Journal of Mining and Environment (JME) Vol. 17, No. 5, 2026, 1649-1674

Shahrood University of Technology

Journal of Mining and Environment (JME)

Journal homepage: www.jme.shahroodut.ac.ir

Iranian Society of Mining Engineering (IRSME)

Mitigating the Productivity–Efficiency Trade-off in Heavy-Duty Mining Trucks Using Deep Reinforcement Learning and Digital Twin Surrogate Modeling

Kusnawi Kusnawi[1,2]*, Mochamad Agung Wibowo[3], and Ridwan Sanjaya[4]

1. Doctoral Program of Information Systems, Universitas Diponegoro, Semarang, Indonesia
2. Informatics, Universitas Amikom Yogyakarta, Yogyakarta, Indonesia
3. Postgraduate School, Universitas Diponegoro, Semarang, Indonesia
4. Department of Information System, Universitas Katolik Soegijapranata, Semarang, Indonesia

Article Info

Received 14 July 2026

Received in Revised form 11 August 2025

Accepted 5 September 2026

Published online 5 September 2026

DOI: 10.22044/jme.2026.17993.3582

Keywords

Digital Twin Surrogate

Deep Reinforcement Learning

Fuel Intensity

Haul-Cycle Productivity

Mining Haul Truck

Abstract

Open-pit haulage must reduce fuel use without sacrificing material movement or operational safety. This study examined whether a telemetry-driven, task-specific Digital Twin surrogate could support vehicle-level eco-driving policies that improve fuel intensity while preserving haul-cycle productivity. The framework combined full haul-cycle reconstruction, an ensemble XGBoost fuel-rate surrogate, nonlinear state-transition models, equal-budget multi-seed Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC), reward-sensitivity screening, a rule-based benchmark, and uncertainty- and out-of-distribution-aware evaluation. One-day telemetry from 71 Komatsu HD785-7 trucks contained 1,927,867 records and yielded 2,386 validated haul cycles. The fuel surrogate achieved a temporal-test MAE of 27.33 L/h, RMSE of 37.75 L/h, and R^2 of 0.6608. Across 26 complete, independent test cycles, common-estimator rescoring showed that PPO reduced fuel intensity by 8.79% relative to the surrogate-human baseline (0.2854 versus 0.3163 L/t; p=0.0029), with a 4.91% mean throughput decrease that was not statistically significant. SAC reduced fuel intensity by 9.53% (0.2820 L/t; p=0.0032) and increased mean throughput by 3.16%, although the throughput change was not significant, and loaded-route safety-envelope exceedance increased significantly. PPO and SAC did not differ significantly in fuel intensity. Policy-generated states also showed substantial out-of-distribution exposure. The framework, therefore, identified efficiency–productivity–safety trade-offs rather than universal algorithm superiority and should be interpreted as an offline policy-screening tool pending operator-in-the-loop and controlled field validation.

1. Introduction

1.1. Broad Background

Open-pit haulage is among the most energy-intensive and cost-sensitive components of surface mining. Heavy-duty mining trucks operate under changing payload, haul-road grade, rolling resistance, route geometry, engine condition, and operator-control inputs, causing fuel demand to emerge from interacting operational factors rather than a single isolated variable [1-6]. Energy and exergy assessments further indicate that avoidable operating losses present significant opportunities for fuel and emission reduction [2], while rolling resistance and route conditions directly affect tractive demand and reliability [3, 4].

Mine production sets the boundary for any fuel-saving claim. Slower travel may reduce instantaneous demand but lengthens the cycle and decreases hourly payload delivery. Conversely, faster travel can shorten the controlled segment, but this benefit may be offset by aggressive acceleration, high engine speeds, braking, or poor momentum management. Eco-driving must therefore coordinate speed with control behavior

Corresponding author: khusnawi@amikom.ac.id (Kusnawi)

rather than pursue speed reduction as an isolated objective [7-9].

The control problem unfolds over time because each decision alters the truck's state, the remaining distance, the fuel required at the next step, and the feasible actions. Reinforcement learning (RL) is suitable for this sequential structure, although its application outside simulation is limited by unsafe exploration, misspecified rewards, model errors, and distribution shifts [10-14]. Exploration on production trucks was avoided because unverified actions could impact output, equipment, or operator safety.

A bounded Digital Twin surrogate provides an intermediate environment for offline policy screening. In this study, the term Digital Twin does not imply a complete physics-based replica of the truck, haul road, drivetrain, traffic system, and physical asset connectivity. Instead, it refers to a task-specific, telemetry-driven surrogate that estimates selected fuel and short-horizon state responses for counterfactual controller evaluation. This scope aligns with the range of Digital Twin representations described in mining and industrial literature and with learned-environment integration into reinforcement learning [15-19].

The study thus evaluates whether policies can reduce full-cycle fuel intensity while maintaining payload throughput under matched historical route contexts. Productivity is primarily measured in t/h, efficiency in L/t and L/(t·km), and speed or L/km only as secondary travel diagnostics.

The proposed method begins by reconstructing complete loading-to-loading haul cycles from raw truck telemetry and separating the cycle-linked loaded and empty travel segments used for longitudinal control modeling. Three bootstrap XGBoost models estimate fuel rate, while nonlinear transition models predict the next vehicle speed and engine speed. PPO and SAC policies are then trained under matched historical route contexts using equal training budgets and multiple random seeds. Their route-level trajectories are reintegrated with the unchanged loading, dumping, queueing, waiting, and other non-driving components of each cycle. The resulting policies are evaluated using complete-cycle productivity and fuel-intensity metrics, statistical comparisons, empirical safety-envelope indicators, predictive uncertainty, and out-of-distribution diagnostics.

1.2. Related Work

Mine-haulage fuel research has increasingly employed nonlinear machine-learning models. XGBoost and related tree ensemble methods can capture interactions among payload, road conditions, speed, engine state, and operator inputs [1, 5, 6, 20]. These studies provide a predictive foundation but generally address the amount of fuel a truck is expected to consume rather than how to select sequential control actions throughout loaded and empty travel.

Digital Twin research in mining has focused on monitoring, equipment representation, simulation, and operational decision support [15-19]. The literature also indicates that Digital Twins vary in physical fidelity, connectivity, and synchronization. A learned surrogate can therefore support a specific analytical task without claiming full bidirectional replication, provided that its scope and uncertainty are explicitly stated.

RL-based eco-driving has been applied to automated and connected road vehicles, including energy-aware control under traffic and intersection uncertainty [8, 21-24]. Mining-oriented RL has more frequently addressed fleet dispatching, truck–shovel allocation, system-level emissions, or hybrid power management [25-29]. Vehicle-level longitudinal eco-driving for heavy haul trucks remains less developed, especially when complete-cycle payload throughput, safety behavior, and surrogate support must be considered jointly.

The two algorithms were selected to demonstrate different learning behaviors within a single experimental design. PPO limits changes between successive policies through clipped on-policy updates [30], whereas SAC learns off-policy from replayed transitions and maintains entropy-driven exploration [31]. Both algorithms were tested under the same environment, routes, reward profile, interaction budget, and seed values; the comparison aimed to reveal configuration-specific behaviors rather than to establish a universally superior algorithm.

Collectively, the literature provides three discrete building blocks: fuel prediction, virtual experimentation, and sequential control. What remains uncertain is whether these components can be integrated into a mining-truck study without interpreting model exploitation, faster travel, or incomplete cycle accounting as evidence of operational improvement.

1.3. Research Gap and Problem Statement

The resulting research question was: under matched historical route conditions, can a telemetry-driven Digital Twin surrogate and RL identify vehicle-level eco-driving policies that

lower full-cycle fuel intensity without a material loss of payload throughput, while remaining acceptable in control behavior, empirical safety-envelope compliance, predictive confidence, and distributional support?

This question was examined offline through counterfactual simulation. The controllers changed only the loaded-travel and empty-return portions linked to each cycle. Observed durations and outcomes for loading, dumping, queueing, waiting, spotting, inactivity, and excluded low-speed records were retained as exogenous components. The comparisons therefore describe route-matched operating alternatives, not savings already demonstrated in production.

The central gap lies in the chain of evidence between raw telemetry and a policy claim, rather than in the absence of another fuel model or another use of PPO. That chain is credible only if the full cycle is retained, measured fuel is separated from modeled fuel, leakage is controlled, production-relevant metrics are used, baselines are transparent, training is repeated across seeds, and departures from historical data support are reported.

1.4. Purpose, Novelty, and Contributions

The purpose was to test an auditable chain from complete haul-cycle records to uncertainty-aware evaluation of continuous control. The contribution lies in that chain—cycle reconstruction, a bounded Digital Twin surrogate, RL policy testing, and evidence limits—rather than in changing the PPO or SAC algorithms.

1. The first contribution is a two-level data structure: full telemetry preserves the loading-to-loading production cycle, whereas a linked travel subset supports longitudinal-control modeling.
2. The second is a bounded surrogate composed of three cycle-bootstrap XGBoost fuel models, nonlinear next-speed and next-engine-speed models, split-conformal calibration, state-dependent error estimates, and multivariate OOD checks.
3. The third is a production-centered assessment built around payload throughput (t/h), fuel intensity (L/t), and transport-normalized fuel intensity [L/(t·km)]; speed and L/km are retained only as route diagnostics.
4. The fourth contribution is an equal-budget comparison across three seeds in which PPO and SAC share the same nine state variables, scalar action, reward profile, route pools, and validation-based checkpoint rule.
5. The fifth contribution is a multi-baseline assessment of measured human operation, surrogate-reconstructed human operation, rule-based eco-driving, PPO, and SAC using statistical results together with safety, uncertainty, and OOD evidence.

2. Materials and Methods

2.1. Overall Research Framework and Scope of the Digital Twin Surrogate

The implementation developed in this study is not a complete physical or operational Digital Twin of the haul truck or the mine-haulage system. Instead, it is a task-specific, telemetry-driven Digital Twin surrogate developed for the offline counterfactual evaluation of longitudinal driving policies. It represents only the relationships required to estimate fuel rate and short-horizon changes in vehicle speed and engine speed under alternative control actions and matched historical operating conditions. It does not reproduce the complete vehicle powertrain and dynamics, continuously synchronize with a production truck, or provide bidirectional real-time control. Accordingly, the resulting policy outcomes are interpreted as surrogate-estimated counterfactual results rather than field-validated responses or guaranteed production-level savings.

The research procedure comprised nine linked stages: telemetry acquisition and audit; operational-phase mapping; full haul-cycle reconstruction; derivation of a cycle-linked driving subset; fuel and transition surrogate modeling; fixed-distance reinforcement learning (RL) environment design and reward screening; equal-budget multi-seed PPO and SAC training; route-level and complete-cycle evaluation; and statistical, uncertainty, out-of-distribution (OOD), and claim-boundary interpretation. Figure 1 summarizes the workflow.

Full telemetry supplied cycle duration, total fuel, delivered payload, loading, dumping, queueing, waiting, and other non-driving intervals. The RL policies modified only loaded and empty travel. This separation prevented a controller from appearing artificially productive because non-driving portions of the haul cycle were omitted.

Three predictive components serve distinct roles. The fuel ensemble supplies the instantaneous fuel estimate; the next-speed model advances distance and determines travel time; and the next-engine-speed model provides powertrain context for the following fuel query. Their accuracy and uncertainty were consequently evaluated separately.

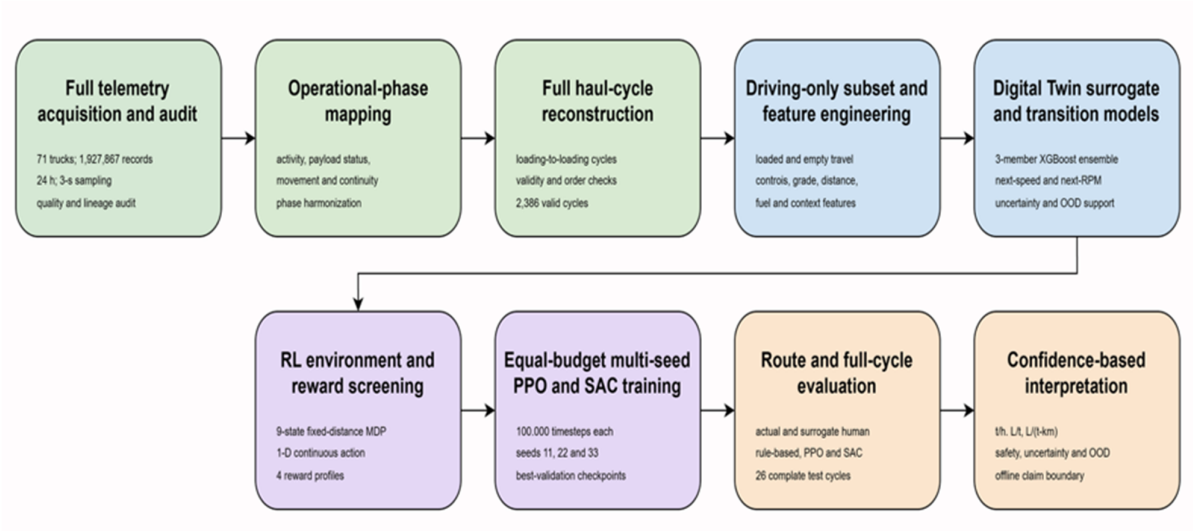

Figure 1. Overall workflow of the telemetry-driven Digital Twin surrogate-assisted reinforcement-learning framework.

