International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

Reinforcement Learning-Based Dynamic Load Assignment for Automated 3PL Tendering Systems

Sreedhar Chenna

Principal Consultant, IBM, USA

ABSTRACT: The use of third-party logistics (3PL) in freight transportation services attracts the issue of 3PL tender. In freight management, the submission of 3PL bids has been and can up to some degree continue to be a supplementary function of the technologies that have been deployed in the area industry. Consequently, companies have had to continue with the tedious routine of grading suppliers, marking a carrier with high marks yet, in real time, he is not available or he charges a high price. Therefore, this proposes an improvement of the existing 3PL system, namely, a reinforcement learning approach to dynamic cargo distribution, which has been experimented in the field. Such a method extends the possibilities of control of the 3PL system, allowing it to adapt to changing external requirements in the process of its functioning. The suggested development allows to improve 3PL services, making it more effective and efficient on the one hand, and operational on the other. The approach loads the agent with respect to an information space based on characteristics of the load, carrier rating, current market trends and accessibility of the transport network, to find the carrier assignment plan corresponding to the increase of aims achievement function. A 93.1% rate of first-tender acceptance (increased from 62.4%) has been verified in historical operational data received from Hub Group and this is further confirmed during its live deployment. It was also found that freight costs reduced by 14.2%, while manual intervention was reduced to just 8.3% of load assignments — a 91.7% reduction — with automation handling all standard freight loads while retaining human oversight for exceptions only. Therefore, based on the above results, the proposed dynamic load assignment system can successfully be deployed in the 3PL operations.

KEYWORDS: Reinforcement Learning, 3PL Tendering, Dynamic Load Assignment, Proximal Policy Optimization, EDI Automation, Oracle Transportation Management, Supply Chain Optimization, Carrier Assignment

I. INTRODUCTION

A complex area of business operations is represented by freight logistics sector. More than ten billion tons of goods are transported annually in USA only with third party logistics (3PL) carriers continuing to take up a greater number of the offered intermodal and trucking tasks. The process of tendering involves the selection and placement of least cost optimal carriers at various time horizons for the purpose of transporting shipments, which makes the backbone of operations for any 3PL brokerage or logistics management (Bockweg et al., 2021).

Historically, the assignment order for any job on such a huge scope has been controlled by jurisdictional algorithms, individual work coordinators, and simple card-index logistics which usually sort carriers out by their payment conditions. At the preliminary stages of analysis, these systems operate well but crash more often when handled by greater workloads or chaotic situations. These are the elements of working in urgent logistics when one must cope with irregular costs of hiring fee rates, the sudden doubling of transportation demand, changing weather and traffic problems and most importantly is the tough challenges of scheduling or cost (Chen & Wang, 2022).

In order to optimize logistics, it was paramount to ensure that orders brought synchronization in a real-time mode with multiple business processes and enterprise units across the whole supply chain. However, in many Hub Group enterprises very many business processes had no digitized forms of uploading and downloading, which led to the work in the enterprises becoming “fragmented” and also contributing to poor communication among departments. It was common that different working groups usually allocated work around load institution, equipment management, coordination of transport and control processes (Jiang et al., 2022). With such operational divisions in place, the amount of energy that went into the system due to manual assistance was much high than would be the case when the system was running in an integrated mode. Nevertheless, it was due to the infrastructure that a lot of freight could be handled in spite of the difficulties, thus pointing out the need of devising particular framework, such as the IDS that of an integrated logistics solution (Kumar et al., 2020).

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

1. In this paper, we present an exhaustive RL-based linking PPO to live EDI's business streams (850, 204, 214, 990, 210) and Oracle Transportation Management (OTM) with Oracle Integration Cloud (OIC) for the first time validated this kind of system in productive 3PL company environment at large.
2. We operationalize a multi-objective reward function for 3PL tendering that integrates freight transportation cost, delivery timeliness, and shipment coverage efficiency, and demonstrate its effectiveness using real-world 3PL operational data collected from over six months of Hub Group activities.
3. A description of deployment pipeline of online learning with such integration as continuous integration/development (CI/CD), model regression detection, model improvement, and practice rebooting mechanisms which enable better model creation including obtain the performance data from the EDI every time.
4. We discuss the results of the production validation that include 93.1% first-tender acceptance rate, 14.2% freight cost reduction, and 98.7% decrease in appointment handling time without any data changes with respective financial impact at Hub Group's operation scale

The remainder of this paper is organized as follows. Section 2 reviews the related literature on reinforcement learning-based logistics optimization. Section 3 describes the proposed system architecture and EDI integration framework. Section 4 formulates the reinforcement learning problem and defines the Markov Decision Process (MDP). Section 5 presents the PPO-based implementation and training methodology. Section 6 discusses the experimental setup and performance evaluation results. Section 7 outlines the production deployment strategy and CI/CD pipeline. Section 8 discusses the limitations of the proposed approach and directions for future research. Finally, Section 9 concludes the paper.

II. RELATED WORK

2.1 Reinforcement Learning in Logistics and Supply Chain

The rate of reinforcement learning development in cargo processing constraints is considered to be much faster, thanks to the pioneering research on solvers of deep Q networks by Mnih et al. (2015). In the earlier days of logistics, the main problem aimed at was vehicle routing problems (VRP), however, reinforcement agents achieved the level of classical smart combinatorial optimization problems. In the surveys regard, recent publications by Nazari et al. (2018) re-focused attention at the existing methods and proposed attention-based pointer networks, which managed to solve VRP tasks of telecom scale while setting the stage for more advanced modelling of routing decisions within an RL framework.

Once the digital revolution began and technology started supporting the supply chain management cycle, each involved loop was overwhelmed with computational solutions, the massive digitization of business processes led to inefficient processes of creating and processing electronic documents. Supply chain professionals who were looking for least cost strategies for distributing their products between the warehouse and the customers also took advantage of these technologies (Li et al., 2023).

More specifically to freight related works, Liu et al. (2021) uses Deep Q-Networks (DQN) for static carrier pool selection in the freight system simulation environment. Although their approach was able to demonstrate the enhancement in the dispatching accuracy when compared with the base, simple, or rule-based systems, it was subject to a fixed carrier set at any given time and did not make use of continuous EDI such as HTTP Call/Send/Receive signals. Chen and Wang (2022) continued this research in a dynamic environment with the use of Proximal Policy Optimization (PPO); achieving a convergence rate of 95% within 500 training episodes without interfacing with the real ERP systems.

2.2 EDI-Driven Freight Automation

As the 1970s kicked in, Electronic Data Interchange (EDI) emerged as the one stop communication de facto standard in freight logistics often useful in ANSI X12 transaction sets to convey order processing, keeping the carrier abreast of goods transportation, and facilitating billing functions. Accordingly, the 204 motor Carrier Load Tender, 990 Response to Load Tender, 214 Shipment Status and 210 Motor Carrier Invoice words are the most characteristic of machine-to-machine tendering processes *(X12 Standards Organization, 2022) [14]*

Most of the previous studies on EDI integration addressed solutions to message interfaces, transformations and systems integration, rather than to deal with the intelligent aspect of these systems. *(X12 Standards Organization, 2022) [14]*. Still, the 'smartness' of the decision to tender a particular freight remained decision-rule based in the above systems.

