Heliyon 10 (2024) e30697

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Systematic review and meta-analysis

Deep deterministic policy gradient algorithm: A systematic review

Ebrahim Hamid Sumiea [a,b,*], Said Jadid Abdulkadir [a,b], Hitham Seddig Alhussian [a,b], Safwan Mahmood Al-Selwi [a,b], Alawi Alqushaibi [a,b], Mohammed Gamal Ragab [a,b], Suliman Mohamed Fati [c]

[a] Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Perak, Malaysia
[b] Center for Research in Data Science (CeRDaS), Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Perak, Malaysia
[c] Information Systems Department, Prince Sultan University, Riyadh, Saudi Arabia

ARTICLE INFO

Dataset link: https://github.com/SafwanAlselwi/DDPG

Keywords:
Deep deterministic policy gradient (DDPG)
Deep reinforcement learning (DRL)
Hyperparameter
Optimization

ABSTRACT

Deep Reinforcement Learning (DRL) has gained significant adoption in diverse fields and applications, mainly due to its proficiency in resolving complicated decision-making problems in spaces with high-dimensional states and actions. Deep Deterministic Policy Gradient (DDPG) is a well-known DRL algorithm that adopts an actor-critic approach, synthesizing the advantages of value-based and policy-based reinforcement learning methods. The aim of this study is to provide a thorough examination of the latest developments, patterns, obstacles, and potential opportunities related to DDPG. A systematic search was conducted using relevant academic databases (Scopus, Web of Science, and ScienceDirect) to identify 85 relevant studies published in the last five years (2018-2023). We provide a comprehensive overview of the key concepts and components of DDPG, including its formulation, implementation, and training. Then, we highlight the various applications and domains of DDPG, including Autonomous Driving, Unmanned Aerial Vehicles, Resource Allocation, Communications and the Internet of Things, Robotics, and Finance. Additionally, we provide an in-depth comparison of DDPG with other DRL algorithms and traditional RL methods, highlighting its strengths and weaknesses. We believe that this review will be an essential resource for researchers, offering them valuable insights into the methods and techniques utilized in the field of DRL and DDPG.

1. Introduction

Reinforcement Learning (RL) is an artificial intelligence domain, which focuses on making decisions through learning the optimal behavior in environments to maximize a rewards signal [1]. In RL, an agent interacts with an environment and receives feedback in the form of rewards, which in turn is used to update the decision-making policy [2] [3] [4]. This approach has been successfully applied to a wide range of applications such as game development [5], anomaly detection [6], robotics [7], and autonomous control [8].

* Corresponding author at: Department of Computer and Information Sciences, Universiti Teknologi PETRONAS, Seri Iskandar, 32610, Perak, Malaysia.
E-mail addresses: ebrahim_22006040@utp.edu.my (E.H. Sumiea), saidjadid.a@utp.edu.my (S.J. Abdulkadir), seddig.alhussian@utp.edu.my (H.S. Alhussian), safwan_21002827@utp.edu.my (S.M. Al-Selwi), alawi_18000555@utp.edu.my (A. Alqushaibi), mohd.gamal_20497@utp.edu.my (M.G. Ragab), sgaber@psu.edu.sa (S.M. Fati).

https://doi.org/10.1016/j.heliyon.2024.e30697
Received 31 January 2024; Received in revised form 26 April 2024; Accepted 2 May 2024

Available online 7 May 2024
2405-8440/

To enhance the performance of RL in the situation of high-dimensional state space, the integration of DL and RL whereby Deep Neural Network (DNN) represents the agent's decision-making policy [9]. Such an integration, which is named Deep Reinforcement Learning (DRL), has led to a significant advancement in the field and has enabled the progress of highly effective algorithms for decision-making in complex environments [9,10]. Deep Deterministic Policy Gradient (DDPG) is One of the popular DRL algorithms [11], which merges the strengths of both value-based and policy-based RL, following an actor-critic approach [12].

In DDPG, the actions are generated using the actor network, and the generated actions will be evaluated using the critic network [11,13]. Both actor and critic networks are trained simultaneously to optimize the policy and value-based functions [14]. Thus, the algorithm has been shown to be highly scalable, with the ability to handle problems with millions of parameters and complex non-linear dynamics. Additionally, due to its ability to deal with high-dimensional state and action spaces and its stability and convergence properties, DDPG has been widely used in various domains, including robotics [15], simulation-based tasks [16], energy management [17], and control problems [18,19]. Notably, one of the critical assets of DDPG is its ability to learn deterministic policies, which are functions that map states to specific actions [11]. Deterministic policies are desirable in many real-world applications, as they provide more interpretable and reliable control compared to stochastic policies [20]. Another advantage of DDPG is its ability to learn from raw sensory inputs, such as images, without needing hand-engineered features or representation learning [21]. This has enabled the algorithm to be applied to a wide range of challenging tasks, including video games, robotic manipulation, and autonomous navigation.

Therefore, the motivation for writing this systematic review paper is the importance of DDPG, as an emerging and powerful tool, for decision-making in complex environments due to its ability to handle high-dimensional states and action spaces, and its stability and convergence properties. Although it is a fact that DDPG has proven useful in a wide range of industrial applications, it still suffers from some drawbacks, such as overestimation bias, overly sensitive parameters, and exploration versus exploitation dilemmas. This review aims to investigate these drawbacks and conduct a comprehensive and up-to-date analysis of the available solutions in the literature. To the best of the authors' knowledge, this is the first systematic review about the DDPG and the main contributions can be summarized as follows:

1. Conducting an extensive and revised study mapping process that consists of five steps (shown in Fig. 4) to write this systematic review.
2. Summarizing the state-of-the-art in the field of DDPG, highlighting its key contributions, advantages, and limitations.
3. Providing a comprehensive overview of the various applications of DDPG, including its successes and challenges in different domains, such as robotics, game-playing, and autonomous control.
4. Highlighting the recent advances and developments in the field of DDPG, including new algorithms and techniques that have been proposed to address its limitations and improve its performance.
5. Providing insights into the future directions of the field, including potential new applications, improvements to the existing algorithms, and the integration of DDPG with other machine learning techniques.

We strongly believe that the latest progress in this area serves as a strong motivation to examine and discuss current methodologies, applications, patterns, research issues, and future research directions. Therefore, we identified specific research questions (RQs) (Sub-section 3.2) that are closely linked to the core objective, which allowed us to structure our study around a compelling central idea.

The rest of this research work is structured as follows. Section 2 highlights the literature's related studies. Section 3 explains the approach utilized to perform this systematic literature review. Section 4 provides results using synthesized data from the included research, and examines each RQs. Finally, Section 5 discusses the study's merits and limitations and brings the research to a conclusion.

2. Related work

DDPG is a model-free, off-policy RL algorithm that adopts an actor-critic approach to solving continuous control problems in which the action space is continuous. Fig. 1 illustrates the DDPG structure. It was introduced in 2015 by Lillicrap et al. [22] and builds upon the Q-learning and actor-critic algorithms. It memorizes the Q-function utilizing off-policy data point, the Bellman equation, and then employs the Q-function to learn the policy. This method is precisely related to Q-learning and is defined similarly as the most desirable action-value function. Then, in each given state, the best action is $a*(s)$ refers to the optimal action-value function that can be learned by resolving [23,22].

$$a^*(s) = \arg\max_a Q^*(s,a) \tag{1}$$

DDPG interrupts the learning approximator toward $Q^*(s,a)$ along with learning an approximator toward $\pi^*(s)$. It performs this in a manner that is particularly well-suited for environments with continuous action spaces. However, it is common knowledge that DDPG is well-suited for environments with continuous action spaces. It calculates the maximum over actions as $\max_a Q^*(s,a)$. When dealing with a reduced number of discrete actions, finding the maximum poses no dilemma, as it can simply compute the Q-values for each action individually and directly compare these Q-values. Thus, it instantly provides the action that maximizes the Q-value. However, when the action space is continuous, exhaustively exploring the space and solving the optimization problem becomes highly non-trivial [22]. Using a standard optimization method to calculate $\max_a Q^*(s,a)$ would be prohibitively costly.

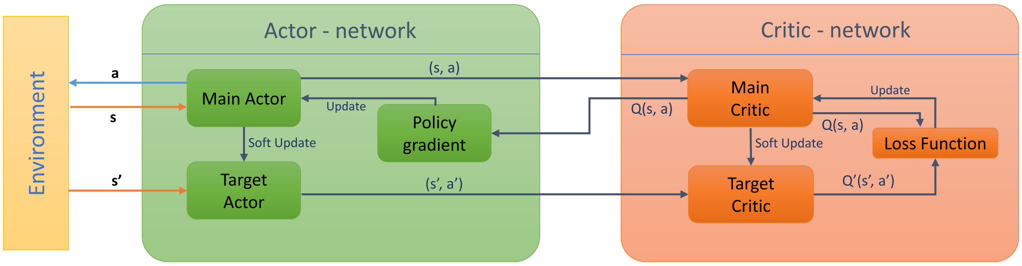

Fig. 1. DDPG algorithm structure.

Consequently, it would be expected to run this costly function every time agents need to take action in the environment, which is intolerable. Given that the action spaces are continuous, the function $Q(s,a)$ is expected to be differentiable with respect to the action argument. This enables the establishment of an efficient, gradient-based learning rule for a policy $\pi(s)$. As a result, instead of performing costly optimization functions every time, $\max_a Q(s,a)$ can be approximated using available information [24].

$$\max aQ(s,a) = Q(s,\mu(s)) \tag{2}$$

Through this paper, we aim to provide a comprehensive and systematic understanding of DDPG and its variants, which could serve as a valuable resource for researchers and practitioners in the field of RL. Hence, the upcoming subsections provide a comprehensive overview of DDPG and its variants. Specifically, we present a detailed discussion of the Q-learning sides of DDPG (Sub-section 2.1), which includes the actor-critic architecture, the target networks, and the experience replay mechanism. We then delve into the policy learning sides of DDPG (Sub-section 2.2), which involves the use of the actor network to learn the optimal policy through gradient ascent. Finally, we review some extensions and modifications of DDPG (Sub-section 2.3), such as Prioritized Experience Replay (PER) and Twin Delayed DDPG (TD3), which have been proposed to enhance the performance and stability of the algorithm.

2.1. The Q-learning sides of DDPG

The Bellman-based formula explaining the optimal actions and value functions, $Q*(s,a)$ is presented via the following equation:

$$Q^*(s,a) = \underset{s'\sim P}{\mathrm{E}}\left[r(s,a) + \gamma \max a' Q^*\left(s',a'\right)\right] \tag{3}$$

The notation "$s' \sim P$" means that the next state, s', is generated by the environment from a probability distribution P given the current state s and action a. The Bellman-based formula is the starting point for developing an approximate function for $Q^*(s,a)$. Assuming that the approximator is a neural network $Q^{\phi}(s,a)$ with parameters ϕ and a set of transitions $\mathcal{D} = (s,a,r,s',d)$, where d denotes whether the state s' is terminal, a mean-squared Bellman error (MSBE) function can be constructed to measure how well Q^{ϕ} satisfies the Bellman-based equation:

$$L(\phi,\mathcal{D}) = \mathbb{E}\left[(s,a,r,s',d)\sim\mathcal{D}\right]\left[\left(Q_{\phi}(s,a) - \left(r + \gamma(1-d)\max_{a'} Q_{\phi}(s',a')\right)^2\right]\right. \tag{4}$$

Deep Q-Network (DQN) and all of its variants, as well as DDPG, are Q-learning-based algorithms for function approximators that are primarily based on minimizing MSBE loss functions [25]. Here are two primary techniques utilized by Schulman [24] that are worth explaining, and then a precise detail for DDPG.

First Technique: Replay Buffer. The standard algorithms utilized for training a DNN to approximate $Q*(s,a)$ all rely on an experiences replay buffer, which comprises a set of experiences denoted as $\mathcal{D}$. To make the algorithm function reliably, the replay buffer needs to be of sufficient size to encompass a broad range of experiences. However, retaining all experiences within the buffer may not always be beneficial. Simply relying on the most recent data will result in overfitting and ultimately break the system, while using too much experience may hinder the learning process. Therefore, fine-tuning may be necessary to achieve optimal balance.

Second Technique: The target value. Q-learning algorithm makes utilization of networks that are targeted. This target Network is given as follows:

$$r + \gamma(1-d)\max a' Q_{\phi}\left(s',a'\right) \tag{5}$$

This function is referred to as the target to reduce the MSBE loss, the Q-function is modeled after it. Inconveniently, the aim is dependent on the same parameters used to train. Thus, MSBE reduction becomes unstable. The idea is to utilize a set of parameters that approaches, although with a time delay, to respond to a second network, known as the target network, which lags behind the first. The target network's specifications are shown ϕ_{targ}.

In DQN algorithms, the target-based network is replicated from the original network after a predetermined number of steps. Polyak averaging is employed to modernize the target network once every main network update in DDPG-style algorithms [26].

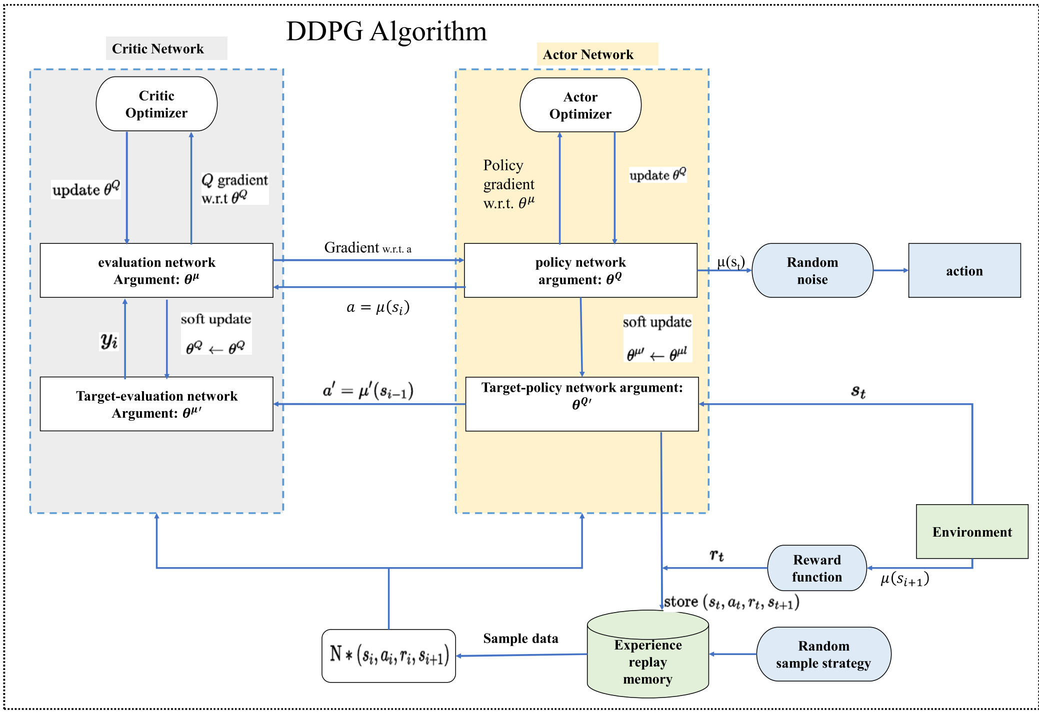

Fig. 2. The learning policies of the DDPG algorithm.

$$\phi_{\text{targ}} \leftarrow \rho\phi_{\text{targ}} + (1-\rho)\phi \tag{6}$$

Where p is a hyperparameter between 0 and 1 that is often close to 1, this hyperparameter is named polyak. The DDPG Feature calculates the MaxQv er actions in the Targets. As discussed previously, calculating the maximum over actions in the targets is a hard task in continuous action spaces. DDPG deals with this by way of utilizing the target policy network to calculate actions that approximately maximize $Q_{\phi_{\text{targ}}}$.

