JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(3), SEPTEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

Short-term Efficiency and Adaptability Assessment of a Modified PPO Model for Pipeline Monitoring

[1]Kefas Yunana, [2]Ishaq Oyebisi Oyefolahan, [3]Sulaimon Adebayo Bashir
[1]*Department of Computer Science, Nasarawa State University, Keffi, Nigeria,*
[2]*Department of Information Technology, Federal University of Technology, Minna, Nigeria*
[3]*Department of Computer Science,*
Federal University of Technology, Minna, Nigeria

ABSTRACT

Abstract: Critical infrastructure of oil pipelines needs robust monitoring techniques to promptly detect and address potential leaks, faults, or anomalies that can lead to severe economic and environmental consequences. Traditional monitoring techniques often rely on static or rule-based approaches, which may struggle with the dynamic and complex operational nature of pipeline environments. This study introduces a Modified Proximal Policy Optimization-based Pipeline network Monitoring Agent (MPPOPNMA) specifically designed to enhance the short-term efficiency and adaptability of pipeline monitoring through advanced reinforcement learning techniques. The MPPOPNMA model has been optimized with tailored reward structures and hyperparameter tuning to enable rapid and accurate decision-making in response to state changes characterized by sensor data, including pressure, flow rate, and velocity. In contrast to standard PPO and DQN models, MPPOPNMA demonstrated superior performance in terms of average reward and convergence rate, indicating its ability to both maximize immediate gains and achieve efficient learning outcomes with fewer iterations. The average reward metric highlights MPPOPNMA effectiveness in consistently selecting optimal actions, enhancing its responsiveness to dynamic shifts within the pipeline system. Meanwhile, the high convergence rate emphasizes the model's ability to quickly adapt and stabilize, underscoring its potential for real-time application where immediate action is paramount. This study thus positions MPPOPNMA as a viable and powerful tool for critical infrastructure monitoring, suggesting that reinforcement learning can advance safety, efficiency, and responsiveness within oil pipeline operations.

ARTICLE INFO
Article History
Received: July, 2024
Received in revised form: August, 2024
Accepted: August, 2024
Published online: September, 2024

KEYWORDS
Oil Pipeline Monitoring, Proximal Policy Optimization, Reinforcement Learning, Critical Infrastructure, Average Reward, Convergence Rate, Adaptive Monitoring Systems, Machine Learning, Pipeline Management, Pipeline Integrity Management

INTRODUCTION

Oil and gas pipelines are among the most vital infrastructure components worldwide, conveying critical energy resources over vast geographical regions to meet energy requirements. The monitoring of these pipelines is crucial due to the risks associated with leaks, ruptures, and external damages that could result in serious environmental damage, safety hazards, and financial losses. Traditional monitoring methods, including manual inspections, acoustic sensors, and pressure-wave systems, have long been used in the industry. While these methods offer some level of abnormality detection, they tend to rely heavily on predefined rules and thresholds thereby restraining their efficiency in handling complex and evolving operational conditions, especially in real-time (Reddy & Xie, 2017) (Y. Yang et al., 2022) (Huang et al., 2023).

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

Recent developments in artificial intelligence (AI), particularly machine learning (ML) and reinforcement learning (RL), have introduced new possibilities for pipeline monitoring (S. Zhang et al., 2024).Reinforcement learning allows for continuous learning and adaptation based on direct interaction with the environment, making it well-suited for dynamic, complex applications such as pipeline monitoring (Boudlal et al., 2024). Proximal Policy Optimization (PPO), a robust RL algorithm developed by OpenAI, has shown promising results in various high-dimensional and real-time applications (Silva et al., 2024)(S. Zhang et al., 2024). However, existing versions of PPO are not optimized for oil pipeline environments where conditions change frequently, and the need for immediate and precise decision-making is paramount (Saha et al., 2024). These limitations make it crucial to adapt PPO for pipeline monitoring through modifications that improve its learning speed and decision accuracy.

More so, the notion that an agent can learn to make decisions by interacting with its environment, getting feedback in the form of rewards and updating its strategies to maximize cumulative rewards over time is a central element of reinforcement learning (RL) which is a subfield of artificial intelligence (X. Yang et al., 2023) (Kaloev & Krastev, 2021). Resource management, robotics, gaming, curriculum learning, and autonomous driving are just a few of the domains where reinforcement learning has lately shown promise (Kiefer & Dorer, 2023) (X. Wang & Guo, 2021)

Zhang et al. (2023). Due to its ability to learn complex behaviours and make the best choices when faced with ambiguity, reinforcement learning (RL) is a promising alternative for oil pipeline management. The Proximal Policy Optimisation (PPO) and Deep Q-Network (DQN) algorithms are two of the many RL techniques that have demonstrated remarkable efficacy in the optimisation of intricate, dynamic, and unpredictable systems (Khalid et al., 2023) (Zeng et al., 2023) (Vieira et al., 2023).

LITERATURE REVIEW

Related Work

Developing fast and dependable algorithms for learning optimal policies is crucial in the rapidly evolving field of reinforcement learning (RL). The use of reinforcement learning in a range of domains, such as games and autonomous vehicles, has been studied by numerous scholars. In the study by (Zhu et al., 2024), a closed-loop scheduling method in an autonomous inter-terminal system that employs unmanned shipment vessels (USVs) to transport containers among operational berths in seaport terminals was presented. The study developed a scheduling model by considering energy replenishment, time windows, and berth restrictions, aiming to obtain cost-saving USV transportation solutions and conflict-free paths. The proposed multi-attention reinforcement learning (MARL) algorithm integrated with an encoder and decoder structure with unsupervised auxiliary network offers prompt problem solving abilities and benefits from extensive offline training. While the application domain was in unmanned shipment vessels, the use of average reward and convergence rate was not analysed.

In Wang & Wang (2024), a study of an adaptive financial trading strategy, called proximal policy optimization based on financial signal representation trading strategy (FSRPPO) was proposed. The strategy uses financial signal representation technology (FSR) that merges complete ensemble extreme-point symmetric mode decomposition with adaptive noise (CEEMDAN) and modified rescaled range analysis (MRS) to precisely and strongly represent dynamic market states thereby improving efficiency. However, the domain of applications was the financial trading process.

A system using wireless power transfer and unmanned aerial vehicles (UAVs) was introduced in (Lee et al., 2023) to obtain new data from sensor nodes. Moving, sending or aggregating data, and transferring energy are among the decisions that the system optimizes using a deep reinforcement learning algorithm (DeepUAV). The controller outperforms comparative schemes in terms of gathered new data by continuously learning online and refining these decisions via trial-and-error. Even though

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

the study emphasized the advantages of reinforcement learning for UAVs, the suggested DeepUAV algorithm was not clearly analysed instead, the average reward was the only metric considered.

