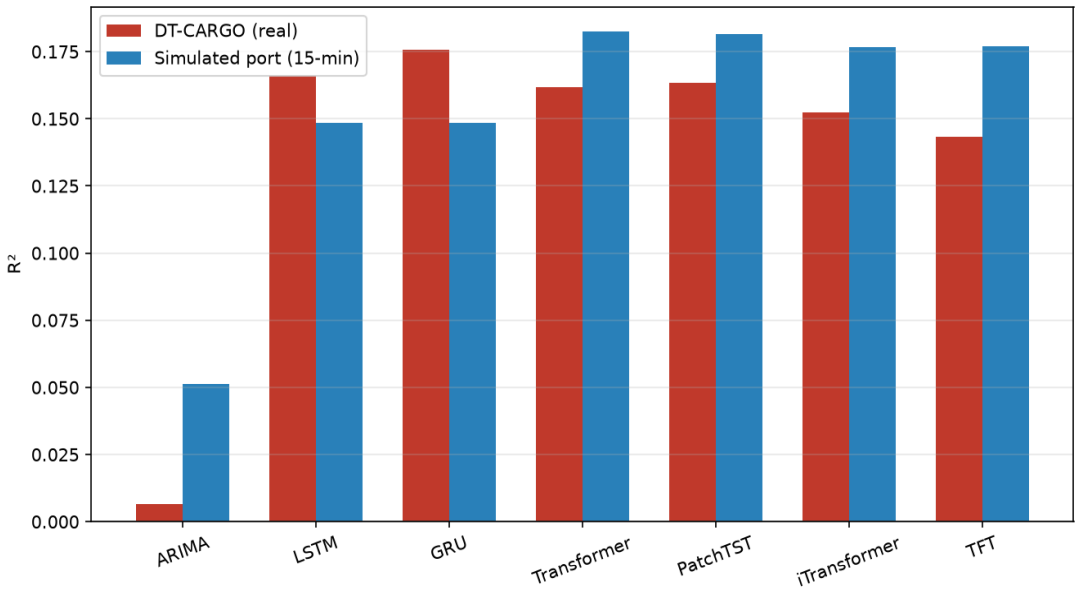
Table 6

Real-world (DT-CARGO) vs. simulated-port avg_speed forecasting, all seven models.

Model	DT-CARGO R^2	Simulated Port R^2 (15-min)
ARIMA	0.0066	0.0514
LSTM	0.1659	0.1486
GRU	0.1756	0.1484
Transformer	0.1618	0.1826
PatchTST	0.1635	0.1816
iTransformer	0.1523	0.1766
TFT	0.1434	0.1770

Figure 15

R^2 comparison: DT-CARGO (real-world) vs. simulated port, avg_speed target.