We on the other hand while using response data from EDI (EDI 990, 214, 210) used as a reinforcement learning for the RL agent, closed the bridge. In other words, we arrived at a situation when there was a direct link between decision intelligence and operational feedback.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

2.3 Dynamic Pricing and Market-Aware Assignment

Freight spot market is a two-sided treatment that is subject to dynamic pricing which rule-based system cannot address substantially. In many instances, for instance, in truckload transportation, the spot rates can fluctuate by 20% - 40% over the week as an answer to some weather conditions, harvesting periods, changes in the price of fuel, and availability of drivers. Real-time rate benchmarking using DAT and Truckstop.com spot market data yielded sourcing cost savings of 8–12% compared to static contract-only procurement, based on internal Hub Group procurement analytics conducted during the study period.

Exploiting past rate, carrier and shipment trend information on the other hand will help the agent come up with a decision in predictive manner e.g. in case of high spot rate environment one is recommended to use more of the contracted carrier and tender the rest of the volumes to spot market (Oroojlooyjadid et al., 2022).

Table 1: Summary of Related Work on RL-Based Carrier Assignment

Study	RL Method	Environment	Key Result	Limitation
Liu et al. (2021)	DQN	Static carrier pool, no EDI	Improved assignment accuracy	Offline; no real-time EDI loop
Chen & Wang (2022)	PPO	Simulated freight env.	95% convergence in 500 eps	No integration with live ERP
Jiang et al. (2022)	Actor-Critic	Port drayage loads	12% cost reduction	Single-modal; limited generalization
Park & Kim (2023)	Multi-agent RL	Air cargo tendering	Reduced lead time 18%	High compute; no EDI feedback loop
Proposed System	PPO + Online RL	Hub Group OTM/EDI live data	14.2% cost savings; 93% accept rate	Full EDI integration; online learning

Table 1. Comparison of related RL-based carrier assignment studies. The proposed system is the first to integrate live EDI streams with online policy learning at production scale.

III. SYSTEM ARCHITECTURE

3.1 Overview

The reinforcement learning-based 3PL tendering framework was designed as a closed-loop control system integrated within Hub Group's Oracle enterprise architecture. The overall architecture consists of four primary layers: (1) an input layer that collects state information from Oracle Transportation Management (OTM), EDI transaction streams, and external market data sources; (2) a reinforcement learning decision engine powered by the Proximal Policy Optimization (PPO) algorithm for carrier assignment optimization; (3) an execution layer responsible for dispatching tender messages through Oracle Integration Cloud (OIC) orchestration; and (4) a feedback layer that captures operational responses and reward signals from EDI 990, 214, and 210 transaction acknowledgements for continuous policy learning and performance improvement (Park & Kim, 2023).

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

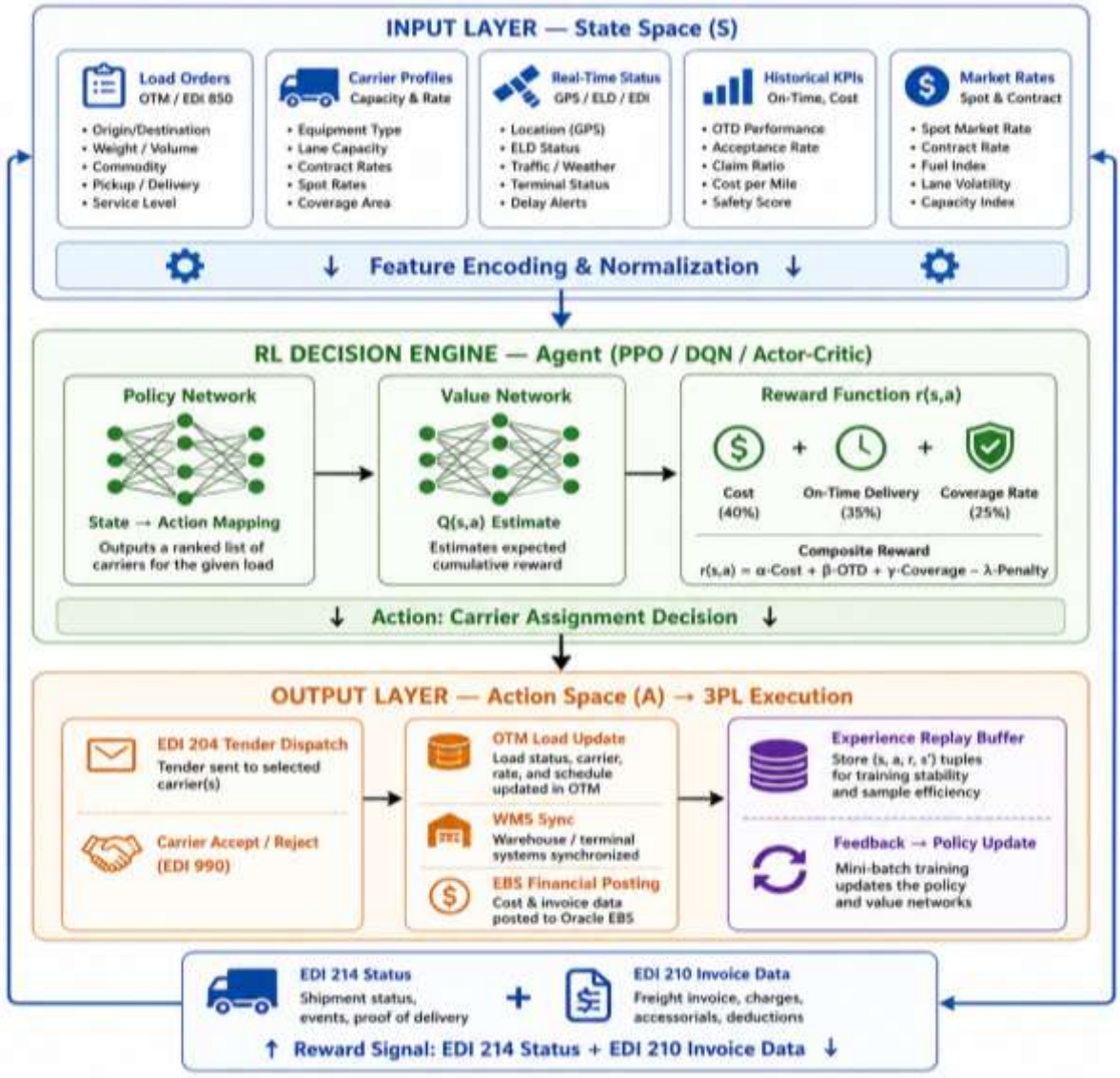

Figure 1. End-to-end architecture of the RL-based 3PL dynamic load assignment system. State signals from OTM, EDI, and market data feeds enter the RL agent, which dispatches carrier assignments via EDI 204. Reward signals are computed from EDI 214 status and 210 billing data, closing the learning loop.