The target policy-based network is found to be similar to the target Q-function: it is achieved by taking a polyak average of the policy parameters during training. Simultaneously, in DDPG, Q-learning is performed by minimizing the MSBE losses using stochastic gradient descent, as described in the following equation:

$$L(\phi, \mathcal{D}) = \underset{(s,a,r,s',d)\sim\mathcal{D}}{\mathrm{E}} \left[\left(Q_\phi(s,a) - \left(r + \gamma(1-d) Q_{\phi_{\text{targ}}}\left(s', \mu_{\theta_{\text{targ}}}(s')\right)\right)\right)^2 \right] \tag{7}$$

Where $\mu_{\theta_{\text{targ}}}$ is the targets policy.

2.2. The policies learning sides of DDPG

Policy acquisition with DDPG is straightforward. To discover a deterministic strategy (s) that maximizes $Q^\theta(s,a)$. Assuming the Q-functions are differentiable regarding action and that the actions, and spaces are continuous, gradient ascent concerning policy parameters alone may be used to solve the problem.

$$\max \theta E_{s\sim\mathcal{D}}\left[Q_\phi\left(s, \mu_\theta(s)\right)\right] \tag{8}$$

Fig. 2 represents the DDPG algorithm learning policies. The algorithm of DDPG involves two neural networks: The actor and the critic networks and they have distinct roles.

The actor network is in charge of acquiring knowledge about the best policy function, which is used to map states to actions. The policy function is a representation of map states to actions. The actor network receives the current state as input and produces an action as output, which is then transmitted to the environment.

In contrast, the critic network has the responsibility of learning the Q-value function, which represents the expected total reward that results from taking a specific action in a particular state and following the policy thereafter. The critic network takes in the current state and action as input and produces the corresponding Q-value as output. The Q-value function is utilized to assess the quality of the action performed in a particular state. By calculating the gradient of the Q-value function with respect to the actions

Table 1
Developments and expansions of DDPG algorithms.

Authors and year	Motivations	Contributions	Advantages	Limitations
(Fujimoto et al., 2018a) [27]	TD3 addressed the overestimating of DDPG by using two value functions and by delaying the policy updates.	TD3 helps to prevent overestimating action values.	Solving a wide range of DRL problems, especially in the domain of continuous control.	Computational complexity, sensitivity to hyperparameters, exploratory behavior, limited sample efficiency.
(Haarnoja et al., 2018) [28]	SAC was proposed to overcome the limitations of the traditional actor-critic algorithms.	SAC addresses unstable or struggle to find good policies in complex environments.	The maximum entropy objective helps to encourage exploration.	Difficulty in handling Partial Observability.
(Lowe et al., 2017) [29]	MADDPG was proposed to address the challenges posed by multi-agent reinforcement learning.	MADDPG introduces a centralized training and decentralized execution framework for multi-agents.	MADDPG can handle multi-agents reinforcement learning scenarios with a large number of agents.	MADDPG increases the complexity of the algorithm.
(Barth-Maron et al., 2018) [30]	D4PG was proposed to develop a distributed framework for off-policy learning and improve the performance of DRL algorithms on various control tasks.	D4PG combines the distributional perspective on reinforcement learning with a distributed framework for off-policy learning.	Improves the stability and convergence of the algorithm.	The D4PG algorithm is computationally expensive due to its distributed nature.
(Du et al., 2019) [31]	D3PG modification was proposed to increase the convergence rate of both the actor and the critic.	The use of true gradients from a differentiable physical simulator.	Improve the performance without sacrificing its robustness.	Physical simulators may introduce additional computational overhead.

executed by the actor network, the critic network provides feedback to it. The DDPG algorithm's learning procedures entail modifying the weights of both the actor network and the critic network via gradient descent.

2.3. Extensions and modifications of DDPG

DDPG has several limitations, such as the tendency to get stuck in local optima, sensitivity to hyperparameters, and difficulty handling continuous action spaces [22]. Hence, there are several extensions and modifications of DDPG as shown in Table 1. By extending and modifying DDPG, researchers aim to overcome these limitations and improve the performance of the algorithm. Moreover, some other reasons for doing this are:

- **Improving robustness**: DDPG can be sensitive to the choice of hyperparameters and the initialization of the neural networks [27]. Extensions and modifications of DDPG aim to improve the robustness of the algorithm and make it more stable and reliable in different environments.
- **Handling new challenges**: The field of DRL is constantly evolving, with new challenges and applications being proposed. Extensions and modifications of DDPG aim to address these new challenges and make the algorithm better suited to these new domains.
- **Incorporating new ideas**: DRL is an interdisciplinary field, with ideas from machine learning, control theory, and artificial intelligence being incorporated into the development of new algorithms. Extensions and modifications of DDPG aim to incorporate these new ideas and make the algorithm more effective and efficient [27–31].

In Fig. 3, DDPG is the central node, and the other algorithms are arranged in a tree-like structure around it, which explains each algorithm and its relationship to DDPG. DDPG is the base algorithm, and all of the other algorithms in the diagram are extensions of DDPG. Overestimation of Q-values in algorithms like DDPG arises from factors like function approximation and max operator bias.

- **TD3** algorithm by Fujimoto et al. [27] mitigates this by employing two critic networks, using the minimum Q-value of both for updates. This approach reduces the overestimation bias, ensuring more stable and accurate value estimations.
- **SAC** algorithm by Haarnoja et al. [28] extends DDPG by using entropy regularization to encourage exploration and prevent premature convergence. It also uses a soft Q-function instead of a deterministic Q-function, allowing for more policy optimization flexibility. TD3-SAC algorithms [27,28] combine the best parts of TD3 and SAC to achieve improved stability and better exploration.
- **D4PG** algorithm by Barth-Maron et al. [30] extends DDPG by using a distributional critic to estimate the value distribution, improving stability and reducing bias.
- **D3PG** algorithm by Dong et al. [31] extends DDPG by using a separate network to learn the policy directly rather than using the Q-value to derive the policy. This helps the agents to learn from a broader range of experiences and can lead to faster learning.
- **MADDPG** algorithm by Lowe et al. [29] extends DDPG to multi-agent settings, where multiple agents learn concurrently in a cooperative or competitive environment.

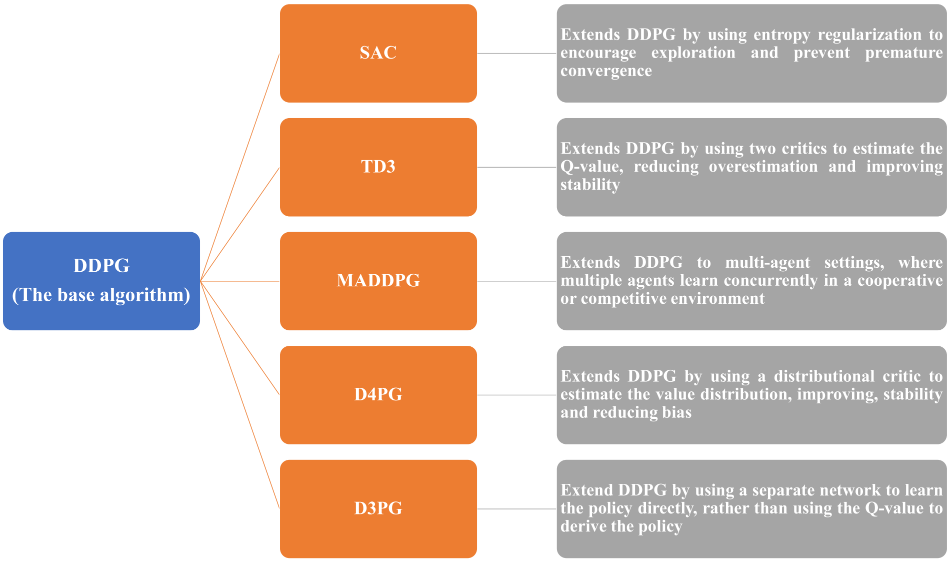

Fig. 3. Extensions and Modifications of DDPG.

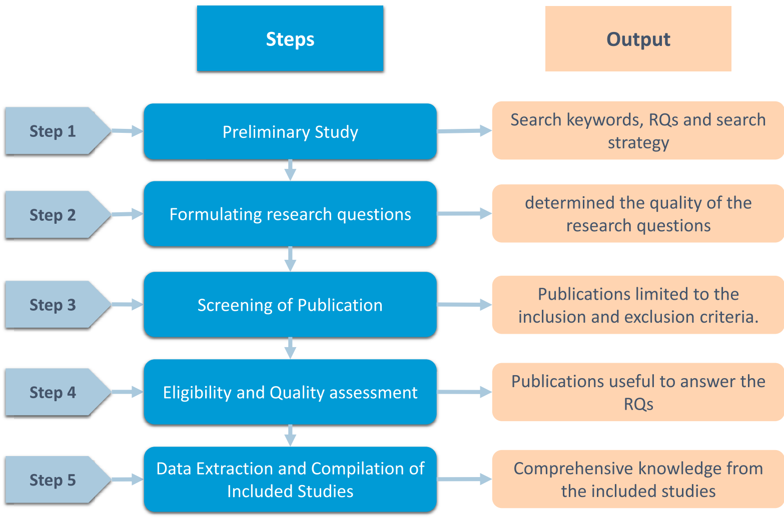

Fig. 4. The literature review study mapping process.

3. Methodology

This section explains the approach utilized to perform this systematic literature review. The PRISMA SLR recommendations by Page et al. [32] served as the basis for the methods employed in this study. Fig. 4 illustrates the improved mapping method used in this work which, in the rest of this section, consists of five steps: preliminary study (3.1), formulating research questions (3.2), screening of publication (3.3), eligibility and quality assessment (3.4), and data extraction and compilation of included studies (3.5).

3.1. Step 1: preliminary study

Preliminary research was conducted before the literature review to better grasp the primary issue under discussion. This stage also acts as the authors' "kick-off" to uncover pertinent topics, keywords, and the scope of their systematic study. Next, two sub-tasks were conducted: finding keywords, and figuring out the search criteria and strategies.

3.1.1. Keywords identification

Using the keyword "DDPG" as the starting point of our search, we conducted a Google Scholar search with the keyword to see how many studies are available on this topic before selecting keywords and keyword variations. Following this step, we found four studies that may provide insight into the RQs [14,22,30,31]. In addition to identifying a few relevant keywords, the retrieved studies

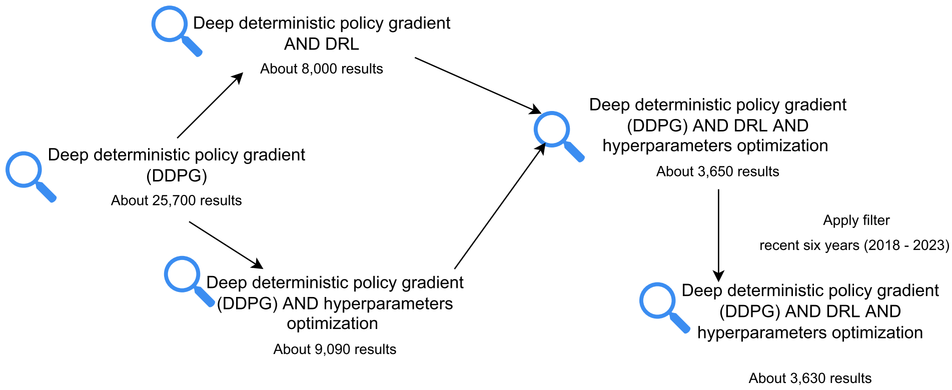

Fig. 5. A summary of Google Scholar's keyword combination survey.

Table 2
Search criteria and strategies.

Domain focus	Keywords	Generalizing search string
Deep Reinforcement Learning, Deep Deterministic Policy Gradient Algorithm.	DDPG, deep deterministic policy gradient, deep reinforcement learning, actor-critic, continuous control, DDPG optimization, hyperparameter tuning.	(DDPG OR "deep deterministic policy gradient") AND (hyperparameter OR optimize*))

provided a general idea of the search venue. In order to determine the most appropriate keyword combinations for a search string, keywords were analyzed to determine which combinations returned the most relevant articles related to DDPG. A summary of the keyword combination survey results is presented in Fig. 5, along with the total number of publications retrieved from Google Scholar during the literature search.

3.1.2. *Identify search criteria and strategies*

To proceed further, we narrowed down the search by using specific criteria to focus on particular areas. The most recent database search was completed on 03/03/2023 on three established databases, namely Scopus, Web of Science, and ScienceDirect. The search was conducted by searching for titles, abstracts, and author keywords to retrieve relevant studies. Only publications such as journal articles, conference proceedings, and book chapters published between 2018 and 2023 were included in the search. A total of 986 studies were identified and to ensure that the search results were pertinent, these studies were screened to verify that they contained information that answered the RQs stated in Sub-section 3.2. Advanced search operators and wildcards were used with the search keywords, following the manual for each database (Appendix A). The keywords, generalized search strings, and domain focus used for the search are outlined in Table 2.

A PRISMA flowchart depicts the flow of information during the different phases of this SLR is illustrated in Fig. 6. This flowchart was designed using the PRISMA tool by Haddaway et al. [33].

3.2. *Step 2: formulating research questions*

In this step, we have created the primary research inquiries that will guide our research and writing processes. These inquiries were assessed based on their constructive nature, level of focus, and relevance to a particular area or issue. As a result, we have conducted extensive research on the advancement of the DDPG algorithm and its extensions. However, it has been difficult to find adequate literature on the novel DDPG, making it a challenging task. Therefore, the primary objective of this study is to present a comprehensive review of the development of the DDPG, its optimization methods, and the areas of its application.

The formulated RQs for this study are as follows, along with their justifications:

- **RQ1:** What are the current applications and domains into which the DDPG algorithm has been proposed in the literature?
  Motivation: Classification of selected studies is a way to organize research papers based on their contribution to a particular field. This classification can help researchers and readers quickly understand the nature and significance of the study and identify trends and patterns in the research.
- **RQ2:** What are the commonly applied techniques with DDPG in DRL applications?
  Motivation: To provide a comprehensive overview of the techniques and algorithms that are used with the DDPG algorithm to develop DRL applications. Understanding these techniques can help practitioners and researchers understand the strengths and limitations of the algorithm and make informed decisions about its suitability for a particular problem.
- **RQ3:** What are the optimization methods used to overcome the instability of hyperparameters in DDPG?
  Motivation: To provide a comprehensive overview of the various approaches that have been proposed to address the challenges associated with hyperparameter optimization in DRL algorithms.

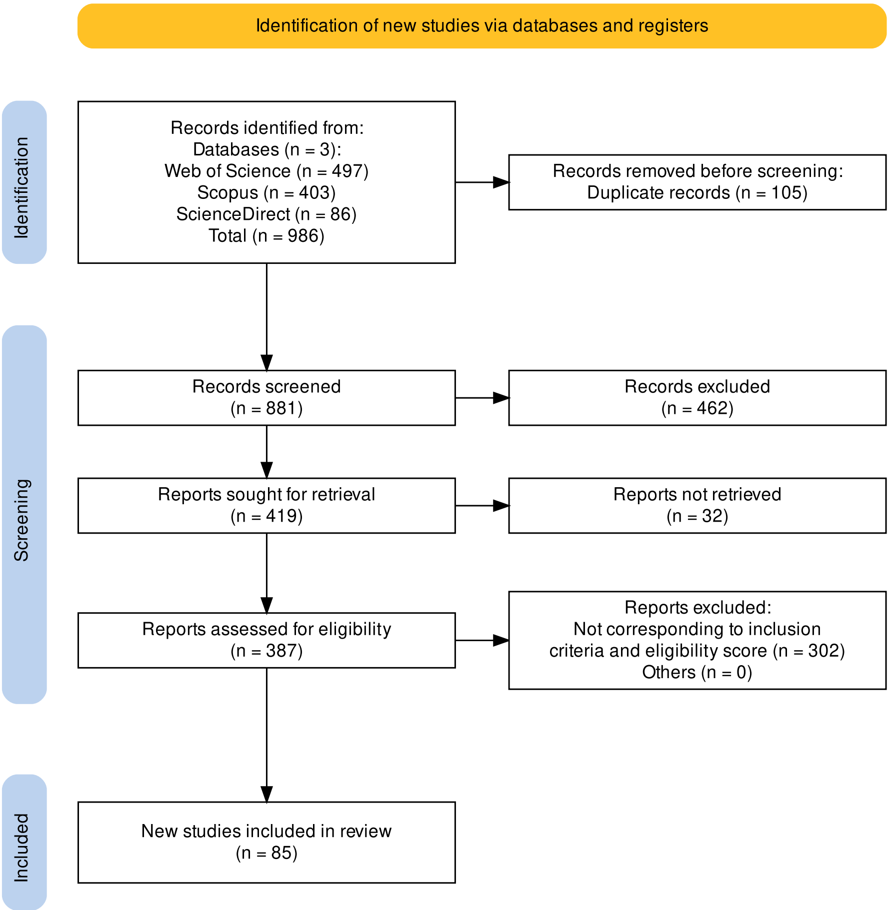

Fig. 6. PRISMA flowchart of the systematic literature process.