Using Soft Actor-Critic and PyTorch frameworks, the study in (Chatterjee et al., 2023) compares model predictive control (MPC) and RL-trained UAV controllers. The findings indicate that RL controllers outperform MPC in nonlinear UAV models, highlighting the promise of RL in analytical control techniques. Even though the study shows that RL can beat MPC, the application domain was in UAVs and no metrics were used to evaluate the agents' performance.

The use of DRL in reconfigurable intelligent surface (RIS) Unmanned aerial vehicle (UAV) communication systems was investigated in (Nguyen et al., 2023), with particular attention paid to goals, parameters, deployment situations, and techniques. It draws attention to areas of research that need improvement to improve (RIS)-UAV networks. The study was more of a review of previous research with potential uses in UAV networks. The application of customized Deep Q-Network (DQN) and Soft Actor-Critic (SAC) algorithms to identify the best oil well control is described in (Ding et al., 2023).

The importance of geological development features such reservoir heterogeneity is considered by the writers. More so, the study techniques, findings, conclusion, and introduction are all logically organized. However, there are several spots where the work may be strengthened to make it clearer and more credible. Although the inputs of the algorithm are initially explained, it is still unclear which characteristics are removed or how they are handled, and no RL standard metrics were developed to assess the asserted methods. Additionally, it was unclear how the personalized DQN and SAC algorithms differed from the conventional ones, what their structure and details were, and why they were chosen.

The study in Gao et al. (2023) proposes a Q-learning algorithm-based method for efficiently scheduling Coalbed Methane (CBM) well maintenance activities to boost production efficiency. The interactive reinforcement learning environment was constructed, and the proposed method is independent of the task size and works more efficiently than traditional algorithms. The observed issue was the use of the convergence curve as the only performance metric.

In Jayanetti et al. (2022), Deep Reinforcement Learning was used to create a workflow scheduling framework that effectively addresses these problems in an edge computing environment. The framework consists of a hierarchical action space and a hybrid actor-critic-based scheduling framework with proximal policy optimization techniques. In terms of execution time and energy consumption, the proposed Deep Reinforcement Learning approach outperforms baseline techniques, demonstrating its superiority in finding a balance between the two. The only metrics visible are execution time and energy consumption.

To enhance federated learning performance and efficiency in Sensor cloud systems (SCSs), a deep reinforcement learning (DRL) based scheduling technique was proposed in (T. Zhang et al., 2023). Proximity policy optimization was used to carefully design and train the DRL environment, which includes the state space, action space, and reward function. According to experimental results, the suggested approach performs better than alternative baselines on datasets that are identically distributed (IID) and those that are not. The optimization methods that produced the suggested model's accuracy and cumulative reward performance were chosen without any rationale.

To determine an opportunistic Condition-based maintenance (CBM) strategy for a multi-unit system with economic dependency across an indefinite horizon, the study in (Najafi & Lee, 2023) suggests a modified deep reinforcement learning (DRL) algorithm for semi-Markov decision processes (SMDP). The technique uses the proportional hazards model and system reliability features to consider the simultaneous effects of age and variables. When the suggested algorithm was used on a multi-unit hydroelectric power system with damage self-

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

healing capabilities, it outperformed competing strategies in terms of cost reduction and demonstrated how improving system dependability lowers costs over time. The hydroelectric power system was the study's application domain, and the cost reduction metric was considered.

The authors of Huan et al. (2021) address a real-world issue in industry by investigating non-invasive, affordable IoT-based methods for pipeline temperature sensing. It offers thorough theories and simulation findings, defining fundamental models for heat transport in rectangular and cylindrical pipes and suggesting a method for multilayer non-metallic pipelines. The accuracy of the analytical answers in comparison to traditional approaches was uncertain, and the study lacked precise machine learning models or techniques for temperature prediction, anomalies, or optimization. Furthermore, it is uncertain if the model can be applied to multilayer non-metallic pipelines additional information regarding tests and applicability may be required.

A high-fidelity digital twin (DT) pipeline that uses deep learning techniques to adjust operational situations and show its potential is presented by the authors in (Liang et al., 2023). With its implementation in a natural gas pipeline and addressing next work in expanding the diversity of abnormal conditions, it exhibits practical importance. Although deep learning is used in the work, it is a more robust decision-making system such as reinforcement learning methods was not considered.

A methodology for monitoring critical infrastructure that makes use of Deep Reinforcement Learning (DRL) was presented in a prior study by (Yunana et al., 2022). In the study, communication networks, power plants, electricity grids, gas pipelines, and transportation systems are considered examples of essential infrastructure that is networked with the Internet of Things (IoT). The advantages of using Deep Reinforcement Learning (DRL) to monitor important observable parameters, like flow rate, temperature, and pressure sensor data, to attain optimal management was also examined in the study. It is important to note that the suggested technique has not yet been put into use based on the study carried.

Using an enhanced Proximal Policy Optimization (PPO) algorithm, the work by (Lü et al., 2024) suggests an energy management plan for fuel cell hybrid electric vehicles with the goals of lowering hydrogen use, halting fuel cell deterioration, and increasing efficiency. To improve the algorithm's convergence time, the study makes changes such as using a clipping technique rather than a divergence penalty. The study does, however, concentrate on fuel cell management, which allows Modified PPO to be applied to vital physical infrastructure, such as oil pipelines. Managing real-time leaks, ensuring operational safety, and managing complicated, continuous variables like pressure, flow rate, and velocity are some of the difficulties faced by the oil pipeline monitoring industry.

According to the study of Kahnamouei and Moallem (2024), welding automation has advanced, with a focus on the incorporation of AI and control systems in welding robots for arc, laser, and friction stir welding techniques. It talks on how important sensor technologies such as force, temperature, and vision systems are to real-time monitoring and guaranteeing the best possible weld quality. Furthermore, the use of machine learning (ML) is investigated, allowing for the detection of weld defects, process optimization, and predictive maintenance. This shows how artificial intelligence (AI) can improve the flexibility and effectiveness of welding robots, particularly in demanding settings like ocean pipeline welding. To improve welding accuracy and fault identification, the study focuses on automated welding using machine learning techniques. The goal of the current work is to assess the short-term efficiency and adaptability of a modified PPO-based model for oil pipeline infrastructure monitoring that addresses leak detection and ongoing critical system surveillance. In contrast to current AI systems in welding automation, Modified PPO can offer more sophisticated control in the oil pipeline domain, which involves dynamic real-time data.