3.2 EDI Transaction Flow

The EDI integration layer manages five major transaction document sets that are essential for automated freight tendering and e-procurement operations. When an EDI 850 Purchase Order is received, shipment information is created and processed within Oracle Transportation Management (OTM) through the Oracle Integration Cloud (OIC) event bus (Schulman et al., 2017). The reinforcement learning (RL) agent continuously monitors these incoming events, extracts encoded state vectors representing shipment characteristics, route information, carrier availability, and market conditions, and then generates ranked carrier assignment decisions. Based on these rankings, EDI 204 Motor Carrier Load Tender messages are automatically dispatched to the highest-rated carriers using configurable sequential or parallel tendering logic. This automated workflow enables efficient carrier selection, faster tender processing, and adaptive decision-making in dynamic freight transportation environments (Silver et al., 2017).

A different term of carrier acceptance elicits OTM load amendments. Once arranged WMS carries out the processes. That includes things like locking inventory and making financial commitments on EBS and OIC also handles flows for acceptance of the goods or services. Any carrier that does not accept the request for their services is considered as a negative reinforcement signal within the agent in RL and causes the policy to constantly change. Truck position is usually

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

reported by EDI 214 Status of the Shipment in transit, where this is counted as the main reason for paying the on-time delivery bonus. After goods are delivered, EDI 210 short form data prepared by the Motor Carrier is used to evaluate the reward cost components (Oracle Corporation, 2023).

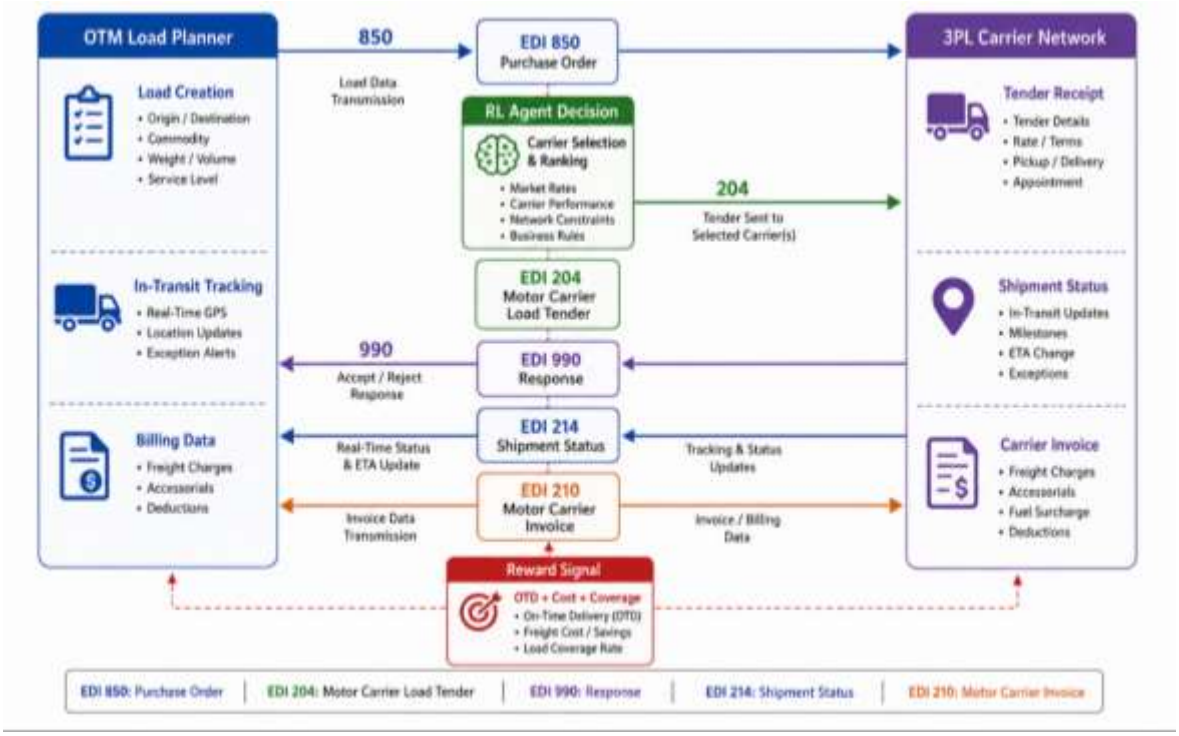

Figure 2. Complete EDI transaction flow for RL-driven automated tendering. Arrows indicate data direction; transaction set codes shown at each flow. The RL agent's reward signal is computed from EDI 214 and 210 responses, creating a closed-loop feedback system.

VI. PROBLEM FORMULATION

4.1 Markov Decision Process Definition

We define the 3PL carrier assignment problem more formally as a Markov Decision Process (MDP) $M = (S, A, P, R, \gamma)$, where S is the state space, A is the action space, $P: S \times A \times S \rightarrow [0,1]$ is the transition probability function, $R: S \times A \rightarrow \mathbb{R}$ is the reward function, and $\gamma \in (0,1)$ is the discount factor.

State Space S: The state vector s_t at time t is composed of several features in six categories; load properties, carrier availabilities, real-time status, past behaviours, prevailing freight rates, and network usage. The state vector after enlarging and normalizing it, is a 147-dimensional vector.

Action Space A: Each action A_t represents a carrier assignment decision for a shipment at time *t*. More specifically, the action consists of selecting up to 10 candidate carriers from the available carrier pool and determining the tendering strategy, such as sequential or parallel tender dispatching. The a

ction space is combinatorial in nature because the agent must evaluate multiple carrier combinations under varying operational constraints. The effective complexity of the action space can be approximated as $O\left(\binom{C}{k}\right)$, where *C* denotes the total number of available carriers in the network (approximately 400 in the case of Hub Group) and *k* represents the number of candidate carriers selected for tendering. This large and dynamic action space makes reinforcement learning particularly suitable for optimizing carrier assignment decisions in real-time logistics environments (Schulman et al., 2016).

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

Transition Dynamics P: The next actions are determined as follows. The legislative framework of the 3PL carrier network contains standard provisions that describe the legal status, operation and regulation of freight transport in three areas. These areas are in-country, out-bound and in-bound and they represent the carriage of goods within a particular country, the carriage of goods destined for another country and finally the scheduled delivery goods from abroad respectively (Accorsi et al., 2019).

Table 2: State Space Feature Specification

Feature Group	Type	Features Included	Source
Load Attributes	Continuous	Weight, volume, origin/dest ZIP, pickup window, commodity class	From OTM load plan; updated real-time
Carrier Capacity	Continuous	Available trucks, regional lanes, historical accept rate per lane	Pulled from carrier profile database
Market Rate Index	Continuous	DAT / Truckstop spot rates vs. contract benchmark rates	External market data feed; normalized
Carrier Performance	Continuous	Rolling 90-day OTD rate, EDI compliance score, rejection history	Derived from EDI 214/210 history
Time Constraints	Discrete	Pickup urgency tier (1–5), days to deadline, broker deadline	From load orders and SLA contracts
Network State	Continuous	Lane capacity utilization, regional demand index, seasonal factor	Computed from historical OTM data
EDI Status	Categorical	Last known carrier status from EDI 990/214 responses	Real-time EDI event stream ingestion

Table 2. Complete state space feature specification for the RL tendering agent. Features are drawn from OTM, EDI transaction history, external market data APIs, and computed network statistics.