Final ranking did not rely on accumulated RL reward. Measured-human, surrogate-human, rule-based, PPO, and SAC trajectories covered the same route distance and were then returned to their original complete cycles. Inferential tests used cycles as the independent units; step-level safety, uncertainty, and out-of-distribution (OOD) summaries were treated as diagnostics rather than independent replications.

2.2. Study Equipment, Data Acquisition, and Dataset Architecture

2.2.1. Equipment and operational context

The anonymized sample comprised Komatsu HD785-7 rigid mining trucks from an Indonesian open-pit operation. Published specifications list a 91.7 t nominal payload, 879 kW net power, and a maximum travel speed of 65 km/h [32]. The 110 t analytical cutoff was a data-quality bound, not a revised rated payload. No production year was available in the research file, nor was one inferred.

2.2.2. Telemetry acquisition

Telemetry captured truck and equipment identifiers, timestamps, operating activities, payload state, vehicle and engine speed, fuel rate, payload, inclination, accelerator, service brake, retarder, transmission indicators, location, shift, and assignments. The study was observational: no commands were sent to the trucks, and identifying information for equipment or operators was anonymized or handled as confidential.

The 24-hour window ran from 13 August 2025 at 09:23:19 to 14 August 2025 at 09:23:19 and contained 1,927,867 rows across 45 recorded fields. The mine's telemetry and vehicle-monitoring infrastructure generated the records without intervention by the researchers.

2.2.3. Two-level dataset architecture

The full-telemetry dataset was used for quality auditing, phase mapping, cycle reconstruction, complete-cycle time and fuel integration, payload throughput, and full-cycle fuel intensity. The driving-only analytical subset retained moving loaded and empty travel for fuel-surrogate development, transition modeling, safety-envelope estimation, rule-based control, and RL simulation. Each driving record originated from the audited full-telemetry master data and was linked to a validated haul cycle where possible.

Table 1. Summarizes the equipment and telemetry-acquisition characteristics used in this study.

Item	Description
Equipment	Komatsu HD785-7 rigid mining haul truck
Manufacturer specification	Nominal payload 91.7 t; net power 879 kW; maximum travel speed 65 km/h
Mining environment	Indonesian open-pit mine
Observation period	13 August 2025 09:23:19 to 14 August 2025 09:23:19
Observation duration	24 h
Nominal sampling interval	3 s
Raw telemetry	1,927,867 records; 71 trucks; 45 recorded variables
Production year	Unavailable in the anonymized telemetry

Table 2. Traces the data lineage from raw telemetry to the analytical datasets.

Data stage	Records	Trucks	Interpretation
Raw telemetry	1,927,867	71	Complete source data
Clean full telemetry	1,927,857	71	After invalid-key and non-target-equipment removal
Driving-only analytical subset	788,817	70	Moving loaded and empty travel
Cycle-linked driving observations	469,221	64	Used for surrogate and transition modeling
Candidate haul cycles	2,584	–	Loading-to-loading candidates
Validated haul cycles	2,386	–	92.34% passed all validity criteria

Each driving observation remained traceable to its truck, timestamp, continuity block, load state, and reconstructed cycle identifier when available. This lineage allowed simulated loaded and empty routes to be reattached to the correct historical cycle, rather than borrowing payload, grade, or exogenous time from another operation.

2.2.4. Data-quality screening and treatment of outliers

Outlier treatment was based on domain-informed plausibility rules rather than blanket statistical trimming. Measurements that violated applicable physical or operational validity limits were excluded from the corresponding analytical dataset. These rules included an engine-speed range of 500–2,500 rpm, a payload ceiling of 110 t, and a road-inclination range of −25° to 25°. Records reporting engine speeds above 1,200 rpm and a fuel rate below 1.0 L/h were also treated as internally inconsistent. The 110 t payload ceiling was used exclusively as a data-quality threshold and did not replace the manufacturer's nominal payload rating of 91.7 t [32].

No general interquartile range, z-score, or percentile-based removal was applied to observations that remained physically plausible. Operationally meaningful extremes, such as high fuel rates, steep inclinations, high engine speeds, or aggressive control inputs, were retained because they may represent genuine operating conditions relevant to policy robustness and safety. Non-driving observations were likewise not classified as outliers; they remained in the clean full-telemetry layer for complete haul-cycle reconstruction but were excluded from the driving-only dataset because those phases were outside the control scope of the longitudinal driving policy. Table 2 reports the records and trucks retained at each processing stage.

2.3. Full Haul-Cycle Reconstruction, Data Preprocessing, and Feature Engineering

2.3.1. Temporal continuity and data-quality audit

Records were ordered within the truck by timestamp. A gap greater than 15 s opened a new continuity block, so distance or fuel was not integrated across a logging interruption. Cleaning removed seven records with invalid keys and three records from non-target equipment; no duplicate record ID or truck–timestamp pair remained.

Plausibility screening was applied to individual sensor values: speed 0–70 km/h, engine speed 0–2,500 rpm, fuel rate 0–500 L/h, payload 0–110 t, grade −25° to +25°, and accelerator, brake, and retarder positions 0–100%. A value outside its range was set to missing, while the other valid fields in that row were retained.

The audit identified 21 invalid speed values, two engine-speed values, 1,938 payload values, 29 grade values, and one retarder value. Only the affected sensor entry was set to missing; identifiers, activities, and unaffected measurements in the same row remained available for phase and cycle checks.

2.3.2. Operational phases and cycle reconstruction

Activity identifiers, payload-status codes, movement status, payload trends, and speed were harmonized into loading, loaded travel, dumping, empty travel, queue/waiting, delay, and inactive phases. Activity–payload-status agreement reached 96.13%. A complete haul cycle began at one validated loading run and ended at the start of the next validated loading run for the same truck and continuity block.

A candidate cycle was retained when its duration was 5–180 min, both loaded and empty travel exceeded 1 min, the delivered payload was 10–110 t, temporal coverage was at least 80%, loading–loaded travel–dumping–empty return occurred in the correct order, and the cycle-quality score was at least 80%. The procedure produced 2,386 valid cycles from 2,584 candidates.

Table 3. Summarizes the principal preprocessing and haul-cycle validity rules.

Component	Rule
Continuity block	New block when inter-record gap >15 s
Driving subset	Loaded or empty travel; moving status; speed ≥5 km/h
Engine and terrain plausibility	500–2,500 rpm; grade −25° to +25°
Cycle boundaries	Start of one loading run to start of the next loading run
Cycle duration	5–180 min
Required travel	Loaded travel ≥1 min and empty travel ≥1 min
Delivered payload	10–110 t
Temporal coverage	≥80%
Required order	Loading → loaded travel → dumping → empty return
Cycle quality	≥80%

2.3.3. Distance, fuel, productivity, and efficiency metrics

Distance and fuel consumption were integrated using the actual valid time interval between consecutive telemetry records rather than assuming perfectly uniform sampling. For observation i, the incremental travel distance was calculated as:

$$\Delta d_i = \frac{v_i \Delta t_i}{3600} \quad (1)$$

and the incremental fuel consumption was calculated as:

$$\Delta F_i = \frac{FR_i \Delta t_i}{3600} \quad (2)$$

where Δd_i is the incremental distance in km, v_i is vehicle speed in km/h, Δt_i is the valid interval between consecutive observations in seconds, ΔF_i is incremental fuel consumption in L, and FR_i is the measured or surrogate-estimated fuel rate in L/h.

Complete-cycle fuel consumption and duration were obtained by summing all valid observations belonging to haul cycle c, such that $F_c = \sum_{i \in c} \Delta F_i$ and $T_c = \sum_{i \in c} \Delta t_i$. The nominal three-second interval was used only within the simulated reinforcement-learning environment after temporal-continuity screening.

Payload throughput was defined as the delivered payload divided by the complete-cycle duration:

$$\mathrm{P_c} = \frac{\mathrm{M_c}}{\mathrm{T_c}/3600} \quad (3)$$

where P_c is the payload throughput in t/h, M_c is the delivered payload in t, and T_c is the complete-cycle duration in seconds.

Complete-cycle fuel intensity was expressed as:

$$\mathrm{FI_{c,L/t}} = \frac{\mathrm{F_c}}{\mathrm{M_c}} \quad (4)$$

To account jointly for fuel use, transported payload, and haul distance, transport-normalized fuel intensity was calculated using Eq. (5). This production-centered efficiency formulation follows the haulage-efficiency assessment principles discussed in [1], 5, 6].

$$\mathrm{FI_{c,L/(t\cdot km)}} = \frac{\mathrm{F_c}}{\mathrm{M_c D_{c,loaded}}} \quad (5)$$

where F_c is the complete-cycle fuel consumption in L and $D_{c,\mathrm{loaded}}$ is the loaded-travel distance in km.

Payload throughput was therefore derived from the amount of material delivered per unit of complete-cycle time rather than from average vehicle speed. Fuel intensity in L/t represented the fuel required per unit of delivered payload, while L/(t·km) additionally accounted for the loaded haul distance. Fuel consumption in L/km and mean travel speed were retained only as secondary route-level diagnostics rather than as direct measures of mine productivity. This distinction is consistent with production-centered evaluations of mining-haulage performance [1, 5, 6].

2.3.4. Feature engineering and terrain classification

Control inputs, including throttle, service-brake, and retarder positions, were normalized to [0,1]. Engineered variables included control intensity, signed and absolute speed change, load state, night-shift indicator, cleaned gear and invalid-gear flag, lock-up and eco-mode indicators, transmission-output speed, distance and fuel steps, remaining route distance, empirical target speed, contextual safe speed, and grade category. Control intensity summarized the bounded combined demand of the three control inputs, while signed and absolute speed changes captured short-horizon vehicle response. Speed changes were calculated within the same haul-cycle and load-state sequence to avoid linking unrelated operational segments.

Road inclination was retained as a continuous variable for surrogate modeling and reinforcement-learning state representation. For terrain-stratified performance analysis, the observed grade signal was additionally grouped into four operational categories: downhill ($g < -0.5°$), flat ($-0.5° \leq g < 0.5°$), uphill ($0.5° \leq g < 3.0°$), and steep uphill ($g \geq 3.0°$), where g denotes the road-inclination signal. The ±0.5° interval around level terrain was used as a near-zero dead band to prevent small sensor fluctuations from causing frequent switching among terrain categories. The 3.0° boundary separated the higher-demand climbs observed in the dataset from moderate positive grades, reflecting the increasing tractive effort and fuel demand associated with gravitational resistance [3, 4].

2.4. Digital Twin Surrogate Modeling

2.4.1. Scope and 17-variable fuel model

The fuel component was implemented as a task-specific data-driven surrogate rather than a complete physics-based replica. XGBoost was selected for its ability to represent nonlinear relationships in tabular telemetry, support scalable tree construction, and handle missing feature values through learned split directions [20]. For observation i and ensemble member m, the predicted fuel rate was formulated as:

$$\widehat{FR}_{i,m} = f_{\theta_m}(\boldsymbol{x}_i) \quad (6)$$

The temporal test assessed predictive performance on later haul cycles rather than on a completely external mine or fleet. Because the same truck identities could appear in more than one temporal partition, grouped truck cross-validation and temporal testing provided complementary forms of internal validation rather than external validation.

Within this validation framework, the 17 predictors were limited to variables available in the cycle-linked driving data and reproducible during counterfactual simulation. Report hour, day of week, region identifiers, event alarms, and route labels used in earlier drafts were removed because they could not be reconstructed consistently for policy-generated states or could create leakage from outcome-related operational coding.