METHODOLOGY

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(3), SEPTEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

The methods used to implement and assess the suggested Modified Proximal Policy Optimization based Pipeline Network Monitoring Agent (MPPOPNMA) model for the optimal management of oil pipelines are described in the next section. Additionally, a comparison with other cutting-edge models is conducted. In a similar vein, the study created a customized environment known as the oil pipeline environment that replicates the operational characteristics of an oil pipeline inside the framework of reinforcement learning. A typical OpenAI Gym setting is formed in this environment (Jaensch et al., 2022), (Gléau et al., 2021) (Christian et al., 2022) (Badakhshan et al., 2023). With properties determined by states like pressure, flow rate, and velocity, the environment is a representation of the oil pipeline system. Random states within a certain range, including the ideal ranges for each state, are used to initialize it.

Action Space

The action space encompasses possible operational adjustments that can be made to maintain optimal conditions. The action space A in the oil pipeline reinforcement learning environment is discrete and consists of three possible actions denoted closing the valve (CV), do nothing (DN), and stand by (SB). The execution of actions changes the environment's state which is categorised by the pressure, flow rate, and velocity within the oil pipeline environment. The following section shows the mathematical illustration of how each action impacts the environment states.

Closing Valve (CV) Action

Given the incidence of leak, the valve closing action is triggered where the flow rate decreases, pressure increases, and velocity decreases which is described by the equations. (3.1) to (3.3).

$F_{Next} = F_{Current}(1-x)$ (3.1)
$P_{Next} = P_{Current}(1+y)$ (3.2)
$V_{Next} = V_{Current}(1- z)$ (3.3)

Do Nothing (DN) Action

In the case of no leakage along the pipeline the do-nothing action is initiated, where the flow rate, pressure, and velocity all naturally increase slightly due to ongoing processes of fluid movement, slight thermal expansions, or minor build-ups which is described by the following equations (3.4) to (3.6).

$F_{Next} = F_{Current}(1+x)$ (3.4)
$P_{Next}=P_{Current}(1+y)$ (3.5)
$V_{Next}= V_{Current}(1+z)$ (3.6)

Standby (SB) Action

In in the case of a partial variation in the flow characteristics, the action of partial closure of the pipeline valve might be initiated to prepare for rapid changes but does not necessarily change dramatically depending on the nature of activities along the pipeline. Hence, it implies a small decrease in flow and pressure, while velocity decrease slightly if less fluid is moving through. This is modelled in equations (3.7) to (3.9).

$F_{Next} = F_{Current}(1-x)$ (3.7)
$P_{Next} = P_{Current}(1-y)$ (3.8)
$V_{Next}= V_{Current}(1-z)$ (3.9)

The x, y and z variables used in the equations (3.1) to (3.9) denote proportionate variations of the state observations for respective actions.

State Space

The environment's state S is a three-dimensional vector space of continuous values which includes pressure, velocity, and flow rate. The selection of these characteristics as state variables is based on their importance in describing the operational condition of an oil pipeline. These variable values were chosen to be within certain ranges to ensure the system's safe and effective operation. Equations (3.10) define these ranges.

$S = \{s \in R^3 \mid P\_Min \leq s_P \leq P\_Max, F\_Min \leq s_F \leq F\_Max, V\_Min \leq s_V \leq V\_Max\}$ (3.10)

Reward Function

The reward function been a numerical value to incentivise the agent to promote optimal operation by offering negative rewards for circumstances that can cause pipeline failure and

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(3), SEPTEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

positive rewards for preserving ideal conditions. The agent gets a positive reward each time the state values are within the optimal range and a negative reward if it is not in the optimal range. Equations (3.12) provide the reward function complete representation.

Reward_{Custom} = { $\text{R}_{positive}$ if s is Within_{Range} and $\text{R}_{negative}$ if s is $\text{Not}_{in-Range}$ and the episode ends} (3.12)

Policies

The MPPOPNMA agent's policies (π) is represented by a deep neural network that has been modified to have higher layers, parameterized by θ. The function θ takes the current state (s) as input and outputs a probability distribution over the action space (A). The MPPOPNMA algorithm optimizes θ to maximize the expected cumulative reward as shown in equation (3.13) which is aimed at finding the parameters θ of the policy π that maximize the sum of discounted rewards over the course of an episode.

$\max E [t=0\sum T\gamma^{t}\cdot R_{Custom} (st, \pi\theta(st))]$ (3.13)

MPPOPNMA Model and Environment Interaction

The model was created especially for the task with adjustments made to the normal PPO architecture and other hyper parameters tuning to enhance its suitability to the unique oil pipeline domain. The MPPOPNMA utilised a customize policy that was inherited from the actor critic policy of the stable baseline 3 class library. The actor critic is a kind of policy consisting of both a policy function, the actor and a value function often referred to as the critic. The actor part is used to select actions and the value function for evaluating the selected actions which are both updating while training is going on.

The neural network architecture used for the policy and value functions comprises two hidden layers of 128 neurons each with a hyperbolic tangent (Tanh) function for the neural networks. The activation function is often used entry wise to the outputs of the neurons in the network which also regulates the output ranges of each neuron. In discrete time steps, the agent interacts with the environment where at each time step t, the agent observes the state s_t, takes an action a_t based on policy π ($a_t|s_t$; θ) and then receives a reward r_t and transitions to a new state s_{t+1}. The interaction is denoted mathematically as shown in Eq. (3.14).

$([P, F, V]_t, a_t, r_t, [P, F, V]_{t+1}) \sim P_r(\cdot|[P, F, V]_t, a_{t-1}), \pi(a_{t-1}|[P, F, V]_t; \theta)$ (3.14)

More so, during the agent and environment interaction, the sum of all rewards received by an agent during the episode can computed as shown in equation (3.15).

$G_t = R_{t+1} + R_{\{t+2\}} + R_{\{t+3\}} + \ldots + R_T$ (3.15)

Where G_t signifies the total reward at time step t, R_{t+1}, $R_{\{t+2\}}$, ... R_T are the rewards acquired at each time step from t+1 until T at the end of episode, and T is the episode's last time step.

The MPPOPNMA model architecture involving the agent observing the environment states and selecting appropriate action based on its policy is as shown in Figure. 1. The various experimental parameters used is as shown in table 1.