- **RQ4:** What are the evaluation measures used to evaluate the performance of the DDPG algorithm?
  Motivation: To provide insight into the criteria used to assess the effectiveness and accuracy of the DDPG algorithm in various DRL applications.
- **RQ5:** What is the intensity of publications related to DDPG?
  Motivation: To provide visualization insights on the intensity of DDPG publications based on yearly published papers and featured journals, so researchers know the trend line of DDPG and find the best journals to fill their knowledge.

3.3. Step 3: screening of publication

After identifying 986 studies and removing a total of 105 duplicates, a two-step screening process was used to screen publications. The first step was to screen the studies based on inclusion and exclusion criteria (Table 3). A screening process was conducted to exclude any unrelated works, 462 studies, that have not met the selection criteria from the identified papers. The second step was using the titles, abstracts, and conclusions of the studies, the literature was evaluated by a formal analytical and data curation team. To ensure that no important studies were missed, each literature piece was thoroughly examined based on its relevance to the topic. RQs and themes of the study were taken into consideration when selecting the studies. This step was carried out by E.H. Sumiea, and S.M. Al-Selwi and they worked together to finalize it. Finally, the screening phase ended up with 387 unique studies for the next phase, the eligibility and quality assessment.

3.4. Step 4: eligibility and quality assessment

The SLR investigation involved evaluating the eligibility and quality of the remaining 387 studies that were screened, using specific criteria outlined in Table 4. These criteria included score values of 1 (indicating agreement), 0.5 (partially agreeing), and 0 (disagreeing), and were used to assess whether the studies had clearly defined constraints, procedures, goals, and aims. This was done to ensure that only the most suitable research was included in the final selection.

Table 3
Inclusion criteria and exclusion criteria.

Inclusion Criteria	Exclusion Criteria
❖ Studies published between 2018–2023.	❖ Published before 2018.
❖ The study is available in full text.	❖ Unavailable in full text.
❖ The study is Written in the English language.	❖ Written in a different language than English.
❖ Relevant to the RQs.	❖ Irrelevant to the RQs.
❖ Papers that are more than five pages.	❖ Less than five pages.

Table 4
The set of standards used to evaluate the eligibility and quality of assessment.

Criteria	Score	Description
Does the study have a defined set of aims and objectives?	**1**	There is a clear objective and goal presented in the study.
	0.5	The study's objectives are defined, but the study's goals are not.
	0	Objectives and goals are not clearly stated in the study.
Is the methodology of the study presented clearly?	**1**	The methodology presented in this study is clear, systematic, and well-documented.
	0.5	Incomplete/non-systematic methodology documentation is present.
	0	There is no documentation of the methodology in the study.
Does the study disclose any of the work's limitations?	**1**	Yes, the report does include a widely noted weakness.
	0.5	The research briefly mentions its drawbacks.
	0	No, the study does not include a statement of the research's limitations.
Does the study effectively articulate its findings?	**1**	Yes, the study presents clear, comprehensive, and well-presented research findings. The findings/results are presented using appropriate visualizations.
	0.5	Although the study's findings were reported, more context should be given. The results are related to the data supplied.
	0	No, the study does not clearly communicate its research findings, and/or no more explanation is given. The results are given in a random sequence and do not affect the study's aims and objectives.

Table 5
Studies obtained at each stage of the systematic literature process.

Databases of academics	Studies Identified	Studies Screened	Studies Passed EA	Studies Included
Scopus	403	45	52	52
ScienceDirect	86	10	1	1
Web of Science	497	50	32	32
Total	986	105	85	85

Each study's eligibility and quality assessment phases were conducted using the grading criteria listed in Table 4. To guarantee that only high-quality research was included in the final list, only papers with a minimum score of 3.0 were chosen. A total of 302 research papers were removed in this phase, leaving 85 eligible studies to be included in the qualitative synthesis of this SLR study.

3.5. Step 5: data extraction and compilation of included studies

Once the eligibility and quality evaluations were finished, a spreadsheet was created using metadata obtained from scholarly databases to list the selected studies. These publications include details such as the title, authors, and publication venue. Additionally, information was included about the type of paper (i.e., review, application, or survey), the year of publication, the digital object identifier (DOI), the study type, and the scholarly database from which it was retrieved. The EndNote software has been utilized to automatically remove duplicate studies and keep offline copies for future reference and citation. This step allowed for the extraction of knowledge that could contribute to answering the RQs (Sub-section 3.2) and achieving the study's objectives. The extraction process was done by E.H. Sumiea, and S.M. Al-Selwi and they worked together independently to extract the data, and then they reviewed each other work to make sure the extracted data was accurate. In total, 85 studies have been included and analyzed to answer the established RQs in this systematic review. The most valuable data obtained from these studies pertained to existing DDPG techniques and their enhanced methods, applications, research trends, and current challenges in the DDPG and DRL domains. Table 5 shows the number of studies collected from scholarly databases at each level of the SLR process.

For consideration for inclusion in the finalized selected papers, our SLR only examined studies that had defined aims and goals, offered clear methodological justifications, acknowledged their limits, and presented unambiguous study findings. As a consequence, Table 5 data draws us to the conclusion that most studies were found in Scopus (N = 52), then Web of Science (N = 32), and finally ScienceDirect (N = 1).

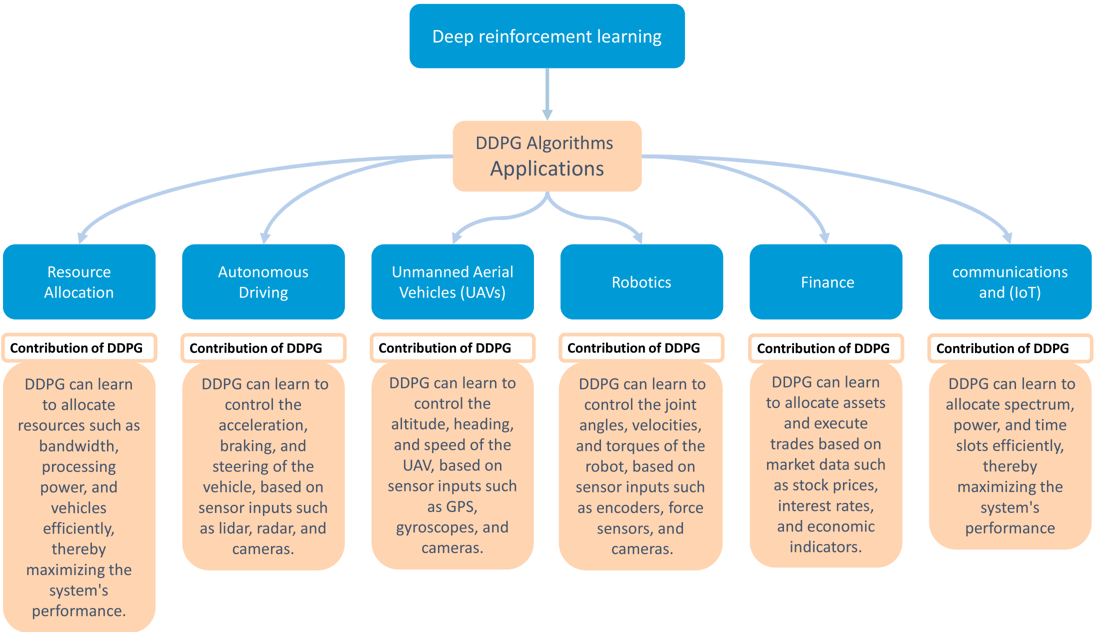

Fig. 7. DDPG applications.

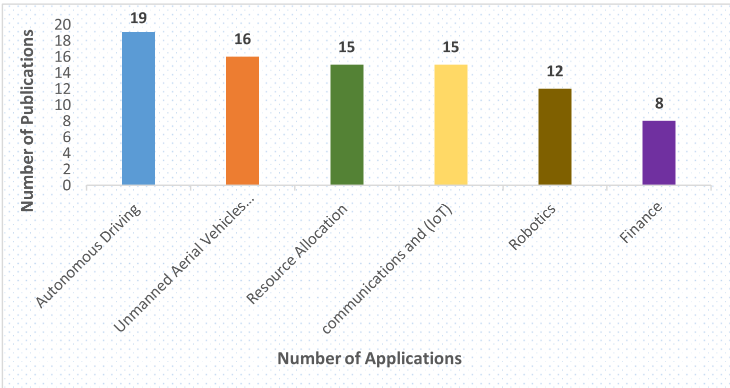

Fig. 8. Classification of the included studies based on their utilization of the DDPG algorithm.

4. Synthesis of data and analysis

Using visualization aids and answers to this study's RQs (3.2), this section synthesizes and summarizes the data collected from the included works. It presents evidence related to recent applications, current research, and challenges associated with DDPG-DRL to deliver to both novice and experienced researchers in order to understand the current state of DDPG applications, current research, and challenges.

4.1. RQ1: what are the current applications and domains into which the DDPG algorithm has been proposed in the literature?

A comprehensive classification and overview of the 85 selected studies is presented as an answer to this question. Classification of selected studies is a way to organize research papers based on their contribution to a particular field. Fig. 7 shows that the DDPG algorithm has found many applications in different fields. This classification can help researchers and readers quickly understand the nature and significance of the study and identify trends and patterns in the research.

Also, Fig. 8 displays the classification of the chosen studies according to their application area. The largest proportion of studies was focused on Autonomous Driving (N = 19), followed by Unmanned Aerial Vehicles (UAVs) (N = 16), Resource Allocation (N = 15), communications and the Internet of Things (IoT) (N = 15), Robotics (N = 12), and Finance (N = 8). This suggests that, compared to other applications that utilized the DDPG algorithm, these six areas have received more attention in terms of exploring the potential of AI techniques to enhance DRL and optimization methods.

For organizational clarity, these studies have been systematically classified into six distinct sub-sections, all predicated on the utilization of the DDPG algorithm. Each group encompasses details such as the respective authors, publication years, employed techniques, adopted methodologies, and specific applications. The sub-sections are Resource Allocation (4.1.1), Autonomous Driving (4.1.2), Unmanned Aerial Vehicles (UAVs) (4.1.3), Robotics (4.1.4), Communications and IoT (4.1.5), as well as Finance (4.1.6).

4.1.1. Summary of studies classified as resource allocation

DDPG is widely employed for resource allocation problems in cloud computing [34] and energy management [35] to optimize resource allocation, such as computing resources, memory, storage, and bandwidth. In cloud computing, DDPG can be used to allocate virtual machines to different physical servers based on the current load and demand to ensure efficient utilization of the resources and minimize the cost. In energy management, DDPG can allocate energy resources such as batteries and generators to different loads based on the current demand and availability of the resources to ensure a reliable and cost-effective power supply.

DDPG findings demonstrate that it outperforms traditional optimization methods in resource allocation problems, as it can learn complex policies that consider the nonlinear relationships between the system parameters and the rewards [36]. Additionally, DDPG can adapt to changes in the system and learn from experience, which makes it suitable for dynamic environments such as cloud computing and energy management.

DDPG effectively solves resource allocation problems in cloud computing and energy management. By learning from experience, DDPG can adapt to changing conditions and optimize the allocation of resources in real time. The use of DNNs allows DDPG to learn complex policies that can consider the nonlinear relationships between the system parameters and the rewards, making it a powerful tool for solving resource allocation problems in dynamic environments. Table 6 summarizes the studies classified as Resource Allocation Based on DDPG algorithms.

4.1.2. Summary of studies classified as autonomous driving

DDPG algorithm has made significant contributions to autonomous driving applications by enabling the development of more sophisticated and adaptive driving policies. As can be seen in the summarized studies in Table 7, DDPG can be used to learn a driving policy that can control the vehicle based on sensor data [37], such as lidar data, camera images, and other sensors data. Using DDPG, the autonomous driving system can learn to navigate complex environments, make decisions in real time, and adapt to changing conditions. The algorithm can be used to optimize driving behavior, including following traffic rules, avoiding collisions, and minimizing the distance to other vehicles [38]. Overall, DDPG offers the potential to create safer and more efficient autonomous driving systems that can adapt to a wide range of driving scenarios and conditions [39].

In addition, DDPG has also been used to overcome some of the challenges associated with traditional autonomous driving algorithms, such as the need for a pre-defined set of rules and difficulty adapting to changing road conditions. With DDPG, the driving policy is learned through trial and error, which allows the system to adapt to new situations and learn from experience.

4.1.3. Summary of studies classified as unmanned aerial vehicles (UAVs)

One of the primary advantages of DDPG is its ability to handle continuous control problems, which are common in the context of UAVs [40] [41]. DDPG has been utilized in several UAV applications, including obstacle avoidance [42], path planning [43], and formation control [44]. In obstacle avoidance, the algorithm can learn a policy that enables the UAV to navigate around obstacles while still reaching its destination. Path planning involves finding the optimal path for the UAV to take to reach a particular goal, while formation control involves coordinating the movement of multiple UAVs to achieve a particular objective.

Table 8 summarizes the studies of classified UAVs based on the DDPG algorithm and shows how the DDPG algorithm has made significant contributions to the UAVs field, particularly in the area of autonomous control.

4.1.4. Summary of studies classified as robotics

DDPG algorithm has been successfully applied to Robotics and motion control problems [45,46]. DDPG can be used for a variety of tasks, such as manipulation, locomotion, and navigation [47]. For example, DDPG can learn an optimal policy for manipulation tasks, such as picking and placing objects in a specific location. Similarly, DDPG can learn an optimal policy for locomotion tasks, such as walking or running, by controlling a robot's joint angles and velocities. Moreover, DDPG can be used for learning complex motor skills, such as acrobatic maneuvers, by learning an optimal policy for controlling the movements of a robot. This is useful in fields such as aerial Robotics, where precise and agile control is required. The studies that are relevant to Robotics based on DDPG are summarized in Table 9.

4.1.5. Summary of studies classified as communications and IoT

DDPG can be used to learn policies that optimize the use of resources, such as bandwidth or power, while maximizing performance metrics, such as throughput or reliability [48] [49]. In communication systems, DDPG can be used to optimize the transmission parameters of wireless networks, such as the transmission power, modulation scheme, and channel allocation. By learning policies that consider the quality of the wireless channel, interference from other devices, and the traffic demand, DDPG can optimize the use of resources and improve network performance. In IoT applications, DDPG can be used to optimize the operation of IoT devices, such as sensors and actuators, in a way that maximizes the performance of the system. For example, DDPG can learn policies that optimize the sampling rate of a sensor or the activation time of an actuator, considering factors such as energy consumption, data quality, and the desired performance metric. Table 10 summarizes the studies of Communications and IoT based on the DDPG Algorithm.

Table 6
Summary of studies classified as Resource Allocation based on DDPG algorithm.

