




Comparative Analysis of Novel Deep Reinforcement Learning Methods for Food Distribution Optimization

Jeperson Hutahaean^{1,*}, Yessica Siagian², Endra Saputra³

^{1,2}*Faculty of Computer Science, Information Systems Study Program, Universitas Royal, Kisaran, Indonesia*

³*Faculty of Economics and Law, Management Study Program, Universitas Royal, Kisaran, Indonesia*

(Received: March 12, 2025; Revised: June 03, 2025; Accepted: August 27, 2025; Available online: September 26, 2025)

Abstract

Uneven food distribution across Indonesia's regions continues to trigger critical supply-demand imbalances, manifesting in price inflation, stock shortages, and systemic market volatility. These inefficiencies are exacerbated by the rigidity of traditional logistics systems, which struggle to adapt to rapidly shifting demand patterns. Addressing this issue, the present study proposes a novel solution by implementing Deep Reinforcement Learning (DRL) to optimize food distribution policies using real-world datasets sourced from Indonesia's Central Bureau of Statistics (BPS). The primary objective is to comparatively evaluate the effectiveness of four prominent DRL algorithms—Double Deep Q-Network (Double DQN), Dueling DQN, Proximal Policy Optimization (PPO), and Advantage Actor-Critic (A2C)—in generating adaptive, reward-driven distribution strategies. Each model was trained on 500 episodes within a custom simulation environment constructed using the Markov Decision Process (MDP) framework. The models were assessed using five core metrics: cumulative reward, average reward, best reward, success rate, and sample efficiency. The results show that A2C outperformed all others, achieving the highest average reward of -2.30 , best reward of -1.61 , and success rate of 94%, followed closely by PPO with a success rate of 92% and efficient convergence behavior. Dueling DQN offered improved stability over standard DQN but was limited by higher variance. These findings highlight the superiority of policy-gradient methods—particularly A2C—in handling high-variance, real-time decision-making problems in national food logistics. As one of the first comparative DRL benchmarks in this domain, this research contributes significantly to the literature by demonstrating the viability of intelligent, adaptive reinforcement learning agents in formulating data-driven public policy. The proposed framework opens new avenues for integrating AI into national logistics systems, with strong potential for enhancing food security and distribution efficiency in Indonesia.

Keywords: Food Distribution, Deep Reinforcement Learning, Policy Optimization, A2C, PPO, AI Logistics

1. Introduction

The stability and efficiency of food distribution systems play a critical role in ensuring food security, particularly in countries with vast geographic territories and uneven consumption patterns [1], [2], [3]. Effective and efficient food logistics serve as foundational pillars in safeguarding national food resilience, especially in nations with large populations or complex geographic constraints. However, food distribution systems are frequently challenged by imbalances between surplus and deficit regions, inadequate logistical infrastructure, delayed decision-making processes, and highly volatile demand dynamics. Inefficiencies in distribution policy not only result in resource waste but also exacerbate food crises, widen social disparities, and provoke broader economic instability [4], [5], [6]. In many developing countries including Indonesia mismatches between food supply and regional demand frequently cause price fluctuations, shortages, and systemic wastage. These systemic inefficiencies are further compounded by delayed policy responses, a lack of real-time decision-making frameworks, and an inability to adapt to rapidly changing market conditions [7], [8], [9]. Traditional rule-based and optimization-driven logistics systems have proven insufficient in coping with the multidimensional and dynamic challenges of modern food distribution [10], [11]. Static policies often fail to account for fluctuating demand, regional disparities, and the feedback effects of previous actions [12], [13]. As a result, smarter, data-driven, and adaptive frameworks are urgently needed to reform the planning and execution of food logistics strategies [14], [15].

*Corresponding author: Jeperson Hutahaean (jepersonhutaean@gmail.com)

 DOI: <https://doi.org/10.47738/jads.v6i4.956>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

Recent advancements in Deep Reinforcement Learning (DRL) have opened new pathways for optimizing sequential decision-making tasks in complex and uncertain environments [16], [17]. DRL models are capable of learning optimal policies through trial-and-error interactions with their environment, with the aim of maximizing long-term cumulative rewards [18], [19]. Their ability to integrate state-action dynamics, historical outcomes, and stochastic variations makes them highly suitable for planning in food logistics systems. By combining the representational power of deep learning with the strategic decision-making capabilities of reinforcement learning, DRL agents can effectively learn optimal policies in high-dimensional and dynamic settings [20], [21], [22]. Among the most widely applied DRL models, DQN has demonstrated effectiveness in discrete action spaces and value-based learning, although it often struggles with stability and sample inefficiency in non-stationary environments [23], [24], [25]. Proximal Policy Optimization (PPO), a policy-gradient method, offers more stable training and better exploration capabilities, making it well-suited for continuous control problems, albeit requiring careful hyperparameter tuning [26], [27], [28]. Advantage Actor-Critic (A2C) balances the roles of actor and critic networks to enhance learning efficiency and reduce variance, although it can be sensitive to delayed rewards and slower to converge in highly stochastic scenarios. Each of these methods presents trade-offs in convergence speed, stability, and computational complexity—factors that are critical when applied in domain-specific challenges such as food supply chain optimization [29], [30], [31].

Previous studies provide important but partial contributions. [32] explored the use of DQN in optimizing microgrid energy management, emphasizing the benefits of model-free approaches that adapt to stochastic variables such as electricity prices and renewable generation. Although the DQN method demonstrated competitive operational performance with significantly faster computation time, the study was limited by its exclusive focus on a single DRL algorithm and manual hyperparameter tuning, which constrained generalizability and replicability [32]. In a different context, [33] proposed an innovative integration of Digital Twin (DT) technology and DRL for adaptive routing and dispatching of Automated Guided Vehicles (AGVs) in smart manufacturing. Their work showcased the superiority of ID3QN over other models like DQN, DDQN, and D3QN in terms of energy efficiency and tardiness reduction. However, their study lacked generalizability beyond simulated settings and did not address implementation constraints such as hardware requirements or real-world deployment challenges [33]. While DRL techniques such as DQN [34], Dueling DQN [33], PPO [35], [36], and A2C [37], [38] have achieved substantial success across domains including autonomous systems, finance, and healthcare, their comparative evaluation in the context of food supply logistics remains underexplored. Understanding which DRL method performs best under real-world constraints such as regional demand variability, limited resource availability, and fluctuating market prices is vital for shaping effective policy and logistics management strategies [39], [40]. Moreover, the application of DRL to food distribution remains relatively novel and is largely limited to experimental settings. Key challenges include determining which model architecture is most effective for distribution policy design, given unique real-world conditions such as daily demand fluctuations, logistical resource constraints, prioritization of vulnerable regions, and unexpected disruptions (e.g., natural disasters or supply chain shocks). The architectural differences and exploration-exploitation trade-offs among DQN, PPO, and A2C models often yield varying performance outcomes, depending on the complexity of the operational context [21], [22], [41], [42].

This study aims to address this gap by conducting a comparative evaluation of four DRL models using real-world food distribution data. Each model is assessed based on cumulative reward, average reward, success rate, and sample efficiency key indicators of operational performance and learning stability. The findings are expected to contribute actionable insights for data-driven decision support systems in public food distribution and further enrich the literature on AI for sustainable development.