Table 1. Experimental Parameters

Parameters	Value	Description
N steps	512	Number of steps to run for each environment per update
Entropy Coefficient	0.1	Entropy coefficient for the loss calculation
Learning Rate	0.0003	Learning rate for the model
Batch Size	64	Number of training examples used in one iteration of model training
Gae Lambda	0.95	Factor for trading-off bias with variance for generalized advantage estimator
Gamma	0.99	Discount factor
N Epochs	15	Number of epochs for optimizing the surrogate loss

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

Parameters	Value	Description
Clip Range	0.2	Clipping range utilized to limit the variability of the updates during the policy optimization
Value Function Coefficient	0.6	Value function coefficient for the loss calculation
Total Timesteps	2500	The total number of samples (timesteps) to train on for each learning iteration
N evaluation episodes	25	The number of episodes to use for calculating the mean reward
Log interval	100	The number of timesteps after which to log files
Evaluation Frequency	500	The frequency (number of timesteps) after which the evaluation callback is called
Saving Frequency	1000	The frequency (number of time steps) after which a checkpoint is saved
Training Iterations	400	The number of iterations to train for.
Simulation number	06	The simulation number to repeat the experiments

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(3), SEPTEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

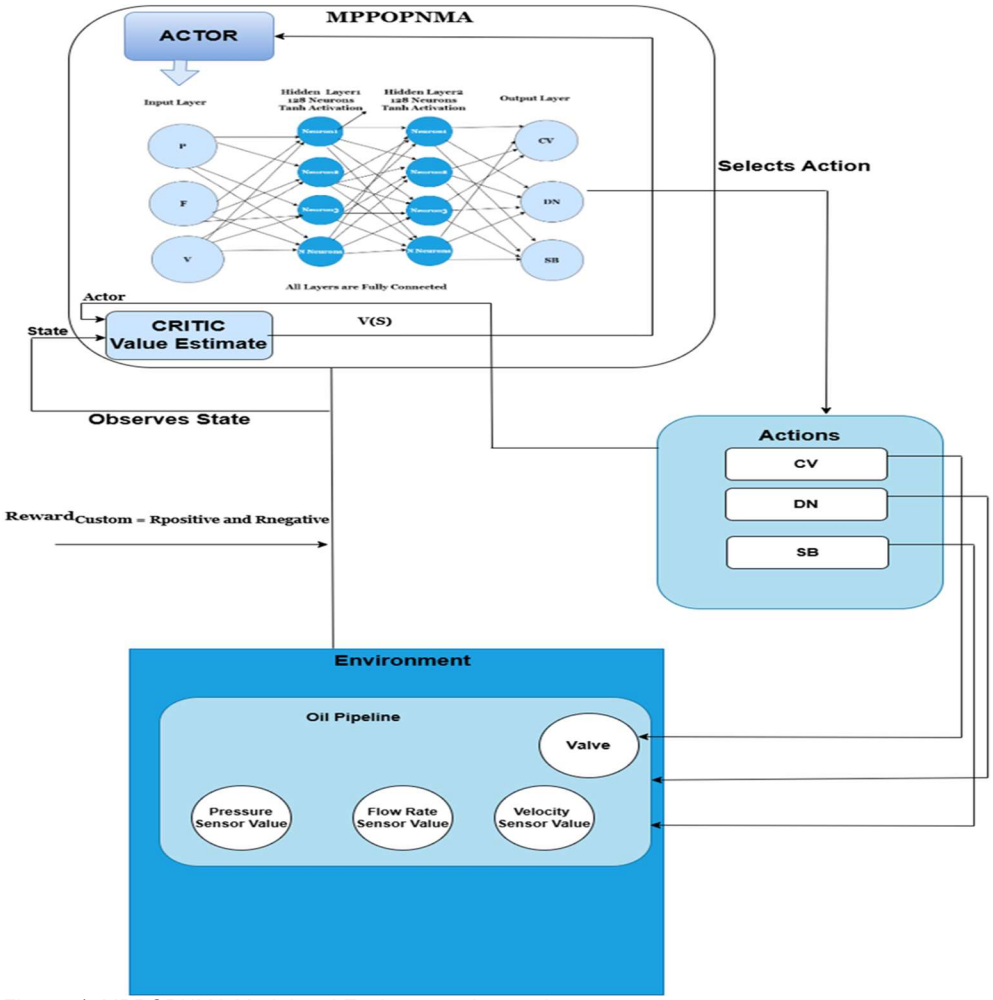

Figure. 1: MPPOPNMA Model and Environment Interaction

MPPOPNMA Initialization and Training Algorithm

The different initialization and training steps for the MPPOPNMA reinforcement learning model for the oil pipeline environment monitoring are described in the algorithm that follows.

Algorithm: MPPOPNMA Initialization and Training Algorithm

Step 1: Inputs: Environment, States (S), Actions (A), Policy π (a|s), equations (3.10) to (3.12), experimental parameters (Table 1).

Step 2: Initialization:

- *Initialize the policy network with random weights θ, and the value function network with weights ϕ*
- *Initialize the optimizer with given learning rate, clip range, and other hyperparameters (Table 1.)*

Step 3: For each episode do:

- *Initialize the environment and get the initial state* s_0
- *For each time step do:*
  - *Choose an action* a_t *from the policy network* $\pi(a_t | s_t; \theta)$ *based on equation. (3.13) and (3.14)*

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

- Execute action a_t in the environment, get the next state s_{t+1} and reward r_t, equation. (3.1) to (3.9) and equation (3.13) and (3.14)
- Store transition (s_t,a_t,r_t,s_{t+1}) in the memory D
- Update the current state s_t, to s_{t+1}, based on equation (3.14)
- At the end of the episode, compute returns G_t, from the collected rewards r_t, equation. (3.15)
Step 4: After N episodes or time steps, perform an update:
- For each (s_t,a_t,r_t,s_{t+1}) in D, compute the advantage estimate A_t = G_t - V (s_t; ϕ; PR, FR, VL)
*- Update policy by ascending gradients*Et [π (at/st; θ,PR,FR,VL)/πold(at/st) ∗At+R_{custom}]
- Update value function by descending gradients ∇$_\phi$ 1/2 E_t [(G_t - V (s_t; ϕ; PR, FR, VL)) ^2 + R_{custom}]
Step 5: Repeat steps 3-4 until convergence or maximum number of episodes or time steps reached
*Step 6: Output: Optimized policy parameters θ**

The first phase in the given MPPOPNMA initialization and training algorithm comprises inputs such as the environment, a set of states S, a set of actions A, and a policy (a|s) that indicates the probability of performing an action in each state s. The value function network, which produces the expected return or value V(s) for each state s, and the policy network, which outputs (a|s), are initialized with random weights as part of the initialization stage. Additionally, the optimizer is configured with hyperparameters, such as clip range and learning rate.

The step 3 is the process for each episode or run of the environment where the environment is reset to its initial condition. The current state is then inputted to the policy network at each timestep, an action is chosen based on the network's output, and that action is carried out in the environment. The ensuing next state, reward, and action are saved in a memory buffer D. The computation of the returns follows at the end of the episode where the total return G_t is determined from the rewards. The return is used to calculate the advantage of each action, which is the difference between the return and the value anticipated by the value function network. After N episodes, the policy network is updated to make the selected actions more likely if their advantage is positive, and less likely otherwise. This is accomplished by computing a ratio between the new and old policy probability for each action and utilising this to weight the advantage in the gradient ascent step. The value network is also modified to better forecast the returns, with a mean squared error loss between the expected and actual returns.