2.4.2. Temporal splitting, grouped cross-validation, and ensemble construction

Validated haul cycles were ordered by start time and divided chronologically into 60% training, 20% validation, and 20% untouched test partitions. All observations belonging to the same haul cycle remained within one partition, thereby preventing row-level leakage across the same operational sequence.

Hyperparameters were selected from 12 randomly sampled configurations using three-fold cross-validation grouped by truck, preventing a truck from appearing in both fitting and holdout folds. Table 5 reports the search ranges, selected configuration, 60,000-observation training folds, 10,000-observation fold-validation subsamples, three 600-tree cycle-bootstrap members, and their common 269,332-row training pool. Candidate settings were ranked by grouped-fold MAE, with RMSE secondary.

Table 4. Lists the input variables of the Digital Twin fuel-rate surrogate.

No.	Variable	Description
1	speed_kmh	Vehicle speed
2	eng_speed	Engine speed
3	plm_payload	Measured payload
4	plm_inc	Road inclination
5	accel_norm	Normalized accelerator input
6	brake_norm	Normalized service-brake input
7	retarder_norm	Normalized retarder input
8	control_intensity	Combined bounded control intensity
9	speed_change	Signed speed difference
10	abs_speed_change	Absolute speed difference
11	load_flag	Loaded versus empty indicator
12	is_night_shift	Night-shift indicator
13	gear_clean	Cleaned gear value
14	gear_invalid_flag	Invalid original gear indicator
15	lu_on_clean	Cleaned lock-up indicator
16	tm_out_speed_clean	Cleaned transmission-output speed
17	eco_on_clean	Cleaned eco-mode indicator

Table 5. Surrogate-data partitions, XGBoost hyperparameter search, and selected ensemble configuration.

Component	Search specification or data partition	Selected value / final configuration
Training / validation / test	Chronological cycle-level split: 60% / 20% / 20%	269,332 / 98,693 / 101,196 observations; 1,431 / 477 / 478 cycles
Grouped cross-validation	60,000 cycle-stratified training observations; three folds grouped by truck identity	12 randomly sampled configurations; seed = 2026
n_estimators	randint(300, 801); effective range 300-800	397
max_depth	randint(4, 9); effective range 4-8	7
learning_rate	loguniform[0.02, 0.12]	0.02423
min_child_weight	randint(3, 16); effective range 3-15	12
subsample	uniform[0.70, 1.00]	0.92675
colsample_bytree	uniform[0.70, 1.00]	0.91179
reg_alpha	loguniform[0.0001, 0.5]	0.000189
reg_lambda	loguniform[0.5, 5.0]	0.63958
Objective / tree method	Fixed during search	reg:squarederror / hist
Selection criterion	Mean fold RMSE (primary); mean fold MAE (secondary)	Lowest mean RMSE, with MAE used as a secondary criterion
Ensemble	Cycle-level bootstrap resampling	Three XGBoost members

Three final XGBoost members were trained using cycle-level bootstrap resampling. Ensemble averaging reduced dependence on a single bootstrap sample, while between-member standard deviation provided a model-disagreement diagnostic. The ensemble-mean fuel-rate prediction for observation i was calculated as:

The final ensemble prediction was subsequently calculated using Eq. (7) as the arithmetic mean of the three cycle-bootstrap XGBoost predictions. This equation defines the study-specific aggregation rule used in the present framework and does not represent a newly proposed machine-learning algorithm. The bootstrap-aggregation rationale for averaging predictions from models fitted to bootstrap replicates follows Breiman [33].

$$\widehat{\mathrm{FR}}_{\mathrm{i}} = \frac{1}{\mathrm{M}} \sum_{\mathrm{m}=1}^{\mathrm{M}} \widehat{\mathrm{FR}}_{\mathrm{i,m}}, \ \mathrm{M} = 3 \tag{7}$$

where $\widehat{FR}_{i,m}$ is the fuel-rate prediction produced by ensemble member m, $\widehat{FR}_i$ is the ensemble-mean prediction for observation i, and M is the number of ensemble members.

A Dummy mean regressor and a Histogram Gradient Boosting regressor served as benchmark models. MAE, RMSE, and R^2 were evaluated using the same validation and untouched temporal-test partitions to determine whether the selected XGBoost ensemble improved materially over a constant baseline and an alternative nonlinear tree ensemble.

2.4.3. Conformal uncertainty and state-dependent error

Split-conformal calibration used the absolute ensemble residuals from the temporal-validation set with a target miscoverage level of $\alpha = 0.05$. The corrected empirical 95th-percentile residual quantile defined the following global symmetric prediction set:

$$\mathcal{C}_{0.95}(x) = [\overline{FR}(x) - \hat{q}_{0.95}, \overline{FR}(x) + \hat{q}_{0.95}] \tag{8}$$

where $\widehat{FR}(x)$ is the ensemble-mean fuel-rate prediction for input vector x, and $\hat{q}_{0.95}$ is the corrected 95th-percentile quantile of the absolute calibration residuals.

This global interval was interpreted as a calibration diagnostic rather than a guarantee of conditional accuracy because the telemetry was temporally dependent and future policies could induce covariate shift [34]. The conformal procedure provided a distribution-free global residual interval without assuming normally distributed prediction errors.

A second XGBoost model estimated state-dependent absolute residuals using the same 17 predictor variables. Its regression target was defined as:

$$e_i = |FR_i - \widehat{FR}_i| \tag{9}$$

where e_i is the absolute prediction residual for observation i, FR_i is the observed fuel rate, and $\widehat{FR}_i$ is the ensemble-mean predicted fuel rate.

This model did not create a formal prediction interval. It supplied a relative uncertainty signal so that policy states expected to have larger surrogate errors could be penalized and reported separately. Ensemble disagreement, conformal residual magnitude, and predicted state-dependent error represent different uncertainty components. Low ensemble disagreement does not imply low total error because ensemble members can share the same feature limitations and model bias.

2.4.4. OOD support assessment

Ten continuous surrogate inputs were standardized using the means and standard deviations estimated from the training data. A Ledoit–Wolf shrinkage covariance estimator was used to obtain a well-conditioned covariance matrix and its corresponding precision matrix for Mahalanobis-distance calculation [35]. This regularization reduced numerical instability associated with the direct inversion of the empirical covariance matrix.

Mahalanobis distance was used as a global multivariate measure of departure from the training-data distribution. For a standardized policy-query vector z, it was calculated as:

$$D_M(z) = \sqrt{(z-\mu)^\top \Sigma^{-1}(z-\mu)} \quad (10)$$

where μ and Σ denote the location vector and Ledoit–Wolf shrinkage covariance matrix estimated from the standardized training observations. Mahalanobis-distance-based OOD detection evaluates whether a new feature vector departs from the multivariate structure of the reference training distribution [36].

A five-nearest-neighbor distance was used as a complementary non-parametric measure of local training-data support:

$$D_{kNN}(z) = \frac{1}{k}\sum_{j=1}^{k} \left\| z - z_{(j)} \right\|_2, \; k = 5 \quad (11)$$

where $z_{(j)}$ denotes the j-th nearest standardized training observation. Unlike covariance-based scoring, the kNN measure does not require fitting a global parametric covariance representation. A larger mean neighbor distance indicates that the evaluated state is located in a locally sparse region of the historical training distribution. Nearest-neighbor-based methods have been used to identify OOD observations whose feature representations remain distant from available in-distribution reference samples [37].

The support models were constructed using up to 25,000 training observations, while 10,000 temporal-validation observations were used to calibrate the thresholds. The 95th percentile of each validation-distance distribution defined the corresponding OOD threshold. Binary and low-cardinality variables were excluded from the geometric distance calculation.

OOD indicators were not used to remove unfavorable policy outcomes. Mahalanobis OOD entered the training reward only as a soft penalty, whereas both Mahalanobis and kNN indicators were reported during final evaluation. This design avoided post-hoc selection of only well-supported or favorable policy states.

2.4.5. Common-estimator final rescoring

During training, one fixed bootstrap member supplied stepwise fuel estimates to control computational cost and keep the environment stationary. After checkpoint selection, all saved rule-based, PPO, and SAC queries—and the surrogate-human baseline—were rescored with the same three-member ensemble mean. This audit changed only fuel quantities and related statistics; actions, speed, RPM, completion time, throughput, safety, predicted error, and OOD trajectories were unchanged.

2.5. Reinforcement Learning Environment Design

A Gymnasium-compatible environment represented a fixed-distance longitudinal-control problem. Each three-second simulation step provided a nine-variable state, accepted one continuous action, predicted the next vehicle speed and engine speed, reconstructed the 17-feature fuel-query vector, estimated the fuel rate, updated the cumulative route distance, and returned a multi-objective reward. Figure 2 illustrates the interaction among the route context, state, agent, control allocation, transition models, fuel surrogate, reward function, and route-progress update.

The environment was formulated as the following Markov Decision Process:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma) \quad (12)$$

where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ is the continuous action space, $\mathcal{P}$ is the data-driven transition mechanism, $\mathcal{R}$ is the reward function, and γ is the discount factor.

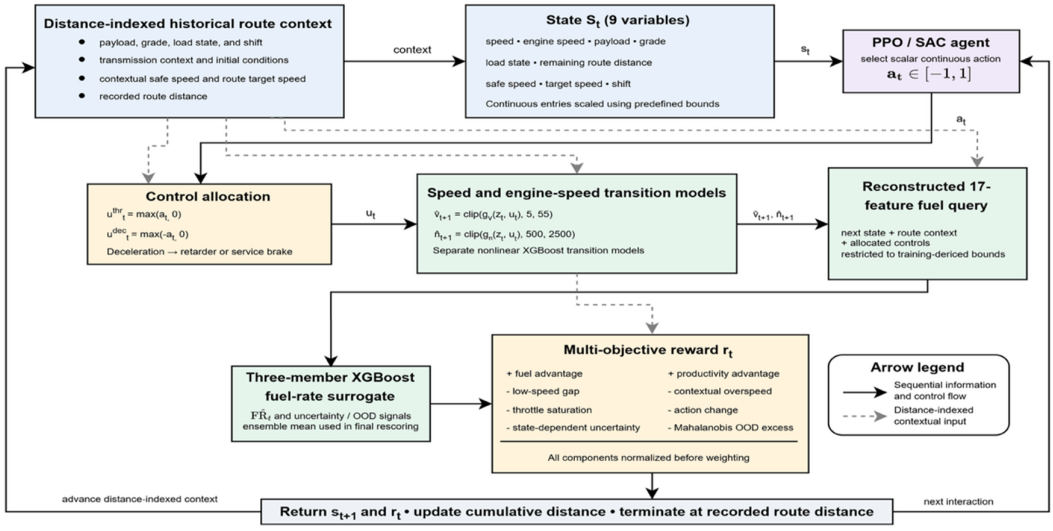

Figure 2. State–action–transition–reward interaction in the fixed-distance Digital Twin surrogate-assisted reinforcement-learning environment.

Table 6. Nine-variable RL state representation and normalization bounds.

State variable	Meaning	Lower	Upper
Current speed	Simulated vehicle speed	5 km/h	55 km/h
Current engine speed	Simulated engine speed	500 rpm	2,500 rpm
Payload	Distance-indexed route payload	0 t	110 t
Road grade	Distance-indexed inclination	−25°	+25°
Load state	Empty = 0; loaded = 1	0	1
Remaining distance	Fraction of route remaining	0	1
Contextual safe speed	Empirical load–grade–payload limit	5 km/h	50 km/h
Target speed	Route-specific productivity target	5 km/h	50 km/h
Shift context	Day = 0; night = 1	0	1

This design differed from fixed-step evaluation because every controller was required to complete the same historical route distance. A slower controller therefore required more simulation steps and a longer completion time rather than being evaluated over a shorter traveled distance.

A route ended when the simulated cumulative distance reached the recorded route length. If this had not occurred after 1,500 steps, the episode was marked as truncated and incomplete rather than counted as a completed route.