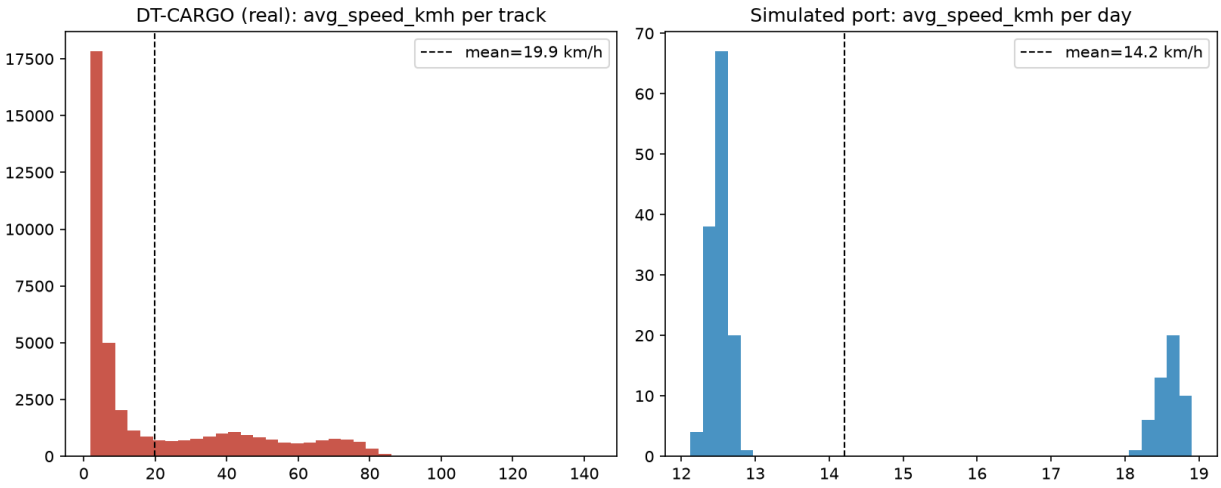
Two findings emerge. First, the core claim replicates and more starkly than in simulation: ARIMA collapses to near-zero predictive power on real data (R^2 = 0.0066), while every deep architecture achieves R^2 between 0.14 and 0.18. The gap separating the worst-performing deep model (TFT, R^2 = 0.1434) from ARIMA (0.1368 R^2 points) is more than four times larger than the spread separating the best and worst deep models from each other (0.032 R^2). Second, and reported honestly rather than smoothed over: unlike the simulated results, where transformer variants generally led, the two recurrent architectures achieved the highest R^2 on real data (GRU = 0.1756, LSTM = 0.1659), narrowly ahead of every transformer variant (best: PatchTST, R^2 = 0.1635). Given the single-seed evaluation and the narrow 0.032 spread among all six deep models, we treat this specific ordering as suggestive rather than conclusive, and discuss it further in Section 5.5.

Figure 16 illustrates the differences between the real-trip and simulated-day speed distributions.

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

Figure 16

Distribution of avg_speed_kmh: DT-CARGO (per real trip) vs. simulated port (per simulated day)

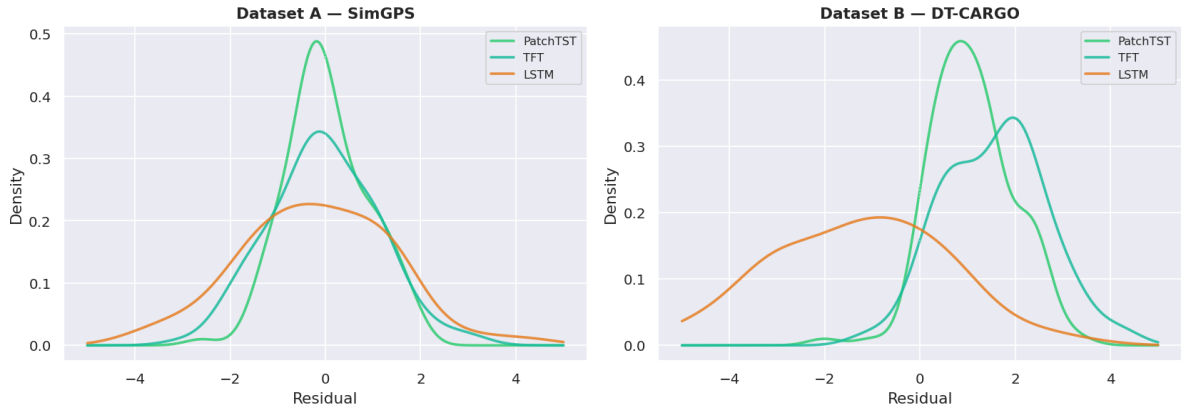


The underlying speed distributions (Figure 16) differ substantially in shape. This is not a like-for-like statistical comparison one is per-trip, the other per-day but the shape mismatch itself is informative and is discussed as a simulation-realism limitation in Section 5.5.

Residual-density comparisons for the evaluated datasets are presented in Figure 17.