The Steps 3-5 are repeated for a set number of iterations, or until the policy's performance stops improving, and finally the algorithm returns the final optimised policy parameters θ*. These parameters can then be used in the policy network to select actions in the oil pipeline environment.

Evaluation Models

The selection of Proximal Policy Optimisation (PPO) and Deep Q-Network (DQN) as comparison points for MPPOPNMA can be supported by their advantages, unique learning methodologies, and widespread use in the reinforcement learning space. PPO has been considered a standard in policy optimization because of its clipped objective function, which is noteworthy for its ability to balance sample efficiency and implementation convenience while still permitting sufficient exploration. Conversely, DQN is an off-policy technique that has demonstrated exceptional performance in playing Atari games directly from pixel inputs, marking a significant turning point in the integration of deep learning and reinforcement learning. As a result, these algorithms provide strong benchmarks for evaluating MPPOPNMA's effectiveness.

Performance Metrics

Each model's performance was assessed using two common measures for reinforcement learning. The following section provides a brief analysis of the metrics.

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

Average Reward

The mean value of the rewards that the reinforcement learning model receives during training is shown by the average reward. It is calculated by adding up all the rewards earned in each step or episode, then dividing that total by the number of episodes. The algorithm's overall performance in terms of obtaining rewards in the specified task or environment is shown by the average reward. The mean reward is the mean of all the awards earned in every episode. It is calculated mathematically as indicated by Equation (3.16).

$$Avg_{Reward} = \Sigma(G_t)/N_{Episodes} \quad (3.16)$$

Where $\Sigma(G_t)$ is the sum of all cumulative rewards earned across all episodes and $N_{Episodes}$ is the total number of episodes.

Average Convergence Rate (ACR)

The ACR calculates the rate at which a reinforcement learning model approaches a stable or ideal course of action. It measures how rapidly a model settles into a consistent policy that produces good rewards and is calculated using variations in the cumulative reward over iterations. The average convergence rate has several significant ramifications in the context of the oil pipeline setting, where the objective is to optimize actions based on pressure, flow, and velocity to prevent leaks. These include learning speed, model stability, and efficiency. As shown in equation (3.17), it is frequently calculated using variations in the cumulative reward over iterations.

$$ACR = \frac{1}{n-1}\sum_{i=1}^{n-1} | C_{i+1}\text{-}C_i| \quad (3.17)$$

Where n is the number of episodes, C_i is the cumulative reward at episode i and $|C_{i+1}\text{-}C_i|$ is the absolute difference in cumulative rewards between episodes.

Data collection

The data used for this research was from the Federal University of technology anti pipeline vandals pipeline test bed. It consists of Ultrasonic flow sensors, which were used to measure the parameters of pressure, flow rate and velocity of the fluid. The pipeline is characterised by 100 meter in length and 3 inches in diameter. The sample sensor data is as shown in table 2.

Table 2: Sample Sensor Data

Pressure (psi)	Flow (M^3)	Velocity(m/s)	Target Variables
6.870	29.215	2.161	Not Leaking
6.887	29.142	2.154	Not Leaking
6.914	29.028	2.143	Not Leaking
6.863	29.242	2.189	Not Leaking
6.740	29.779	2.200	Not Leaking
6.747	29.746	2.205	Not Leaking
6.756	29.707	2.195	Not Leaking
6.904	29.070	2.150	Not Leaking
6.769	29.651	2.196	Not Leaking
6.767	29.660	2.193	Not Leaking
6.778	29.611	2.189	Not Leaking

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

Pressure (psi)	Flow (M^3)	Velocity(m/s)	Target Variables
6.796	29.534	2.188	Not Leaking
6.790	29.559	2.191	Not Leaking
6.782	29.592	2.187	Not Leaking
7.446	26.955	1.994	Leaking
7.444	26.962	1.996	Leaking
7.662	26.193	1.935	Leaking
7.690	26.098	1.930	Leaking
7.731	25.961	1.920	Leaking
7.736	25.945	1.920	Leaking
7.566	26.526	1.987	Leaking
7.772	25.822	1.867	Leaking
8.057	24.910	1.831	Leaking
8.108	24.753	1.828	Leaking
8.153	24.618	1.824	Leaking
8.151	24.624	1.819	Leaking
8.163	24.588	1.826	Leaking
8.127	24.694	1.822	Leaking
8.145	24.642	1.822	Leaking
8.145	24.642	1.823	Leaking
8.156	24.609	1.822	Leaking
8.151	24.622	1.822	Leaking
8.146	24.638	1.822	Leaking

Experimental Setup

The purpose of the experiment was to train and evaluate the three reinforcement learning models which includes MPPOPNMA, PPO, and DQN. A terminal with an Intel Core i5 CPU and 16GB of RAM was used for the experiment. Python 3.7.6, which comes with the Jupyter notebook IDE and the installed Stable Baselines3 library, was used to run all the models. Using the different hyperparameters and environmental parameters, individual models were trained on the customized oil pipeline environment for 400 iterations. The experiment was reproduced 6 times due to the variability in the learning process, so as obtain a more robust estimation of each model's assessment. The Average Reward and convergence rates metrics for each model was computed for comparison.

RESULTS AND DISCUSSION

The results from this study revealed substantial insights into the short-term efficiency and adaptability of the MPPOPNMA when applied to dynamic oil pipeline monitoring tasks. Considering the two-core metrics involving Average Reward and Convergence Rate, the study evaluated MPPOPNMA's capacity for real-time responsiveness and adaptability which are essential attributes for critical infrastructure where rapid detection of anomalies can prevent costly damages. The results indicate that MPPOPNMA achieves a higher Average Reward than both the baseline PPO and DQN models. Specifically, MPPOPNMA's elevated Average Reward reveals its proficiency in making optimal decisions within each iteration thereby ensuring a responsive

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

short-term approach to unusual activities detection and mitigation.

This high average reward implies that the model can rapidly recognize and respond to shifts in pipeline conditions, such as pressure, flow rate, or velocity fluctuations. In contrast, the PPO and DQN models displayed lower Average Rewards, with DQN showing the least efficient short-term performance, suggesting that these baseline models struggle to consistently achieve optimal responses in the fluctuating pipeline environment. Additionally, the Convergence Rate for MPPOPNMA was superior to that of PPO and DQN, indicating a faster and more stable adaptation to the dynamic pipeline environment. A high Convergence Rate is particularly advantageous for real-time monitoring as it shows the model's ability to learn effective decision-making policies more quickly, thus stabilizing performance sooner. This rapid convergence not only allows MPPOPNMA to reach an optimal state faster but also minimizes the risk of prolonged suboptimal behaviour, which could be critical in scenarios where immediate responses are required. PPO, while somewhat effective, exhibited a slower convergence rate compared to MPPOPNMA, suggesting that it requires more time to stabilize, which could lead to delayed responses in real-world applications. DQN, with the slowest convergence rate, demonstrated limitations in achieving prompt learning stability, likely due to its simpler architecture that is less suited for complex, dynamic environments.