2.5.1. State representation, action space, contextual speed limits, and transition mechanism

The state contained nine contemporaneous variables: vehicle speed, engine speed, payload, road grade, load state, remaining-route-distance fraction, contextual safe speed, route target speed, and shift context, as summarized in Table 6. Each entry was transformed to [0, 1] using fixed bounds established before final training and was clipped to that interval; no future route observation was exposed to the agent.

$$\tilde{x}_t = \mathrm{clip}\left(\frac{x_t - x_{\min}}{x_{\max} - x_{\min}}, 0,1\right) \tag{13}$$

The scalar action $a_t \in [-1,1]$ represented longitudinal control tendency. Positive values were mapped to throttle, whereas negative values were mapped to deceleration. The corresponding propulsion and deceleration demands were defined as:

$$u_t^{\mathrm{thr}} = \max(a_t, 0), \quad u_t^{\mathrm{dec}} = \max(-a_t, 0) \tag{14}$$

where u_t^{thr} is normalized throttle demand and $u_t(\mathrm{dec})$ is normalized deceleration demand. When deceleration was requested, it was allocated deterministically to the retarder if the road grade was below −0.5° or if the current speed exceeded the contextual safe speed; otherwise, it was allocated to the service brake:

$$\left(u_t^{\mathrm{brk}}, u_t^{\mathrm{ret}}\right) = \begin{cases} \left(0, u_t^{\mathrm{dec}}\right), & g_t < -0.5^\circ \text{ or } v_t > v_t^{\mathrm{safe}} \\ \left(u_t^{\mathrm{dec}}, 0\right), & \text{otherwise} \end{cases} \tag{15}$$

This rule prevented simultaneous positive throttle and braking commands and prevented simultaneous service-brake and retarder commands. It remained a simplified longitudinal-control abstraction and did not reproduce independent gear commands, brake pressure, retarder levels, steering, or engine-torque control.

A contextual safe-speed envelope was estimated from the training data as the 95th percentile of historically non-overspeed speed for combinations of load state, grade category, and payload tertile group. Hierarchical fallback was applied for sparsely represented cells, and the resulting threshold was subject to a 50 km/h cap and a 0.5 km/h safety margin. The resulting value represented an empirical, site-specific operating envelope rather than a manufacturer-certified, regulatory, or geotechnically verified safety limit.

The route-level productivity target was calculated once when the route object was constructed. A productivity-oriented candidate combined 90% of the mean historical route speed with the first quartile of the historical contextual safe speed, after which the candidate was capped by the median historical safe speed:

$$v_{\mathrm{cand}} = \max\left(0.90\overline{v}^{\mathrm{hist}}, Q_{0.25}\left(v^{\mathrm{safe,hist}}\right)\right) \tag{16}$$

$$v^{\mathrm{tar}} = \min\left(v_{\mathrm{cand}}, Q_{0.50}\left(v^{\mathrm{safe,hist}}\right)\right) \tag{17}$$

Here, $\overline{v}^{\mathrm{hist}}$ is the mean historical route speed and Q denotes a quantile of the historical safe-speed distribution. Thus, 90% of the historical mean speed served as a candidate rather than a guaranteed minimum; the final target could be lower when constrained by the route-specific safe-speed distribution.

Separate XGBoost transition models predicted next vehicle speed and next engine speed from the current state, allocated controls, and transmission context. Let z_t denote the current state and route-context vector, and let u_t denote the allocated throttle, service-brake, and retarder controls. The transitions were calculated as:

$$\hat{v}_{(t+1)} = clip(g_v(\mathrm{z}_t, \mathrm{u}_t), 5{,}55), \qquad \hat{n}_{(t+1)} = clip(g_n(\mathrm{z}_t, \mathrm{u}_t), 500{,}2500) \tag{18}$$

where $\hat{v}_{t+1}$ is the predicted next vehicle speed in km/h, $\hat{n}_{t+1}$ is the predicted next engine speed in rpm, and $g_v(\cdot)$ and $g_n(\cdot)$ are the respective nonlinear transition models. The clip(·) operator constrained the predictions to the predefined operational intervals. Transmission context was an internal environment input and was not an additional agent-observed state.

Predicted next speed determined the distance traveled during the three-second sampling interval and cumulative route progress:

$$\Delta d_t = \hat{v}_{t+1}\frac{\Delta t}{3600}, \qquad d_{t+1} = \min(d_t + \Delta d_t, D), \qquad \Delta t = 3\ \mathrm{s} \tag{19}$$

The environment mapped d_{t+1} to the first historical route record whose cumulative distance was equal to or greater than the simulated distance using searchsorted (..., side="left"); the index was capped at the final record. Grade, payload, load state, shift, transmission context, and contextual safe speed were then read from that distance-indexed record. Remaining-route-distance fraction was updated as $q_{t+1} = \max\left(0, \frac{D-d_{t+1}}{D}\right)$. Inputs to the transition and fuel models were restricted to percentile bounds learned from training data. This clipping stabilized recursive simulation but did not by itself establish physical validity within those bounds.

After constructing the next-state and allocated-control features, the XGBoost ensemble produced a policy-conditioned fuel-rate estimate $\widehat{FR}_t^{\mathrm{pol}}$. A matched reference estimate $\widehat{FR}_t^{\mathrm{ref}}$ was obtained under the historical control at the same route context. Step-level policy fuel consumption was calculated as:

$$\hat{f}_t^{\text{pol}} = \widehat{FR}_t^{\text{pol}} \frac{\Delta t}{3600} \tag{20}$$

where fuel rate was expressed in L/h and step-level fuel in liters. This sequence linked action allocation, learned state transition, route advancement, fuel estimation, and reward calculation within each environment step.

2.5.2. Reward function

The reward encompasses eight concerns: normalized fuel and productivity advantages, low-speed deviation, contextual overspeed, throttle saturation, action smoothness, state-dependent uncertainty, and Mahalanobis OOD support. Its stepwise definition is as follows:

$$r_t = w_f A_t^{\text{fuel}} + w_p A_t^{\text{prod}} - w_l L_t^2 - w_o O_t^2 - w_s S_t^2 - w_{\Delta a}(a_t - a_{t-1})^2 - w_D D_t - w_U U_t \tag{21}$$

All reward components are dimensionless. Fuel and productivity advantages and the four direct operating gaps are defined as follows, with ε denoting a small positive numerical safeguard:

$$A_t^{\text{fuel}} = \frac{\widehat{FR}_t^{\text{ref}} - \widehat{FR}_t^{\text{pol}}}{\max\left(\widehat{FR}_t^{\text{ref}}, \varepsilon\right)} \tag{22}$$

$$A_t^{\text{prod}} = \text{clip}\left(\frac{\hat{v}_{t+1} - v^{\text{tar}}}{\max(v^{\text{tar}}, \varepsilon)}, -1,1\right) \tag{23}$$

$$L_t = \max\left(0, \frac{v^{\text{tar}} - \hat{v}_{t+1}}{\max(v^{\text{tar}}, \varepsilon)}\right); \qquad O_t = \max\left(0, \frac{\hat{v}_{t+1} - v_t^{\text{safe}}}{\max(v_t^{\text{safe}}, 1.0)}\right) \tag{24}$$

$$S_t = \max\left(0, \frac{u_t^{\text{thr}} - 0.90}{1.00 - 0.90}\right); \ \Delta a_t = a_t - a_{t-1} \tag{25}$$

Fuel advantage is positive when the policy-conditioned surrogate estimate is below the matched reference estimate. Productivity advantage compares the predicted next speed with the route target and is clipped to [−1, 1]. Low-speed, overspeed, throttle-saturation, and action-change terms penalize productivity loss, empirical-envelope exceedance, throttle above 0.90, and abrupt control changes, respectively.

OOD and uncertainty penalties are activated only when the ratios exceed the validation-calibrated threshold of one:

$$D_t = \max\left(0, \rho_t^{\text{OOD}} - 1\right), \rho_t^{\text{OOD}} = \frac{d_t^{\text{Mah}}}{Q_{0.95}\left(d_{\text{val}}^{\text{Mah}}\right)} \tag{26}$$

$$U_t = \max(0, \rho_t^{\text{unc}} - 1), \qquad \rho_t^{\text{unc}} = \frac{\hat{e}_t}{Q_{0.95}\left(\left|\widehat{FR}_{\text{val}} - FR_{\text{val}}\right|\right)} \tag{27}$$

Here, d_t^{Mah} is the Mahalanobis distance of the policy query in the proxy-feature space, and $\hat{e}_t$ is the state-dependent absolute-error proxy. Both denominators were computed exclusively from validation data: the 95th percentile of validation Mahalanobis distance and the 95th percentile of ensemble absolute error, respectively. Consequently, D_t and U_t were zero within their calibrated support and increased only beyond those thresholds. The complete weighted reward is given in Eq. (21). The rationale, sensitivity screening, and final selection of the reward-weight coefficients are described in Section 2.5.3.

2.5.3. Reward-weight sensitivity screening and final selection

The reward coefficients were determined through a validation-based sensitivity screening rather than adopted as fixed or universally optimal constants. Because the reward components defined in Eqs. (21)-(27) are dimensionless, their coefficients represent relative operational priorities rather than compensating for differences in physical units or numerical magnitude. Four predefined reward-weight profiles were evaluated before final training, as reported in Table 7.

Table 7. Reward-weight profiles evaluated before final training.

Profile	w_f	w_p	w_l	w_o	w_s	w_Δa	w_D	w_U
Balanced	1.00	1.00	2.00	5.00	0.75	0.20	0.30	0.30
Fuel dominant	1.50	0.75	2.50	5.00	0.75	0.20	0.30	0.40
Productivity dominant	0.75	1.50	3.00	6.00	1.00	0.25	0.30	0.30
Safety/support constrained	1.00	1.00	2.50	10.00	1.50	0.35	0.75	0.75

The four predefined profiles represent Balanced, Fuel-dominant, Productivity-dominant, and Safety/support-constrained priorities. For sensitivity screening, PPO and SAC were trained separately under each profile for 20,000 timesteps using seed 7 and subsequently evaluated deterministically on the same validation routes. Independent test routes and test cycles were excluded from reward-weight selection.

Profiles were ranked using a composite validation score spanning fuel efficiency, productivity retention, speed-envelope and full-throttle behavior, OOD exposure, predicted surrogate error, and route completion. The mean cross-algorithm score was primary, with the lower algorithm-specific score as the tie-breaker.

The Productivity-dominant profile achieved the highest mean validation score (6.4189) and worst-algorithm score (5.5985) and was frozen before final training: wf = 0.75, wp = 1.50, wl = 3.00, wo = 6.00, ws = 1.00, wΔa = 0.25, wD = 0.30, and wU = 0.30. Because screening used one seed and 20,000 timesteps, it was treated only as transparent model selection.

2.6. Policy Optimization Algorithms

PPO and SAC were compared because they represent complementary mechanisms for continuous-control learning. PPO uses current-policy rollouts and clipped updates, whereas SAC reuses replay-buffer transitions and optimizes a maximum-entropy objective [30, 31]. The study did not assume that either algorithm was intrinsically superior.

2.6.1. Proximal Policy Optimization

PPO is an on-policy actor–critic algorithm that restricts the benefit of excessive policy updates through a clipped probability-ratio objective [30]:

$$L^{(CLIP)}(\theta) = \mathbb{E}_t\left[\min\left(\rho_t(\theta)\hat{A}_t, \mathrm{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t\right)\right] \tag{28}$$

where

$$\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{(\theta_o ld)}(a_t|s_t)} \tag{29}$$

is the probability ratio between the updated and rollout policies, $\widehat{A}_t$ is the estimated advantage at time step t, and ϵ is the clipping parameter.

PPO used Generalized Advantage Estimation, a 0.20 clipping range, ten optimization epochs per 1,024-step rollout, and mini-batches of 64 observations. Clipping limits the improvement credited to large policy-probability changes, but it is not a hard safety constraint and cannot prevent entry into unsupported or empirically unusual states.

2.6.2. Soft Actor-Critic

SAC is an off-policy actor–critic method that reuses replay buffer transitions and maximizes cumulative reward together with policy entropy [31]. Its maximum-entropy objective is expressed as:

$$J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{T}\gamma^t\left(r_t + \alpha\mathcal{H}\left(\pi(\cdot|\, s_t)\right)\right)\right] \tag{30}$$

where τ denotes a trajectory generated by policy π, r_t is the reward at time step t, γ is the discount factor, α is the entropy-temperature coefficient, and $\mathcal{H}\left(\pi(\cdot|\, s_t)\right)$ is the entropy of the policy distribution at state s_t.

SAC employs a stochastic actor, two action-value critics, slowly updated target critics, and automatic entropy-temperature optimization. Twin critics reduce positive value-estimation bias, while entropy regularization maintains exploration within the continuous one-dimensional action space [31].

During training, actions were sampled from the entropy-regularized stochastic policy. Deterministic evaluation used the mean action of the learned policy distribution rather than a randomly sampled action. This separation preserved exploration during training while enabling reproducible route-level comparisons.

Table 8. Comparison of the algorithmic structures of PPO and SAC.