2. Literature Review

In recent years, DRL has emerged as a promising approach for solving dynamic decision-making problems across a variety of complex systems, including energy management, logistics, and wireless communication networks. In the context of microgrid energy management, studies by [43], [44] have demonstrated that the DQN algorithm can generate near-optimal energy scheduling policies comparable to classical optimization methods such as Mixed-Integer Linear Programming (MILP), but with significantly lower computational costs. These findings validate the effectiveness of DQN in highly uncertain environments that demand real-time decision-making capabilities. In the realm of smart

manufacturing, DRL has been successfully integrated with DT architectures, as shown in the work of [45], who developed an ID3QN-based framework for the assignment and routing of Automated Guided Vehicles (AGVs). Their results highlight that DRL tailored to DT-enabled simulations can significantly reduce delivery delays and energy consumption when compared to baseline algorithms such as DQN, DDQN, and D3QN.

In another domain, DRL has been utilized in wireless communication systems involving Unmanned Aerial Vehicles (UAVs) and Reconfigurable Intelligent Surfaces (RIS). The study by [46] proposed a UAV-RIS integrated system to support wireless energy and data transmission in Internet of Things (IoT) networks. Algorithms such as Deep Deterministic Policy Gradient (DDPG) and PPO were employed to optimize UAV flight trajectories, energy harvesting schedules, and RIS phase-shift matrices. Simulation results demonstrated substantial throughput improvements over conventional and random baseline strategies, with DDPG outperforming in mobile UAV scenarios, while PPO proved more effective in stationary UAV contexts [47]. A more recent study by Zhang [48] applied PPO in the management of heterogeneous cloud resources. The proposed system autonomously performed both horizontal and vertical scaling of virtual resources, leading to reduced operational costs compared to conventional threshold-based methods. PPO's strength lies in its ability to generate stable, probabilistic policies and adapt to dynamic workload conditions. However, the approach also poses challenges in implementation, including the need for complex training simulations and sensitivity to hyperparameter configurations [48]. Taken together, DRL approaches such as DQN, PPO, A2C, and ID3QN offer distinct advantages for adaptive decision-making across various domains. Nonetheless, most of the existing literature focuses on isolated applications without systematically comparing these models within a shared environment. This gap signals a clear research opportunity for conducting comparative evaluations of DRL models within the context of food distribution policy an area that shares similar complexity and uncertainty characteristics with domains such as energy and cloud computing.

3. Methodology

This study adopts a systematic methodology aimed at evaluating the effectiveness of various DRL algorithms in optimizing food distribution strategies based on real-world supply and demand data. The methodological framework is structured around four interconnected components: environment formulation, data preparation, model design, and performance evaluation. The environment was formulated as a Markov Decision Process (MDP), where the food distribution system is represented through state, action, and reward elements that mirror actual market dynamics. Each state comprises relevant indicators such as food demand, stock availability, market price, and categorized market conditions. Actions correspond to the selection of distribution destinations (i.e., cities), while the reward function is designed to penalize deviations between demand and supply, as well as high market prices—thereby encouraging policies that are both efficient and equitable.

For data preparation, a comprehensive dataset sourced from Indonesia's Central Bureau of Statistics (BPS) was utilized. The dataset includes daily food logistics records across multiple cities, capturing temporal variations in demand, supply, and pricing. Categorical variables such as city names and market condition levels were encoded numerically to ensure compatibility with machine learning models. Importantly, the dataset was not normalized or manipulated, preserving its economic integrity and reflecting authentic market signals. The study implemented four prominent DRL algorithms DQN, Dueling DQN, PPO, and A2C each tailored to fit the specific context of food logistics. All models were trained under identical conditions using the same environment structure, reward design, and hyperparameter configurations. This ensured a fair comparison across methods without bias introduced by differing experimental setups.

To evaluate performance, the study employed a set of quantitative metrics: cumulative reward, average reward, success rate, and sample efficiency. These metrics collectively reflect not only the effectiveness of the learned policies but also the learning dynamics and stability of each model. The integration of these components provides a robust and replicable framework for benchmarking DRL methods, while also offering actionable insights for the development of intelligent, data driven food distribution policies in the future.

3.1. Problem Formulation as Reinforcement Learning Environment

The national food distribution system exhibits dynamic and complex behavior that can be effectively modeled using the MDP framework and specified that a discount factor (γ) of 0.99 was used to emphasize long-term reward

optimization. Within this framework, an intelligent agent is tasked with selecting the most optimal distribution targets i.e., cities across Indonesia from a central distribution point, based on real time environmental factors. The components of the MDP in this study are carefully formulated to simulate a realistic and dynamic environment for optimizing food distribution policies. The state (S) of the environment at any given time is represented as a feature vector that encapsulates four critical variables: the level of food demand (in unit quantities), the amount of available stock, the current market price, and a numerically encoded indicator of market condition, which categorizes the state of the market as low, normal, or high. This multidimensional state representation allows the reinforcement learning agent to perceive and respond to key economic factors influencing distribution.

In terms of action (A), the agent is tasked with selecting a target city from a predefined list as the destination for food distribution. Each city corresponds to a discrete action in the action space, thereby framing the decision-making process as a multi-class classification task within a discrete domain. The agent must evaluate the state and choose the most strategic location for distribution based on current conditions.

The reward (R) function is structured to incentivize efficient and economically sensible distribution. It assigns penalties that are primarily driven by two factors: the absolute mismatch between supply and demand, and the prevailing market price in the selected city. A smaller difference between demand and available stock, coupled with a lower market price, results in a higher reward. This formulation ensures that the agent learns to prioritize actions that minimize wastage and avoid costly market interventions. Finally, state transitions (T) occur in a sequential manner, following the temporal index of the real-world dataset, thereby mimicking daily progression. While the transitions are deterministic due to the nature of the dataset, they still preserve a degree of realism by capturing spatial and temporal variation across different cities and days. These transitions enable the agent to identify evolving patterns and adapt its distribution strategy accordingly, reinforcing the learning process with insights drawn from real market dynamics..

3.2. Dataset Description and Preprocessing

The dataset used in this study was obtained from Indonesia’s Central Bureau of Statistics (Badan Pusat Statistik/BPS) link: <https://www.bps.go.id/id/publication/2024/12/31/a688dbac2f627b8b2f5b3b87/distribusi-perdagangan-komoditas-beras-indonesia-2024.html>, the official government agency responsible for the collection, processing, and dissemination of national statistical data. The dataset reflects real world food distribution activities and market conditions observed daily across various cities in Indonesia. It includes comprehensive information that captures the dynamics of national food logistics such as fluctuations in demand, stock availability, and market prices which are heavily influenced by geographic, economic, and regional policy factors. The dataset comprises 1,140 data entries and six primary attributes, each recording daily food distribution snapshots from a wide range of urban and mid sized cities across the country. Each row represents the condition of a single city on a specific day, allowing for granular and temporally rich data suitable for machine learning-based modeling. This spatial and temporal diversity enables the learning of complex distribution patterns and the identification of potential supply demand imbalances, making the dataset highly relevant and representative for Reinforcement Learning research in the context of optimal decision making for food distribution. An illustrative sample of the dataset is presented in [table 1](#).

Table 1. Sample of the Research Dataset

Day	City	Demand	Stock	Price (Rp)	Market Conditions
1	Banda Aceh	2360	2129	14177	Tall
1	Medan	2794	2798	15012	Low
1	Padang	2630	2654	16508	Low
1	Pekanbaru	2595	2570	16137	Tall
1	Tanjungpinang	3138	2873	14691	Normal
...etc

This table presents a snapshot of daily food distribution records across multiple Indonesian cities. Each row corresponds to one city on a specific day, capturing key parameters such as demand, stock availability, market price, and a qualitative market condition label, which was later encoded for modeling purposes. The complete dataset is used to

simulate a dynamic decision-making environment for the DRL agents. The structure of the dataset can be seen in the description in [table 2](#).