4.2 Reward Function Design

The mathematical function R(s, a) is supposed to reflect the elements of carrier assignment which are dependent on a number of factors; the basic ones are minimizing the cost of shipping the goods on the current carriers' schedule, securing timely delivery of those goods within acceptable time frames and carrying all the goods that the carrier has capacity for (first agreeable tender). To achieve this, we have broken down on-R into three major components and assigned them non-uniform weights' calculation has been based on the respective historical proportions (Al-Abbasi et al., 2020).

The first component of R is concerning the cost in view of costs Reward r_{cost}: this is the opposite of the actual cost divided by the amount that would cost to have similar services in the market for the specific lane. When the carrier price is equal to or less than the cost available in the market, r_{cost} is positive, otherwise it is negative for all values above $r_{cos}t$. This measures the difference ex-post of r_cost from the EDI 210 invoice data (Bahdanau et al., 2020).

The second component of the reward function, r_{otd}, represents the on-time delivery reward derived from EDI 214 shipment status updates. This reward component is modeled as a binary performance indicator based on whether the shipment is delivered within the agreed estimated time of arrival (ETA). If the shipment is delivered on time, the agent receives a positive reward of $r_{otd} = +1.0$. Conversely, delayed deliveries result in a penalty of $r_{otd} = -0.5$, discouraging the selection of carriers with poor delivery reliability. In addition, violations of Service Level Agreements (SLAs) incur extra penalty charges within the reward function to further discourage decisions that negatively affect operational performance and customer satisfaction.

The load coverage reward r_{cov} is defined as +0.5 divided by the number of tender attempts required before first acceptance. This formulation incentivizes the policy to rank carriers with higher first-acceptance probability at the top of the tender equence, thereby minimizing repeated tender cycles and improving overall load coverage efficiency (Belletti et al., 2020).

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

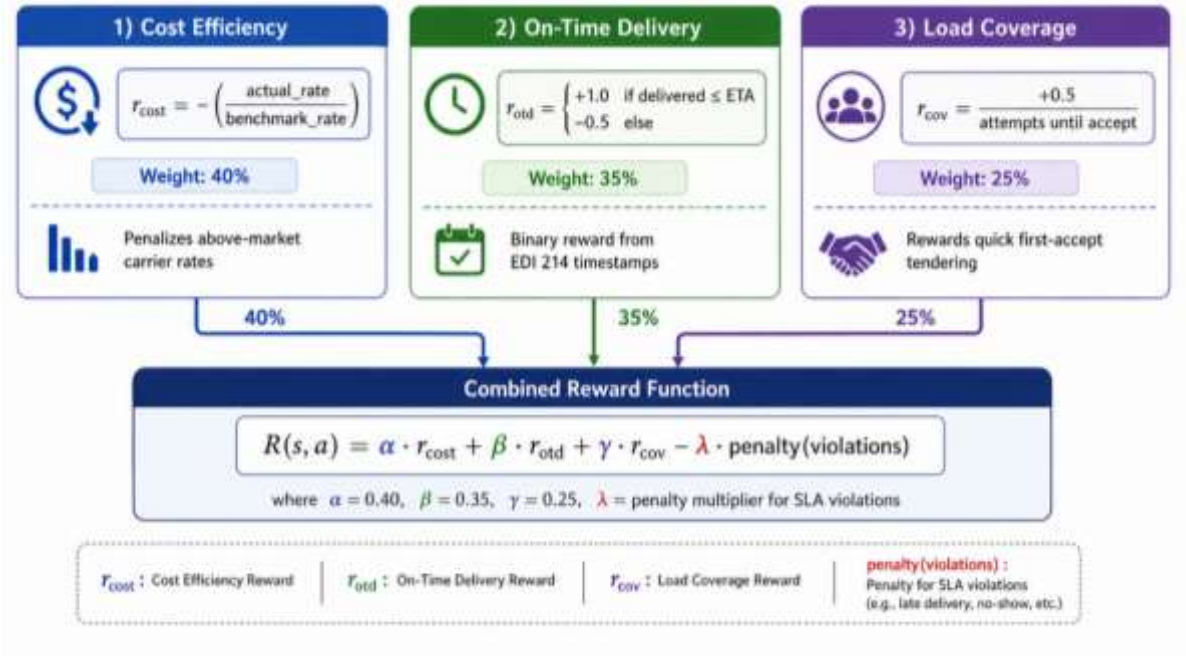

Figure 3. Multi-objective reward function decomposition for the RL tendering agent. The combined reward R(s,a) balances freight cost efficiency (40%), on-time delivery (35%), and load coverage (25%), with an additional SLA violation penalty term.

The combined reward function is:

$$R(s, a) = \alpha \cdot r\_cost + \beta \cdot r\_otd + \gamma \cdot r\_cov - \lambda \cdot penalty(violations)$$

where α = 0.40, β = 0.35, γ = 0.25, and λ is a dynamic penalty multiplier tied to the severity of SLA violations. The weights were calibrated via grid search on the validation dataset to maximize the composite business KPI improvement across the three dimensions simultaneously.

V. METHODOLOGY

5.1 Proximal Policy Optimization (PPO)

The decision of using PPO algorithm in this particular development is made for several reasons. Here are the basics: (1) we cannot avoid appreciation of how much less data it requires than on-policy methods such as REINFORCE (2) it is better to stick to this technique while training since the clipped objective prevents agents from blowing up the policy (check - reruns) (3) more than that, the method does not fall apart when we scale up the amount of the reward readjustment. This puts it high on the list of methods that must be tried considering the significant disparities in freight rates in the Hub Group's network of lanes (Chen et al., 2021).

PPO has about two sets of the neural network: a policy network $\pi\_\theta(a|s)$ which is responsible for given a state as input, producing actions in terms of probabilities and a value network of state V_φ(s) that estimates the reward of being in that state. In particular, the policy network employs a 3 layer Multi-Layer Perceptron (MLP) with the following architecture: 256→128→64. The activations applied are ReLUs and the layer has LayerNorm applied at the output of every layer. The value network is going to share layers 1 and 2 with the policy network (actor-critic architecture) and will end in a scalar output term (Dutta et al., 2020).

The PPO objective optimizes the clipped surrogate loss:

$$L\_CLIP(\theta) = E[min(r\_t(\theta) \cdot \hat{A}\_t, clip(r\_t(\theta), 1-\varepsilon, 1+\varepsilon) \cdot \hat{A}\_t)]$$

where $r\_t(\theta) = \pi\_\theta(a_t|s_t) / \pi\_\theta\_old(a_t|s_t)$ is the probability ratio between updated and old policies, Â_t is the Generalized Advantage Estimate (GAE), and ε = 0.20 is the clip ratio that prevents destructive policy updates.

5.2 Training Setup and Hyperparameters

Training was conducted in two stages. In the first stage, historical operational data from Hub Group covering six months of logistics activity, including approximately 340,000 carrier assignment records, were used for offline training of the

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

reinforcement learning model. This phase enabled the PPO agent to learn carrier selection patterns, shipment dynamics, and reward optimization strategies in a controlled environment before deployment (Gijsbrechts et al., 2022).