Characteristic	PPO	SAC
Learning type	On-policy	Off-policy
Policy objective	Clipped policy gradient	Maximum-entropy actor objective
Experience reuse	Current rollout only	Replay buffer
Critic	State-value function	Twin action-value functions
Stabilization	Probability-ratio clipping	Twin critics, target networks, entropy
Training action	Stochastic	Stochastic
Final evaluation	Deterministic	Deterministic
Action interval	(−1,1)	(−1,1)

2.7. Training, Hyperparameter, and Reproducibility Configuration

The selected Productivity-dominant reward profile was frozen before final training. PPO and SAC used the same training and validation route pools, nine-variable state, action interval, fuel and transition surrogates, safety envelope, target-speed logic, nominal budget of 100,000 interaction timesteps, and seeds 11, 22, and 33. Validation occurred every 10,000 timesteps, and the best operational-validation checkpoint was retained for the untouched test cycles.

Randomness was controlled in Python, NumPy, PyTorch, CUDA, Stable-Baselines3, and environment resets. Experiments ran in Google Colab with Python 3.12.13, Stable-Baselines3 2.9.0, PyTorch 2.11.0/CUDA 12.8, Gymnasium 1.3.0, NumPy 2.0.2, XGBoost 2.0 or later, and an NVIDIA A100 GPU. Runs retained models, learning curves, seeds, reward profiles, budgets, scores, and elapsed times.

Table 9. Details the final PPO and SAC training configuration.

Parameter	PPO	SAC
Policy / hidden layers	MLP; 256–256 ReLU	MLP; 256–256 ReLU
Learning rate	3×10^{-4}	3×10^{-4}
Discount factor γ	0.99	0.99
Nominal budget / seeds	100,000; 11, 22, 33	100,000; 11, 22, 33
Validation interval	10,000 timesteps	10,000 timesteps
Rollout length / epochs	1,024 / 10	Not applicable
Batch size	64	256
GAE λ / clip range	0.95 / 0.20	Not applicable
Entropy	Coefficient 0.005	Automatic; target entropy −1
Replay buffer / learning starts	Not applicable	200,000 / 5,000
Train frequency / gradient steps	Rollout update	1 step / 1 update
Target coefficient τ	Not applicable	0.005
Checkpoint used for test	Best validation	Best validation

2.7.1. Training and validation route pools

The earliest training cycles generated 2,716 eligible route objects: 1,342 loaded and 1,374 empty. A fixed, diversity-selected validation set represented payload, shift, route pairing, grade variability, and uphill/downhill composition. Validation outcomes did not produce gradient updates.

2.7.2. Equal-budget final training

After reward selection, PPO and SAC were each trained using a nominal budget of 100,000 environment interactions per seed, with seeds 11, 22, and 33. Both algorithms used the same route pools, state and action definitions, reward coefficients, a learning rate of 3×10^{-4}, a discount factor of 0.99, and two 256-unit ReLU hidden layers. Algorithm-specific rollout and replay mechanisms were retained because they define the methods.

2.7.3. Checkpoint selection and seed control

Deterministic validation was performed every 10,000 timesteps. The best operational validation checkpoint, which was not automatically the terminal model, was used for test evaluation. Python, NumPy, PyTorch, CUDA, the environment, and Stable-Baselines3 seeds were set for each run. Data partitions, route inventories, support thresholds, normalization bounds, and reward weights were frozen before the final experiment.

2.8. Evaluation Protocol and Baseline Controllers

2.8.1. Independent test cycles and controllers

The chronologically latest 20% of validated cycles formed the untouched test pool. Thirty diverse cycles were preselected based on payload, route pair, shift, grade variability, and uphill/downhill composition. After minimum-distance and component-completeness checks, 26 cycles contained both eligible loaded and empty routes and served as the primary complete-cycle unit.

Historical Human-Actual represented measured telemetry. Historical Human-Surrogate replayed the same historical trajectory through the common fuel ensemble. Rule-Based Eco-Driving provided a transparent, non-learning benchmark. PPO and SAC were represented by three independently trained, best-validation policies each.

Table 10. Lists the controllers and their roles in the final evaluation.

Controller	Trajectory	Fuel source	Analytical role
Historical Human—Actual	Recorded	Measured telemetry	Observed operational reference
Historical Human—Surrogate	Recorded	Three-member ensemble mean	Matched modeled reference
Rule-Based Eco-Driving	Simulated deterministic rules	Three-member ensemble mean	Transparent non-learning benchmark
PPO	Three best-validation seed policies	Three-member ensemble mean	On-policy DRL controller
SAC	Three best-validation seed policies	Three-member ensemble mean	Off-policy DRL controller

2.8.2. Fixed-distance routes and complete-cycle reconstruction

Every simulated controller was required to complete the same historical loaded- or empty-route distance. Payload, grade, load state, shift, initial state, and the distance-indexed exogenous route context were held constant across controllers, whereas speed, engine speed, control actions, fuel consumption, and completion time were generated counterfactually.

The simulated loaded and empty-travel components were subsequently combined with the same historical exogenous residual time and fuel to reconstruct the complete haul cycle:

$$T_{c,k}^{\text{cycle}} = T_c^{\text{exo}} + T_{c,L,k} + T_{c,E,k,}$$
$$F_{c,k}^{\text{cycle}} = F_c^{\text{exo}} + F_{c,L,k} + F_{c,E,k} \quad (31)$$

where $T_{c,k}^{\text{cycle}}$ and $F_{c,k}^{\text{cycle}}$ denote the reconstructed duration and fuel consumption of cycle c under controller k; T_c^{exo} and F_c^{exo} denote the historical exogenous residual time and fuel; and subscripts L and E denote the simulated loaded-travel and empty-return components, respectively.

The historical exogenous residuals included loading, queueing, waiting, spotting, dumping, inactive periods, and low-speed portions excluded from the controllable route model. These operational phases were neither removed nor optimized by the simulated controllers.

A deterministic, rule-based controller tracked the lower of the route target speed and the contextual safe-speed envelope, applied capped throttle in response to a positive speed error, and preferred service-brake or retarder-based deceleration during overspeed and downhill operation. It served as a transparent, non-learning benchmark rather than a manufacturer-certified controller.

2.8.3. Outcome measures and statistical testing

Primary outcomes were complete-cycle payload throughput in t/h, fuel intensity in L/t, and transport-normalized fuel intensity in L/(t·km). Secondary diagnostics included L/km, route completion time, productivity retention, mean and maximum speed, control intensity, contextual safe-envelope exceedance, nominal overspeed above 50 km/h, full-throttle rate > 0.90, harsh service-braking rate > 0.80, predicted absolute fuel error, and Mahalanobis and kNN OOD rates.

Fuel savings were calculated relative to the Historical Human–Surrogate baseline because the policy and baseline fuel values were evaluated using the same frozen ensemble-mean estimator. For cycle $c$and controller k, the percentage fuel saving was defined as:

$$S_{c,k}^{\text{fuel}} = 100\left(\frac{FI_{c,\text{human,surr}} - FI_{c,k}}{FI_{c,\text{human,surr}}}\right) \quad (32)$$

where $FI_{c,\text{human,surr}}$ is the complete-cycle fuel intensity of the Historical Human–Surrogate baseline, and $FI_{c,k}$ is the corresponding fuel intensity under controller k. A positive value

indicates lower fuel intensity than the surrogate-human baseline, whereas a negative value indicates higher fuel intensity.

Throughput changes were calculated relative to the Historical Human–Actual baseline because the measured historical cycle time represented observed production. The percentage throughput change was defined as:

$$\Delta P_{c,k} = 100\left(\frac{P_{c,k} - P_{c,\text{human,actual}}}{P_{c,\text{human,actual}}}\right) \quad (33)$$

where $P_{c,k}$ is the reconstructed payload throughput under controller k, and $P_{c,\text{human,actual}}$ is the observed historical-human throughput for the same cycle. A positive value indicates increased throughput, whereas a negative value indicates decreased throughput.

The complete haul cycle was treated as the primary independent statistical unit. PPO- and SAC-specific results from the three training seeds were first averaged within each cycle before paired controller comparisons. Two-sided Wilcoxon signed-rank tests were then applied to the common cycle-level outcomes using $\alpha = 0.05$.

Controller means and hierarchical 95% confidence intervals were estimated using 10,000 bootstrap replicates. The bootstrap procedure resampled complete haul cycles and, for PPO and SAC, training seeds to represent both cycle-to-cycle and seed-to-seed variability.

The reported p-values were unadjusted for multiple comparisons. They were interpreted jointly with mean paired differences, hierarchical confidence intervals, seed variability, safety behavior, surrogate uncertainty, and OOD exposure. Statistical significance was not treated as evidence of physical causality, verified field effectiveness, deployment readiness, or universal algorithm superiority.

2.9. Confidence and Claim-Boundary Protocol

Evidence was classified as observed operational evidence, surrogate-reconstructed evidence, or offline counterfactual policy evidence. Only historical measured fuel and timing were observed physical outcomes. Surrogate-human results assessed matched model fidelity, while PPO, SAC, and rule-based outcomes were counterfactual. No operator-in-the-loop, controlled field, or production-deployment evidence was generated.

Statistical repeatability within the simulation was distinguished from surrogate predictive uncertainty. A narrow cycle-and-seed confidence interval could coexist with a wide fuel-surrogate conformal radius or high OOD exposure. A controller was therefore not considered preferable based on fuel, throughput, or reward alone. Interpretation required joint consideration of fuel intensity, productivity, safety-envelope behavior, action saturation, predicted error, distributional support, and seed stability.

Statements such as 'PPO reduced surrogate-estimated fuel intensity within the evaluated offline environment' were permitted. Statements such as 'PPO reduced actual production fuel', 'the policy was proven safe', or 'the framework is deployment-ready' were prohibited. Controller preference was based on the joint profile of fuel, throughput, safety, action behavior, uncertainty, OOD support, and repeatability.

3. Results

Results follow the analytical sequence: dataset and cycle integrity, surrogate and transition fidelity, reward selection, multi-seed learning, full-cycle controller outcomes, statistical comparison, safety and control behavior, uncertainty and OOD exposure, and integrated interpretation. All final modeled-controller fuel values utilize a common three-member ensemble rescoring.

3.1. Dataset Characteristics, Data Quality, and Haul-Cycle Reconstruction

3.1.1. Operational coverage and quality audit

The clean full-telemetry dataset contained 1,927,857 observations. Loaded and empty travel accounted for 55.54%, leaving 44.46% for loading, dumping, queueing/waiting, inactive, and disposal-queue activities. The large non-driving share confirms that average travel speed cannot represent complete-cycle mine productivity.

The driving subset comprised 788,817 observations from 70 trucks (466,646 loaded; 322,171 empty). Means ± SD were 24.38 ± 8.89 km/h for speed, 1,593.12 ± 206.91 rpm for engine speed, 128.29 ± 64.26 L/h for fuel rate, 53.79 ± 42.73 t for payload, 0.57 ± 1.67° for grade, and 70.17 ± 29.09% for accelerator position. Because empty and loaded travel are pooled here, the observation-level payload mean does not reflect the delivered payload per cycle.

All audit gates were satisfied. Activity and payload status agreed for 96.13% of observations; no duplicate identifiers or truck–timestamp pairs remained. Payload generated the largest invalid count, but 1,938 affected rows represented only about 0.10% of the cleaned telemetry.

3.1.2. Haul-cycle reconstruction results

The procedure reconstructed 2,584 candidate loading-to-loading cycles, of which 2,386 (92.34%) satisfied all validity criteria. The accepted cycles had a median quality score of 100%, a median duration of 19.12 minutes, and a median delivered payload of 89.50 tonnes. Table 12 summarizes the complete-cycle distribution.

Table 11. Presents the operational-state distribution and the principal data-quality findings.

Category	Records	Percentage / result
Inactive	409,683	21.25%
Empty travel	506,134	26.25%
Front queue or waiting	231,948	12.03%
Loading	146,777	7.61%
Loaded travel	564,559	29.28%
Disposal queue	32,382	1.68%
Dumping	36,374	1.89%
Activity–payload-status agreement	–	96.13%
Out-of-range speed / engine speed	21 / 2	Set to missing
Out-of-range payload / grade / retarder	1,938 / 29 / 1	Set to missing

Table 12. Descriptive statistics of the 2,386 validated haul cycles.

Cycle variable	Minimum	Mean	SD	Maximum
Cycle duration (min)	5.42	19.49	6.37	116.79
Delivered payload (t)	18.00	89.40	7.25	109.00
Total cycle fuel (L)	3.87	23.82	7.83	59.65
Total travel distance (km)	0.55	5.44	1.93	13.63
Payload throughput (t/h)	49.83	301.47	97.21	935.98
Fuel intensity (L/t)	0.0445	0.2686	0.0945	1.5611
Fuel intensity [L/(t·km)]	0.0457	0.1123	0.0339	0.7137

Note: The maximum throughput reflects a small number of unusually short-duration cycles and is not interpreted as a sustained production rate.