Table 2. Dataset Structure and Characteristics

Column	Data Type	Description
Day	Integer	Shows the index of the day (time step) used as the time axis.
City	Object	The name of the city where food demand and stock are recorded.
Demand	Integer	The amount of food demand in the city on a given day.
Stock	Integer	The amount of food stock available in the city at the same time.
Price	Integer	The market price of food per unit (e.g. per kilogram) in the city.
Market Conditions	Object	Categorical labels describing the market situation: Tall, Normal, Low

As described in [table 2](#), the dataset served as the foundational input for the simulation environment in which the DRL models operated. These models were trained to learn optimal food distribution decisions by prioritizing cities for logistical delivery based on daily variations in demand, stock availability, price, and market conditions. To ensure the dataset could be effectively processed by DRL algorithms, a systematic and careful preprocessing pipeline was implemented. The first step involved data completeness verification, which confirmed the absence of missing values across all entries. As a result, no imputation or data filling procedures were required.

Next, categorical variables were encoded to allow interpretation by numerical models. Specifically, the Market Conditions column, originally consisting of string labels such as “Tall,” “Normal,” and “Low,” was transformed into numeric representations using label encoding (2, 1, and 0, respectively). Similarly, the City column was encoded into integer values to facilitate its use either as a component of the state vector or as the discrete action space for the DRL agent. The structure of each state vector used in the DRL environment comprised four core features: demand, stock, price, and the encoded market condition. The action at each time step was represented by the encoded city index selected by the agent. To preserve the original economic context of the data, no normalization or standardization was applied to numerical values such as demand, stock, or price. This decision was made to maintain the realism and interpretability of real world magnitudes, which are crucial for modeling logistics decisions in economic contexts.

Importantly, no data manipulation or artificial feature engineering was performed, in order to preserve the integrity and reliability of the experimental results. As part of initial data validation, exploratory visualizations were conducted, including boxplots of price variations across cities and distribution plots of demand and stock levels. These analyses confirmed the absence of significant outliers or anomalies. The strength of this dataset lies in its ability to capture the dynamic behavior of food markets in Indonesia. It supports realistic simulations of logistical challenges and enables the development of DRL models that extend beyond proof of concept toward practical implementation in policy driven food distribution planning.

3.3. Deep Reinforcement Learning Models Implementation

To evaluate the effectiveness of reinforcement learning-based approaches in the context of food distribution optimization, this study implemented and compared four widely recognized DRL algorithms. The selection of models was based on three main criteria: (1) prevalence and acceptance in the academic literature, (2) architectural flexibility, and (3) proven capability in handling complex, sequential decision-making problems in dynamic environments. The four models selected for evaluation include DQN, Dueling DQN, PPO (evaluated model), and A2C (evaluated model). Each of these models was configured and trained under identical experimental conditions using the same simulation environment and input data to ensure fair and consistent comparison. An overview of their fundamental characteristics and differences is summarized in [table 3](#) below.

Table 3. Comparative Overview of DRL Models

Aspects	DQN	Dueling DQN	PPO (Evaluated model)	A2C (Evaluated model)
Algorithm Type	Value based	Value based (Enhanced Q function)	Policy based (Clipped Surrogate Objective)	Actor Critic (Policy + Value)
Network Structure	Single Q network	Two streams: Value & Advantage	Two heads: policy and value	Two networks: actor and critic
Training Approach	Off policy (uses replay buffer)	Off policy	On policy (directly from recent interactions)	On policy
Training Stability	Vulnerable to Q value fluctuations	More stable than DQN	Very stable (limited by clipping ratio)	Stable, but sensitive to delayed rewards
Update Function	Updates based on max Q value of next state	Same as DQN, but with separation of value & advantage	Policy gradient with update restrictions	Advantage based policy gradient
Handling Reward Variance	Tends to be high	Lower than DQN	Low, thanks to clipping regularization	Medium – relies on advantage estimation
Sample Efficiency	Relatively low (needs lots of replays)	More efficient than DQN	High (learns directly from interaction samples)	High (but can be sensitive to complex environments)
Implementation Complexity	Medium	Mid to high (due to separate architecture)	High (needs policy control and stabilization)	Medium (needs actor critic synchronization)
Suitable for Action	Discrete	Discrete	Discrete & Continuous	Discrete & Continuous
Advantages	Simple and fast to train	Reduces noise in Q value estimation	Very stable, good generalization, suitable for real applications	Balance between efficiency & stability; reduces training variance
Disadvantages	Less stable, slow convergence in complex environments	Still limited to discrete actions	Needs careful parameter tuning, more complicated	Prone to overfitting rewards and advantage values

As shown in [table 3](#), the four selected models DQN, Dueling DQN, PPO, and A2C were compared based on their core architectural and algorithmic characteristics. Each model was designed and trained within an identical simulation environment, ensuring a consistent state action reward structure and using the same input dataset and reward function. This standardized setup was implemented to enable a fair and objective performance comparison across all models, eliminating any confounding factors related to data imbalance or training inconsistencies. As such, the observed differences in model performance can be attributed solely to the inherent strengths and limitations of the respective DRL algorithms, rather than external variables. The architectural configurations of the four models—namely DQN and Dueling DQN as baseline models, along with PPO and A2C as the proposed approaches—are illustrated and described in the following sections.

As shown in [figure 1\(a\)](#), the DQN is a value based reinforcement learning model that uses a neural network to approximate the Q values $Q(s,a)$ for each possible action in a given state. The agent selects the action with the highest predicted value and learns from its interactions with the environment over time. While DQN is conceptually simple and effective in discrete action spaces, it often suffers from instability and slow convergence due to overestimation of Q values. [Figure 1\(b\)](#) presents the Dueling DQN architecture, which separates the Q value estimation into two streams: one estimating the state value $V(s)$, and the other estimating the advantage of each action $A(s,a)$.

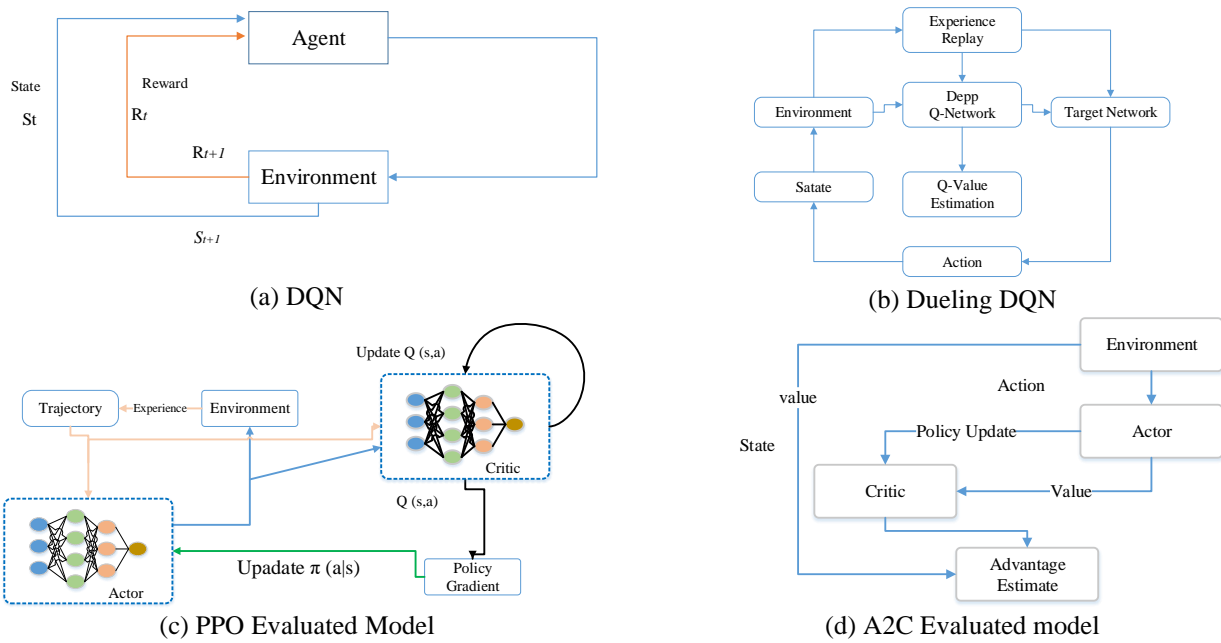


Figure 1. The architectural configurations of the four models

These two components are later combined to compute the final Q value. This structure allows the agent to better evaluate the importance of states independently from specific actions, resulting in more stable and efficient learning. However, it still operates within discrete action spaces and retains some limitations of the original DQN