Author and Year	Description	Application
(Zheng and Liu, 2019) [55]	**Techniques:** Multi-agent (MADDPG) algorithms for path planning-based crowd simulations. **Methodology:**The proposed algorithm uses several metrics, including the average velocity, flow rate, and collision rate of the crowd	Crowd simulation environment.
(Guo et al., 2020) [35]	**Techniques:**The higher tier utilizes (DDPG) algorithm for emergency medical services (EMSs) training at varying speed ranges. The lower levels incorporate the transfer learning (TL) technique to adapt pre-existing neural networks for a new driving cycle. **Methodology:**The fuel efficiency and the battery's state-of-charge (SOC) were two measures that the authors used to assess how well the suggested technique performed.	Hybrid-tracked vehicle simulator.
(Meng et al., 2020) [56]	**Techniques:** Deep Q-network (DQN) and (DDPG). **Methodology:** (DQN) learns to allocate power to each user based on the channel conditions and user demand and (DDPG) learns a continuous policy for power allocation.	Multi-user cellular networks
(Zheng et al., 2023) [57]	**Techniques:** (DDPG) algorithm for joint time and energy management. **Methodology:** The DDPG algorithm suggested intends to enhance the balance between energy consumption and the delay in data transmission.	Radio networks
(Zhao et al., 2021) [54]	**Techniques:** Proposes a DRL-based approach using (DDPG) algorithm for dynamics power allocation in cell-free massive systems. **Methodology:** The approach uses a DNN and (DDPG) algorithm to learn the optimal power allocation policy.	Cell-free massive multiple-input multiple-output (MIMO) systems.
(T. Zhang et al., 2021) [58]	**Techniques:** DDPG, Mode selection, Device-to-device (D2D). **Methodology:** Optimize mode selection and resource allocation for device-to-device (D2D)-enabled heterogeneous networks.	Heterogeneous networks
(B. Zhang et al., 2022) [59]	**Techniques:** The authors developed intelligent EMS by combining (DDPG)-(PER). **Methodology:** The proposed approach uses (DDPG)-(PER) to train an energy management system (EMS) for (HETVs) in an online and adaptive manner.	Series Hybrid Electric Tracked Vehicle (SHETV)
(Xia et al., 2021) [34]	**Techniques:** DRL algorithm based on (DDPG) framework. **Methodology:** A model for satellite resource allocation is developed that optimizes multiple objectives. A technique involving the division of regions and a feature extraction network is employed to reduce the dimensionality of input data. The resource allocation for satellites transmitting short messages is examined within an area that is serviced by multiple satellites.	Brief messages satellites.
(Wei et al., 2022) [60]	**Techniques:** Enhanced ability to perceive the environment (DDPG) algorithm with priority experience replay **Methodology:** Formulation of a multi-objective optimization problem	Lithium-ion batteries in electric vehicles.
(Chen et al., 2022) [36]	**Techniques:** Two DDPG algorithms are combined to allocate the amplitude and phase shift of individual reflecting elements for an ideal intelligent reflecting surface (IRS) in a dynamic manner. **Methodology:** Dynamic optimization is formulated for sub-carrier assignment, power allocation, amplitude control, and phase shift design.	Intelligent reconfigurable surface (IRS)
(Chen et al., 2021) [61]	**Techniques:** Temporal attentional deterministic policy gradient (TADPG) based on (DDPG). **Methodology:** Joint optimization problem of computation offloading and resource allocation (JCORA)	Mobile edge computing (MEC)
(J. Wang et al., 2022) [62]	**Techniques:** Decentralized multi-agent (De-DDPG) algorithms based on DDPG. **Methodology:** The paper proposes a joint optimization scheme for computation offloading decisions and computing resource allocation on the Internet of Vehicles (IoV) scenario, where the computing capacity of vehicles is limited, and computation tasks are offloaded to nearby multiaccess edge computing (MEC) servers.	Internet of Vehicle (IoV)
(Z. Wang et al., 2022) [63]	**Techniques:** (twin-actor DDPG) **Methodology:** The method suggested involves optimizing communication, computing, and caching resources simultaneously in multi-access edge network slicing using (DRL) approach.	Caching resources in multi-access edge network slicing
(Qu et al., 2022) [64]	**Techniques:** An extended DDPG algorithm with multi-objective reward. **Methodology:** The edge server's level of trust is determined by the computation offloading success rate, while the system model is constructed based on the user's standpoint, with delay and energy consumption being the two multi-objective tasks that are jointly optimized.	Mobile edge computing (MEC) in the 5G era.
(B. Liu, B. W. Xu, et al., 2022) [65]	**Techniques:** The proposed technique in this paper is a hybrid (HDRL) that combines dueling double deep Q networks (D3QN) algorithms and (DDPG) algorithm. **Methodology:** The approach consists of formulating the issue as a multi-objective optimization problem with both integer and continuous variables while creating an energy management framework for the MEA power system.	Real-time power management in electric aircraft

4.1.6. Summary of studies classified as finance

DDPG is a versatile DRL algorithm that has shown great potential in the field of Finance [50,51]. Table 11 summarizes the studies that explain how DDPG can be used for a variety of Finance tasks, including portfolio optimization, algorithmic trading, risk management, and fraud detection. First, in portfolio optimization, DDPG can learn to allocate assets to maximize returns while minimizing risk based on historical market data and risk preferences [52,53]. Also, DDPG can analyze real-time market data in

Table 7
Summary of studies classified as Autonomous Driving based on DDPG algorithm.

Author and Year	Description	Application
(Zhu et al., 2018) [37]	**Techniques:** The proposed approach uses the DDPG algorithm for DRL-based training of the autonomous car-following model. **Methodology:** The DRL algorithm used in this work is the (DDPG) algorithm, which is evaluated on a simulated highway scenario with varying traffic conditions.	Autonomous driving systems
(Guo and Wu, 2019) [39]	**Techniques:** The proposed approach utilizes (DDPG) to learn the optimal policy for controlling the car's actions in a racing environment. **Methodology:** The methodology involves implementing the DDPG algorithm and experimenting with two strategies to improve performance: action punishment and multiple exploration.	Autonomous driving systems
(Chen et al., 2020) [66]	**Techniques:** The proposed approach utilizes (DDPG) framework and integrates a progressively optimized reward function (PORF). **Methodology:** Constructing a PORF-DDPG algorithm that gradually trains a DNN-based reward model with input from front-view images through human supervision and intervention.	Autonomous driving systems
(Fu et al., 2020) [67]	**Techniques:** (DDPG) algorithm, Collective Learning, Vehicular, Blockchain, Knowledge Transfer. **Methodology:** Formulate the lane-changing issue as (DRL) process and acquire the self-governing lane-changing technique using the (DDPG) algorithm.	Connected and Autonomous Vehicles (CAVs).
(Ashraf et al., 2021) [68]	**Techniques:** (DDPG) algorithm and Whale Optimization Algorithm (WOA) **Methodology:** The study focuses on optimizing the hyperparameters of (DDPG) algorithm to attain the best possible control approach in an autonomous driving control dilemma.	Autonomous driving systems
(Alomari et al., 2021) [69]	**Techniques:** DDPG, Path-following, Control Theory **Methodology:** (DRL) agent is educated in a 3D simulated setting and interfaces with the unfamiliar surroundings to revise the DNN.	Autonomous vehicles
(Y. Zhang et al., 2022) [70]	**Techniques:** (TD3), Multi-objective Optimization **Methodology:** Creating a (TD3)-based energy management system (EMS) that incorporates the durability data of both proton exchange membrane fuel cell (PEMFC) stacks and lithium-ion batteries (LIBs) to manage the hybrid power train based on the vehicle's operating conditions.	Fuel cell vehicle energy management
(M. Li et al., 2020) [38]	**Techniques:** (DDPG), Simulation-based testing **Methodology:** Suggesting (DDPG) algorithm driving plan tailored to each vehicle to lessen fluctuations and boost traffic safety during stop-and-go traffic patterns.	Driving strategy to individual vehicles
(He and Huang, 2021) [71]	**Techniques:** The algorithm consists of two main components: the rule-based speed planning algorithm and the DDPG energy management algorithm. **Methodology:** The speed planning algorithm generates a speed trajectory for the HEV based on all traffic information in a connected environment, while the DDPG algorithm optimizes fuel economy in real-time and satisfies the constraints of driving safety and driving time.	Hybrid electric vehicle (HEV)
(Wang et al., 2023) [72]	**Techniques:** Optimal control, Simplified (S-DDPG) algorithm. **Methodology:** The research article suggests utilizing the S-DDPG algorithm to devise a path-tracking control technique for self-governing underwater vehicles (AUVs).	Autonomous Underwater Vehicles (AUVs)
(Sun et al., 2021) [73]	**Technique:** (DDPG) algorithm **Methodology:** The proposed algorithm classifies and stores experience samples using the SumTree structure to improve the speed of convergence and optimize the AUV's path planning in uncertain and complex underwater environments.	Autonomous Underwater Vehicles (AUVs)
(Yao et al., 2021) [74]	**Technique:** TD3, DDPG, and A-ECMS **Methodology:** The study compares the performance of (TD3),(DDPG), and adaptive equivalent consumption minimization strategy (A-ECMS) algorithms.	Hybrid electric vehicles (HEVs)
(Syavasya and Muddana, 2022) [75]	**Technique:** (DDPG) algorithm, SHapley Additive exPlanations (SHAP) algorithm. **Methodology:**The main aim of this research is to present a DRL framework that can enhance the longitudinal control of autonomous vehicles.	Longitudinal control of autonomous vehicles
(Hu and Li, 2022) [76]	**Technique:** DDPG and ECMS **Methodology:** The article suggests an energy management strategy for hybrid electric vehicles called adaptive hierarchical EMS. This approach integrates both heuristic equivalent consumption minimization strategy (ECMS) and (DDPG) knowledge.	HEVs
(S. C. Li et al., 2022) [77]	**Technique:** (DDPG) algorithm and modified LSTM neural network. **Methodology:** (DDPG) algorithm is used to solve the MDP with continuous action spaces.	Electric vehicle
(Tang et al., 2022) [78]	**Technique:** DDPG and DQN algorithms for the DDRL-based EMS. **Methodology:** The article proposes two energy management strategies (EMSs) for hybrid electric vehicles (HEVs) based on (DRL) and rule-based controls.	HEVs
(Huo et al., 2022) [79]	**Technique:** (DDPG) and (DQL) algorithms with priority experience replay. **Methodology:** The study integrates the factors of power variation and fuel economy into the multi-objective reward functions to diminish fuel usage and extend the lifespan of the fuel cell stack.	Hybrid Vehicles
(Hu and Li, 2022) [76]	**Technique:** DDPG and ECMS **Methodology:** The article suggests an energy management strategy for hybrid electric vehicles called adaptive hierarchical EMS. This approach integrates both heuristic equivalent consumption minimization strategy (ECMS) and (DDPG) knowledge.	HEVs

Table 8
Summary of studies classified as UAVs based on DDPG algorithm.

Author and Year	Description	Application
(Zhou et al., 2022) [80]	**Techniques:** (DDPG) algorithm **Methodology:** The proposed algorithm optimizes the performance of the MEC network by jointly optimizing the task offloading, resource allocation, and flight trajectory of UAVs.	UAVs maritime networks
(C. H. Liu et al., 2020) [81]	**Techniques:** The DRL-EC3 approach uses the DDPG algorithm along with three other baseline approaches. **Methodology:** Creating a completely decentralized control system to guide a swarm of UAVs in order to offer sustained communication coverage for people on the move on the ground.	UAV swarm control
(Samir et al., 2020) [44]	**Techniques:** The proposed technique leverages (DDPG) algorithm to minimize the Expected Weighted Sum Age of Information (EWSA). **Methodology:** The problem of optimization is transformed into a Mixed Integer Non-Linear Program (MINLP), and then the DDPG algorithm is applied to determine the flight paths of the UAVs and decrease the EWSA.	Intelligent transportation systems
(Zhang and Cao, 2022) [82]	**Techniques:** Attentional (ATDDPG) algorithms **Methodology:** The algorithm suggested aims to simultaneously optimize three objectives that may contradict each other: increasing the system's throughput, maximizing the energy gathered, and reducing the total energy consumed by the UAV.	UAV-enabled wireless
(Yu et al., 2021) [43]	**Techniques:** The proposed algorithm MODDPG (multi-objective DDPG) **Methodology:** The rotary-wing (UAV) utilizes a protocol that involves flying, hovering, and communicating to reach (IoT) devices that are in need of attention.	UAV-assisted wireless powered IoT networks
(Y. Li et al., 2020) [83]	**Technique:** (DDPG) **Methodology:** The paper proposes a new (DDPG) strategy based on DRL for the attacking area fitting of unmanned combat aerial vehicles (UCAVs).	Unmanned combat aerial vehicles (UCAVs)
(Cui et al., 2021) [84]	Wireless communication using UAVs **Methodology:** Formulate the problem of maximizing UAV service time and downlink throughput **Techniques:** (DDPG) algorithm for trajectories design and power allocation (TDPA)	
(Zhang et al., 2020) [40]	**Techniques:** Multi-agent(MADDPG) algorithm. **Methodology:** The MADRL method being suggested involves centralized training and decentralized execution. The agents, which include both UAV transmitters and UAV jammers, acquire knowledge on how to effectively optimize their flight paths and power usage in order to enhance the system's secure capacity.	UAVs
(Ho et al., 2021) [41]	**Techniques:** (DDPG) and (TRPO) **Methodology:** Formulate nonconvex optimization problem to maximize UAV energy efficiency while meeting communication requirements and backhaul link constraints.	UAV
(M. Zhang et al., 2021) [85]	**Techniques:** The proposed approach uses a UAV-BS system and employs (DDPG). **Methodology:** The process includes investigating the unbroken ranges of possible states and actions to acquire knowledge about the most effective location for hovering and how much power should be allocated, utilizing the DDPG-based algorithm proposed.	UAV-BS systems
(Xu et al., 2022) [42]	**Techniques:** DDPG algorithm **Methodology:** The proposed methodology includes using a basic controller for stability control of the UAV, reducing the search space, designing an external sparse reward to enhance learning efficiency, and using a DDPG-based controller to compensate for unknown external disturbances and optimize control accuracy.	Quadrotor UAVs
(Barnawi et al., 2022) [86]	**Techniques:** (DDPG) and (POPT) **Methodology:** The task is to reduce the energy used by the UAV-Magnetometer-LM system by choosing time slots and organizing the order in which the UAV visits different locations to uncover the LM hidden underneath dirt or sand.	Landmine detection using UAV-mounted magnetometer sensing.
(Gao et al., 2021) [87]	**Techniques:** The algorithm involves two separate problems: potential game for service assignment and (DDPG) for trajectory planning. **Methodology:** The research paper breaks down the multi-UAV-assisted offloading system into two distinct problems and utilizes a combination of potential game and RL algorithms to optimize them.	Multi-UAV assisted offloading system
(Wang et al., 2021) [88]	**Techniques:** (DDPG)-based cache placement optimizing algorithm **Methodology:** The paper recommends creating a problem-focused on minimizing file access delay through the optimization of cache placement. It also suggests a cache placement optimization algorithm that employs (DDPG).	UAV-relaying networks with D2D communication
(Guo et al., 2021) [89]	**Techniques:** The paper proposes a twin-DDPG deep reinforcement learning (TDDRL) algorithm based on the DDPG framework for the joint design of active beamforming, coefficients of RIS elements, and UAV trajectory. **Methodology:** The TDDRL algorithm is proposed to solve the formulated problem using the DDPG framework.	UAV trajectory
(Din et al., 2022) [90]	**Techniques:** (DDPG) algorithm, modified for learning architecture to handle continuous state and control space domains **Methodology:** Designing an intelligent control architecture for an experimental Unmanned Aerial Vehicle (UAV) (DDPG) algorithm	Airborne UAV applications

algorithmic trading to make trading decisions based on market trends, volatility, and news events. Moreover, in risk management, DDPG can identify potential risks and take actions to mitigate them, based on market data and other relevant factors. Finally, in fraud detection, DDPG can analyze large amounts of transaction data to identify patterns and anomalies that may indicate fraudulent behavior. While DDPG has the potential to be a powerful tool for finance professionals, it is important to carefully select the model,

Table 9
Summary of studies classified as Robotics based on DDPG algorithm.