Across the validated cycles, the mean throughput was 301.47 tonnes per hour, but variation was substantial because cycle duration included queues, waiting, and other delays. A few very short cycles produced extreme maxima; the paired controller analysis therefore used the distribution of complete cycles rather than treating those maxima as sustainable production.

A total of 2,386 validated cycles were divided chronologically into 1,431 training, 477 validation, and 478 test cycles. Thirty test cycles were preselected to capture variation in payload, route, shift, and terrain characteristics. Of these, 26 cycles included complete eligible loaded- and empty-travel routes and constituted the held-out full-cycle evaluation set.

Truck identities recur across the chronological partitions. The test therefore measures transfer to later cycles from the same fleet, not external generalization to unseen equipment, another truck family, or a different mine.

3.2. Fuel-Surrogate Performance, Uncertainty, and Error Stratification

3.2.1. Cross-validation, temporal testing, and benchmarks

Grouped-truck cross-validation yielded MAE of 27.53 L/h, RMSE of 38.04 L/h, and R^2 of 0.6400. Similar values for temporal validation and untouched test data indicate no overall deterioration in the later period examined here.

Table 13. Reports the predictive performance of the fuel surrogate.

Model / stage	N	MAE (L/h)	RMSE (L/h)	R^2
XGBoost grouped truck CV	60,000	27.53	38.04	0.6400
XGBoost ensemble—validation	98,693	28.23	38.62	0.6439
XGBoost ensemble—test	101,196	27.33	37.75	0.6608
Dummy mean—test	101,196	57.04	64.84	−0.0005
Histogram Gradient Boosting—test	101,196	27.56	38.00	0.6564

Compared to the Dummy model, XGBoost reduced test MAE by 52.09% and RMSE by 41.78%. Histogram Gradient Boosting was close in performance, suggesting that the evidence favors nonlinear boosting generally rather than a uniquely large advantage for XGBoost. The test R^2 of 0.6608

indicates moderate task-specific fidelity and cannot be equated with physical replication.

The temporal-test ensemble achieved MAE of 27.33 L/h, RMSE of 37.75 L/h, and R^2 of 0.6608. It reduced MAE by approximately 52% and RMSE by approximately 42% relative to the Dummy mean predictor. Histogram Gradient Boosting performed similarly, with XGBoost providing only a modest incremental advantage.

3.2.2. Uncertainty and condition-specific error

Bootstrap members differed by only about 1.85 L/h on average, yet the test MAE was 27.33 L/h. Their agreement therefore did not eliminate common model error. The split-conformal 95% residual radius was 80.66 L/h, and the state-dependent error reference was approximately 80.62 L/h; instantaneous predictions remain materially uncertain.

Table 14. Breaks down the temporal-test MAE by principal operating condition.

Stratification	Category	N	MAE (L/h)	Mean ensemble SD (L/h)
Load state	Empty	40,885	28.20	1.91
Load state	Loaded	60,311	26.74	1.81
Grade	Downhill	3,680	32.31	3.22
Grade	Flat	77,924	28.24	1.73
Grade	Uphill	9,223	29.86	2.50
Grade	Steep uphill	10,369	16.46	1.71
Payload	Low / medium / high	35,541 / 33,717 / 31,938	28.17 / 27.42 / 26.30	1.92 / 1.75 / 1.88
Speed	5–15 / 15–25 / 25–35 / >35 km/h	16,968 / 34,437 / 35,071 / 14,720	26.42 / 27.05 / 26.72 / 30.47	2.27 / 1.71 / 1.62 / 2.24

The largest aggregate errors occurred downhill and above 35 km/h. Downhill operation involves coasting, retarder use, gear and brake choices, and engine-management effects that the 17 inputs only partially represent. Policy findings in these regimes were therefore checked against terrain-specific safety, predicted-error, and out-of-distribution (OOD) evidence.

Loaded observations did not exhibit worse aggregate error than empty observations, and high-payload groups had slightly lower MAE than low-payload groups. This does not imply that payload is unimportant; it indicates that including payload as a predictor maintained broadly comparable error across empirical payload strata.

Sparse joint cells were retained in the supplementary output but were not used for strong condition-specific claims. Sample count was treated as part of evidential strength, particularly for unusual loaded-low-payload or rare terrain-speed combinations.

3.3. Transition-Model Performance and Environment Fidelity

One-step speed predictions were strong, whereas next-engine-speed predictions were only moderate. The closeness of validation and test errors suggests temporal stability for one-step prediction, not long-horizon physical fidelity.

Table 15. Reports the one-step vehicle-state transition performance.

Predicted state	Split	N	MAE	RMSE	R^2
Next speed	Validation	89,948	1.67 km/h	2.39 km/h	0.9178
Next speed	Test	92,579	1.67 km/h	2.41 km/h	0.9230
Next engine speed	Validation	89,948	99.02 rpm	142.28 rpm	0.4967
Next engine speed	Test	92,579	98.87 rpm	143.42 rpm	0.4822

Speed drives simulated distance, completion time, and reconstructed throughput, so its high one-step R^2 supports matched fixed-distance comparisons. Recursive error can still accumulate. The next-engine-speed model explained 48.22% of test variation; omitted torque, shift, converter, and control dynamics probably account for part of the residual. Since RPM is also a fuel-surrogate input, this error can propagate into reward and policy outcomes.

Accordingly, the environment offers strong short-horizon data fidelity for speed and moderate fidelity for engine speed, but limited explicit physics. Tire–road friction, changing rolling resistance, torque maps, brake and retarder temperature, detailed transmission behavior, weather, traffic, and degradation are not modeled.

3.4. Reward-Profile Sensitivity and Selection

Changing reward coefficients affected fuel, throughput retention, envelope exceedance,

saturation, and OOD exposure. The productivity-dominant profile achieved the highest mean validation score across PPO and SAC and was fixed before final training.

SAC under the safety/support-constrained profile produced the lowest screening fuel value but also 115.98% productivity retention, 9.20% envelope exceedance, and 65.57% Mahalanobis OOD. Balanced PPO had very low exceedance but only 96.18% productivity retention. Thus, larger stated penalties did not reliably yield conservative behavior. This one-seed, 20,000-step screen selected a reward profile; it was not used for inference.

Every screened policy entered substantial OOD regions. The soft support penalty was therefore insufficient to confine all queries to historical support, and final evaluation reports those departures rather than deleting them.

3.5. Multi-Seed Training Stability and Best-Checkpoint Selection

Each algorithm was trained for 100,000 timesteps under seeds 11, 22, and 33. The best validation checkpoint was not consistently the terminal model, demonstrating that equal budget does not imply monotonic convergence.

The mean best-validation score was 7.6936±0.2248 for PPO and 6.5818±1.0592 for SAC. Under this configuration, SAC was more seed-sensitive; the result does not establish an intrinsic stability ranking beyond the present experiment.

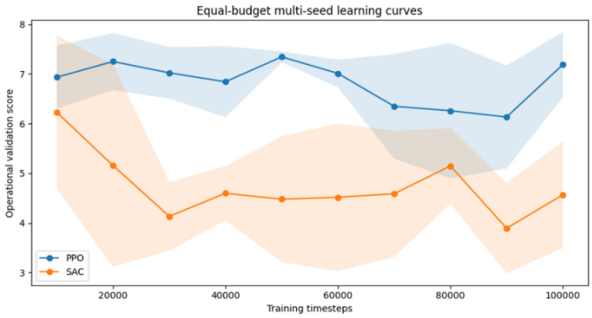

Figure 3. Multi-seed validation learning curves for PPO and SAC under equal 100,000-timestep training budgets. Shaded bands show ±1 standard deviation across seeds.

Table 16. Summarizes the validation-based reward-profile screening.

Profile	Mean score	Worst-algorithm score	Fuel (L/km)	Productivity retention (%)	Safe-envelope exceedance (%)
Productivity dominant	6.4189	5.5985	4.1373	105.90	4.46
Fuel dominant	6.1933	5.6192	4.1499	102.37	4.06
Balanced	6.1130	4.9666	4.2756	96.13	2.76
Safety/support constrained	5.0941	3.0348	4.6482	118.11	11.58

Table 17. Lists the best-validation checkpoints and the corresponding training time.

Algorithm	Seed	Best checkpoint	Best score	Training time
PPO	11	100,000	7.7744	41.90 min
PPO	22	50,000	7.4396	45.81 min
PPO	33	20,000	7.8668	44.41 min
SAC	11	20,000	7.4492	59.65 min
SAC	22	10,000	6.8949	64.47 min
SAC	33	100,000	5.4013	64.91 min

Table 18. Presents the seed-specific complete-cycle outcomes after common-estimator rescoring.

Algorithm	Seed	Throughput (t/h)	L/t	Fuel saving (%)	Loaded envelope (%)	Full throttle (%)	kNN OOD (%)
PPO	11	224.88	0.2892	8.09	8.14	0.07	36.76
PPO	22	224.71	0.2791	10.01	15.37	0.00	36.07
PPO	33	238.00	0.2879	8.26	14.39	0.00	30.13
SAC	11	260.84	0.2493	18.70	26.92	6.57	38.50
SAC	22	243.79	0.2767	10.69	19.75	0.53	35.26
SAC	33	243.99	0.3200	−0.79	15.45	32.92	49.45

PPO remained within 0.2791–0.2892 L/t and almost never used full throttle. SAC ranged from 0.2493 to 0.3200 L/t; seed 33 delivered no mean fuel saving and reached 32.92% full throttle. The average SAC outcome therefore masks material variation among its seeds.

The correct inference is configuration-specific: PPO was more stable under the implemented

environment, reward, hyperparameters, and seeds. Equal interaction budgets improved fairness but did not prove convergence or universal algorithm ranking.

3.6. Full-Cycle Productivity and Fuel-Efficiency Results

All simulated controllers completed 100% of the eligible loaded- and empty-travel routes. Complete haul cycles were reconstructed by combining the simulated travel components with the same cycle-specific historical loading, dumping, queueing, waiting, inactive, and other non-driving residuals. Consequently, differences in reconstructed payload throughput resulted only from changes in loaded- and empty-route completion times. Table 19 presents the final common-estimator results.

Note: Confidence intervals are hierarchical 95% bootstrap intervals. PPO and SAC results summarize 26 cycles and three independently trained seeds. Policy- and rule-based fuel values were calculated using the common three-member ensemble-mean estimator. Historical Human–Actual fuel was integrated from measured telemetry. Fuel-saving percentages are relative to Historical Human–Surrogate, whereas throughput changes are relative to Historical Human–Actual.

Historical Human–Actual and Historical Human–Surrogate had identical throughput because both retained the observed cycle and travel times. Their mean fuel intensities were 0.3083 and 0.3163 L/t, respectively. The mean cycle-wise relative discrepancy was 1.83%, indicating that the surrogate reproduced the aggregate scale of historical-human fuel consumption reasonably closely while retaining the instantaneous prediction limitations reported in Section 3.2.

The rule-based benchmark averaged 0.3208 L/t, 1.51% worse than Historical Human–Surrogate, and 227.63 t/h, 5.43% below Historical Human–Actual. Conservative throttle and target-speed rules alone therefore did not improve full-cycle efficiency.

PPO reduced mean fuel intensity to 0.2854 L/t, representing an 8.79% reduction relative to Historical Human–Surrogate. Its transport-normalized fuel intensity was 0.1116 L/(t·km). Mean throughput decreased to 229.20 t/h, corresponding to a 4.91% reduction relative to observed historical operation.

SAC produced a mean fuel intensity of 0.2820 L/t and a transport-normalized value of 0.1104 L/(t·km), corresponding to a 9.53% fuel-intensity reduction. Its mean throughput increased to 249.54 t/h, or 3.16% above Historical Human–Actual. Figure 4 visualizes the complete-cycle productivity–fuel-efficiency operating points of the historical-human, rule-based, PPO, and SAC controllers, together with their hierarchical 95% confidence intervals.

PPO and SAC therefore occupied different productivity–efficiency operating regions. Their mean L/t difference was only 0.0034 L/t, whereas SAC produced approximately 20.34 t/h greater mean throughput than PPO. The formal paired tests in Section 3.7 show that their fuel-intensity difference was not statistically significant, while their throughput difference was statistically supported. The higher SAC throughput must nevertheless be interpreted jointly with its greater safety-envelope exceedance, control saturation, seed variability, and OOD exposure reported in Sections 3.8 and 3.9.