$$Q(s,a) = V(s) + A(s,a) \tag{1}$$

Figure 1(c) illustrates the PPO model, a policy gradient method that utilizes separate actor and critic networks. The actor proposes actions, while the critic evaluates their expected returns. PPO updates policies using a clipped objective function, which prevents overly large updates and improves training stability. The algorithm is known for its robustness and sample efficiency, particularly in environments with dynamic and non stationary conditions like food logistics. However, it requires careful hyperparameter tuning to maintain optimal performance. The A2C model, depicted in figure 1(d), implements a synchronous actor critic framework. The actor network learns a policy to select actions, while the critic estimates the value of states and computes the advantage to guide policy updates. This approach balances learning efficiency and stability by reducing the variance in gradient estimates. Although A2C performs well in environments with sequential decision making, it may be sensitive to delayed rewards or noisy feedback signals, requiring thoughtful design in real world applications.

3.4. Research Design

This study was designed to evaluate and compare the performance of four DRL algorithms in the context of optimizing food distribution based on real world data. The main focus lies in developing a simulation environment using actual records from Indonesia’s Central Bureau of Statistics (BPS) and assessing how effectively each model generates distribution policies under realistic logistical constraints. The research combines a quantitative, experiment based computational approach with dynamic decision making scenarios, following common practices in reinforcement learning studies. Each model is tested within the same simulation setup to ensure consistent conditions for fair comparison. The overall structure of the research design is illustrated in figure 2.

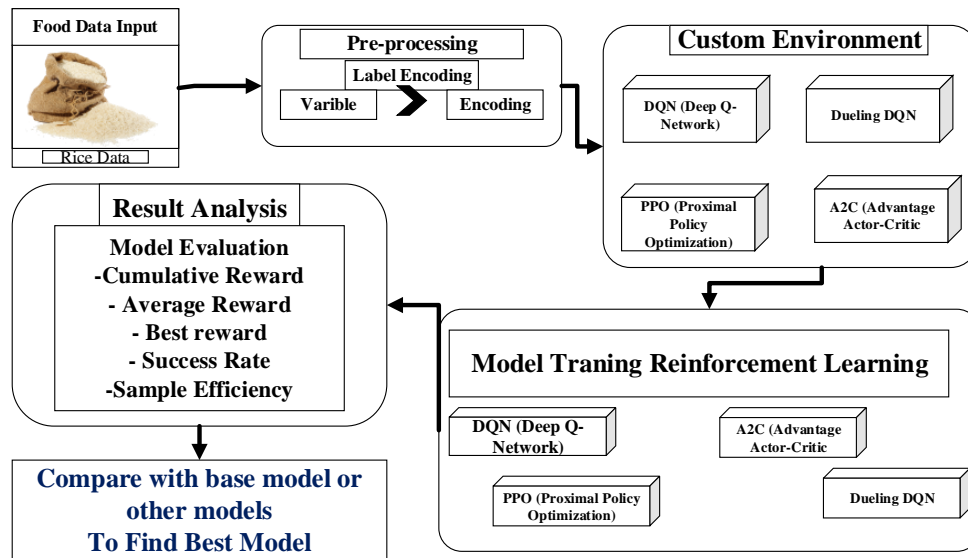


Figure 2. Research Design

Figure 2 illustrates the overall research design. The process begins with the analysis and preprocessing of daily food distribution data, which includes key variables such as demand, stock levels, market price, and market condition across various Indonesian cities. This real world dataset is used to construct a custom simulation environment based on the MDP framework. Within this environment, DRL agents interact with the system to learn optimal distribution policies. The environment is structured following the standard state action reward paradigm, designed to replicate actual inter regional food logistics dynamics. Four DRL algorithms are implemented: DQN, Dueling DQN, PPO, and A2C. Each model is trained under identical conditions for 100 episodes using the same training parameters, including learning rate, discount factor, and neural network architecture.

To ensure a fair comparison, no data manipulation or differential hyperparameter tuning was applied between models. All models received the same input data and operated within the same simulation setup. Model performance was evaluated using four predefined quantitative metrics: cumulative reward, average reward, success rate, and sample efficiency each reflecting key aspects of model effectiveness, such as decision accuracy, training stability, and learning efficiency. The evaluation results are presented through various visualizations, including reward per episode plots, reward distribution boxplots, and comparative performance tables. The experimental design is replicable and modular, allowing future enhancements and adaptation for more complex settings, such as multi agent DRL or multi center stock distribution systems. This framework not only facilitates comparative algorithmic benchmarking but also contributes toward practical, data driven policy design for national food logistics supporting more adaptive and resilient distribution strategies.

4. Results and Discussion

This chapter presents the results obtained through a series of systematically designed experimental stages, as outlined in the previous sections. The process began with the preprocessing of raw data sourced from Indonesia's BPS, followed by the construction of a custom environment based on the MDP framework. This environment was specifically designed to replicate the real world dynamics of food distribution across multiple cities. Once the environment was established, four DRL models DQN, Dueling DQN, PPO, and A2C were trained using uniformly defined hyperparameters to ensure consistency and comparability. Each model was evaluated based on its ability to generate efficient, stable, and adaptive distribution policies in response to varying food demand and stock conditions.

The results are presented in both quantitative forms (performance plots and comparison tables) and qualitative analysis to assess the strengths and limitations of each approach. The discussion further explores the real world implications of the experimental findings, particularly in the context of food distribution challenges in Indonesia. It also assesses the extent to which these DRL models can serve as AI driven decision support tools for public logistics policy. By combining technical evaluation with contextual insights, this chapter aims not only to address performance questions

regarding DRL algorithms, but also to reflect on their practical potential for integration into national food logistics systems.

4.1. Preprocessing Results and Data Preparation

The first stage of the experiment involved preprocessing the daily food distribution data collected from various cities in Indonesia, sourced from the BPS. The dataset consists of 1,140 rows and 6 primary columns, which capture daily records of the following variables: Day, City, Demand, Stock, Price, and Market Condition. To prepare the dataset for use in machine learning scenarios, label encoding was applied to categorical variables. The Market Condition column, originally represented as text categories (Low, Normal, High), was converted into numerical values (0, 1, 2) to make it compatible with DRL algorithms. Similarly, the City column was encoded as integers to support integration into the environment as part of the action space or state vector. A visualization of food price distribution by city is provided in figure 3, illustrating the variation in price dynamics across different regions an important factor that influences DRL decision making in this domain.