Author and Year	Description	Application
(Sehga et al., 2022) [91]	**Techniques:** The GA+DDPG+HER approach uses a combination of GA for parameter tuning, DDPG for DRL, and HER for improved sample efficiency. **Methodology:** Use the GA+DDPG+HER approach to perform DRL on the robotic manipulation tasks of FetchReach, FetchSlide, FetchPush, FetchPick Place, DoorOpening, and AuboReach environments.	Robotic manipulation
(H. X. Zhang et al., 2021) [45]	**Techniques:** Importance-Weighted Autoencoder (IWAE) and Gaussian parameter (Gaussian-DDPG) **Methodology:** Addition of Gaussian parameters to DDPG algorithm for better exploration and optimization of grasping position control using torque information.	Robotic surgery and Object manipulation
(Yang and Peng, 2021) [92]	**Techniques:** (DDPG) algorithm, Meta-learning-based experience replay buffer separation (MSER), and neural network design **Methodology:** Experiments are conducted in a simulation environment for trajectories planning of a robot manipulator in V-REP.	Robot manipulator
(Rajendran and Zhang, 2022) [93]	**Techniques:** The paper proposes a learning-based control design that employs the use (DDPG) algorithm to train the agent. **Methodology:** The mathematical model of the soft robotic fish is based on a 3-link representation and includes both geometric and dynamic aspects.	Soft robotic fish
(Min et al., 2019) [47]	**Techniques:** The paper proposes a method that combines Hindsight Experience Replay (HER) and Twin Delayed DDPG (TD3) algorithms. **Methodology:** The proposed method is evaluated in a simulated environment using a 7-DoF manipulator.	Manipulation of multi-DOF manipulators in 3D space
(Q. Liu et al., 2021) [94]	**Techniques:** The proposed algorithm is intrinsic reward (IRDDPG) and combines the DDPG algorithm. **Methodology:** formulates the safe human-robot collaboration manufacturing problem.	Industrial human-robot collaboration
(X. Li et al., 2020) [95]	**Techniques:** (DDPG) is used to optimize robot control policy. **Methodology:** The method is implemented within ROS for controlling a Baxter robot in a simulation environment.	Robot control
(Dankwa and Zheng, 2019) [96]	**Techniques:** Twin-Delayed DDPG (TD3), and automatic feature engineering. **Methodology:** The TD3 algorithm is used to reduce overestimation bias in Deep Q-Learning with discrete actions in an Actor-Critic domain setting.	4-Ant-legged robot
(Hao et al., 2022) [46]	**Techniques:** (DDPG), replay buffer optimization, and weighted training samples. **Methodology:** The method interferes with network exploration by utilizing the fundamental relationship between pedal opening and vehicle speed.	Vehicle speeding tracking control with a robotic driver
(Z. Li et al., 2022) [97]	**Techniques:** Powell Deep Deterministic Policy Gradient (PDDPG) **Methodology:** The technique considers each agent as a single-dimensional variable, and it represents the learning of multi-robot collaboration as an optimal vector search.	Multi-robot cooperation learning
(P. Li et al., 2021) [98]	**Techniques:** (DDPG) algorithm **Methodology:** The proposed algorithm is tested and compared with the original DDPG algorithm through simulation experiments in a cloud robot path planning system.	Mobile robots
(Jiang et al., 2022) [99]	**Techniques:** The paper proposes an integrated tracking control approach for a continuum robot in space capture missions using (DDPG) and rolling optimization method. **Methodology:** Combining (DDPG) with the rolling optimization method.	Continuum robot in space capture missions

engineer features appropriately, and validate results thoroughly to ensure that it is effectively addressing the problem at hand. Furthermore, DDPG can also be used in credit scoring, where it can learn to predict the creditworthiness of an individual or a company based on their financial history and other relevant factors [54].

4.2. RQ2: what are the commonly applied techniques with DDPG in DRL applications?

Based on the 85 included studies, various techniques have been observed to be applied alongside DDPG as shown in Fig. 9. The most predominant technique is DDPG itself, mentioned in approximately 80 studies. There are other techniques, though less prevalent, which have been combined with DDPG. These include DQN with 7 studies, TD3 with 6 studies, A2C and MADDPG each with 4 studies, and PPO appears in 4 studies. Moreover, techniques such as HER, LSTM, and PER have been applied in 3, 2, and 2 studies respectively. Furthermore, D3PG, D4PG, DQL, GA, POPT, SAC, TRPO, and WOA each have a single study mention (N = 1 for each). This distribution highlights the prominence of DDPG and its adaptability to a wide range of other methods in the scope of DRL applications.

4.3. RQ3: what are the optimization methods used to overcome the instability of hyperparameters in DDPG?

DDPG algorithm can suffer from instability and slow convergence when dealing with complex environments. There are several techniques that can be used to solve this:

Table 10
Summary of studies classified as communications and IoT Based on DDPG algorithm.

Author and Year	Description	Application
(Zou et al., 2020) [100]	**Techniques:** The authors create a DRL method that adjusts the beamforming and relaying strategies in real-time to maximize the overall Signal-to-Noise Ratio (SNR). **Methodology:** The authors introduce a new hierarchical (H-DDPG) approach that incorporates model-based optimization into the conventional DDPG framework, with a focus on optimization-driven solutions.	Wireless communication systems
(Hu et al., 2023) [101]	**Techniques:** To optimize the covert rate while adhering to covert constraints, a model-free and off-policy algorithm based on (DDPG) is employed. **Methodology:** The DDPG algorithm suggested in the study generates the transmit beamformer vector for the authorized transmitter and the phase shifts matrix for the Intelligent Reflecting Surfaces (IRS). These elements are adjusted to maximize the performance of covert communication.	Wireless communication systems
(Z. W. Wang et al., 2022) [102]	**Techniques:** (DDPG) **Methodology:** By maximizing a quadratic reward function determined by state errors and control inputs, wireless vehicle-to-vehicle communication technology is integrated into the system.	Wireless V2V communication
(Saifaldeen et al., 2022) [103]	**Techniques:** (DRL) solution based on (DDPG) algorithm **Methodology:** The study aims to optimize the Secrecy Capacity (SC) of a Visible Light Communication (VLC) system that employs mirror array sheets as Intelligent Reflecting Surfaces (IRS). This optimization involves determining the optimal beamforming (BF) weights for VLC fixtures and mirror orientations for the mirror array sheet.	Visible Light Communication (VLC) systems
(Mlika and Cherkaoui, 2022) [48]	**Techniques:** (DDPG), Non-orthogonal multiple access, mixed-integer nonlinear programming. **Methodology:** Minimizing the Age of Information (AoI) in Cellular Vehicle-to-Everything (C-V2X) communications. The problem involves several aspects, including selecting half-duplex transceivers, optimizing broadcast coverage, allocating power, and scheduling Resource Blocks (RB).	Cellular vehicle-to-everything communications
(Ale et al., 2022) [49]	**Techniques:** The proposed algorithm is called Dirichlet (D3PG), which is built on (DDPG). **Methodology:** A new DRL technique named Dirichlet (D3PG) is introduced for solving the problem.	Multiple IoT devices and multiple edge servers
(Budhiraja et al., 2021) [104]	**Techniques:** Combination of several DRL techniques, including DDPG, CO-DDPG, MARL, MDP, and multi-agent DQN. **Methodology:** The DQN is combined with DDPG to form the distributed DDPG scheme and the conventional optimization scheme is integrated with DDDPG to control the power of both the cellular user equipment (CUE) and D2D transmitters (DTs).	IoT devices
(Chen et al., 2023) [105]	**Techniques:** The proposed algorithm is a deep deterministic policy gradient-based WBAN offloading strategy (DDPG-WOS). **Methodology:** The research paper suggests a joint optimization issue that concerns both computational offloading and resource allocation (JCORA) in Mobile Edge Computing (MEC) designed for healthcare service scenarios.	Mobile edge computing (MEC)
(Lee et al., 2022) [106]	**Techniques:** (DDPG). **Methodology:** jointly optimize the beamforming matrix for the base station (BS) and IRS. Simulation experiments are conducted to evaluate the performance of the proposed algorithm in various scenarios.	mmWave V2I communication systems
(Ciftler et al., 2022) [107]	**Techniques:** Distributed (DDPG) algorithm **Methodology:** The task is to create a problem related to the allocation of downlink power in a distributed manner, aimed at optimizing the transmit power for users, so that they can achieve target data rates in hybrid networks that combine Radio Frequency (RF) and Visible Light Communication (VLC).	Hybrid radio frequency and visible light communication networks
(Shi et al., 2022) [108]	**Techniques:** Distributed multidimensional resource management algorithm based on DRL and DDPG **Methodology:** Proposing a distributed multidimensional resource management algorithm based on DRL to manage spectrum, energy, and time resources. Adopting DDPG algorithm and introducing a simplified AA for improved training efficiency and battery performance protection	Wireless sensor communications in Internet of Things (IoT) systems
(Kwon et al., 2020) [109]	**Techniques:** The proposed algorithm utilizes multiagent (MADDPG) and federated learning (FL). **Methodology:** The proposed algorithm is intended to achieve federated learning (FL) computation using Internet-of-Underwater-Things (IOUT) devices within the oceanic environment.	Internet-of-Underwater-Things (IOUT) devices
(Ma et al., 2023) [110]	**Techniques:** DDPG-DQN algorithm **Methodology:** To optimize voltage regulation over both short and long time frames, the proposed approach employs a hybrid of two algorithms (DQN) and (DDPG) algorithm. The DQN algorithm is used for longer time periods, while the DDPG algorithm is used for shorter ones.	Distribution networks
(Baktayan et al., 2022) [111]	**Technique:** The DDPG algorithm is utilized to implement dynamic pricing for computation offloading in UAV-MEC for vehicles. **Methodology:** (DDPG) algorithm solves the optimization problem formulated as (MDP).	5G cellular networks
(Y. N. Liu et al., 2021) [112]	**Techniques:** (DDPG) and (LSTM) algorithms **Methodology:** A cellular scenario in which down-link transmission is split into different time slots. Its aim is to satisfy a range of real-time quality of service (QOS) demands. It also includes a Multi-Dimensional Intelligent Multiple Access (MD-IMA) scheme to optimize the use of available radio resources across diverse domains.	Beyond 5G and 6G wireless networks.

Table 11
Summary of studies classified as Finance based on DDPG algorithm.

Author and Year	Description	Application
(Ye et al., 2020) [52]	**Techniques:** DDPG, prioritized experience replay, bi-level optimization. **Methodology:** The proposed methodology combines DRL with a PER strategy to model strategic bidding decisions in deregulated electricity markets.	Strategic bidding
(Yang et al., 2020) [113]	**Techniques:** The proposed three actor-critic-based algorithms: (DDPG), (PPO), and (A2C) **Methodology:** The ensemble strategy is created by combining the best features of the three algorithms to the load-on-demand technique employed to process the large amount of data required for training networks with continuous action space.	Financial Instruments
(Vishal et al., 2021) [114]	**Techniques:** (DDPG), (A2C), (TD3) and (PPO) **Methodology:** The trading agent is developed using the Actor-Critic reinforcement learning approach, which involves training two DNNs simultaneously, one for the actor and one for the critic.	Stock Market scenario
(Sagiraju and Mogalla, 2022) [50]	**Techniques:** (DDPG), (A2C), (PPO), and (DQN) **Methodology:** The agent learns to predict stock market investments and returns and develops a trading strategy based on the given environment.	Stock market
(Y. K. Liu et al., 2022) [115]	**Techniques:** (DDPG) algorithm **Methodology:** The proposed approach uses (DDPG) algorithm to learn the optimal policy for service composition in a dynamic cloud manufacturing environment.	Cloud manufacturing environment
(H. F. Li et al., 2021) [116]	**Techniques:** (DDPG) and portfolio optimization. **Methodology:** The average logarithmic cumulative returns and the cumulative Sharpe Ratio are used as the reward functions.	Stock portfolio
(Kong and So, 2023) [51]	**Techniques:** DDPG, A2C, PPO, ACKTR, SAC, TD3, and TRPO **Methodology:** The paper describes an experimental study to evaluate the performance of the proposed automated stock trading system.	Algorithmic trading
(Chau et al., 2022) [53]	**Techniques:** DDPG, PPO, ACKTR, and TD3 **Methodology:** The study involves training and testing the DRL agents on 30 different Forex currencies.	Automation Forex Trading

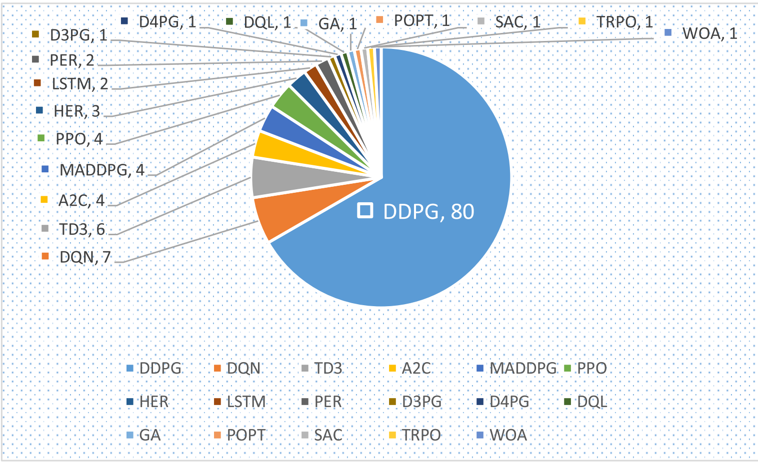

Fig. 9. Commonly applied techniques with DDPG in DRL applications.

- **Experience replay**: This method involves storing experiences in a replay buffer and randomly sampling them to update the network. By doing so, the network is trained on a wider range of experiences, which can help stabilize learning and reduce the effects of any correlations between consecutive samples.
- **Target networks**: In DDPG, the target Q-network and target policy network are used to calculate the target Q-values and target actions, respectively. These target networks are slowly updated with the weights of the online networks, which can help stabilize learning and prevent overestimation of the Q-values.
- **Batch normalization** is another technique that can be used to stabilize the learning process. It involves normalizing the inputs to each layer of the network, which can help prevent vanishing or exploding gradients and improve the stability of the training process [117].
- **Hyperparameter noise** can also be added to the exploration policy during training to encourage exploration and prevent overfitting. This can help the agent learn more robust policies that are less sensitive to small environmental variations.
- **Gradient Clipping:** This method involves constraining the magnitude of the gradients during backpropagation to prevent them from becoming too large. This can help prevent the network from diverging and improve the stability of the training process.

- **Learning Rate Scheduling:** This technique involves adjusting the learning rate of the optimizer during training to improve the convergence rate and prevent overfitting. One common method is to gradually reduce the learning rate over time, which can help the network to settle into a stable policy.
- **Exploration Strategies:** DDPG can suffer from an overestimation of the Q-values, which can lead to poor exploration and slow convergence. One way to address this issue is to use alternative exploration strategies, such as adding noise to the actions during training or using Bayesian optimization to select actions.
- **Prioritized Experience Replay:** This method involves prioritizing experiences in the replay buffer based on their importance for learning. By sampling experiences with higher priority more frequently, the network can focus on the most important experiences and improve the convergence rate.

In a study by Lillicrap et al. [22], the authors proposed the DDPG algorithm and demonstrated its effectiveness on various continuous control tasks, such as reaching, grasping, and locomotion. To improve the stability of the algorithm, they used experience replay, target networks, and gradient clipping.

Also, in a research paper by Fujimoto et al. [27], the authors proposed a modification to the DDPG algorithm called Twin Delayed DDPG TD3. This algorithm uses two Q-networks to reduce the overestimation of the Q-values and includes several optimization techniques, such as target networks, delayed updates, and clipped double Q-learning.

In the proposal made by the authors Pinto et al. [118], the authors proposed an extension to the DDPG algorithm called Robust Adversarial Reinforcement Learning (RARL). This algorithm includes several optimization techniques, such as parameter noise, adversarial exploration, and prioritized experience replay, to improve the stability and robustness of the learning process.

The authors put forth a proposal in a study by Haarnoja et al. [28], the authors proposed a modification to the DDPG algorithm called Soft Actor-Critic (SAC). This algorithm includes several optimization techniques, such as target networks, entropy regularization, and automatic temperature tuning, to improve the stability and robustness of the learning process.

Moreover, in a study by Duan et al. [119], the authors compared the performance of several reinforcement learning algorithms, including DDPG, TD3, and SAC, on a suite of continuous control tasks. They found that TD3 and SAC were the most effective algorithms due in part to their use of optimization techniques such as target networks and prioritized experience replay.

A proposition was put forward by the authors in a research article by Silver et al. [120], the authors proposed a modified DDPG algorithm that uses residual policy learning and a value function to improve stability and convergence. They demonstrated the effectiveness of their algorithm on several continuous control tasks, including locomotion and manipulation.