Graphical analyses further visualise the model's performance considering the line plot of Average Reward over iterations as shown in Figure 2 where the MPPOPNMA reliably outperforms PPO and DQN models thereby maintaining a higher average reward throughout. This reliable performance suggests that MPPOPNMA is better equipped for scenarios requiring immediate reward optimization. Similarly, a bar chart as shown in Figure 3 comparing Average Convergence Rates (ACR) across models visually strengthens MPPOPNMA's advantage in adaptability, as it reached stable learning states faster than the other models. This implies that MPPOPNMA's customizations such as its tailored reward function and fine-tuned hyperparameters have a significant influence on its short-term efficiency and convergence abilities, allowing it to optimize action-selection strategies for both immediate and ongoing conditions in the pipeline system.

In terms of practical implications, the higher Average Reward and faster Convergence Rate of MPPOPNMA offer considerable aids for real-time pipeline monitoring. Enhanced short-term responsiveness means that the model can detect and respond to anomalies faster, reducing the likelihood of undetected leaks or blockages and, consequently, minimizing operational risks. Furthermore, the model's high adaptability makes it suitable for real-world applications where rapid changes in environmental parameters are common, contributing to both improved safety and efficiency in pipeline management. The 3 models Average Reward and Convergence Rate across simulations for the individual models beginning with MPPOPNMA, PPO and DQN are shown in Table 3, Table 4 and Table 5 respectively.

Lastly, the results highlighted MPPOPNMA's effectiveness as a monitoring tool, capable of achieving high short-term rewards while adapting quickly to the dynamic nature of pipeline environments. Its excellent performance in both average reward and convergence rate emphasizes its potential as a reliable, adaptive solution for monitoring critical infrastructure.

Table 3: MPPOPNMA Average Reward and Convergence Rate across simulations

Average Reward	Convergence Rate
29704.9454	29655.49138
33680.83706	33628.48527
32925.11536	32870.48708

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

32497.43229	32451.11006
31827.11989	31783.85628
33049.63799	32996.43282

Table 4: PPO Average Reward and Convergence Rate across simulations

Average Reward	**Convergence Rate**
26696.03222	26674.93581
29392.1524	29351.38611
27451.93297	27422.53105
29653.84189	29609.67132
26289.08783	26267.58504
29529.56406	29497.9978

Table 5: DQN Average Reward and Convergence Rate across simulations

Average Reward	**Convergence Rate**
3701.854202	3701.732533
9533.427563	9533.483772
11760.62439	11760.64024
3141.76788	3141.798376
23535.31271	23535.3972
5622.481901	5622.628723

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 12(3), SEPTEMBER, 2024
ISSN: 2277-0011; Journal homepage: www.atbuftejoste.com.ng