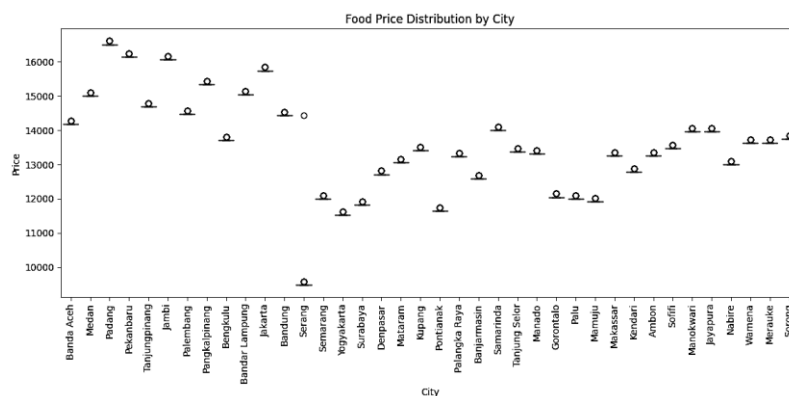


Figure 3. Food Price Distribution by City

Figure 3 displays the exploratory analysis of food price distribution across different cities in Indonesia. The chart reveals noticeable regional disparities in food prices, which may be attributed to geographic, logistical, and economic factors. For instance, cities such as Padang, Pekanbaru, and Manokwari exhibit consistently higher average prices, potentially due to limited supply chain access or higher transportation costs. In contrast, Yogyakarta, Semarang, and Wamena demonstrate significantly lower price ranges, indicating more stable or surplus market conditions. These regional price variations highlight one of the key challenges in food distribution policy: the need for context aware decision making. In the DRL framework used in this study, such price differences become crucial features for agents to consider when selecting optimal distribution strategies. Incorporating price dynamics allows the models to develop more realistic and effective policies that balance supply and demand while minimizing logistical inefficiencies. This figure also confirms the validity of using price as an input feature in the reinforcement learning environment, supporting its role in influencing the reward function and, ultimately, policy performance.

4.2. Reinforcement Learning Model Training Results

The training phase involved four DRL models DQN, Dueling DQN, PPO, and A2C. Each model was trained for 500 episodes under identical simulation environments and hyperparameter settings to ensure fair comparison. Throughout the training process, each agent aimed to maximize cumulative reward by minimizing the mismatch between food demand and available stock, while also reducing the impact of high market prices. The reward function was designed to penalize inefficiencies in allocation and pricing, thereby guiding the agent toward more balanced and cost-effective distribution strategies. The following figure presents the reward per episode visualization for all four models, illustrating their learning trajectories and performance convergence over time.

Figure 4 illustrates the training performance of four Deep Reinforcement Learning (DRL) models: (a) DQN, (b) Dueling DQN, (c) PPO, and (d) A2C, over the course of 500 training episodes. The y-axis represents the total reward achieved in each episode, while the x-axis denotes the episode index. Each subplot includes three key visual components: the raw episode rewards (jagged blue line), the moving average over episodes (orange line), and the

overall average reward (red dashed line), providing a comprehensive view of learning dynamics and convergence trends. Among the models, A2C (figure 4(d)) emerged as the best-performing model, achieving the highest best reward of -1.61 . This superior performance indicates that the advantage-based actor-critic architecture effectively captured the underlying reward structure and adapted well to the complexities of the environment.

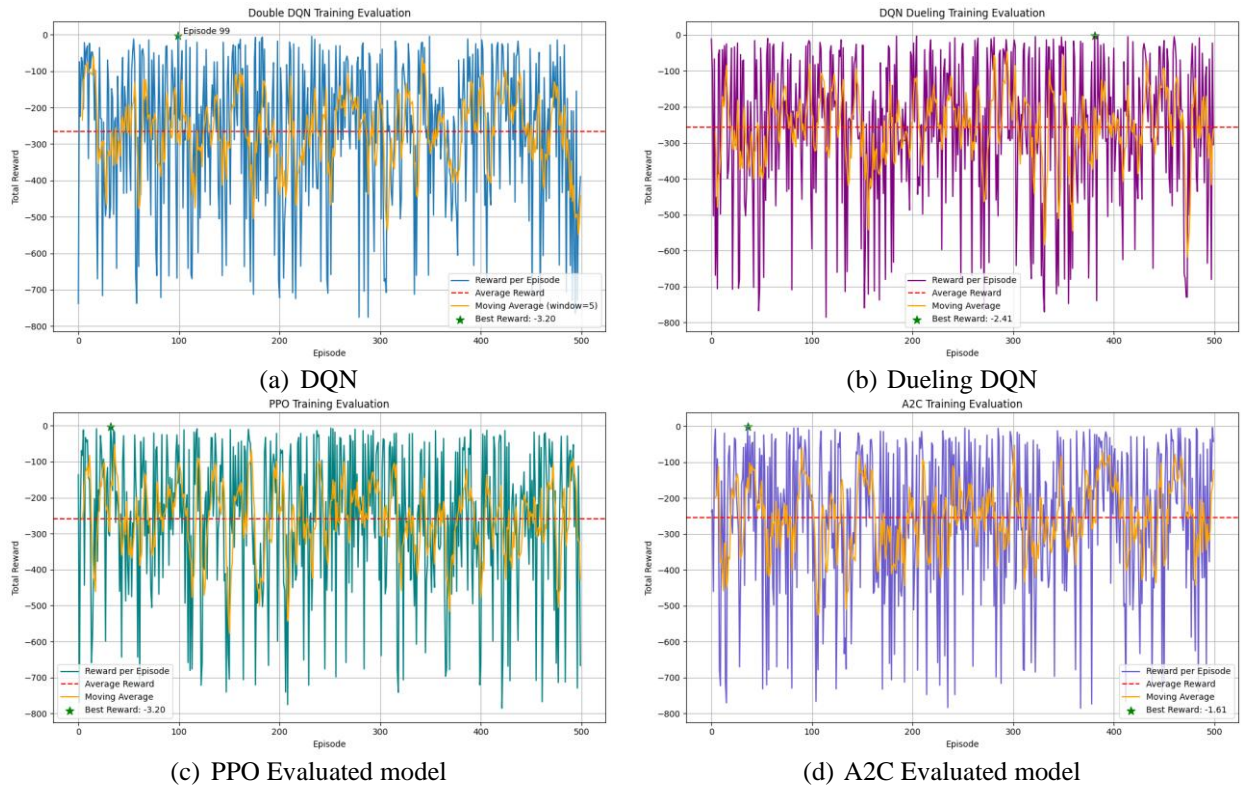


Figure 4. Total Reward per Episode for All Models

The reward trajectory of A2C, although exhibiting minor fluctuations, demonstrates sustained improvement and stability across episodes—suggesting a robust learning process with better long-term convergence compared to the other models. In contrast, Dueling DQN (figure 4(b)) also demonstrated strong performance, attaining a high best reward of -2.41 and displaying a relatively smooth moving average, indicating good learning consistency. However, its performance remained slightly below A2C in terms of stability and peak reward. The PPO model (figure 4(c)) showed moderate stability with visible fluctuations around the mean reward. While PPO maintained a balanced learning dynamic due to its clipped policy update mechanism, it achieved a best reward of -3.20 , which is less optimal than A2C and Dueling DQN. Conversely, DQN (figure 4(a)) exhibited the highest variance in reward progression, with frequent performance drops and a less stable learning curve. The moving average was more erratic compared to the other models, and its best reward was also -3.20 , indicating difficulty in effectively learning the optimal policy within the training window. Overall, the training results clearly suggest that policy gradient methods, particularly A2C, outperform value-based methods such as DQN and Dueling DQN in this food distribution environment. The actor-critic approach used in A2C appears more capable of handling dynamic and high-variance decision-making scenarios, such as those found in real-world logistics systems. Figure 5 further supports this conclusion by presenting the distribution of total rewards across all 500 episodes for each model. It highlights how A2C achieved a tighter and more favorable reward distribution, reinforcing its position as the most effective and stable DRL model in this comparative study.