Figure 17

Error Distribution - Density vs Probability



5. Discussion

The single most consequential methodological finding of this study is not which architecture performs best, but that the answer to that question depends entirely on training data volume. This is consistent with the well-documented data-hungry nature of attention-based architectures relative to classical statistical models (Vaswani, 2017), but it carries a sharper methodological implication for comparative forecasting studies specifically: an architecture comparison conducted on an inadequately sized dataset does not measure architectural merit it measures data sufficiency, and can produce directionally misleading conclusions if reported without that caveat.

AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

5.1. *The Non-Monotonic Relationship Between Forecast Horizon and Accuracy*

Mean R^2 across model/target pairs peaked at the Mid tier (0.517) rather than at either the Short tier, which has the largest training set (0.290), or the Long tier, which has explicit weekly calendar features (0.325). We note, however, that a Spearman correlation test found no statistically significant monotonic relationship between horizon-tier order and R^2 ($\rho = -0.038$, $p = 0.809$); with only 14 model/target pairs carrying complete three-tier coverage, this test is likely underpowered, and the Mid-tier optimum should be treated as a descriptive pattern warranting confirmation with a longer simulation horizon (beyond 180 days) rather than an established result.

5.2. *Feature Selection versus Feature Abundance*

The ablation results (Section 4.5) refine, rather than confirm, the intuitive expectation that more engineered features improve forecasting accuracy. Critically, the univariate configuration was the single weakest performer at every tier without exception, meaning the practical implication is not that fewer features are categorically better, but that feature selection should be matched to available sample size rather than defaulting to maximal feature inclusion a consideration likely to generalize to other low-data forecasting settings beyond bulk port operations.

5.3. *On FusionNet, and the Revised Role of DT-CARGO*

The original research design for this study included fusing real-world heavy-vehicle telemetry (DT-CARGO; (Balke & Adenaw, 2023)) with the simulated port data through a dedicated cross-attention architecture (FusionNet). Using that field rather than an undocumented threshold on the corrupted variable directly DT-CARGO could be cleaned properly and repurposed for a simpler, arguably more valuable role: an independent real-world benchmark of the same seven architectures, reported in Section 4.6. We regard the original outlier discovery as a data-quality safeguard that correctly blocked an unreliable fusion result, not as grounds to abandon a real-world dataset that, once properly cleaned, still had legitimate scientific use.

5.4. *External Validity: Testing the Core Finding Against Real-World Data*

The DT-CARGO benchmark (Section 4.6) provides an opportunity to test this paper's central claim deep learning architectures outperform ARIMA once given adequate data against data this study did not generate or control. A plausible explanation is that real operational trip sequences lack the smooth, partially engineered periodicity of the simulated series (Section 3.2), leaving ARIMA's linear autoregressive structure with very little to latch onto, while nonlinear deep architectures recurrent or attention-based can still extract usable structure from noisier, more irregular real data.

The second finding requires more caution. We do not treat this as evidence that transformers are unsuited to real-world traffic forecasting; the DT-CARGO benchmark used a single training seed rather than the five-seed averaging applied everywhere else in this paper, and the entire spread among all six deep architectures (0.032 R^2) is narrow enough that a different random seed could plausibly reorder it. We consider this the more defensible framing of this paper's contribution, and recommend future work treat the transformer-versus-recurrent ordering specifically as an open question rather than a settled one.

The distributional mismatch shown in Figure 16 also identifies a concrete simulation-realism gap worth naming directly. This is a direct consequence of the simulation's two-state (weekday/weekend) congestion model (Section 3.2), which is sufficient for the daily/weekly aggregate forecasting task this paper addresses, but would need substantial revision modeling idling, loading, and urban stop-and-go conditions explicitly before the simulation could be used for finer-grained, trip-level speed modeling.