These studies highlight the ongoing efforts to improve the stability and convergence of the DDPG algorithm through the use of various optimization techniques. By incorporating these techniques, researchers are making progress toward developing more effective and robust reinforcement learning algorithms that can solve increasingly complex problems.

4.4. RQ4: what are the evaluation measures used to evaluate the performance of the DDPG algorithm?

While DDPG has shown impressive results in various applications, evaluating its performance is critical to ensure its effectiveness in solving real-world problems. In this regard, this answer will discuss the evaluation measures commonly used to evaluate the DDPG algorithm's performance. Researchers commonly use several evaluation measures, such as reward, convergence, exploration, and robustness, and several environments, such as OpenAI Gym, MuJoCo, and TORCS, to test the algorithm's ability to learn an optimal policy.

To evaluate the performance of DDPG, researchers commonly use several environments to test the algorithm's ability to learn an optimal policy. These environments come in different forms, such as simulated environments, real-world environments, and game environments. Here are some commonly used ones:

1. **OpenAI Gym:** It is a popular toolkit that provides a wide range of simulated environments for testing DRL algorithms' performance. The toolkit includes a collection of classic control tasks, Atari games, and robotics tasks. OpenAI Gym is widely used to evaluate the DDPG algorithm's performance due to its flexibility and ease of use [121].
2. **MuJoCo:** It is a physics engine that simulates a range of robotics tasks. The engine provides a high-fidelity simulation environment that enables testing DRL algorithms' performance in realistic scenarios. MuJoCo is commonly used to evaluate the DDPG algorithm's performance on robotics tasks [122].
3. **The Open Racing Car Simulator (TORCS):** It is a popular racing game utilized to evaluate the DRL algorithms' performance. TORCS provides a challenging environment that requires the agent to learn a complex policy to win the game. It is commonly used to evaluate the DDPG algorithm's performance in game environments [123].

Then, there are several evaluation measures used to assess the performance of DRL algorithms, including DDPG, in these environments. These measures aim to quantify the algorithm's ability to learn an optimal policy and generalize it to unseen environments. Here are some of the commonly used evaluation measures in the included studies:

1. **Reward:** It is the primary measure used to evaluate the DDPG algorithm's performance. It is the feedback signal that the agent receives from the environment based on its actions. In this regard, a high reward indicates that the agent has learned an optimal policy.

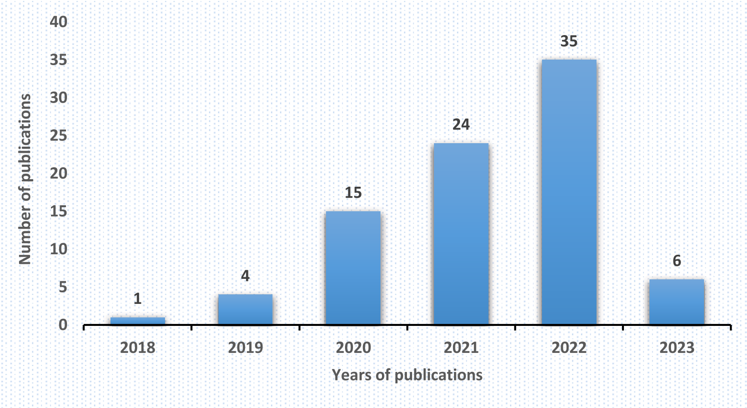

Fig. 10. Classification of the included studies by publication year.

2. **Convergence:** It refers to the rate at which the algorithm learns an optimal policy. In DDPG, convergence is typically evaluated by monitoring the change in the Q-value or the policy over time. A fast convergence rate indicates that the algorithm can learn an optimal policy quickly.
3. **Exploration:** It measures the algorithm's ability to explore the environment and find an optimal global policy. In DDPG, exploration is typically evaluated by monitoring the agent's action diversity and how often it visits new states.
4. **Robustness:** It measures the algorithm's ability to generalize its learned policy to unseen environments. In DDPG, robustness is typically evaluated by testing the learned policy on different environments with varying degrees of complexity.

4.5. *RQ5: what is the intensity of publications related to DDPG?*

This question's answer includes characteristics of the 85 selected studies based on their year of publication (Fig. 10), and the journals in which they were featured (Fig. 11).

To begin with, the distribution of studies based on the years from 2018 to 2023 is presented in Fig. 10. A noticeable upward trend is observed in the number of publications, culminating in a significant spike in 2022 with 35 publications indicating that DRL-DDPG is gaining popularity and receiving more attention from scholars. This was followed by 24 publications in 2021 and 15 in 2020. Earlier years, such as 2019 and 2018, witnessed a limited number of studies, with 4 and 1 publications respectively. The count for 2023 stands at 6 studies, but it is imperative to note that the research was conducted in March 2023, which likely accounts for the reduced number of publications for that year.

Next in the classification of the included studies based on their respective journals with more than 2 publications (Fig. 11), the *IEEE Internet of Things* Journal emerges as the predominant source with 5 publications. This is closely followed by *IEEE Access* and *IEEE Transactions on Vehicular Technology*, each accounting for 4 publications. Several journals, including *IEEE Transactions on Communications*, *IEEE Transactions on Transportation Electrification*, *IEEE Transactions on Wireless Communications*, and the collective of *Lecture Notes in Computer Science*, have contributed 3 studies each. Furthermore, a number of journals have been the source of 2 studies each, namely *ACM International Conference Proceeding Series*, *IEEE Communications Letters*, *IEEE Journal on Selected Areas in Communications*, *IEEE Photonics Journal*, and *IEEE Wireless Communications Letters*. The remaining 50 studies have distinct journals. This distribution underscores the diversity of publication sources and the interdisciplinary nature of the subject matter under review.

5. Conclusion

This DDPG SLR is concluded based on its future research directions, limitations, and closing remarks in the sub-sections 5.1, 5.2, and 5.3 respectively.

5.1. *Future research direction*

DDPG has shown promising results in domains such as resource allocation, autonomous driving, unmanned aerial vehicles (UAVs), robotics, finance, communications and Internet of Things (IoT), game playing, recommendation systems, and energy management. However, there are still many domains in which the DDPG algorithm can be improved, and future research can focus on addressing these limitations.

One direction for future research in DDPG could be to improve its sample efficiency. The current implementation of DDPG requires a large number of samples to converge to the optimal policy. This is due to the use of experience replay and the exploration strategy, which are necessary but can be inefficient. Recent advancements in sample-efficient reinforcement learning, such as the use of model-based methods or meta-learning, can be applied to DDPG to improve its sample efficiency. For example, researchers can explore the use of model-based DDPG, which combines the benefits of model-based approaches with the advantages of DDPG to improve sample efficiency.

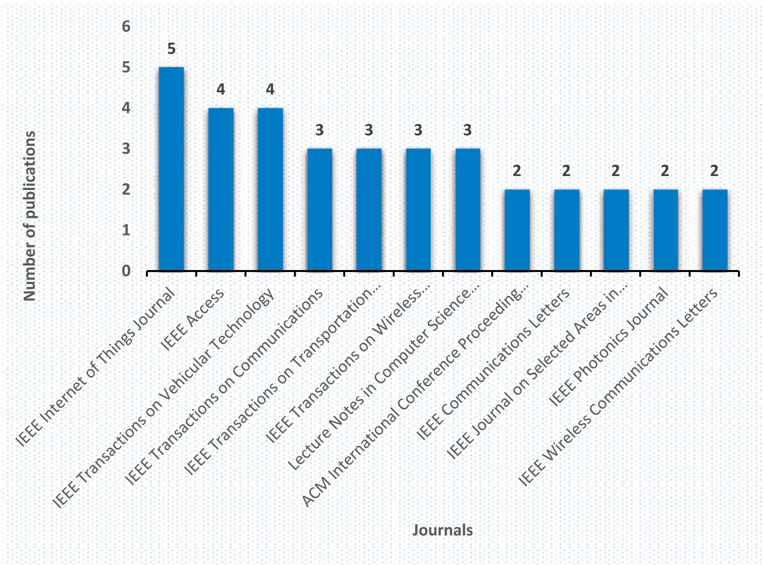

Fig. 11. Classification of the included studies by journals.

Another one is to incorporate multiple objectives into DDPG. In many real-world problems, there are multiple objectives that need to be optimized simultaneously. For example, in autonomous driving, the vehicle needs to avoid collisions while reaching the destination in the shortest time possible. DDPG can be extended to handle multi-objective optimization by using methods such as Pareto optimization or weighted sum optimization [1]. Researchers can also explore the use of multi-agent DDPG to handle scenarios where multiple agents need to coordinate to achieve a common objective.

A third one is to improve the robustness of DDPG to changes in the environment. DDPG is sensitive to changes in the environment, such as changes in the dynamics of the system or the introduction of new obstacles. One approach to improving the robustness of DDPG is to use techniques such as domain randomization or transfer learning [1]. Domain randomization involves training the agent in a variety of environments with different parameters to make it more robust to changes in the environment. Transfer learning involves training the agent in one environment and then transferring the learned policy to a new environment with similar dynamics.

Finally, there is a need for research on the theoretical properties of DDPG. The current understanding of DDPG is mostly empirical, and there is a lack of theoretical analysis of the algorithm. Researchers can explore the convergence properties of DDPG, its sensitivity to hyperparameters, and the relationship between DDPG and other reinforcement learning algorithms. This research can help to provide a better understanding of the algorithm and improve its performance in various real-world scenarios.

In summary, there are many future research directions in DDPG, ranging from improving its sample efficiency to incorporating multiple objectives and improving its robustness to changes in the environment. Theoretical analysis of the algorithm can also provide a better understanding of its properties and limitations. These research directions can lead to the development of more efficient and robust reinforcement learning algorithms that can solve a wide range of real-world problems.

5.2. Limitation

While this SLR of the DDPG algorithm offers valuable insights into the latest developments, applications, and comparative analyses of DDPG in the field of DRL, several limitations must be acknowledged. First, our review focuses on studies published within the last five years (2018-2023). The rapidly evolving nature of DRL suggests that newer developments may have arisen subsequent to our search period, potentially rendering our findings incomplete regarding the most recent advancements in DDPG. Second, this study only focuses on 6 main domains as listed in section 4.1 (RQ1). Third, language and geographic bias may exist as our search was primarily conducted in English, potentially overlooking valuable research in other languages and non-Western research communities. Furthermore, our inclusion and exclusion criteria are subject to a degree of subjectivity and may introduce bias during the study selection process. Variability in the quality of the included studies can also affect the overall quality of the review. In addition, our comparative analysis of DDPG with other DRL algorithms and traditional RL methods is susceptible to inherent biases and subjective judgments. The heterogeneity in evaluation metrics and experimental settings in the selected studies poses challenges in making direct and conclusive comparisons. Lastly, the generalizability of our findings across different applications and domains of DDPG may be limited as the algorithm's performance can significantly vary depending on the specific problem and environment. Despite these limitations, we believe this systematic review serves as a valuable resource for researchers in the field of DRL, providing a comprehensive overview of the current state of DDPG and its applications, as well as highlighting avenues for future research and exploration.

5.3. Closing remarks

To conclude, our SLR has provided a comprehensive overview of the key components, modifications, domains, optimization methods, decision environments, and evaluation measures of the DDPG algorithm. We have found that various modifications, such as adding a noise process and incorporating distributed and parallel computing techniques, have been proposed to improve the algorithm's performance. Additionally, we have identified various optimization methods, including different learning rates and prioritized experience replay, that have been used to overcome the algorithm's instability. Our SLR also highlighted the diverse decision environments in which the DDPG algorithm has been applied, such as robotic control, game playing, and finance. We found that DDPG has shown promising results in various applications and has the potential to be applied in even more domains. Ultimately, our SLR's analysis of the evaluation measures and environments used to assess DDPG's performance indicates that different metrics have been used, such as cumulative reward and success rate, and a range of environments have been utilized, including OpenAI Gym and Atari games. This provides valuable insight into how to evaluate the performance of DDPG in different applications. The findings of this SLR provide researchers and practitioners with important information and guidance for applying and improving the DDPG algorithm in various applications. Further research in this area can contribute to developing more robust and efficient DDPG algorithm variations.

Declaration of generative AI

During the preparation of this work, the authors used different AI tools to improve the language and readability. After using these tools/services, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

CRediT authorship contribution statement

Ebrahim Hamid Sumiea: Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Said Jadid Abdulkadir:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Hitham Seddig Alhussian:** Writing – review & editing, Resources. **Safwan Mahmood Al-Selwi:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Conceptualization. **Alawi Alqushaibi:** Writing – review & editing, Formal analysis, Conceptualization. **Mohammed Gamal Ragab:** Writing – review & editing, Formal analysis. **Suliman Mohamed Fati:** Writing – review & editing, Formal analysis.

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.

Data availability

The data used to conduct this SLR is available via this link: https://github.com/SafwanAlselwi/DDPG.

Acknowledgements

The authors would like to thank the Ministry of Higher Education (MOHE), Malaysia for providing financial assistance under the Fundamental Research Grant Scheme (FRGS/1/2022/ICT02/UTP/02/4) and Universiti Teknologi PETRONAS under the Yayasan Universiti Teknologi PETRONAS (YUTP-FRG/015LC0-521) for providing the required facilities to conduct this research work. The authors would like also to thank Prince Sultan University for their generous support.

Appendix A

Different search schemes were employed by all literature search databases. Across all databases, we either used operators and wildcards to create advanced search queries, or we used the advanced search form provided on each database's website. On March 6, 2023, the following links provide advanced search guidelines, which can be accessed on the latest search date.

Scopus: (TITLE-ABS-KEY (ddpg OR "deep deterministic policy gradient") AND TITLE-ABS-KEY (drl OR "Deep reinforcement learning") AND TITLE-ABS-KEY (optimiz*) OR TITLE-ABS-KEY (hyperparameter)) AND PUBYEAR > 2017 AND PUBYEAR < 2024 AND (LIMIT-TO (PUBSTAGE, "final")) AND (LIMIT-TO (DOCTYPE, "ar") OR LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "ch") OR LIMIT-TO (DOCTYPE, "re")) AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (EXACTKEYWORD, "Deep Deterministic Policy Gradient") OR LIMIT-TO (EXACTKEYWORD, "Deep Reinforcement Learning") OR LIMIT-TO (EXACTKEYWORD, "DDPG") OR LIMIT-TO (EXACTKEYWORD, "Deep Reinforcement Learning (DRL)") OR LIMIT-TO (EXACTKEYWORD, "Deep Deterministic Policy Gradient (DDPG)"))

Sciencedirect: ddpg OR "deep deterministic policy gradient") AND (drl OR "Deep reinforcement learning") AND (optimization OR optimized OR hyperparameter)

Table 12
PRISMA 2020 checklist.

Section/Topic	Item	Location
Title	1	Page 1: Title
Abstract	2	Page 1: Abstract
Introduction	3	Page 1: Section 1
	4	Page 7: Section 3.2
Methods	5	Page 9: Table 3
	6	Pages (7, 9): Table 5
	7	Pages (8, 7): Sections 3.3, Table 2
	8	Page 8: Section 3.3, Table 3
	9	Page 9: Section 3.5
	10-15	Not applicable
Synthesis of Data and Analysis	16a	Page 8: Flowchart 6
	16b	Not applicable
	17	Pages (12-17): Section 4
	18-22	Not applicable
Other Information	24	Not applicable
	25	Page 21: Acknowledgments
	26	Page 21: Declarations of Competing Interests
	27	Page 21: Data Availability

Web of science: TS = ((ddpg OR “deep deterministic policy gradient”) AND (drl OR “Deep reinforcement learning”) AND (optimization OR optimized OR hyperparameter)) and 2023 or 2022 or 2021 or 2020 or 2019 or 2018 (Publication Years) and English (Languages) and Early Access or Proceeding Paper or Article (Document Types).

Appendix B

The PRISMA 2020 Checklist is taken from http://prisma-statement.org/PRISMAStatement/Checklist (last accessed 16/03/2024). Table 12 shows the full PRISMA checklist applied in our SLR with checklist section, number, and location in our paper.

Abbreviations

All abbreviations used in this manuscript are defined in Table 13.