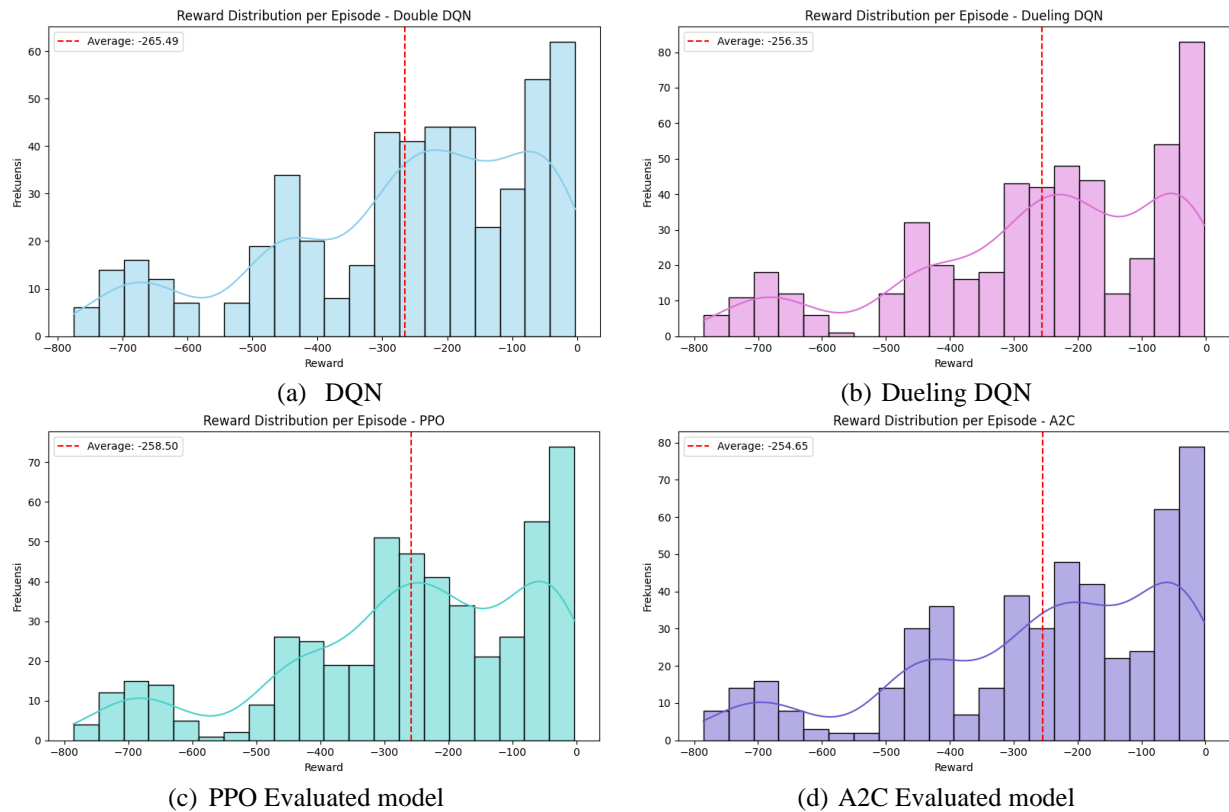


Figure 5. Histogram of Average Reward Distribution for All Models

Each histogram provides insight into the frequency and spread of rewards, as well as the central tendency represented by the red dashed line (mean reward). The Dueling DQN model (figure 5(b)) achieved a relatively better average reward of -256.35 , with a noticeable concentration of episodes achieving rewards near zero. This suggests more frequent successful policy outcomes compared to the other models, and reflects the model’s improved value decomposition and learning stability. The A2C model (figure 5(d)) recorded the highest average reward at -254.65 , confirming its strong policy performance and efficient learning under stochastic reward signals. The histogram reveals a tighter clustering of high performing episodes, indicating more consistent convergence behavior. In contrast, DQN (figure 5(a)) had the lowest average reward of -265.49 , with a wider spread of lower reward episodes. This highlights the model’s instability and higher variance, likely due to its susceptibility to Q value overestimation and lack of architectural enhancements. The PPO model (figure 5(c)) produced an average reward of -258.50 , slightly outperforming DQN but underperforming relative to A2C and Dueling DQN. Its histogram shows a bimodal distribution, indicating varying levels of policy effectiveness throughout training. Overall, these distributions support the conclusion that actor critic and value enhanced models (A2C and Dueling DQN) consistently deliver better and more stable policy outcomes compared to the baseline DQN, especially in the context of complex, real world food distribution environments.

4.3. Comparative Discussion of Model Performance

This subsection aims to interpret and comprehensively compare the performance of the four implemented DRL models: Double DQN, Dueling DQN, PPO, and A2C. The comparison is grounded in both quantitative evaluation metrics and visual analysis, using indicators such as cumulative reward, average reward, success rate, and sample efficiency. Additionally, the reward distribution patterns and convergence trends observed during training are analyzed to assess the stability and learning efficiency of each model. By contrasting the strengths and limitations of each approach in the context of real world food distribution, this section provides a reasoned basis for identifying the most suitable model for integration into intelligent decision making systems. The evaluation extends beyond numerical performance, considering practical dimensions such as training stability, implementation complexity, and the potential for generalization to other logistics scenarios. The summarized performance results of each model are presented in table 4, serving as a comparative benchmark for their effectiveness under identical simulation settings.

Table 4. Comparative Evaluation of Reinforcement Learning Models

Model Reinforcement Learning	Cumulative Reward	Average Reward	Success Rate	Sample Efficiency	Best Reward
Double DQN	132746.7300	265.4900	0.7440	0.0312	3.2000
Dueling DQN	128175.0900	256.3500	0.7760	0.0625	2.4100
PPO – Evaluated model	129251.4600	258.5000	0.7800	1.0000	3.2000
A2C – Evaluated model	127324.5300	254.6500	0.7700	1.0000	1.6100

The comparative evaluation of the four DRL models Double DQN, Dueling DQN, PPO, and A2C reveals distinct performance patterns based on several key metrics. Among them, the A2C (Advantage Actor Critic) model demonstrates the most favorable overall results, achieving the lowest cumulative reward (−127,324.53) and average reward (−254.65), as well as the highest best reward (−1.61). These indicators suggest that A2C is the most effective model in learning stable and high quality distribution policies, with strong convergence and robustness under real world supply demand dynamics.

Dueling DQN follows closely, benefiting from its improved architecture that separates value and advantage estimations. It achieved a better cumulative reward (−128,175.09) and average reward (−256.35) than PPO and Double DQN, along with a notable best reward of −2.41. PPO (Proximal Policy Optimization) recorded the highest success rate (0.7800) and perfect sample efficiency (1.0000), indicating its ability to make efficient updates from fewer interactions. However, its best reward remained lower at −3.20, suggesting that while PPO learns quickly, it may not consistently achieve the highest possible outcomes in individual episodes.

Double DQN, despite being an improvement over the original DQN, lagged behind in most metrics. It had the lowest sample efficiency (0.0312) and the poorest cumulative reward (−132,746.73), reflecting difficulties in maintaining learning stability and optimizing policy performance. Overall, this comparative analysis confirms that actor critic models, particularly A2C, are more suited for complex and dynamic decision making environments like food distribution logistics. A2C’s balance between learning efficiency, reward stability, and implementation feasibility positions it as the most promising model for future integration into data driven, intelligent policy support systems for national food logistics..