5.5. *Limitations*

Several limitations should be considered when interpreting these results. The Mid- and, especially, Long-horizon tiers rest on small raw sample sizes (146 and 61 windows respectively) before train/validation/test splitting, which limits the statistical power of both the cross-tier comparison (Section 4.4) and the feature ablation study (Section 4.5); results at these tiers, including the ablation configuration exceeding the full-feature baseline at the Long tier, should be treated as suggestive rather than conclusive pending a longer simulation run. Finally, the simulation models a single generic bulk port configuration with a fixed weekly structure; generalization to ports of

AITBSS AI and Tech in Behavioral and Social Sciences E-ISSN: 3041-9433

different scale, cargo composition, or seasonal pattern has not been evaluated.

6. Conclusion

This paper set out to compare transformer-based architectures against classical statistical and recurrent baselines for short-, mid-, and long-horizon traffic forecasting in a bulk port environment, using a validated discrete-event simulation as the data source. Three findings stand out.

First, deep learning superiority over ARIMA is real, conditional on adequate training data, and now supported by evidence beyond simulation alone. The robust, generalizable finding of this paper is the coarser one: deep learning as a class outperforms classical linear forecasting given adequate data, both in simulation and on independent real-world data; the specific ordering among deep architectures appears more setting-dependent and should be treated as such in future comparative work.

Second, forecasting accuracy did not degrade monotonically with horizon length. Mean R^2 peaked at the intermediate (weekly) horizon rather than at the data-richest short horizon or the calendar-feature-equipped long horizon, a pattern we attribute to a trade-off between training sample size and the availability of explicit weekly structure in the feature set though we note this pattern did not reach statistical significance under a Spearman rank-correlation test and should be confirmed with a longer simulation run in future work.

Third, the value of engineered features was found to depend on available sample size rather than being uniformly positive: the full feature set was optimal at the Short and Mid tiers but was outperformed by reduced configurations at the data-scarce Long tier, while a univariate baseline was the weakest configuration at every tier without exception. This suggests that feature selection, not merely feature engineering, deserves explicit attention when applying these architectures to small forecasting datasets.

A planned architectural contribution of this study FusionNet, intended to fuse simulated data with real-world heavy-vehicle telemetry could not be evaluated after the DT-CARGO speed variable was found to contain physically implausible outlier values during preprocessing. We identify constructing a validated real-world bulk-port operational dataset as distinct from the general heavy-vehicle telemetry used here as the most direct next step for extending this work, alongside lengthening the underlying simulation to increase statistical power at the Mid and Long horizon tiers, investigating the anomalous Long-horizon LSTM result identified in Section 5.6, and testing whether the transformer-versus-recurrent ordering observed on DT-CARGO holds across a larger, multi-vehicle real-world sample.

Finally, we note explicitly that this paper is scoped to aggregate operational traffic forecasting vehicle counts and average speeds rather than spatial trajectory forecasting, even though the underlying simulation generates GPS trajectory data for illustrative purposes (Section 3.2). Forecasting spatial position or route-level traffic patterns directly is a distinct modeling problem, requiring different feature representations, model outputs, and evaluation criteria than the scalar time series addressed here. We intentionally leave this direction to a dedicated follow-up study rather than treating it as a minor extension of the present work.

Authors' Contributions

Sanaz Jafarinia: Conceptualization, methodology, software, data curation, formal analysis, visualization, writing—original draft. Mohammad Ali Afshar Kazemi: Conceptualization, supervision, methodology, writing—review and editing. Seyed Abdollah Amin Mousavi: Methodology, validation, writing—review and editing. Hassan Mehrmanesh: Investigation, validation, visualization, writing—review and editing. MohammadAli Keramati: Formal analysis, resources, writing—review and editing.

Declaration

In order to correct and improve the academic writing of our paper, we have used the language model ChatGPT.

Transparency Statement

The DT-CARGO dataset is available through Zenodo (https://doi.org/10.5281/zenodo.7599687). The computational workflow is described in the Methods section.

Acknowledgments

We would like to express our gratitude to all individuals helped us to do the project.

AITBSS AI and Tech in Behavioral and Social Sciences E-ISSN: 3041-9433

Declaration of Interest

The authors report no conflict of interest.

Funding

According to the authors, this article has no financial support.