References

[1] K. Arulkumaran, M.P. Deisenroth, M. Brundage, A.A. Bharath, Deep reinforcement learning: a brief survey, IEEE Signal Process. Mag. 34 (2017) 26–38.
[2] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.
[3] G. Dulac-Arnold, N. Levine, D.J. Mankowitz, J. Li, C. Paduraru, S. Gowal, T. Hester, Challenges of real-world reinforcement learning: definitions, benchmarks and analysis, Mach. Learn. 110 (2021) 2419–2468.
[4] A. Rehman, T. Saba, K. Haseeb, T. Alam, J. Lloret, Sustainability model for the Internet of health things (ioht) using reinforcement learning with mobile edge secured services, Sustainability 14 (2022) 12185.
[5] K. Zhao, J. Song, Y. Luo, Y. Liu, Research on game-playing agents based on deep reinforcement learning, Robotics 11 (2022) 35.
[6] K. Arshad, R.F. Ali, A. Muneer, I.A. Aziz, S. Naseer, N.S. Khan, S.M. Taib, Deep reinforcement learning for anomaly detection: a systematic review, IEEE Access (2022).
[7] B. Singh, R. Kumar, V.P. Singh, Reinforcement learning in robotic applications: a comprehensive survey, Artif. Intell. Rev. (2022) 1–46.
[8] A.A. Shahid, D. Piga, F. Braghin, L. Roveda, Continuous control actions learning and adaptation for robotic manipulation through reinforcement learning, Auton. Robots 46 (2022) 483–498.
[9] K. Arshad, R.F. Ali, A. Muneer, I.A. Aziz, S. Naseer, N.S. Khan, S.M. Taib, Deep reinforcement learning for anomaly detection: a systematic review, IEEE Access (2022).
[10] M.-S. Kim, G. Eoh, T.-H. Park, Decision making for self-driving vehicles in unexpected environments using efficient reinforcement learning methods, Electronics 11 (2022) 1685.
[11] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, M. Riedmiller, Deterministic policy gradient algorithms, in: Proceedings of the 31st International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 32, PMLR, Bejing, China, 2014, pp. 387–395.
[12] H. Alturkistani, M.A. El-Affendi, Optimizing cybersecurity incident response decisions using deep reinforcement learning, Int. J. Electr. Comput. Eng. 12 (2022) 6768.
[13] C. Qiu, Y. Hu, Y. Chen, B. Zeng, Deep deterministic policy gradient (ddpg)-based energy harvesting wireless communications, IEEE Int. Things J. 6 (2019) 8577–8588.
[14] Y. Hou, L. Liu, Q. Wei, X. Xu, C. Chen, A novel ddpg method with prioritized experience replay, in: 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017, pp. 316–321.
[15] J. Xu, Z. Hou, W. Wang, B. Xu, K. Zhang, K. Chen, Feedback deep deterministic policy gradient with fuzzy reward for robotic multiple peg-in-hole assembly tasks, IEEE Trans. Ind. Inform. 15 (2018) 1658–1667.
[16] E.H.H. Sumiea, S.J. Abdulkadir, M.G. Ragab, S.M. Al-Selwi, S.M. Fati, A. AlQushaibi, H. Alhussian, Enhanced deep deterministic policy gradient algorithm using grey wolf optimizer for continuous control tasks, IEEE Access 11 (2023) 139771–139784, https://doi.org/10.1109/ACCESS.2023.3341507.

Table 13
Abbreviations.

Abbreviation	Definition
A2C	Advantage Actor-Critic
ACKTR	Actor-Critic using Kronecker-Factored Trust Region
AI	Artificial Intelligence
CAVs	Connected and Autonomous Vehicles
D2D	Device to Device
D3PG	Deep Differentiable Deterministic Policy Gradients
D3QN	Dueling Double Deep Q Networks
D4PG	Distributed Distributional DDPG
DDPG	Deep Deterministic Policy Gradient
DL	Deep Learning
DNN	Deep Neural Network
DP	Dynamic Programming
DQN	Deep Q-Network
EMS	Energy Management System
GA	Genetic Algorithm
HER	Hindsight Experience Replay
HEV	Hybrid Electric Vehicle
IoT	Internet of Things
IOUT	Internet-of-Underwater-Things Devices
IRS	Intelligent Reflecting Surface
JCORA	Joint Optimization Problem of Computation Offloading and Resource Allocation
LIBs	Lithium-Ion Batteries
LSTM	Long Short-Term Memory
MADDPG	Multi-Agent DDPG
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
MEC	Mobile Edge Computing
PER	Prioritized Experience Replay
PPO	Proximal Policy Optimization
RL	Reinforcement Learning
RNN	Recurrent Neural Network
SAC	Soft Actor-Critic
SARSA	State-Action-Reward-State-Action
TD	Temporal Difference
TRPO	Trust Region Policy Optimization
UAVs	Unmanned Aerial Vehicles
VLC	Visible Light Communication

[17] C. Qiu, Y. Hu, Y. Chen, B. Zeng, Deep deterministic policy gradient (ddpg)-based energy harvesting wireless communications, IEEE Int. Things J. 6 (2019) 8577–8588.
[18] N. Casas, Deep deterministic policy gradient for urban traffic light control, arXiv preprint arXiv:1703.09035, 2017.
[19] K. Li, Y. Emami, W. Ni, E. Tovar, Z. Han, Onboard deep deterministic policy gradients for online flight resource allocation of uavs, IEEE Netw. Lett. 2 (2020) 106–110.
[20] M. Sewak, M. Sewak, Deterministic policy gradient and the ddpg: deterministic-policy-gradient-based approaches, in: Deep Reinforcement Learning: Frontiers of Artificial Intelligence, 2019, pp. 173–184.
[21] A. Gupta, A.S. Khwaja, A. Anpalagan, L. Guan, B. Venkatesh, Policy-gradient and actor-critic based state representation learning for safe driving of autonomous vehicles, Sensors 20 (2020) 5991.
[22] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv preprint arXiv:1509.02971, 2015.
[23] R. Nian, J. Liu, B. Huang, A review on reinforcement learning: introduction and applications in industrial process control, Comput. Chem. Eng. 139 (2020) 106886.
[24] J. Schulman, Optimizing expectations: from deep reinforcement learning to stochastic computation graphs, Thesis, UC Berkeley, 2016.
[25] A.T. Azar, A. Koubaa, N. Ali Mohamed, H.A. Ibrahim, Z.F. Ibrahim, M. Kazim, A. Ammar, B. Benjdira, A.M. Khamis, I.A. Hameed, et al., Drone deep reinforcement learning: a review, Electronics 10 (2021) 999.
[26] C. Tallec, L. Blier, Y. Ollivier, Making deep q-learning methods robust to time discretization, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 6096–6104.
[27] S. Fujimoto, H. van Hoof, D. Meger, Addressing function approximation error in actor-critic methods, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 1587–1596.
[28] T. Haarnoja, A. Zhou, P. Abbeel, S. Levine, Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor, in: Proceedings of the 35th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 80, PMLR, 2018, pp. 1861–1870.
[29] R. Lowe, Y.I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, I. Mordatch, Multi-agent actor-critic for mixed cooperative-competitive environments, Adv. Neural Inf. Process. Syst. 30 (2017).
[30] G. Barth-Maron, M.W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. Tb, A. Muldal, N. Heess, T. Lillicrap, Distributed distributional deterministic policy gradients, arXiv preprint arXiv:1804.08617, 2018.
[31] Y. Dong, C. Yu, H. Ge, D3pg: decomposed deep deterministic policy gradient for continuous control, in: Distributed Artificial Intelligence, Springer International Publishing, Cham, 2020, pp. 40–54.
[32] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, The prisma 2020 statement: an updated guideline for reporting systematic reviews, Int. J. Surg. 88 (2021) 105906.

[33] N.R. Haddaway, M.J. Page, C.C. Pritchard, L.A. McGuinness, Prisma2020: an R package and shiny app for producing prisma 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis, Campbell Syst. Rev. 18 (2022) e1230, https://doi.org/10.1002/cl2.1230.
[34] K. Xia, J. Feng, C. Yan, C. Duan, Beidou short-message satellite resource allocation algorithm based on deep reinforcement learning, Entropy 23 (2021), https://doi.org/10.3390/e23080932.
[35] X. Guo, T. Liu, B. Tang, X. Tang, J. Zhang, W. Tan, S. Jin, Transfer deep reinforcement learning-enabled energy management strategy for hybrid tracked vehicle, IEEE Access 8 (2020) 165837–165848, https://doi.org/10.1109/ACCESS.2020.3022944.
[36] J. Chen, L. Guo, J. Jia, J. Shang, X. Wang, Resource allocation for irs assisted sgf noma transmission: a madrl approach, IEEE J. Sel. Areas Commun. 40 (2022) 1302–1316, https://doi.org/10.1109/JSAC.2022.3144726.
[37] M. Zhu, X. Wang, Y. Wang, Human-like autonomous car-following model with deep reinforcement learning, Transp. Res., Part C, Emerg. Technol. 97 (2018) 348–368, https://doi.org/10.1016/j.trc.2018.10.024.
[38] M. Li, Z.B. Li, C.C. Xu, T. Liu, Deep reinforcement learning-based vehicle driving strategy to reduce crash risks in traffic oscillations, Transp. Res. Rec. 2674 (2020) 42–54, https://doi.org/10.1177/0361198120937976.
[39] F. Guo, Z. Wu, A deep reinforcement learning approach for autonomous car racing, in: E-Learning and Games, Springer International Publishing, Cham, 2019, pp. 203–210.
[40] Y. Zhang, Z.Y. Mou, F.F. Gao, J. Jiang, R.J. Ding, Z. Han, Uav-enabled secure communications by multi-agent deep reinforcement learning, IEEE Trans. Veh. Technol. 69 (2020) 11599–11611, https://doi.org/10.1109/tvt.2020.3014788.
[41] T.M. Ho, K.K. Nguyen, M. Cheriet, Uav control for wireless service provisioning in critical demand areas: a deep reinforcement learning approach, IEEE Trans. Veh. Technol. 70 (2021) 7138–7152, https://doi.org/10.1109/TVT.2021.3088129.
[42] Z. Xu, J. Qi, M. Wang, C. Wu, G. Yang, Compensation control of uav based on deep deterministic policy gradient, in: 2022 41st Chinese Control Conference (CCC), 2022, pp. 2289–2296.
[43] Y. Yu, J. Tang, J. Huang, X. Zhang, D.K.C. So, K.K. Wong, Multi-objective optimization for uav-assisted wireless powered iot networks based on extended ddpg algorithm, IEEE Trans. Commun. 69 (2021) 6361–6374, https://doi.org/10.1109/TCOMM.2021.3089476.
[44] M. Samir, C. Assi, S. Sharafeddine, D. Ebrahimi, A. Ghrayeb, Age of information aware trajectory planning of uavs in intelligent transportation systems: a deep learning approach, IEEE Trans. Veh. Technol. 69 (2020) 12382–12395, https://doi.org/10.1109/tvt.2020.3023861.
[45] H.X. Zhang, F. Wang, J.H. Wang, B. Cui, Robot grasping method optimization using improved deep deterministic policy gradient algorithm of deep reinforcement learning, Rev. Sci. Instrum. 92 (2021) 11, https://doi.org/10.1063/5.0034101.
[46] G. Hao, Z. Fu, X. Feng, Z. Gong, P. Chen, D. Wang, W. Wang, Y. Si, A deep deterministic policy gradient approach for vehicle speed tracking control with a robotic driver, IEEE Trans. Autom. Sci. Eng. 19 (2022) 2514–2525, https://doi.org/10.1109/TASE.2021.3088004.
[47] C.H. Min, J.B. Song, End-to-end robot manipulation using demonstration-guided goal strategies, in: 16th International Conference on Ubiquitous Robots (UR), International Conference on Ubiquitous Robots and Ambient Intelligence, Ieee, New York, 2019, pp. 159–164.
[48] Z. Mlika, S. Cherkaoui, Deep deterministic policy gradient to minimize the age of information in cellular v2x communications, IEEE Trans. Intell. Transp. Syst. 23 (2022) 23597–23612, https://doi.org/10.1109/TITS.2022.3190799.
[49] L. Ale, S.A. King, N. Zhang, A.R. Sattar, J. Skandaraniyam, D3pg: Dirichlet ddpg for task partitioning and offloading with constrained hybrid action space in mobile-edge computing, IEEE Int. Things J. 9 (2022) 19260–19272, https://doi.org/10.1109/JIOT.2022.3166110.
[50] K. Sagiraju, S. Mogalla, Deployment of deep reinforcement learning and market sentiment aware strategies in automated stock market prediction, Int. J. Eng. Trends Technol. 70 (2022) 43–53, https://doi.org/10.14445/22315381/IJETT-V70I1P205.
[51] M. Kong, J. So, Empirical analysis of automated stock trading using deep reinforcement learning, Appl. Sci. (Switzerland) 13 (2023), https://doi.org/10.3390/app13010633.
[52] Y. Ye, D. Qiu, M. Sun, D. Papadaskalopoulos, G. Strbac, Deep reinforcement learning for strategic bidding in electricity markets, IEEE Trans. Smart Grid 11 (2020) 1343–1355, https://doi.org/10.1109/TSG.2019.2936142.
[53] T. Chau, M.-T. Nguyen, D.-V. Ngo, A.-D.T. Nguyen, T.-H. Do, Deep reinforcement learning methods for automation forex trading, in: 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), 2022, pp. 671–676.
[54] Y. Zhao, I.G. Niemegeers, S.M.H. De Groot, Dynamic power allocation for cell-free massive mimo: deep reinforcement learning methods, IEEE Access 9 (2021) 102953–102965, https://doi.org/10.1109/ACCESS.2021.3097243.
[55] S.F. Zheng, H. Liu, Improved multi-agent deep deterministic policy gradient for path planning-based crowd simulation, IEEE Access 7 (2019) 147755–147770, https://doi.org/10.1109/access.2019.2946659.
[56] F. Meng, P. Chen, L.N. Wu, J.L. Cheng, Power allocation in multi-user cellular networks: deep reinforcement learning approaches, IEEE Trans. Wirel. Commun. 19 (2020) 6255–6267, https://doi.org/10.1109/twc.2020.3001736.
[57] K. Zheng, X. Jia, K. Chi, X. Liu, Ddpg-based joint time and energy management in ambient backscatter-assisted hybrid underlay crns, IEEE Trans. Commun. 71 (2023) 441–456, https://doi.org/10.1109/TCOMM.2022.3221422.
[58] T. Zhang, K. Zhu, J.H. Wang, Energy-efficient mode selection and resource allocation for d2d-enabled heterogeneous networks: a deep reinforcement learning approach, IEEE Trans. Wirel. Commun. 20 (2021) 1175–1187, https://doi.org/10.1109/twc.2020.3031436.
[59] B. Zhang, Y. Zou, X. Zhang, G. Du, F. Jiao, N. Guo, Online updating energy management strategy based on deep reinforcement learning with accelerated training for hybrid electric tracked vehicles, IEEE Trans. Transp. Electrif. 8 (2022) 3289–3306, https://doi.org/10.1109/TTE.2022.3156590.
[60] Z. Wei, Z. Quan, J. Wu, Y. Li, J. Pou, H. Zhong, Deep deterministic policy gradient-drl enabled multiphysics-constrained fast charging of lithium-ion battery, IEEE Trans. Ind. Electron. 69 (2022) 2588–2598, https://doi.org/10.1109/TIE.2021.3070514.
[61] J. Chen, H. Xing, Z. Xiao, L. Xu, T. Tao, A drl agent for jointly optimizing computation offloading and resource allocation in mec, IEEE Int. Things J. 8 (2021) 17508–17524, https://doi.org/10.1109/JIOT.2021.3081694.
[62] J. Wang, Y.C. Wang, H.C. Ke, Joint optimization for mec computation offloading and resource allocation in iov based on deep reinforcement learning, Mob. Inf. Syst. 2022 (2022) 11, https://doi.org/10.1155/2022/9230521.
[63] Z. Wang, Y. Wei, F.R. Yu, Z. Han, Utility optimization for resource allocation in multi-access edge network slicing: a twin-actor deep deterministic policy gradient approach, IEEE Trans. Wirel. Commun. 21 (2022) 5842–5856, https://doi.org/10.1109/TWC.2022.3143949.
[64] B. Qu, Y. Bai, Y. Chu, L.E. Wang, F. Yu, X. Li, Resource allocation for mec system with multi-users resource competition based on deep reinforcement learning approach, Comput. Netw. 215 (2022), https://doi.org/10.1016/j.comnet.2022.109181.
[65] B. Liu, B.W. Xu, T. He, W. Yu, F.H. Guo, Hybrid deep reinforcement learning considering discrete-continuous action spaces for real-time energy management in more electric aircraft, Energies 15 (2022) 21, https://doi.org/10.3390/en15176323.
[66] J. Chen, T. Wu, M. Shi, W. Jiang, Porf-ddpg: learning personalized autonomous driving behavior with progressively optimized reward function, Sensors (Switzerland) 20 (2020) 1–19, https://doi.org/10.3390/s20195626.
[67] Y.C. Fu, C.L. Li, F.R. Yu, T.H. Luan, Y. Zhang, An autonomous lane-changing system with knowledge accumulation and transfer assisted by vehicular blockchain, IEEE Int. Things J. 7 (2020) 11123–11136, https://doi.org/10.1109/jiot.2020.2994975.
[68] N.M. Ashraf, R.R. Mostafa, R.H. Sakr, M.Z. Rashad, Optimizing hyperparameters of deep reinforcement learning for autonomous driving based on whale optimization algorithm, PLoS ONE 16 (2021) 24, https://doi.org/10.1371/journal.pone.0252754.
[69] K. Alomari, R.C. Mendoza, D. Goehring, R. Rojas, Path following with deep reinforcement learning for autonomous cars, in: 2nd International Conference on Robotics, Computer Vision and Intelligent Systems (ROBOVIS), Scitepress, SETUBAL, 2021, pp. 173–181.