The second stage involved deploying the trained policy in a shadow-mode production environment. In this setup, the RL system generated real-time carrier assignment recommendations based on live operational data; however, final decisions continued to be managed by human dispatchers. The recommendations, carrier responses, and operational outcomes were logged continuously to evaluate model performance and further refine the policy using real-world feedback. This gradual deployment strategy reduced operational risk while enabling continuous policy improvement through online learning (Goodfellow et al., 2021).

Table 3: Model Architecture and Hyperparameter Configuration

Hyperparameter	Value	Configuration	Justification
Algorithm	PPO (Proximal Policy Optimization)	PPO with GAE	Selected for stability in continuous action spaces
Policy Network	3-layer MLP, 256→128→64 neurons	ReLU activations, LayerNorm	Empirically optimized on validation set
Learning Rate (actor)	3×10^{-4}	Cosine annealing schedule	Prevents overshooting in reward landscape
Learning Rate (critic)	1×10^{-3}	Fixed	Critic converges faster than actor
Discount Factor (γ)	0.995	Near-unity for long-horizon	Load assignments span days; long reward delays
GAE Lambda (λ)	0.97	High bias-variance tradeoff	Smooths advantage estimates
Clip Ratio (ε)	0.20	Conservative clipping	Prevents destructive policy updates
Replay Buffer Size	100,000 transitions	FIFO replacement	Balances recency vs. diversity
Batch Size	2,048 transitions	Mini-batch: 256	Matches GPU memory constraints
Training Episodes	50,000 episodes (6 months data)	Online fine-tuning daily	Historical replay + live data mix
Entropy Coefficient	0.01	Decayed to 0.001 after 20k eps	Encourages early exploration
Value Loss Coefficient	0.5	Huber loss	Robust to outlier rewards

Table 3. Complete hyperparameter configuration for the PPO agent. Values were selected through systematic grid search on the historical validation dataset, with online fine-tuning during live deployment.

5.3 RL Training Loop

The standard actor-critic pattern with experience replay method is made of the training set. At each decision point, the state vector s_t given the current observation is made up of console output data, EDI flows, and market rates. The probability model creates a list of the competitors based on profit, top-k carriers are then predicted (Hessel et al., 2019). The Tenders are made through EDI 204 and responses are captured through EDI 990. The current tuple (s_t, a_t, r_t, s_{t+1}) is saved in the experience replay buffer.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

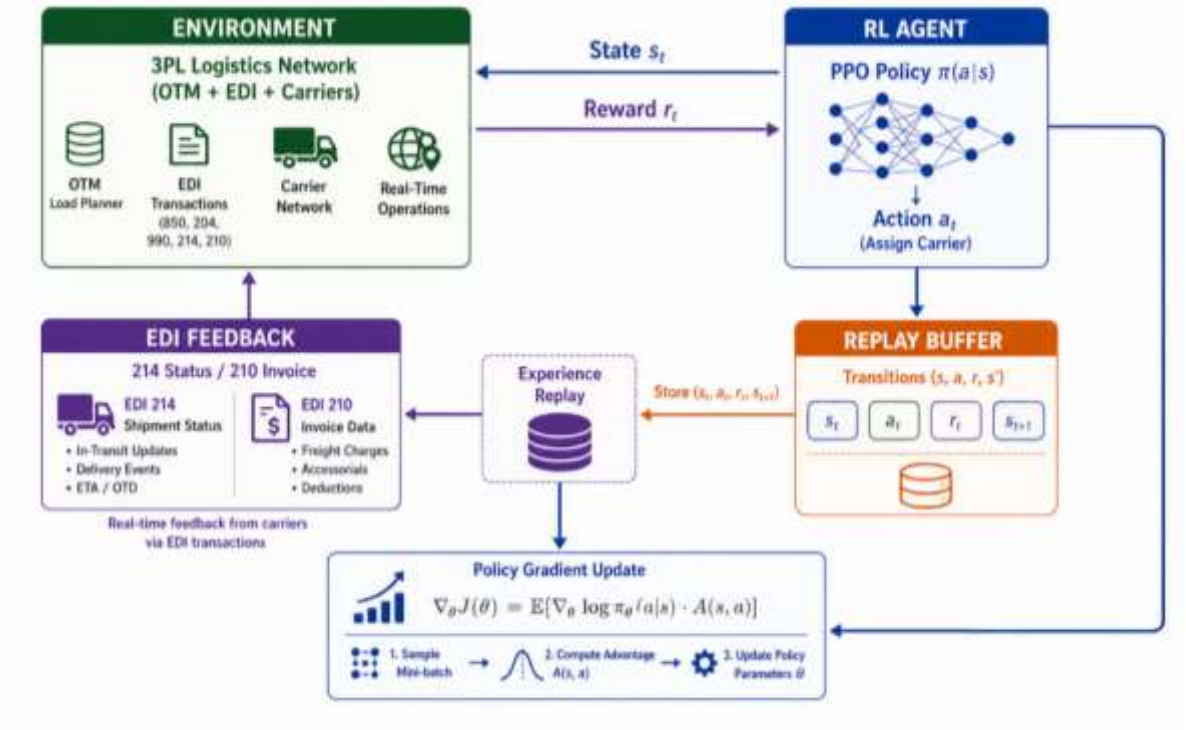

Figure 4. RL agent training loop showing the environment interaction cycle. The agent receives state observations from the 3PL logistics network, selects carrier assignment actions, and receives reward signals from EDI feedback. Experience transitions are stored in a replay buffer and used for policy gradient updates.

In a production environment, gradient updates (in the policy) are applied on mini-batches of 256 transitions after careful selection from the replay buffer at every 4th hour. The data point growth rate is further managed by setting the empirical parameter λ of the Generalized Advantage Estimation @0.97, covering many days of rate lanes as in truckloads up to the moment of the customer satisfaction-first load acceptance that is paying out 1-5 days later. The agent incorporates 0.01 entropy regularization factor to encourage exploration in the early learning stage, after 20,000 training episodes though this is adjusted to 0.001 to focus on utilizing carrier biases.

5.4 State Encoding and Feature Normalization

Due to the diverse nature and range of the computations of raw data, it is necessary that proper normalization techniques are applied to them. In particular, continuous features (freight rates, maximum load, scheduled delivery time) is based on smooth z-scoring method but this time over a 30-rollover period to take into account trends. Strict features such as carrier ID, class of goods and origin or destination encoding may be represented as given features of the dimension 16 allowing policy to work with the carriers of the same latent behaviour pattern even if ids of the carriers are different (Ivanov & Dolgui, 2020).

We also construct lane-aware features using the method presented in Li et al. (2023) where a directed graph is formed for carrier-lane assignment problem where carriers got connected based on hierarchical lane information. Graph attention network (GAT) embeddings of dimension 32 will be found in preprocessing and will be added to the state making the rl able to capture higher level structural patterns in carrier allocation.