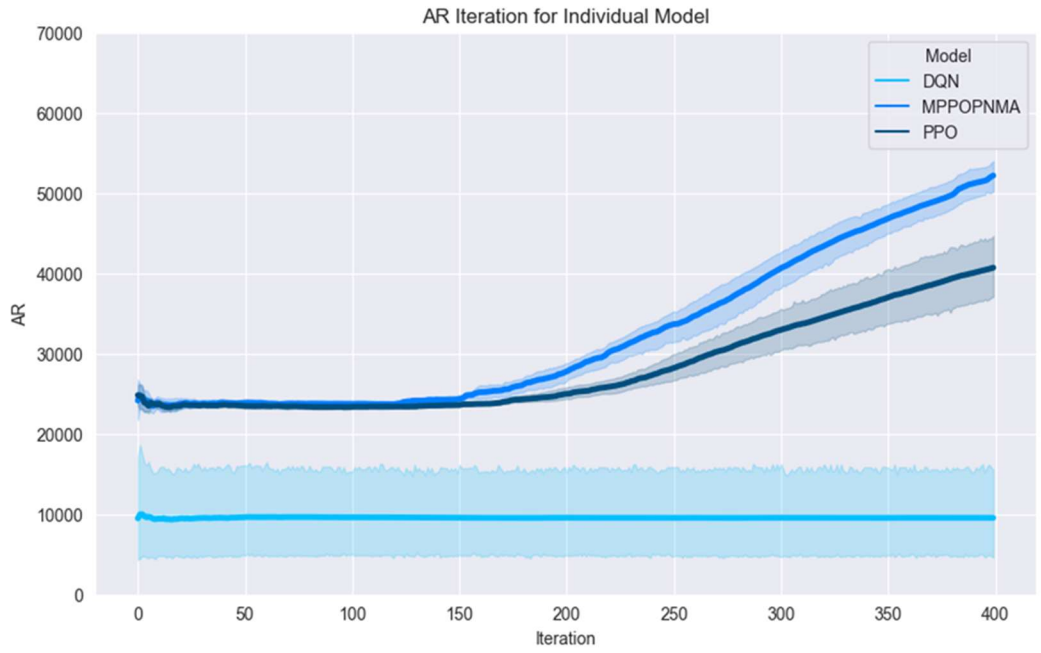

Figure 2: Average Reward for Individual Model

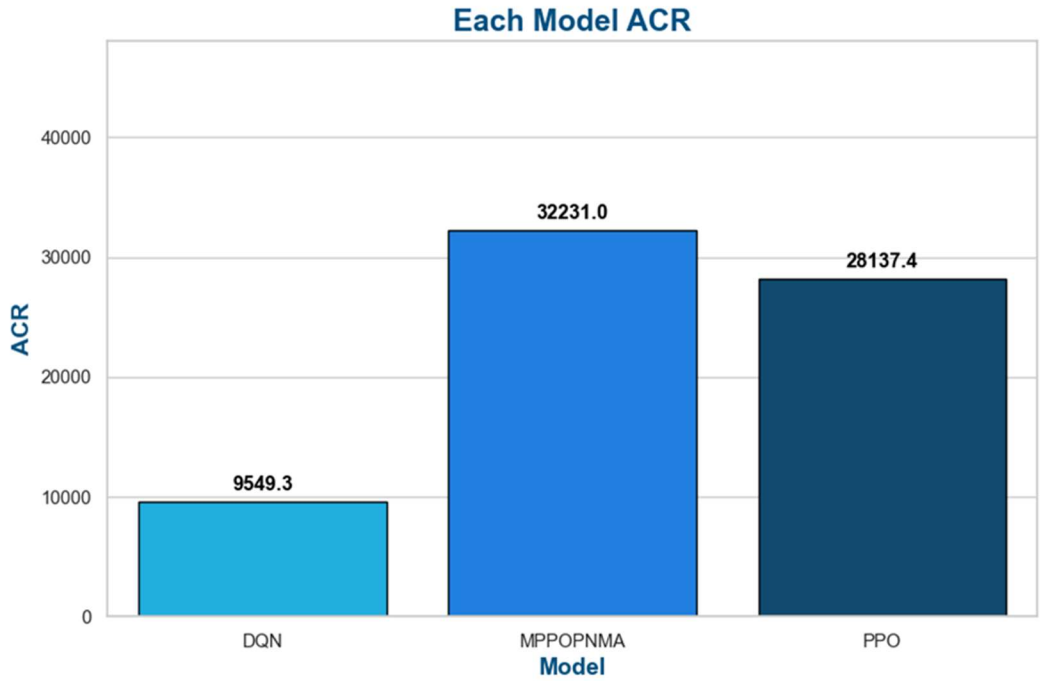

Figure 3: Average Convergence Rate for each Model

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

CONCLUSION

The study shows the ability of the MPPOPNMA in advancing real-time monitoring capabilities for oil pipeline infrastructure and by considering two major reinforcement metrics of Average Reward and Convergence Rate. The model's short-term efficiency and adaptability were assessed which is aimed at providing key insights into its performance in a dynamic environment.

MPPOPNMA consistently outperformed traditional PPO and DQN models by achieving both higher average rewards and faster convergence rates. The high average rewards reflect MPPOPNMA's ability for reliably making optimal decisions, which is crucial for timely and effective anomaly detection in pipeline monitoring. Additionally, its rapid convergence rate reveals the ability to speedily stabilize within a learning environment, allowing it to adapt to shifting conditions in the pipeline infrastructure more efficiently than the baseline models. These results underscore the model's potential as a reliable, responsive solution for pipeline monitoring, with the flexibility to handle both immediate and long-term variations in environmental conditions.

Future Work

The study similarly highlights the benefits of reinforcement learning models with specialized tuning and customized reward functions, as these adaptations allow models like MPPOPNMA to meet the distinct challenges of pipeline monitoring more efficiently. Given the high stakes of oil pipeline management where undetected abnormalities can lead to severe environmental and economic consequences, MPPOPNMA offers a promising, robust solution to increase infrastructure safety, operational efficiency, and response readiness. While MPPOPNMA have performed significantly in this study, numerous avenues for future exploration could further enhance its abilities and expand its applications. First, refining the model with additional environmental factors such as temperature fluctuations, age of the pipeline and material properties may give further insights for further optimizing the model's decision-making capabilities. These additional features could enable MPPOPNMA identify a wider variation of abnormalities and perform consistently across diverse pipeline settings. Additionally, applying transfer learning techniques could enable MPPOPNMA to rapidly adapt to new pipeline configurations or different segments thereby decreasing the need for extensive retraining and enhancing its generalizability.

More so, increasing the model's application to larger pipeline networks is another crucial step, as this would test its scalability and performance in more complex environments. This extension may include combining multi-agent frameworks where multiple MPPOPNMA models work collaboratively to monitor diverse pipeline segments. Integrating predictive maintenance abilities into MPPOPNMA could also be studied to predict potential points of failure based on historical data and observed trends thereby allowing for preventive actions before a real abnormality occurs.

REFERENCES

Badakhshan, S., Jacob, R. A., Li, B., & Zhang, J. (2023). Reinforcement Learning for Intentional Islanding in Resilient Power Transmission Systems. *2023 IEEE Texas Power and Energy Conference, TPEC 2023*, 1–6. https://doi.org/10.1109/TPEC56611.2023.10078568

Boudlal, A., Khafaji, A., & Elabbadi, J. (2024). Entropy adjustment by interpolation for exploration in Proximal Policy Optimization (PPO). *Engineering Applications of Artificial Intelligence*, *133*, 108401. https://doi.org/10.1016/J.ENGAPPAI.2024.108401

Chatterjee, D., Roy, R., & Sengupta, A. (2023). Comparison of Reinforcement Learning controller with a classical controller for an UAV. *2023 Second International Conference on Electrical, Electronics, Information and Communication*

Corresponding author: Kefas Yunana
kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

Technologies (ICEEICT), 1–5. https://doi.org/10.1109/ICEEICT56924.2023.10156999

Christian, A. B., Lin, C.-Y., Tseng, Y.-C., Van, L.-D., Hu, W.-H., & Yu, C.-H. (2022). Accuracy-Time Efficient Hyperparameter Optimization Using Actor-Critic-based Reinforcement Learning and Early Stopping in OpenAI Gym Environment. *2022 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 230–234. https://doi.org/10.1109/IoTaIS56727.2022.9975984

Ding, Y., Wang, X., Cao, X., Hu, H., & Bu, Y. (2023). A reinforcement learning method for optimal control of oil well production using cropped well group samples. *Heliyon*, *9*(7), e17919. https://doi.org/10.1016/j.heliyon.2023.e17919

Gao, X., Peng, D., Kui, G., Pan, J., Zuo, X., & Li, F. (2023). Reinforcement learning based optimization algorithm for maintenance tasks scheduling in coalbed methane gas field. *Computers and Chemical Engineering*, *170*(December 2022), 108131. https://doi.org/10.1016/j.compchemeng.2022.108131

Gléau, T. Le, Marjou, X., Lemlouma, T., & Radier, B. (2021). A Multi-agent OpenAI Gym Environment for Telecom Providers Cooperation. *2021 24th Conference on Innovation in Clouds, Internet and Networks and Workshops (ICIN)*, 28–32. https://doi.org/10.1109/ICIN51074.2021.9385538

Huan, H., Liu, L., Huan, B., Chen, X., Zhan, J., & Liu, Q. (2021). A theoretical investigation of modelling the temperature measurement in oil pipelines with edge devices. *Measurement: Journal of the International Measurement Confederation*, *168*(September 2020), 108440. https://doi.org/10.1016/j.measurement.2020.108440

Huang, Z., Gong, W., & Duan, J. (2023). TBDQN: A novel two-branch deep Q-network for crude oil and natural gas futures trading. *Applied Energy*, *347*(June), 121321. https://doi.org/10.1016/j.apenergy.2023.121321

Jaensch, F., Klingel, L., & Verl, A. (2022). Virtual Commissioning Simulation as OpenAI Gym - A Reinforcement Learning Environment for Control Systems. *2022 5th International Conference on Artificial Intelligence for Industries (AI4I)*, 64–67. https://doi.org/10.1109/AI4I54798.2022.00023

Jayanetti, A., Halgamuge, S., & Buyya, R. (2022). Deep reinforcement learning for energy and time optimized scheduling of precedence-constrained tasks in edge–cloud computing environments. *Future Generation Computer Systems*, *137*, 14–30. https://doi.org/10.1016/j.future.2022.06.012