[70] Y. Zhang, C. Zhang, R. Fan, S. Huang, Y. Yang, Q. Xu, Twin delayed deep deterministic policy gradient-based deep reinforcement learning for energy management of fuel cell vehicle integrating durability information of powertrain, Energy Convers. Manag. 274 (2022), https://doi.org/10.1016/j.enconman.2022.116454.
[71] W. He, Y. Huang, Real-time energy optimization of hybrid electric vehicle in connected environment based on deep reinforcement learning, IFAC-PapersOnLine 54 (2021) 176–181, https://doi.org/10.1016/j.ifacol.2021.10.160.
[72] Z. Wang, Y. Li, C. Ma, X. Yan, D. Jiang, Path-following optimal control of autonomous underwater vehicle based on deep reinforcement learning, Ocean Eng. 268 (2023), https://doi.org/10.1016/j.oceaneng.2022.113407.
[73] Y.S. Sun, X.K. Luo, X.R. Ran, G.C. Zhang, A 2d optimal path planning algorithm for autonomous underwater vehicle driving in unknown underwater canyons, J. Mar. Sci. Eng. 9 (2021) 24, https://doi.org/10.3390/jmse9030252.
[74] Z.Y. Yao, J. Olson, H.S. Yoon, Sensitivity analysis of reinforcement learning-based hybrid electric vehicle powertrain control, SAE Int. J. Commer. Veh. 14 (2021) 409–419, https://doi.org/10.4271/02-14-03-0033.
[75] C.V.S.R. Syavasya, A.L. Muddana, Optimization of autonomous vehicle speed control mechanisms using hybrid ddpg-shap-drl-stochastic algorithm, Adv. Eng. Softw. 173 (2022), https://doi.org/10.1016/j.advengsoft.2022.103245.
[76] B. Hu, J.X. Li, An adaptive hierarchical energy management strategy for hybrid electric vehicles combining heuristic domain knowledge and data-driven deep reinforcement learning, IEEE Trans. Transp. Electrif. 8 (2022) 3275–3288, https://doi.org/10.1109/tte.2021.3132773.
[77] S.C. Li, W.H. Hu, D. Cao, T. Dragicevic, Q. Huang, Z. Chen, F. Blaabjerg, Electric vehicle charging management based on deep reinforcement learning, J. Mod. Power Syst. Clean Energy 10 (2022) 719–730, https://doi.org/10.35833/mpce.2020.000460.
[78] X.L. Tang, J.X. Chen, H.Y. Pu, T. Liu, A. Khajepour, Double deep reinforcement learning-based energy management for a parallel hybrid electric vehicle with engine start-stop strategy, IEEE Trans. Transp. Electrif. 8 (2022) 1376–1388, https://doi.org/10.1109/tte.2021.3101470.
[79] W.W. Huo, D. Chen, S. Tian, J.W. Li, T.Y. Zhao, B. Liu, Lifespan-consciousness and minimum- consumption coupled energy management strategy for fuel cell hybrid vehicles via deep reinforcement learning, Int. J. Hydrog. Energy 47 (2022) 24026–24041, https://doi.org/10.1016/j.ijhydene.2022.05.194.
[80] S. Zhou, S. Fei, Y. Feng, Deep reinforcement learning based uav-assisted maritime network computation offloading strategy, in: 2022 IEEE/CIC International Conference on Communications in China (ICCC), 2022, pp. 890–895.
[81] C.H. Liu, X. Ma, X. Gao, J. Tang, Distributed energy-efficient multi-uav navigation for long-term communication coverage by deep reinforcement learning, IEEE Trans. Mob. Comput. 19 (2020) 1274–1285, https://doi.org/10.1109/TMC.2019.2908171.
[82] S. Zhang, R. Cao, Multi-objective optimization for uav-enabled wireless powered iot networks: an lstm-based deep reinforcement learning approach, IEEE Commun. Lett. 26 (2022) 3019–3023, https://doi.org/10.1109/LCOMM.2022.3210660.
[83] Y. Li, X.H. Qiu, X.D. Liu, Q.L. Xia, Deep reinforcement learning and its application in autonomous fitting optimization for attack areas of ucavs, J. Syst. Eng. Electron. 31 (2020) 734–742, https://doi.org/10.23919/jsee.2020.000048.
[84] Y. Cui, D. Deng, C. Wang, W. Wang, Joint trajectory and power optimization for energy efficient uav communication using deep reinforcement learning, in: IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), 2021, pp. 1–6.
[85] M. Zhang, S. Fu, Q.L. Fan, Joint 3d deployment and power allocation for uav-bs: a deep reinforcement learning approach, IEEE Wirel. Commun. Lett. 10 (2021) 2309–2312, https://doi.org/10.1109/lwc.2021.3100388.
[86] A. Barnawi, N. Kumar, I. Budhiraja, K. Kumar, A. Almansour, B. Alzahrani, Deep reinforcement learning based trajectory optimization for magnetometer-mounted uav to landmine detection, Comput. Commun. 195 (2022) 441–450, https://doi.org/10.1016/j.comcom.2022.09.002.
[87] A. Gao, Q. Wang, K. Chen, W. Liang, Multi-uav assisted offloading optimization: a game combined reinforcement learning approach, IEEE Commun. Lett. 25 (2021) 2629–2633, https://doi.org/10.1109/LCOMM.2021.3078469.
[88] D. Wang, Q. Liu, J. Tian, Y. Zhi, J. Qiao, J. Bian, Deep reinforcement learning for caching in d2d-enabled uav-relaying networks, in: 2021 IEEE/CIC International Conference on Communications in China (ICCC), 2021, pp. 635–640.
[89] X.F. Guo, Y.B. Chen, Y. Wang, Learning-based robust and secure transmission for reconfigurable intelligent surface aided millimeter wave uav communications, IEEE Wirel. Commun. Lett. 10 (2021) 1795–1799, https://doi.org/10.1109/lwc.2021.3081464.
[90] A.F.U. Din, I. Mir, F. Gul, S. Mir, N. Saeed, T. Althobaiti, S.M. Abbas, L. Abualigah, Deep reinforcement learning for integrated non-linear control of autonomous uavs, Processes 10 (2022), https://doi.org/10.3390/pr10071307.
[91] A. Sehgal, N. Ward, H.M. La, C. Papachristos, S. Louis, Ga+ddpg+her: genetic algorithm-based function optimizer in deep reinforcement learning for robotic manipulation tasks, in: 2022 Sixth IEEE International Conference on Robotic Computing (IRC), 2022, pp. 85–86.
[92] J. Yang, G. Peng, Ddpg with meta-learning-based experience replay separation for robot trajectory planning, in: 2021 7th International Conference on Control, Automation and Robotics (ICCAR), 2021, pp. 46–51.
[93] S.K. Rajendran, F.T. Zhang, Design, modeling, and visual learning-based control of soft robotic fish driven by super-coiled polymers, Front. Robot. AI 8 (2022) 13, https://doi.org/10.3389/frobt.2021.809427.
[94] Q. Liu, Z. Liu, B. Xiong, W. Xu, Y. Liu, Deep reinforcement learning-based safe interaction for industrial human-robot collaboration using intrinsic reward function, Adv. Eng. Inform. 49 (2021), https://doi.org/10.1016/j.aei.2021.101360.
[95] X. Li, W.W. Shang, S. Cong, Model-based reinforcement learning for robot control, in: 5th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM), Ieee, New York, 2020, pp. 300–305.
[96] S. Dankwa, W. Zheng, Twin-delayed ddpg: a deep reinforcement learning technique to model a continuous movement of an intelligent robot agent, in: Proceedings of the 3rd International Conference on Vision, Image and Signal Processing, Association for Computing Machinery, New York, NY, USA, 2020.
[97] Z. Li, C. Xiao, Z. Liu, X. Guo, Multi-robot cooperation learning based on Powell deep deterministic policy gradient, in: Intelligent Robotics and Applications, Springer International Publishing, Cham, 2022, pp. 77–87.
[98] P. Li, X. Ding, W. Ren, Research on path planning of cloud robot in dynamic environment based on improved ddpg algorithm, in: 2021 China Automation Congress (CAC), 2021, pp. 3561–3566.
[99] D. Jiang, Z.Q. Cai, Z.Z. Liu, H.J. Peng, Z.G. Wu, An integrated tracking control approach based on reinforcement learning for a continuum robot in space capture missions, J. Aerosp. Eng. 35 (2022) 10, https://doi.org/10.1061/(asce)as.1943-5525.0001426.
[100] Y. Zou, Y. Xie, C. Zhang, S. Gong, D.T. Hoang, D. Niyato, Optimization-driven hierarchical deep reinforcement learning for hybrid relaying communications, in: 2020 IEEE Wireless Communications and Networking Conference (WCNC), 2020, pp. 1–6.
[101] L. Hu, S. Bi, Q. Liu, Y. Jiang, C. Chen, Intelligent reflecting surface aided covert wireless communication exploiting deep reinforcement learning, Wirel. Netw. 29 (2023) 877–889, https://doi.org/10.1007/s11276-022-03037-2.
[102] Z.W. Wang, S.F. Jin, L.H. Liu, C. Fang, M. Li, S. Guo, Design of intelligent connected cruise control with vehicle-to-vehicle communication delays, IEEE Trans. Veh. Technol. 71 (2022) 9011–9025, https://doi.org/10.1109/tvt.2022.3177008.
[103] D.A. Saifaldeen, B.S. Ciftler, M.M. Abdallah, K.A. Qaraqe, Drl-based irs-assisted secure visible light communications, IEEE Photonics J. 14 (2022), https://doi.org/10.1109/JPHOT.2022.3178852.
[104] I. Budhiraja, N. Kumar, S. Tyagi, Deep-reinforcement-learning-based proportional fair scheduling control scheme for underlay d2d communication, IEEE Int. Things J. 8 (2021) 3143–3156, https://doi.org/10.1109/JIOT.2020.3014926.
[105] Y. Chen, S. Han, G. Chen, J. Yin, K.N. Wang, J. Cao, A deep reinforcement learning-based wireless body area network offloading optimization strategy for healthcare services, Health Inf. Sci. Syst. 11 (2023), https://doi.org/10.1007/s13755-023-00212-3.
[106] Y. Lee, J.H. Lee, Y.C. Ko, Beamforming optimization for irs-assisted mmwave v2i communication systems via reinforcement learning, IEEE Access 10 (2022) 60521–60533, https://doi.org/10.1109/ACCESS.2022.3181152.

[107] B.S. Ciftler, A. Alwarafy, M. Abdallah, Distributed drl-based downlink power allocation for hybrid rf/vlc networks, IEEE Photonics J. 14 (2022) 10, https://doi.org/10.1109/jphot.2021.3139678.
[108] Z. Shi, X. Xie, H. Lu, H. Yang, J. Cai, Z. Ding, Deep reinforcement learning-based multidimensional resource management for energy harvesting cognitive noma communications, IEEE Trans. Commun. 70 (2022) 3110–3125, https://doi.org/10.1109/TCOMM.2021.3126626.
[109] D. Kwon, J. Jeon, S. Park, J. Kim, S. Cho, Multiagent ddpg-based deep learning for smart ocean federated learning iot networks, IEEE Int. Things J. 7 (2020) 9895–9903, https://doi.org/10.1109/jiot.2020.2988033.
[110] M. Ma, W. Du, L. Wang, C. Ding, S. Liu, Research on the multi-timescale optimal voltage control method for distribution network based on a dqn-ddpg algorithm, Front. Energy Res. 10 (2023), https://doi.org/10.3389/fenrg.2022.1097319.
[111] A.A. Baktayan, I.A. Al-Baltah, A.A.A. Ghani, Intelligent pricing model for task offloading in unmanned aerial vehicle mounted mobile edge computing for vehicular network, J. Commun. Softw. Syst. 18 (2022) 111–123, https://doi.org/10.24138/jcomss-2021-0154.
[112] Y.N. Liu, X.B. Wang, J. Mei, G. Boudreau, H. Abou-Zeid, A.B. Sediq, Situation-aware resource allocation for multi-dimensional intelligent multiple access: a proactive deep learning framework, IEEE J. Sel. Areas Commun. 39 (2021) 116–130, https://doi.org/10.1109/jsac.2020.3036969.
[113] H. Yang, X.-Y. Liu, S. Zhong, A. Walid, Deep reinforcement learning for automated stock trading: an ensemble strategy, in: Proceedings of the First ACM International Conference on AI in Finance, ICAIF '20, Association for Computing Machinery, 2021.
[114] M. Vishal, Y. Satija, B.S. Babu, Trading agent for the Indian stock market scenario using actor-critic based reinforcement learning, in: 2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), 2021, pp. 1–5.
[115] Y.K. Liu, H.G. Liang, Y.Y. Xiao, H.F. Zhang, J.X. Zhang, L. Zhang, L.H. Wang, Logistics-involved service composition in a dynamic cloud manufacturing environment: a ddpg-based approach, Robot. Comput.-Integr. Manuf. 76 (2022) 14, https://doi.org/10.1016/j.rcim.2022.102323.
[116] H.F. Li, M. Hai, Y.J. Zhang, P.C. Li, A novel stock portfolio model based on deep reinforcement learning, J. Nonlinear Convex Anal. 22 (2021) 1791–1804.
[117] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: International Conference on Machine Learning, PMLR, 2015, pp. 448–456.
[118] L. Pinto, J. Davidson, R. Sukthankar, A. Gupta, Robust adversarial reinforcement learning, in: Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 2817–2826.
[119] Y. Duan, X. Chen, R. Houthooft, J. Schulman, P. Abbeel, Benchmarking deep reinforcement learning for continuous control, in: M.F. Balcan, K.Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 48, PMLR, New York, New York, USA, 2016, pp. 1329–1338.
[120] T. Silver, K. Allen, J. Tenenbaum, L. Kaelbling, Residual policy learning, arXiv preprint arXiv:1812.06298, 2018.
[121] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, W. Zaremba, Openai gym, arXiv preprint arXiv:1606.01540, 2016.
[122] E. Todorov, T. Erez, Y. Tassa, Mujoco: a physics engine for model-based control, in: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 5026–5033.
[123] B. Wymann, E. Espié, C. Guionneau, C. Dimitrakakis, R. Coulom, A. Sumner, Torcs, the open racing car simulator, Software available at http://torcs.sourceforge.net 4 (2000) 2.