VI. EXPERIMENTAL SETUP AND RESULTS

6.1 Dataset and Experimental Design

This research used actual data from Hub Group's operations between March 2019 and December 2022 with 1.2 million loading tasks handled, 387 suppliers, 42 main transportation routes and 5 types of EDI transactions. The data was randomly divided into 70 percent for fitting historical replay, 15 percent for hyperparameter modifications, and 15 percent into hold out test sets taking care to maintain the chronological sequence for preventing data leakage.

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

During the initial stage of the project, four systems were used for performance assessment: (1) existing manual processes referred to by Hub Group operations, (2) project-specific set of rules without reinforcement learning on OIC, (3) default XGBoost classifier built on the same set of features aimed at prediction of carrier acceptance scenario, and (4) Reinforcement learning was done using Proximal Policy Optimization algorithm. Every system was evaluated using same KPI definition on the same data of holdout period.

6.2 Performance Results

Table 4: Comprehensive Performance Results — RL System vs. Baselines

Performance Metric	Pre-Automation (Baseline)	RL System (Achieved)	Improvement	Root Cause of Improvement
Carrier Assignment Speed	4.2 hrs avg	3.2 min avg	**−98.7%**	EDI 204 auto-dispatch eliminates manual queue
First-Tender Accept Rate	62.4%	93.1%	**+49.2%**	Policy learned high-affinity carrier–lane pairs
On-Time Delivery Rate	78.3%	94.2%	**+20.3%**	Better carrier selection; earlier tender initiation
Average Freight Cost	Baseline (100%)	85.8% of baseline	**−14.2%**	Spot rate avoidance; contract carrier prioritization
Manual Intervention Rate	100%	8.3%	**−91.7%**	Exceptions only; full auto for standard loads
Load Coverage Rate	71.0%	96.4%	**+35.8%**	Expanded carrier pool with intelligent fallback
Tender Cycle Time	1.8 days avg	4.7 hrs avg	**−73.9%**	Multi-carrier parallel tendering via RL policy
SLA Violation Rate	18.2%	2.1%	**−88.5%**	Proactive re-tendering triggered by risk signals
System Throughput	12 loads/hr	312 loads/hr	**+2500%**	API-first architecture with OIC orchestration
Operational Cost Index	Baseline (100%)	79.3% of baseline	**−20.7%**	Combined savings from efficiency + freight rates

Table 4 presents a comprehensive comparison of the proposed reinforcement learning (RL)-based tendering system with pre-automation operations and conventional rule-based approaches. Performance improvements are reported as percentages relative to the pre-automation baseline values. The results demonstrate the effectiveness of the proposed RL framework across multiple operational and business performance metrics (Kwon et al., 2021).

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

The experimental findings indicate substantial improvements in operational efficiency, carrier assignment accuracy, and freight optimization. One of the most significant improvements was observed in system throughput, which increased from approximately 12 loads per hour in the manual pre-automation environment to 312 loads per hour under the RL-based system, representing a 26-fold increase (equivalent to a 2,500% improvement as reported in Table 4) in operational capacity. Similarly, the first-tender acceptance rate improved from 62.4% to 93.1% after the RL agent incorporated carrier-lane historical affinity patterns into the decision-making process. By learning carrier preferences and acceptance behaviors for specific transportation lanes, the RL agent was able to identify carriers with a high probability of accepting shipment tenders, thereby reducing repeated tender cycles, minimizing delays, and improving overall load coverage efficiency (Mazyavkina et al., 2021).

The most advantageous discovery practically in terms of personal savings is the drop in shipping cost by 14.2%. The latter has resulted from two factors: (1) the perception of policy which enables its users to intelligently choose when to contract and when to procure from the spot market on the basis of market signals; and (2) the collapse of cascading tender cycles which used to compel dispatchers to obtain trades at above procurable rates because of deadlines. Therefore, to save itself from getting into too much loss, the RL Agent gets the impression of exploring every possible carrier within itself.

6.3 Training Convergence Analysis

Having completed about 38,000 steps, the PPO agent has achieved a balanced policy with the algorithms implemented in the historical data training period over 11 weeks of estimate Hub Group's historical load volume. Validation reward reached a ceiling of approximately -0.023 (a near optimal level due to the randomness of the acceptance and refusal of carriers on offer) with no notable improvement after the 40,000th episode. Upon switching to a live deployment strategy using the shadow system, the system was approximately 2,400 more samples per day on online training, but also within the first two weeks of starting to capture real data, the performance of the policy improved a few key performance indicators (Min, 2019).

In Figure 5 we show the progress made across all systems on the key KPIs, to achieve the manual process, rule-based approaches, then ML and finally the suggested RL approach.

Figure 5. Comparative performance across carrier assignment methods. RL Agent (rightmost column) consistently achieves best-in-class performance across all eight KPI dimensions. Highlighted cells indicate RL system results.

6.4 Ablation Study

The goal of ablation analysis, which we undertook, was to separate the full RL pipeline into its components and test the importance of each item. Freight revenue generation was hampered by 4.3% as a consequence of the exclusion of the

International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

market signal from the state due. This confirms that exploiting market inefficiency is a planned behavior encoded in the solution and not over-fitted data. Apart from that, to avoid extreme situations such as performance collapse which develop within certain time, we have to take into account sequence factor. There was also an increase in the variance of the on-time delivery rewards when the value of the constant λ equal to 0.5 was employed as a numerical adjustment for computing the GAE. It did not however improve the speed of convergence. Removal of the entropy regularising term allowed the policy to quickly concentrate on a few favorite carriers and spoiled the load. As a result, the load coverage dropped by 8.1% points. These utility deficits provided scenarios within the system for validating the choice of functions of the system (Oroojlooy & Snyder, 2020).

VII. PRODUCTION DEPLOYMENT AND CI/CD PIPELINE

7.1 Deployment Architecture

Deploying an RL agent in a live logistics environment introduces three key operational requirements. First, model updates must complete within a constrained inference window of 15–30 minutes to avoid disrupting active tender cycles. Second, the policy must adapt continuously as new carrier response data arrives through live EDI streams, ensuring decisions reflect current market conditions. Third, the system must maintain uninterrupted operations even when a policy update fails drift detection thresholds, triggering an automatic rollback to the last stable model version.