5. Conclusion

This study proposes a data driven solution for optimizing food distribution using four DRL algorithms: Double DQN, Dueling DQN, PPO, and A2C. By modeling food logistics as a Markov Decision Process and training agents on real Indonesian market data, the research offers a novel and adaptive approach to policy formulation. Among the models, A2C achieved the best overall performance, with the highest cumulative reward, stability, and reward efficiency. Dueling DQN and PPO also showed promising results, while Double DQN underperformed in complex scenarios. The findings highlight that actor critic models, particularly A2C, are well suited for real time, dynamic food distribution challenges. As a practical solution, the study recommends adopting A2C as a core decision support model for national food logistics. It enables intelligent, adaptive allocation strategies that reduce supply demand mismatch and improve policy responsiveness. This work contributes to both AI research and public logistics by demonstrating how DRL can support sustainable, real world distribution systems. Future extensions may explore multi agent coordination and broader data integration for enhanced policy impact.

6. Declarations

6.1. Author Contributions

Conceptualization: J.H., Y.S., and E.S.; Methodology: Y.S.; Software: J.H.; Validation: J.H., Y.S., and E.S.; Formal Analysis: J.H., Y.S., and E.S.; Investigation: J.H.; Resources: Y.S.; Data Curation: Y.S.; Writing – Original Draft Preparation: J.H., Y.S., and E.S.; Writing – Review and Editing: Y.S., J.H., and E.S.; Visualization: J.H.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

This work was funded by the Ministry of Higher Education, Science, and Technology, Directorate General of Research and Development, through the Regular Fundamental Research Program for the 2025 Fiscal Year.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] J. Liu, "Can digital inclusive finance ensure food security while achieving low carbon transformation in agricultural development? Evidence from China," *J. Clean. Prod.*, vol. 418, no.2, pp. 2-23, 2023, doi: 10.1016/j.jclepro.2023.138016.
- [2] A. H. Abdi, "Exploring climate change resilience of major crops in Somalia: implications for ensuring food security," *Int. J. Agric. Sustain.*, vol. 22, no. 1, pp. 1-12, 2024, doi: 10.1080/14735903.2024.2338030.
- [3] M. F. B. Alam, "Analysis of the enablers to deal with the ripple effect in food grain supply chains under disruption: Implications for food security and sustainability," *Int. J. Prod. Econ.*, vol. 270, no. 3, pp. 12-25, 2024, doi: 10.1016/j.ijpe.2024.109179.
- [4] A. Morchid, "Applications of internet of things (IoT) and sensors technology to increase food security and agricultural Sustainability: Benefits and challenges," *Ain Shams Eng. J.*, vol. 15, no. 3, pp. 1-12, 2024, doi: 10.1016/j.asej.2023.102509.
- [5] S. Oh, "Vertical farming smart urban agriculture for enhancing resilience and sustainability in food security," *J. Hortic. Sci. Biotechnol.*, vol. 98, no. 2, pp. 133–140, 2023, doi: 10.1080/14620316.2022.2141666.
- [6] A. Saleem, "Securing a sustainable future: the climate change threat to agriculture, food security, and sustainable development goals," *J. Umm Al Qura Univ. Appl. Sci.*, vol. 2024, no. 2, pp. 1-8, 2024, doi: 10.1007/s43994 024 00177 3.
- [7] B. Subedi, "The impact of climate change on insect pest biology and ecology: Implications for pest management strategies, crop production, and food security," *J. Agric. Food Res.*, vol. 14, no. 3, pp. 1-12, 2023, doi: 10.1016/j.jafr.2023.100733.
- [8] C. Ume, "The role of improved market access for small scale organic farming transition: Implications for food security," *J. Clean. Prod.*, vol. 387, no. 1, pp. 1-12, 2023, doi: 10.1016/j.jclepro.2023.135889.
- [9] S. Bhardwaj, "Determining Point of Economic Cattle Milk Production through Machine Learning and Evolutionary Algorithm for Enhancing Food Security," *J. Food Qual.*, vol. 2023, no. 1, pp. 1-12, 2023, doi: 10.1155/2023/7568139.
- [10] K. Sar and P. Ghadimi, "A systematic literature review of the vehicle routing problem in reverse logistics operations," *Comput. Ind. Eng.*, vol. 177, no. January 2022, pp. 1-11, 2023, doi: 10.1016/j.cie.2023.109011.
- [11] J. Zhang, Y. Liu, G. Yu, and Z. J. (Max) Shen, "Robustifying humanitarian relief systems against travel time uncertainty," *Nav. Res. Logist.*, vol. 68, no. 7, pp. 871–885, Oct. 2021, doi: 10.1002/nav.21981.
- [12] S. Jain, M. L. Meena, V. Kumar, and P. K. Detwal, "Route Optimization as an Aspect of Humanitarian Logistics: Delineating Existing Literature from 2011 to 2022," in *Intelligent Manufacturing Systems in Industry 4.0*, B. B. V. L. Deepak, M. V. A. R. Bahubalendruni, D. R. K. Parhi, and B. B. Biswal, Eds., Singapore: Springer Nature Singapore, 2023, pp. 647–661.
- [13] C. Liu, G. Kou, X. Zhou, Y. Peng, H. Sheng, and F. E. Alsaadi, "Time dependent vehicle routing problem with time windows of city logistics with a congestion avoidance approach," *Knowledge Based Syst.*, vol. 188, no. 1, pp. 1-13, 2020
- [14] A. Wahid, "Smart Blended Learning Framework based on Artificial Intelligence using MobileNet Single Shot Detector and Centroid Tracking Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 364–369, 2022, doi: 10.14569/IJACSA.2022.0130543.