Post-hoc common-estimator rescoring generated only minor numerical changes. PPO fuel saving changed from 8.68% to 8.79%, SAC changed from 9.55% to 9.53%, and Rule-Based Eco-Driving changed from −1.16% to −1.51%. Policy actions, state trajectories, completion times, throughput, safety indicators, and OOD diagnostics were unchanged. The estimator-consistency audit therefore did not alter the principal interpretation of the controller comparison.

3.7. Paired Statistical Comparison and Baseline Evaluation

Paired Wilcoxon tests used the 26 common cycles after averaging PPO and SAC outcomes across their three seeds. Table 20 reports the principal comparisons. P-values are unadjusted and are interpreted with effect size, confidence intervals, and operational diagnostics.

Table 19. Complete-cycle controller outcomes across 26 independent test cycles.

Controller	Throughput (t/h), mean±SD	Throughput 95% CI	L/t, mean±SD	L/t 95% CI	L/(t·km)	Fuel saving (%)	Throughput change (%)
Human—Actual	244.73±89.51	212.64–279.97	0.3083±0.0946	0.2734–0.3448	0.1222	1.83	0.00
Human—Surrogate	244.73±89.51	212.50–280.04	0.3163±0.1035	0.2768–0.3555	0.1240	0.00	0.00
Rule-based	227.63±75.53	199.93–256.52	0.3208±0.1232	0.2761–0.3686	0.1240	−1.51	−5.43
PPO	229.20±79.24	199.25–260.22	0.2854±0.1046	0.2484–0.3260	0.1116	8.79	−4.91
SAC	249.54±86.81	217.16–283.69	0.2820±0.1113	0.2371–0.3386	0.1104	9.53	+3.16

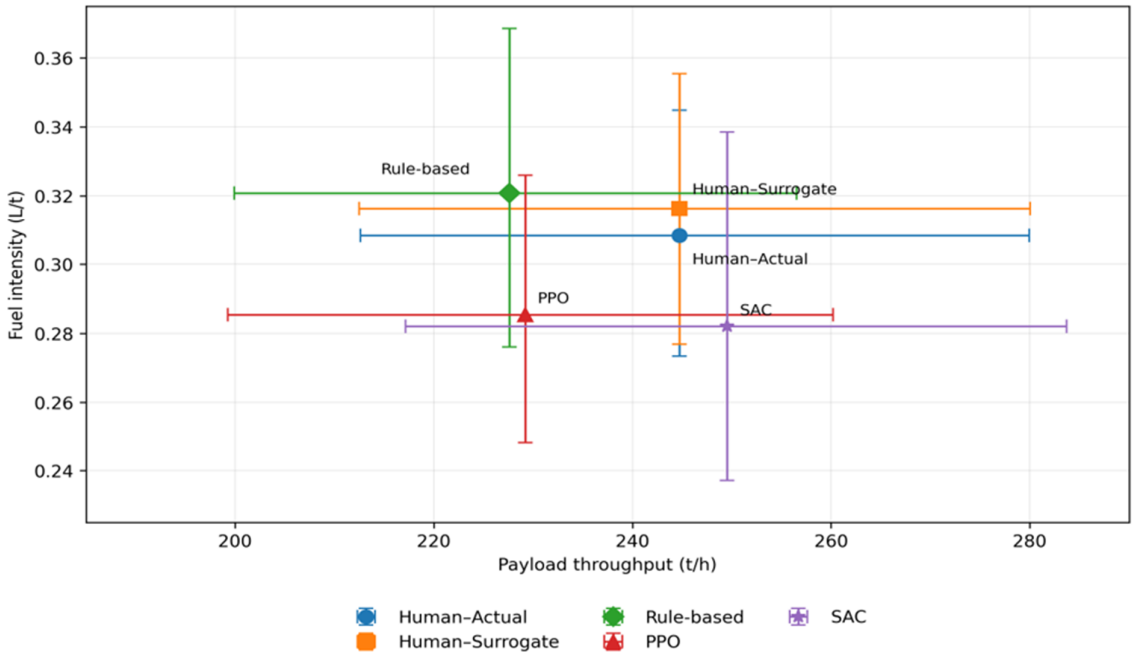

Figure 4. Full-cycle payload-throughput and fuel-intensity operating points across 26 held-out complete test cycles. Horizontal and vertical error bars represent hierarchical 95% confidence intervals for payload throughput and fuel intensity, respectively. PPO and SAC results aggregate three independently trained seeds. Rule-based and policy fuel intensities are common-estimator surrogate predictions, whereas Historical Human–Actual fuel is derived from measured telemetry.

Table 20. Principal paired controller comparisons.

Comparison	Throughput difference; p	L/t difference; p	L/(t·km) difference; p	Loaded envelope difference; p	Full-throttle difference; p
PPO − surrogate human	−15.53; 0.0525	−0.03094; 0.0029	−0.01246; 0.0094	+6.96; 0.1070	−31.74; <0.001
SAC − surrogate human	+4.81; 0.6710	−0.03432; 0.0032	−0.01366; 0.0176	+15.04; 0.0039	−18.42; <0.001
Rule-based − surrogate human	−17.09; 0.0151	+0.00446; 0.7835	−0.00004; 0.8417	−0.76; 0.6012	−31.76; <0.001
PPO − rule-based	+1.56; 0.8809	−0.03540; <0.001	−0.01242; <0.001	+7.72; 0.0045	+0.02; 0.1088
SAC − rule-based	+21.91; <0.001	−0.03878; <0.001	−0.01363; <0.001	+15.80; <0.001	+13.34; <0.001
PPO − SAC	−20.34; <0.001	+0.00338; 0.7835	+0.00120; 0.5651	−8.08; <0.001	−13.32; <0.001

PPO significantly reduced both fuel-intensity metrics relative to surrogate human. Its throughput decrease narrowly missed α=0.05, and the loaded-envelope increase was not statistically established.

SAC also significantly reduced both fuel metrics; its apparent throughput increase was not significant, whereas loaded-route envelope exceedance increased significantly. PPO and SAC did not differ significantly in either fuel-intensity measure. SAC produced significantly higher throughput than PPO but also significantly greater loaded-envelope exceedance and full-throttle behavior.

Because these were planned but unadjusted comparisons, p-values were interpreted with effect size, hierarchical intervals, seed dispersion, and operational diagnostics. They do not establish universal algorithm superiority.

3.8. Safety, Control Feasibility, and Operating Conditions

Historical operation showed a 31.76% full-throttle rate. Rule-based control eliminated full throttle and reduced empirical envelope exceedance but did not improve fuel efficiency or throughput. PPO nearly eliminated full throttle and produced no material nominal overspeed above 50 km/h, yet its contextual loaded-envelope exceedance increased to 12.63%. SAC produced the highest loaded and empty exceedance and retained 13.34% full-throttle behavior.

Table 21. Reports the aggregate safety, control, uncertainty, and OOD indicators.

Controller	Loaded envelope (%)	Empty envelope (%)	Full throttle (%)	Predicted error (L/h)	Mahalanobis OOD (%)	kNN OOD (%)
Historical human	5.67	6.27	31.76	27.19 (surrogate replay)	–	–
Rule-based	4.91	1.09	0.00	36.77	60.85	40.06
PPO	12.63	8.36	0.02	35.47	55.27	34.32
SAC	20.71	14.67	13.34	35.99	58.10	41.07

During loaded downhill operation, no controller exceeded 50 km/h, but contextual envelope exceedance was 0% for rule-based, 2.02% for PPO, and 5.94% for SAC. Mean loaded-downhill speeds were approximately 18.43, 18.22, and 19.41 km/h, respectively. Rule-based and SAC used meaningful retarder input, whereas PPO used almost none. The absence of nominal overspeed therefore does not resolve physical downhill-control concerns, because stopping distance, road condition, braking capacity, and retarder thermal limits were not modeled explicitly [3, 4].

The largest contextual exceedance occurred during loaded steep-uphill operation. This does not necessarily indicate a formal physical safety violation; it denotes departure from the historical 95th-percentile envelope and may signal unusual speed–payload–grade combinations, drivetrain demand, or insufficient support.

3.9. Surrogate Uncertainty, OOD Exposure, and Exploitation Risk

Predicted absolute fuel error for each counterfactual controller was approximately 30–35% above the historical surrogate replay. PPO exhibited the lowest OOD exposure among simulated controllers, yet 55.27% of its states crossed the Mahalanobis threshold and 34.32% crossed the kNN threshold. Corresponding rates were 58.10% and 41.07% for SAC and 60.85% and 40.06% for rule-based control.

For PPO and SAC, deviations centered on speed, payload, accelerator input, control intensity, and transmission-output speed. Individual standardized deviations were generally moderate; many flags arose because the joint feature combination was unusual. Rule-based control also entered OOD regions, so distribution shift was not specific to RL.

An OOD flag does not prove that a state is physically impossible: thresholds describe departure from historical human operation. However, it reduces confidence in the surrogate's favorable response. Because the reward used a soft support penalty, reward gains could outweigh it. The high OOD rates and 80.66 L/h conformal radius are therefore substantive limitations rather than grounds for excluding inconvenient outcomes [12-14].

4. Discussion

4.1. Integrated Interpretation of the Main Findings

Both policies lowered surrogate-estimated full-cycle fuel intensity, but neither dominated all operational criteria. PPO combined a statistically supported reduction with lower saturation and OOD exposure, while its non-significant mean throughput change was negative. SAC also reduced fuel intensity and showed higher mean throughput, but that increase was non-significant and loaded-route envelope exceedance rose significantly. They

remain candidate strategies, not validated replacements.

Because 44.46% of telemetry occurred outside loaded and empty travel, faster controlled segments did not translate directly into equivalent hourly payload gains. Complete-cycle reconstruction therefore prevented travel speed from overstating productivity; SAC's higher reconstructed throughput remained modest and non-significant relative to historical operation.

PPO was more reproducible and less saturated across seeds, with lower OOD exposure but lower mean throughput and some contextual-envelope deviation. SAC was faster but more variable; one seed was favorable while another produced no fuel saving, substantial saturation, and the highest kNN OOD. These configuration-specific patterns are compatible with PPO's clipped updates and SAC's replay-based entropy optimization [30, 31], not universal algorithm properties.

Prior work established nonlinear mine-fuel prediction [1, 5, 6], sequential road-vehicle eco-driving [8, 21-24], and mining RL for dispatch or system optimization [25-29]. This study instead integrates complete-cycle reconstruction, a bounded telemetry-driven Digital Twin surrogate, data-driven transitions, matched multi-seed PPO-SAC evaluation, a non-learning baseline, production metrics, uncertainty, and OOD diagnostics. The contribution is vehicle-level offline policy screening—not a new RL algorithm or a synchronized bidirectional physical twin [15-19].

4.2. Limitations and Practical Implications

Evidence is limited to one mine, one HD785 family, 24 hours of telemetry, 26 complete test cycles, and three seeds per policy. Moderate fuel and RPM fidelity, a scalar action, omitted vehicle and operational dynamics, unadjusted p-values, and a non-certified empirical envelope constrain interpretation. Extensive OOD exposure makes favorable fuel estimates screening signals, not production forecasts. The framework should remain an advisory screening tool until estimator audit, higher-fidelity simulation, hard safety/support constraints, operator-in-the-loop and shadow-mode testing, controlled field trials, and monitored production assessment are completed.

5. Conclusions

This study integrates route-matched complete-cycle reconstruction with a bounded Digital Twin surrogate, equal-budget multi-seed PPO–SAC training, a transparent rule-based baseline, common-estimator rescoring, cycle-level inference, and uncertainty/OOD diagnostics. Joint assessment using payload throughput, L/t, and L/(t·km) prevented favorable travel-only fuel estimates from standing alone.

Both algorithms lowered surrogate-estimated fuel intensity, with no significant difference between PPO and SAC. PPO was less variable, nearly avoided full throttle, and had lower OOD exposure; however, its mean throughput change was negative and not statistically significant relative to historical operation. SAC showed a positive but likewise non-significant mean throughput change relative to historical operation, together with greater saturation, significantly higher loaded-route envelope exceedance than both the surrogate-human baseline and PPO, greater OOD exposure, and higher seed variability. The rule-based controller did not improve full-cycle fuel intensity.

These results identify no universal winner and remain counterfactual surrogate estimates, not field savings or safety evidence. Moderate model fidelity, distribution shift, omitted dynamics, and the limited dataset preclude deployment claims. The framework is an uncertainty-aware screening tool pending longer multisite telemetry, richer vehicle models, explicit safety and support constraints, comparisons with established optimization methods, operator-in-the-loop and shadow-mode testing, controlled field trials, and monitored production evaluation.