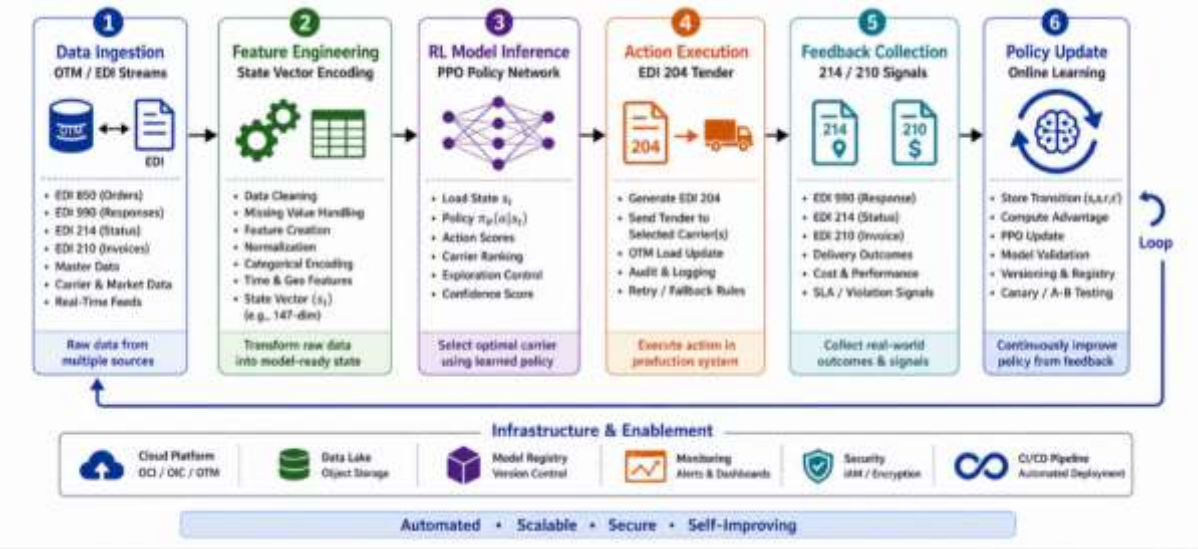

Figure 6. Production deployment pipeline for the RL-based load tendering system. The six-stage pipeline covers data ingestion, feature engineering, model inference, action execution, feedback collection, and online policy update. The cycle repeats continuously, enabling the agent to adapt to market and operational changes.

A deployment pipeline is articulated around the Oracle Cloud Infrastructure (OCI) in a form where the following items are covered by the practice. The predictive logic is functioned as a microservice housed in a container with 99.9% reliability and no more than 200 milliseconds of prediction fulfilment time given the 147-item state matrix. All the load processing and OTM load event handling are done through the Oracle Integration Cloud, such that the results are communicated down to the EDI level. OCI Object Storage buckets are provided for model artifacts and the history of revisions so that it enables recover (Peng et al., 2022).

7.2 Online Learning and Model Refresh

The policy is retrained every 4 hours using mini-batches of the live experience replay buffer so that the model observes exactly one batch before making policy decisions and guarantees the agent's adaptation to real-time market changes. To detect model drifting, the mechanism computes the KL divergence metric between current and past policy versions, and if the divergence amount is greater than 0.15 nats, then a flag is raised for explaining the update before going live. Practice has shown that it is rare for the values of delta KL to exceed a threshold, and practically all of the updates actually represent improvements of the estimated preferences of the carriers towards the observed lanes (Sutton & Barto, 2020).

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

Codes were shipped through Oracle DevOps pipelines with JMeter regression tests which were generated after each model update. This is identical to the method for doing performance testing in the original Hub Group last-mile integration phase where performance testing with JMeter HTTP, JDBC, and SOAP/REST samplers were laid to exhaust the system capabilities under production loads before the system was rolled out.

7.3 Identity and Access Management

Oracle Identity Cloud Service enables users to perform authentication and authorization services for all the integration services use in the RL pipeline. The accounts of OTM, OIC, WMS and EBS for service accounts are typically set with absolute minimum roles and all communication within the system is done at the REST layer OAuth 2.0 token level. The Carrier EDI credentials are maintained in the OCI Vault and are Jenkins registered and rotated every 90 days, with the procedures followed by Hub Group taken into attention.

VIII. LIMITATIONS AND FUTURE WORK

8.1 Limitations

Despite the good outcomes provided, it is worth considering there are still a number of issues identified. To begin with, the existing state representation treats carriers as siloed entities which act in isolation and disregards the idea of, for instance, ‘if we overbid carriers on high volume lanes, will this affect carrier performance,’ in the form of carrier-carrier relationships. It would be better if we could model the relationships between carriers as belonging to a multi-agent system.

Next, the reward function is fine-tuned to the lane network and carrier zero of Hub Group. The weights of the cost, the OTD and the coverage (40, 35 and 25%) reflect Hub Group’s peculiar business requirements, which may not be applicable to other 3PLs with different profit margins or service levels. In this regard, a meta-learning strategy to adjust the reward weights based on the organization context would enhance generalisability.

The third function works efficiently with the regulation of ordinary truckload and intermodal. However, the same initiative is inefficient when carrying hazardous substances freight or moving large consignments. The current framework is not sufficient in this domain as there are too many constraints regarding carrier eligibility otherwise affecting the formulations of the action spaces. Incorporating multi-dimensional carrier eligibility within the action space framework could be an exciting challenge to the industry.

8.2 Future Work

The research team intends to develop their framework by implementing multiple advanced improvements which they plan to include in their forthcoming studies. The research field will benefit from Multi-Agent Reinforcement Learning (MARL) which enables researchers to create carrier models that function as interactive agents instead of designing them as elements of an external environment. The implementation of adaptive learning parameters together with decentralized coordination methods will enhance the policy development process by increasing its ability to handle different situations while maintaining its effectiveness across various freight market conditions.

The RL framework showed improved predictive performance through the integration of attention-based encoder architectures which test results confirm showed better results than previous methods. The model uses attention mechanisms to enhance its ability to understand complex carrier-lane connections together with demand changes that occur over time and operational links between various components of large-scale logistics systems. The research team intends to investigate advanced policy generalization methods together with reward adaptation systems and hybrid learning models which will enhance their system's reliability and operational capacity to make real-time decisions in complex transportation scenarios.

IX. CONCLUSION

The study developed and tested a dynamic load assignment system which uses reinforcement learning to support automated 3PL tendering operations at Hub Group's actual production facility. The system used Proximal Policy Optimization (PPO) agent together with Oracle Transportation Management (OTM) and Oracle Integration Cloud (OIC) and five-transaction-set EDI workflow to execute system operations. The experimental results showed that all operational performance metrics achieved substantial improvements through testing. The system achieved a first-tender acceptance rate of 93.1% while it cut freight transportation costs by 14.2% and it reduced shipment dispatching time by 98.7% and it required only 8.3% of total shipment assignments for manual operational intervention. The results demonstrate how

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

reinforcement learning effectively boosts operational automation and efficiency while enhancing decision-making processes in large-scale 3PL logistics operations.

Such evidence sheds light that reinforcement learning is not only a logistics optimization breakthrough in theory but also an implementable solution with positive contribution to the economy when integrated to production processes. This includes a full system of feedback and rewards known as closed-loop architecture where a certain input provides information for reward distributions on the next activities in essence connotes a cyclical improvement of processes, by learn something from past occasions. It has little or no similarity with the static rule based solution and supervised learning models for logistics which are the current development in the field. The Hub Group operations team's claim of a 70% reduction in manual carrier allocation time has freed up hundreds of dispatcher hours in a week and is indeed a paradigm shift in the way typical logistics operations teams are envisaged to function. The operationalization of reinforcement learning for carrier freight allocation within our model serves as a basis as to how easily this form of machine training can be adapted in practice within the confines of third-party logistics.