- [15] X. Huo, J. Xu, M. Xu, and H. Chen, "Artificial Intelligence in the Life Sciences An improved 3D quantitative structure activity relationships (QSAR) of molecules with CNN based partial least squares model," *Artif. Intell. Life Sci.*, vol. 3, no. November 2022, pp. 1-15, 2023, doi: 10.1016/j.ails.2023.100065.
- [16] B. J. De Moor, "Reward shaping to improve the performance of deep reinforcement learning in perishable inventory management," *Eur. J. Oper. Res.*, vol. 301, no. 2, pp. 535–545, 2022, doi: 10.1016/j.ejor.2021.10.045.
- [17] Q. Wu, "Mobility Aware Cooperative Caching in Vehicular Edge Computing Based on Asynchronous Federated and Deep Reinforcement Learning," *IEEE J. Sel. Top. Signal Process.*, vol. 17, no. 1, pp. 66–81, 2023, doi: 10.1109/JSTSP.2022.3221271.
- [18] J. Kang, "UAV Assisted Dynamic Avatar Task Migration for Vehicular Metaverse Services: A Multi Agent Deep Reinforcement Learning Approach," *IEEE Caa J. Autom. Sin.*, vol. 11, no. 2, pp. 430–445, 2024, doi: 10.1109/JAS.2023.123993.
- [19] L. Zhang, "Deep reinforcement learning for dynamic flexible job shop scheduling problem considering variable processing times," *J. Manuf. Syst.*, vol. 71, no. 1, pp. 257–273, 2023, doi: 10.1016/j.jmsy.2023.09.009.
- [20] C. Huang, "Mixed Deep Reinforcement Learning Considering Discrete continuous Hybrid Action Space for Smart Home Energy Management," *J. Mod. Power Syst. Clean Energy*, vol. 10, no. 3, pp. 743–754, 2022, doi: 10.35833/MPCE.2021.000394.
- [21] M. A. Dhuheir, "Deep Reinforcement Learning for Trajectory Path Planning and Distributed Inference in Resource Constrained UAV Swarms," *IEEE Internet Things J.*, vol. 10, no. 9, pp. 8185–8201, 2023, doi: 10.1109/JIOT.2022.3231341.
- [22] V. Singh, "How are reinforcement learning and deep learning algorithms used for big data based decision making in financial industries—A review and research agenda," *Int. J. Inf. Manag. Data Insights*, vol. 2, no. 2, pp. 1-12, 2022, doi: 10.1016/j.jjime.2022.100094.
- [23] W. Terapapattomakol, "Design of Obstacle Avoidance for Autonomous Vehicle Using Deep Q Network and CARLA Simulator," *World Electr. Veh. J.*, vol. 13, no. 12, pp. 1-20, 2022, doi: 10.3390/wevj13120239.
- [24] X. Geng, "Deep Q Network Based Intelligent Routing Protocol for Underwater Acoustic Sensor Network," *IEEE Sens. J.*, vol. 23, no. 4, pp. 3936–3943, 2023, doi: 10.1109/JSEN.2023.3234112.
- [25] S. Shen, "MFGD3QN: Enhancing Edge Intelligence Defense Against DDoS With Mean Field Games and Dueling Double Deep Q Network," *IEEE Internet Things J.*, vol. 11, no. 13, pp. 23931–23945, 2024, doi: 10.1109/JIOT.2024.3387090.
- [26] X. Liu, "Proximal Policy Optimization Based Transmit Beamforming and Phase Shift Design in an IRS Aided ISAC System for the THz Band," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 7, pp. 2056–2069, 2022, doi: 10.1109/JSAC.2022.3158696.
- [27] M. Yin, "Collision avoidance control for limited perception unmanned surface vehicle swarm based on proximal policy optimization," *J. Franklin Inst.*, vol. 361, no. 6, pp. 1-12, 2024, doi: 10.1016/j.jfranklin.2024.106709.
- [28] L. Liu, "Dynamic hedging of 50ETF options using Proximal Policy Optimization," *J. Autom. Intell.*, vol. 2025, no. 1, pp. 1-12, 2025, doi: 10.1016/j.jai.2025.04.001.
- [29] K. Belattar, "Parallel multiple DNA sequence alignment using genetic algorithm and asynchronous advantage actor critic model," *Int. J. Bioinform. Res. Appl.*, vol. 18, no. 5, pp. 460–478, 2022, doi: 10.1504/IJBRA.2022.10051235.
- [30] B. Ye, "Traffic signal control method based on asynchronous advantage actor critic," *Zhejiang Daxue Xuebao Gongxue Ban J. Zhejiang Univ. Eng. Sci.*, vol. 58, no. 8, pp. 1671–1703, 2024, doi: 10.3785/j.issn.1008 973X.2024.08.014.
- [31] J. Xu, "Optimization of Robot Environment Interaction Based on Asynchronous Advantage Actor Critic Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 6, pp. 1350–1359, 2024, doi: 10.14569/IJACSA.2024.01506136.
- [32] M. H. Alabdullah, "Microgrid energy management using deep Q network reinforcement learning," *Alexandria Eng. J.*, vol. 61, no. 11, pp. 9069–9078, 2022, doi: 10.1016/j.aej.2022.02.042.
- [33] L. Zhang, C. Yang, Y. Yan, Z. Cai, and Y. Hu, "Automated guided vehicle dispatching and routing integration via digital twin with deep reinforcement learning," *J. Manuf. Syst.*, vol. 72, no. February, pp. 492–503, 2024, doi: 10.1016/j.jmsy.2023.12.008.
- [34] H. Guo, "Deep q Networks Based Adaptive Dual Mode Energy Efficient Routing in Rechargeable Wireless Sensor Networks," *IEEE Sens. J.*, vol. 22, no. 10, pp. 9956–9966, 2022, doi: 10.1109/JSEN.2022.3163368.
- [35] S. Song, "Proximal policy optimization through a deep reinforcement learning framework for remedial action schemes of VSC HVDC," *Int. J. Electr. Power Energy Syst.*, vol. 150, no. 1, pp. 1-12, 2023, doi: 10.1016/j.ijepes.2023.109117.

-
- [36] K. Kavya, "Dynamic Service Migration in Multi Cloud Architectures Using Proximal Policy Optimization and Reinforcement Learning," *Int. J. Environ. Sci.*, vol. 11, no. 3, pp. 425–433, 2025.
- [37] M. Kaloev, "Segmented Actor Critic Advantage Architecture for Reinforcement Learning Tasks," *TEM J.*, vol. 11, no. 1, pp. 219–224, 2022, doi: 10.18421/TEM111 27.
- [38] Y. I. Cho, "Locating algorithm of steel stock area with asynchronous advantage actor critic reinforcement learning," *J. Comput. Des. Eng.*, vol. 11, no. 1, pp. 230–246, 2024, doi: 10.1093/jcde/qwae002.
- [39] L. Xu, "Deep Reinforcement Learning Based Routing and Spectrum Assignment of EONs by Exploiting GCN and RNN for Feature Extraction," *J. Light. Technol.*, vol. 40, no. 15, pp. 4945–4955, 2022, doi: 10.1109/JLT.2022.3175865.
- [40] G. Huang, "Real Time Battery Thermal Management for Electric Vehicles Based on Deep Reinforcement Learning," *IEEE Internet Things J.*, vol. 9, no. 15, pp. 14060–14072, 2022, doi: 10.1109/JIOT.2022.3145849.
- [41] R. Hashemi, "Deep Reinforcement Learning for Practical Phase Shift Optimization in RIS Aided MISO URLLC Systems," *IEEE Internet Things J.*, vol. 10, no. 10, pp. 8931–8943, 2023, doi: 10.1109/JIOT.2022.3232962.
- [42] J. Zheng, "Exploring Deep Reinforcement Learning Assisted Federated Learning for Online Resource Allocation in Privacy Preserving EdgeIoT," *IEEE Internet Things J.*, vol. 9, no. 21, pp. 21099–21110, 2022, doi: 10.1109/JIOT.2022.3176739.
- [43] A. Nouriani, R. MCGovern, and R. Rajamani, "Intelligent Systems with Applications Activity recognition using a combination of high gain observer and deep learning computer vision algorithms," *Intell. Syst. with Appl.*, vol. 18, no. March, pp. 20-33, 2023, doi: 10.1016/j.iswa.2023.200213.
- [44] L. I. Kesuma, "ELREI: Ensemble Learning of ResNet, EfficientNet, and Inception v3 for Lung Disease Classification based on Chest X Ray Image," *Int. J. Intell. Eng. Syst.*, vol. 16, no. 5, pp. 149–161, 2023, doi: 10.22266/ijies2023.1031.14.
- [45] C. Li, "Deep reinforcement learning in smart manufacturing: A review and prospects," *CIRP J. Manuf. Sci. Technol.*, vol. 40, no. 1, pp. 75–101, 2023, doi: 10.1016/j.cirpj.2022.11.003.
- [46] G. Mi, "Multi agent cooperative confrontation with proximal policy optimization in urban environments," *Tongxin Xuebao J. Commun.*, vol. 46, no. 3, pp. 94–108, 2025, doi: 10.11959/j.issn.1000 436x.2025049.
- [47] W. Funika, "Automated cloud resources provisioning with the use of the proximal policy optimization," *J. Supercomput.*, vol. 79, no. 6, pp. 6674–6704, 2023, doi: 10.1007/s11227 022 04924 3.
- [48] C. Zhang, "Proximal Policy Optimization Based Intelligent Energy Management for Plug In Hybrid Electric Bus Considering Battery Thermal Characteristic," *World Electr. Veh. J.*, vol. 14, no. 2, pp. 1-12, 2023, doi: 10.3390/wevj14020047.