Kahnamouei, J. T., & Moallem, M. (2024). Advancements in control systems and integration of artificial intelligence in welding robots: A review. In *Ocean Engineering* (Vol. 312). Elsevier Ltd. https://doi.org/10.1016/j.oceaneng.2024.119294

Kaloev, M., & Krastev, G. (2021). Experiments Focused on Exploration in Deep Reinforcement Learning. *ISMSIT 2021 - 5th International Symposium on Multidisciplinary Studies and Innovative Technologies, Proceedings*, 351–355. https://doi.org/10.1109/ISMSIT52890.2021.9604690

Khalid, M., Wang, L., Wang, K., Aslam, N., Pan, C., & Cao, Y. (2023). Deep reinforcement learning-based long-range autonomous valet parking for smart cities. *Sustainable Cities and*

Corresponding author: Kefas Yunana
✉ kefasyunana@nsuk.edu.ng
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

Society, *89*(June 2022), 104311. https://doi.org/10.1016/j.scs.2022.104311

Kiefer, J., & Dorer, K. (2023). Double Deep Reinforcement Learning. *2023 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, 17–22. https://doi.org/10.1109/ICARSC58346.2023.10129640

Lee, J., Seo, S., & Ko, H. (2023). UAV-Based Data Collection and Wireless Power Transfer System with Deep Reinforcement Learning. *2023 International Conference on Information Networking (ICOIN)*, 400–403. https://doi.org/10.1109/ICOIN56518.2023.10048978

Liang, J., Ma, L., Liang, S., Zhang, H., Zuo, Z., & Dai, J. (2023). Data-driven digital twin method for leak detection in natural gas pipelines. *Computers and Electrical Engineering*, *110*(174), 108833. https://doi.org/10.1016/j.compeleceng.2023.108833

Lü, X., Qian, S., Zhai, X. R., Wang, P., & Wu, T. (2024). Adaptive energy management strategy for FCHEV based on improved proximal policy optimization in deep reinforcement learning algorithm. *Energy Conversion and Management*, *321*, 118977. https://doi.org/10.1016/J.ENCONMAN.2024.118977

Najafi, S., & Lee, C. G. (2023). A deep reinforcement learning approach for repair-based maintenance of multi-unit systems using proportional hazards model. *Reliability Engineering and System Safety*, *234*(January 2022), 109179. https://doi.org/10.1016/j.ress.2023.109179

Nguyen, T.-H., Park, H., & Park, L. (2023). Recent Studies on Deep Reinforcement Learning in RIS-UAV Communication Networks. *2023 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 378–381. https://doi.org/10.1109/ICAIIC57133.2023.10067052

Reddy, K. S., & Xie, E. (2017). Cross-border mergers and acquisitions by oil and gas multinational enterprises: Geography-based view of energy strategy. *Renewable and Sustainable Energy Reviews*, *72*(November 2016), 961–980. https://doi.org/10.1016/j.rser.2017.01.016

Saha, U., Jawad, A., Shahria, S., & Rashid, A. B. M. H. U. (2024). Proximal policy optimization-based reinforcement learning approach for DC-DC boost converter control: A comparative evaluation against traditional control techniques. *Heliyon*, *10*(18). https://doi.org/10.1016/j.heliyon.2024.e37823

Silva, J. R., Euzébio, T. A. M., & Braga, M. F. (2024). Control of conventional continuous thickeners via proximal policy optimization. *Minerals Engineering*, *214*, 108761. https://doi.org/10.1016/J.MINENG.2024.108761

Vieira, S., Rocha, A., & Chaimowicz, L. (2023). Towards sample efficient deep reinforcement learning in collectible card games. *Entertainment Computing*, *47*(July), 100594. https://doi.org/10.1016/j.entcom.2023.100594

Wang, L., & Wang, X. (2024). An adaptive financial trading strategy based on proximal policy optimization and financial signal representation. *Engineering Applications of Artificial Intelligence*, *138*, 109365. https://doi.org/10.1016/J.ENGAPPAI.2024.109365

Wang, X., & Guo, H. (2021). Mobility-aware computation offloading for swarm robotics using deep reinforcement

Corresponding author: Kefas Yunana
✉ *kefasyunana@nsuk.edu.ng*
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.

learning. *2021 IEEE 18th Annual Consumer Communications and Networking Conference, CCNC 2021*, 2021–2024. https://doi.org/10.1109/CCNC49032.2021.9369456

Yang, X., Zou, Y., Zhang, H., Qu, X., & Chen, L. (2023). Improved deep reinforcement learning for car-following decision-making. *Physica A: Statistical Mechanics and Its Applications*, *624*(4800), 128912. https://doi.org/10.1016/j.physa.2023.128912

Yang, Y., Liu, Z., Saydaliev, H. B., & Iqbal, S. (2022). Economic impact of crude oil supply disruption on social welfare losses and strategic petroleum reserves. *Resources Policy*, *77*(December 2021), 102689. https://doi.org/10.1016/j.resourpol.2022.102689

Yunana, K., Oyefolahan, I. O., & Bashir, S. A. (2022). A Framework for Critical Infrastructure Monitoring Based on Deep Reinforcement Learning Approach. *Proceedings of the 5th International Conference on Information Technology for Education and Development: Changing the Narratives Through Building a Secure Society with Disruptive Technologies, ITED 2022*, 1–6. https://doi.org/10.1109/ITED56637.2022.10051520

Zeng, J., Gou, F., & Wu, J. (2023). Task offloading scheme combining deep reinforcement learning and convolutional neural networks for vehicle trajectory prediction in smart cities. *Computer Communications*, *208*(July 2022), 29–43. https://doi.org/10.1016/j.comcom.2023.05.021

Zhang, S., Zhang, A., Chen, P., Li, H., Zeng, X., Chen, S., Dong, T., Shi, P., Lang, Y., & Zhou, Q. (2024). Application of artificial intelligence hybrid models in safety assessment of submarine pipelines: Principles and methods. *Ocean Engineering*, *312*, 119203. https://doi.org/10.1016/J.OCEANENG.2024.119203

Zhang, T., Lam, K. Y., & Zhao, J. (2023). Deep reinforcement learning based scheduling strategy for federated learning in sensor-cloud systems. *Future Generation Computer Systems*, *144*, 219–229. https://doi.org/10.1016/j.future.2023.03.009

Zhu, J., Zhang, W., Yu, L., & Guo, X. (2024). A novel multi-attention reinforcement learning for the scheduling of unmanned shipment vessels (USV) in automated container terminals. *Omega*, *129*, 103152. https://doi.org/10.1016/J.OMEGA.2024.103152.

Corresponding author: Kefas Yunana
✉ *kefasyunana@nsuk.edu.ng*
Department of Computer Science, Nasarawa State University, Keffi, Nigeria.