Ethics Considerations

This study used computational simulation and secondary analysis of the publicly available DT-CARGO vehicle telemetry dataset. No human participants were recruited, and no animal experiments were conducted as part of the reported study.

References

Balke, G., & Adenaw, L. (2023). Heavy commercial vehicles' mobility: Dataset of trucks' anonymized recorded driving and operation (DT-CARGO). *Data in Brief*, *48*, 109246. https://doi.org/10.1016/j.dib.2023.109246

Chan, K. C. (2022). Generation of synthetic manufacturing datasets for machine learning using discrete-event simulation. *Production & Manufacturing Research*, *10*, 337-353. https://doi.org/10.1080/21693277.2022.2086642

Dos Santos, R. P., Matos-Carvalho, J. P., & Leithardt, V. R. Q. (2025). Deep learning in time series forecasting with transformer models and RNNs. *PeerJ Computer Science*, *11*, e3001. https://doi.org/10.7717/peerj-cs.3001

Kambale, W. V. (2023). *Transformers in Time Series Forecasting: A Brief Transfer Learning Performance Analysis* 2023 27th International Conference on Circuits, Systems, Communications and Computers (CSCC),

Lamdjad, B. (2025). Comparative forecasting of container throughput for maritime logistics using statistical, machine learning and explainable AI models. *Maritime Business Review*, *11*(3), 202-218. https://doi.org/10.1108/MABR-08-2025-0086

Li, H., Zhao, J., Jia, P., Ou, H., & Zhao, W. (2025). Optimization of Bulk Cargo Terminal Unloading and Outbound Operations Based on a Deep Reinforcement Learning Framework. *Journal of Marine Science and Engineering*, *13*(1), 105. https://doi.org/10.3390/jmse13010105

Lim, B., Arık, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for Interpretable Multi-Horizon Time Series Forecasting. *International Journal of Forecasting*, *37*(4), 1748-1764. https://doi.org/10.1016/j.ijforecast.2021.03.012

Liu, W., Xu, B., & Li, J. (2025). Data-Driven Carbon Emission Dynamics Under Ship In-Port Congestion. *Journal of Marine Science and Engineering*, *13*(4), 812. https://doi.org/10.3390/jmse13040812

Miño Calero, A., Rasheed, A., & Lekkas, A. M. (2026). Unveiling the limitations of transformer models in time series forecasting. *Progress in Artificial Intelligence*, *15*, 1179-1200. https://doi.org/10.1007/s13748-026-00450-y

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2023). *A time series is worth 64 words: Long-term forecasting with transformers* International Conference on Learning Representations (ICLR), https://arxiv.org/abs/2211.14730

Okamoto, Y. (2026). Transformers for time-series forecasting: A comprehensive survey through 2024. *Nonlinear Theory and Its Applications, IEICE*, *17*(1), 260-278. https://doi.org/10.1587/nolta.17.260

Shah, A. B. A., Xu, X., Yongren, Z., & Guo, Z. (2025). A Custom Deep Learning Approach for Traffic Flow Prediction in Port Environments: Integrating RCNN for Spatial and Temporal Analysis. *International journal of advanced computer science and applications*, *16*(2). https://doi.org/10.14569/IJACSA.2025.0160266

Vaswani, A. (2017). *Attention is all you need* Advances in Neural Information Processing Systems,

Wang, M., Guo, X., She, Y., Zhou, Y., Liang, M., & Chen, Z. S. (2024). Advancements in Deep Learning Techniques for Time Series Forecasting in Maritime Applications: A Comprehensive Review. *Information*, *15*(8), 507. https://doi.org/10.3390/info15080507

Zhou, H. (2020). Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *arXiv*, *abs/2012.07436*. https://doi.org/10.1609/aaai.v35i12.17325

Zhu, Y. (2023). DRCNN: decomposing residual convolutional neural networks for time series forecasting. *Scientific Reports*, *13*. https://doi.org/10.1038/s41598-023-42815-6

AITBSS AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433