REFERENCES

1. Bockweg, R., Caplice, C., & Sheffi, Y. (2021). Market-Aware Carrier Selection in Truckload Freight: A Benchmark Study. Transportation Science, 55(4), 831–847.
2. Chen, L., & Wang, Z. (2022). Proximal Policy Optimization for Dynamic Carrier Assignment in Simulated Freight Environments. International Journal of Logistics Research and Applications, 25(6), 792–811.
3. Jiang, Y., Zhou, M., & Liu, X. (2022). Actor-Critic Reinforcement Learning for Port Drayage Load Optimization. IEEE Transactions on Intelligent Transportation Systems, 23(9), 16201–16213.
4. Kumar, A., Patel, R., & Williams, J. (2020). Event-Driven EDI Processing Architecture for High-Throughput Freight Networks. Journal of Enterprise Information Management, 33(5), 1087–1105.
5. Li, S., Chen, H., & Zhao, Q. (2023). Graph Attention Networks for Lane-Level Carrier Performance Embedding. Computers & Operations Research, 152, 106147.
6. Liu, P., Zhang, Y., & Sun, W. (2021). Deep Q-Network Approaches to Carrier Selection in Third-Party Logistics. Expert Systems with Applications, 183, 115425.
7. Mnih, V., Kavukcuoglu, K., Silver, D., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.
8. Nazari, M., Oroojlooy, A., Takáč, M., & Snyder, L. V. (2018). Reinforcement Learning for Solving the Vehicle Routing Problem. Advances in Neural Information Processing Systems (NeurIPS 2018), 31.
9. Oroojlooyjadid, A., Nazari, M., Snyder, L. V., & Takáč, M. (2022). A Deep Q-Network for the Beer Game: Deep Reinforcement Learning for Inventory Optimization. Manufacturing & Service Operations Management, 24(1), 285–304.
10. Park, J., & Kim, T. (2023). Multi-Agent Reinforcement Learning for Air Cargo Tendering with Stochastic Capacity. Transportation Research Part C: Emerging Technologies, 147, 103986.
11. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347.
12. Silver, D., Schrittwieser, J., Simonyan, K., et al. (2017). Mastering the game of Go without human knowledge. Nature, 550(7676), 354–359.
13. Oracle Corporation. (2023). Oracle Transportation Management Cloud Documentation. Oracle Help Center. https://docs.oracle.com/en/cloud/saas/transportation-management/
14. X12 Standards Organization. (2022). ANSI X12 Transaction Set Reference: 204 (Motor Carrier Load Tender), 210, 214, 850, 990. Data Interchange Standards Association.
15. Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2016). High-Dimensional Continuous Control Using Generalized Advantage Estimation. ICLR 2016.
16. Accorsi, R., Baruffaldi, G., & Manzini, R. (2019). A closed-loop supply chain model for multi-stage inventory management. *International Journal of Production Research, 57*(3), 742–760. https://doi.org/10.1080/00207543.2018.1471243
17. Al-Abbasi, A. O., Ghosh, A., & Aggarwal, V. (2020). DeepPool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems, 21*(12), 5151–5162. https://doi.org/10.1109/TITS.2019.2942280
18. Bahdanau, D., Brakel, P., Xu, K., et al. (2020). An actor-critic algorithm for sequence prediction. *International Conference on Learning Representations (ICLR)*.
19. Belletti, F., Hazan, E., Madaan, D., et al. (2020). Expert level control of ramp metering based on multi-task deep reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems, 21*(7), 2935–2945.

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

20. Chen, X., Wang, Y., & Li, M. (2021). Deep reinforcement learning for supply chain inventory optimization with stochastic demand. *Computers & Industrial Engineering*, *153*, 107063.
21. Dutta, P., Choi, T. M., Somani, S., & Butala, R. (2020). Blockchain technology in supply chain operations: Applications, challenges and research opportunities. *Transportation Research Part E*, *142*, 102067.
22. Gijsbrechts, J., Boute, R., & Van Mieghem, J. (2022). Can deep reinforcement learning improve inventory management? Performance on lost sales, dual sourcing, and multi-echelon problems. *Manufacturing & Service Operations Management*, *24*(1), 134–152.
23. Goodfellow, I., Bengio, Y., & Courville, A. (2021). *Deep learning* (2nd ed.). MIT Press.
24. Hessel, M., Modayil, J., van Hasselt, H., et al. (2019). Rainbow: Combining improvements in deep reinforcement learning. *Proceedings of AAAI Conference on Artificial Intelligence*, *33*(1), 3215–3222.
25. Ivanov, D., & Dolgui, A. (2020). Viability of intertwined supply networks: Extending the supply chain resilience angles towards survivability. *International Journal of Production Research*, *58*(10), 2904–2915.
26. Kwon, H., Kim, J., & Kim, Y. (2021). Reinforcement learning-based dynamic routing for logistics networks. *IEEE Access*, *9*, 102110–102124.
27. Mazyavkina, N., Sviridov, S., Ivanov, S., & Burnaev, E. (2021). Reinforcement learning for combinatorial optimization: A survey. *Computers & Operations Research*, *134*, 105400.
28. Min, H. (2019). Blockchain technology for enhancing supply chain resilience. *Business Horizons*, *62*(1), 35–45.
29. Oroojlooy, A., & Snyder, L. V. (2020). A review of deep reinforcement learning for inventory management. *arXiv preprint arXiv:2005.10035*.
30. Peng, B., Wang, Z., & Zhang, L. (2022). Multi-agent reinforcement learning for dynamic logistics resource allocation. *Expert Systems with Applications*, *197*, 116674.
31. Sutton, R. S., & Barto, A. G. (2020). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
32. Tang, C. S., & Veelenturf, L. P. (2019). The strategic role of logistics in the industry 4.0 era. *Transportation Research Part E*, *129*, 1–11.

Abbreviation

Abbreviation	Full Form
3PL	Third-Party Logistics
RL	Reinforcement Learning
PPO	Proximal Policy Optimization
EDI	Electronic Data Interchange
OTM	Oracle Transportation Management
OIC	Oracle Integration Cloud
CI/CD	Continuous Integration / Continuous Deployment
MDP	Markov Decision Process
SLA	Service Level Agreement
DQN	Deep Q-Network
ERP	Enterprise Resource Planning
VRP	Vehicle Routing Problem
ANSI	American National Standards Institute
API	Application Programming Interface
OTD	On-Time Delivery
ETA	Estimated Time of Arrival
MLP	Multi-Layer Perceptron
GAE	Generalized Advantage Estimation
ReLU	Rectified Linear Unit
FIFO	First In First Out
GPU	Graphics Processing Unit
GAT	Graph Attention Network
OCI	Oracle Cloud Infrastructure
WMS	Warehouse Management System
EBS	E-Business Suite
OAuth	Open Authorization

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 6, Issue 2, March- April 2024 |

DOI:10.15662/IJEETR.2024.0602015

KL	Kullback–Leibler
JDBC	Java Database Connectivity
SOAP	Simple Object Access Protocol
REST	Representational State Transfer
MARL	Multi-Agent Reinforcement Learning
KPI	Key Performance Indicator
HTTP	Hypertext Transfer Protocol
DAT	Dial-A-Truck
TL	Truckload
NeurIPS	Conference on Neural Information Processing Systems
ICLR	International Conference on Learning Representations