Conflict of Interest Declaration

The authors declare no financial or personal relationships that could have influenced the research presented in this article.

Data Availability Declaration

The operational telemetry analyzed in this study is not publicly available due to industry confidentiality. Anonymized versions can be requested from the corresponding author, pending approval from the data owner and compliance with applicable data-sharing restrictions.

Funding Declaration

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical Declaration

This study used anonymous operational telemetry and did not involve direct interaction with human participants, clinical data, or animal experiments. No personally identifiable information about the operators was included in the analysis. In accordance with institutional requirements, formal ethics approval was not required.

Acknowledgments

The authors would like to thank Universitas Amikom Yogyakarta and the Doctoral Program of Information Systems, Universitas Diponegoro, for providing the academic environment and institutional support necessary to complete this research.

Declaration on Generative AI in Scientific Writing

During manuscript preparation, AI tools were used for English editing, sentence restructuring, and clarity improvement. These tools were not used to generate or modify the research data, experimental procedures, or results. The authors reviewed and verified the final text and take full responsibility for its accuracy and integrity.

References

[1]. Wang, Q., Zhang, R., Lv, S., & Wang, Y. (2021). Open-pit mine truck fuel consumption pattern and application based on multi-dimensional features and XGBoost. *Sustainable Energy Technologies and Assessments, 43*, 100977.

[2]. Ibañez Noriega, I., Sagastume Gutiérrez, A., & Cabello Eras, J. J. (2024). Energy and exergy assessment of heavy-duty mining trucks: Discussion of saving opportunities. *Heliyon, 10*(3), e25358.

[3]. Alegre, D., de Lemos Peroni, R., da Rosa Aquino, E., & Dille, F. (2021). A method to assess haul roads rolling resistance using dispatch system data. *Mining Technology, 130*(3), 176-187.

[4]. Bodziony, P., Krysa, Z., & Patyk, M. (2025). Operational environment effects on energy consumption and reliability in mine truck haulage. *Energies, 18*(12), 3022.

[5]. Alamdari, S., Hossein Basiri, M., Mousavi, A., & Soofastaei, A. (2022). Application of machine learning techniques to predict haul truck fuel consumption in open-pit mines. *Journal of Mining and Environment, 13*(1), 69-85.

[6]. Tadubana, G., Sigweni, B., & Ayoung, D. A. (2024). Systematic review on prediction of haulage truck fuel consumption in open-pit mines. *International Journal of Mining, Reclamation and Environment, 38*(10), 833-850.

[7]. Huang, Y., Ng, E. C. Y., Zhou, J. L., Surawski, N. C., Chan, E. F. C., & Hong, G. (2018). Eco-driving technology for sustainable road transport: A review. *Renewable and Sustainable Energy Reviews, 93*, 596-609.

[8]. Li, J., Wu, X., Xu, M., & Liu, Y. (2022). Deep reinforcement learning and reward shaping based eco-driving control for automated HEVs among signalized intersections. *Energy, 251*, 123924.

[9]. Vahidi, A., & Sciarretta, A. (2018). Energy saving potentials of connected and automated vehicles. *Transportation Research Part C: Emerging Technologies, 95*, 822-843.

[10]. Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. *Machine Learning, 110*(9), 2419-2468.

[11]. Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine, 34*(6), 26-38.

[12]. Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., & Schoellig, A. P. (2022). Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, and Autonomous Systems, 5*, 411-444.

[13]. Zhao, W., He, T., Chen, R., Wei, T., & Liu, C. (2023). State-wise safe reinforcement learning: A survey. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (pp. 6814-6822).

[14]. Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 11216-11235.

[15]. Hazrathosseini, A., & Moradi Afrapoli, A. (2023). The advent of digital twins in surface mining: Its time has finally arrived. *Resources Policy, 80*, 103155.

[16]. Nobahar, P., Xu, C., Dowd, P., & Shirani Faradonbeh, R. (2024). Exploring digital twin systems in mining operations: A review. *Green and Smart Mining Engineering, 1*(4), 474-492.

[17]. Jones, D., Snider, C., Nassehi, A., Yon, J., & Hicks, B. (2020). Characterising the Digital Twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology, 29*, 36-52.

[18]. Kritzinger, W., Karner, M., Traar, G., Henjes, J., & Sihn, W. (2018). Digital Twin in manufacturing: A

categorical literature review and classification. *IFAC-PapersOnLine, 51*(11), 1016-1022.

[19]. Schena, L., Marques, P. A., Poletti, R., Ahizi, S., Van den Berghe, J., & Mendez, M. A. (2024). Reinforcement twinning: From digital twins to model-based reinforcement learning. *Journal of Computational Science, 82*, 102421.

[20]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).

[21]. Li, J., Fotouhi, A., Pan, W., Liu, Y., Zhang, Y., & Chen, Z. (2023). Deep reinforcement learning-based eco-driving control for connected electric vehicles at signalized intersections considering traffic uncertainties. *Energy, 279*, 128139.

[22]. Jiang, X., Zhang, J., & Li, D. (2023). Eco-driving at signalized intersections: A parameterized reinforcement learning approach. *Transportmetrica B: Transport Dynamics, 11*(1), 1406-1431.

[23]. Wegener, M., Koch, L., Eisenbarth, M., & Andert, J. (2021). Automated eco-driving in urban scenarios using deep reinforcement learning. *Transportation Research Part C: Emerging Technologies, 126*, 102967.

[24]. Zhu, Z., Gupta, S., Gupta, A., & Canova, M. (2024). A deep reinforcement learning framework for eco-driving in connected and automated hybrid electric vehicles. *IEEE Transactions on Vehicular Technology, 73*(2), 1713-1725.

[25]. Huo, D., Sari, Y. A., Kealey, R., & Zhang, Q. (2023). Reinforcement learning-based fleet dispatching for greenhouse gas emission reduction in open-pit mining operations. *Resources, Conservation and Recycling, 188*, 106664.

[26]. Wang, Y., Bian, Y., Hu, M., Tang, Q., & Yang, Z. (2025). Deep reinforcement learning-based energy management strategy for optimizing fuel efficiency and battery life in hybrid mining dump trucks: A comparative study. *Journal of Power Sources, 656*, 238077.

[27]. Chaowasakoo, P., Seppälä, H., Koivo, H., & Zhou, Q. (2017). Digitalization of mine operations: Scenarios to benefit in real-time truck dispatching. *International Journal of Mining Science and Technology, 27*(2), 229-236.

[28]. Chaowasakoo, P., Seppälä, H., Koivo, H., & Zhou, Q. (2017). Improving fleet management in mines: The benefit of heterogeneous match factor. *European Journal of Operational Research, 261*(3), 1052-1065.

[29]. Noriega, R., Pourrahimian, Y., & Askari-Nasab, H. (2025). Deep reinforcement learning based real-time open-pit mining truck dispatching system. *Computers & Operations Research, 173*, 106815.

[30]. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[31]. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research* (Vol. 80, pp. 1861-1870).

[32]. Komatsu. (n.d.). *Komatsu HD785-7—Komatsu official ID*. Retrieved August 4, 2026, from https://www.komatsu.co.id/products/detail/hd785-7-202408151727

[33]. Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140.

[34]. Angelopoulos, A. N., & Bates, S. (2023). Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning, 16*(4), 494-591.

[35]. Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis, 88*(2), 365-411.

[36]. González, C., Gotkowski, K., Fuchs, M., Bucher, A., Dadras, A., Fischbach, R., Kaltenborn, I., & Mukhopadhyay, A. (2022). Distance-based detection of out-of-distribution silent failures for COVID-19 lung lesion segmentation. *Medical Image Analysis, 82*, 102596.

[37]. Galesso, S., Argus, M., & Brox, T. (2023). Far away in the deep space: Dense nearest-neighbor-based out-of-distribution detection. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops* (pp. 4479-4489).

انجمن مهندسی معدن ایران
نشریه مهندسی معدن و محیط زیست
نشانی نشریه: www.jme.shahroodut.ac.ir
دانشگاه صنعتی شاهرود

کاهش معاوضه بهره وری و کارایی در کامیون های سنگین ماینینگ با استفاده از یادگیری تقویتی عمیق و مدل سازی جانشین دوقلوی دیجیتال

کوسناوی کوسناوی[۱,۲*]، موچامد آگونگ ویبوو[۳]، و ریدوان سانجایا[۴]

۱ .برنامه دکتری سیستم های اطلاعاتی، دانشگاه دیپونگورو، سمارنگ، اندونزی
۲ .انفورماتیک دانشگاه آمیکوم یوگیاکارتا، یوگیاکارتا، اندونزی
۳ .دانشکده تحصیلات تکمیلی، دانشگاه دیپونگورو، سمارانگ، اندونزی
۴. گروه سیستم اطلاعات، دانشگاه کاتولیک سوگیجاپراناتا، سمارنگ، اندونزی

اطلاعات مقاله

تاریخ ارسال: ۲۰۲۶/۰۷/۱۴
تاریخ داوری: ۲۰۲۶/۰۸/۱۱
تاریخ پذیرش: ۲۰۲۶/۰۹/۰۵

DOI: 10.22044/jme.2026.17993.3582

کلمات کلیدی

مدل جایگزینِ همزاد دیجیتال
یادگیری تقویتی عمیق
شدت مصرف سوخت
بهره‌وری چرخه حمل‌ونقل
کامیون معدنی (کامیون حمل مواد معدنی)

چکیده

حمل و نقل روباز باید مصرف سوخت را بدون به خطر انداختن جابجایی مواد یا ایمنی عملیاتی کاهش دهد. این مطالعه بررسی کرد که آیا یک جایگزین دوقلوی دیجیتال مبتنی بر تله‌متری و مختص وظیفه می‌تواند از سیاست‌های رانندگی سازگار با محیط زیست در سطح خودرو که شدت سوخت را بهبود می‌بخشد و در عین حال بهره‌وری چرخه حمل و نقل را حفظ می‌کند، پشتیبانی کند. این چارچوب، بازسازی کامل چرخه حمل و نقل، یک جایگزین نرخ سوخت XGBoost گروهی، مدل‌های انتقال حالت غیرخطی، بهینه‌سازی سیاست پروگزیمال چند دانه‌ای با بودجه برابر (PPO) و منتقد نرم بازیگر (SAC)، غربالگری حساسیت به پاداش، یک معیار مبتنی بر قانون و ارزیابی عدم قطعیت و آگاه از توزیع را ترکیب کرد. تله‌متری یک روزه از ۷۱ کامیون کوماتسو HD785-7 شامل ۱۹۲۷۸۶۷ رکورد بود و ۲۳۸۶ چرخه حمل و نقل معتبر را به همراه داشت. جایگزین سوخت به MAE آزمون زمانی ۲۷.۳۳ لیتر بر ساعت، RMSE 37.75 لیتر بر ساعت و R^2 0.6608 دست یافت. در ۲۶ چرخه آزمایش کامل و مستقل، امتیازدهی مجدد تخمین‌گر مشترک نشان داد که PPO شدت سوخت را نسبت به پایه جایگزین-انسان ۸.۷۹٪ کاهش داده است (۰.۲۸۵۴ در مقابل ۰.۳۱۶۳ لیتر بر تن؛ p=0.0029)، با کاهش میانگین توان عملیاتی ۴.۹۱٪ که از نظر آماری معنی‌دار نبود. SAC شدت سوخت را ۹.۵۳٪ کاهش داد (۰.۲۸۲۰ لیتر بر تن؛ p=0.0032) و میانگین توان عملیاتی را ۳.۱۶٪ افزایش داد، اگرچه تغییر توان عملیاتی معنی‌دار نبود و تجاوز از پوشش ایمنی مسیر بارگذاری شده به طور قابل توجهی افزایش یافت. PPO و SAC از نظر شدت سوخت تفاوت معنی‌داری نداشتند. حالت‌های ایجاد شده توسط سیاست نیز مواجهه قابل توجهی با خارج از توزیع نشان دادند. بنابراین، این چارچوب، به جای برتری جهانی الگوریتم، بده‌بستان‌های کارایی-بهره‌وری-ایمنی را شناسایی کرد و باید به عنوان یک ابزار غربالگری سیاست آفلاین در انتظار اعتبارسنجی میدانی کنترل‌شده و اپراتور در حلقه تفسیر شود.

نویسنده مسئول مکاتبات: *khusnawi@amikom.ac.id* (ک. کوسناوی)