IJRRS
International Journal of Reliability, Risk
Saftey: Theory and Application

Online ISSN: 2676-3346

Vol. 4/ Issue 2/ 2021/ pp. 81-89
DOI : 10.30699/IJRRS.4.2.9
Received: 2022.04.19, Accepted: 2022.09.04

Review Article

An overview of reinforcement learning and deep reinforcement learning for condition-based maintenance

Zahra Dehghani Ghobadi[1]*, Firoozeh Haghighi[1], Abdollah Safari[1]

1. School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran

* **ghobadi.dehghani@ut.ac.ir**

Abstract

Condition-based maintenance (CBM) involves making decisions on maintenance based on the actual deterioration conditions of the components. It consists of a chain of states representing various stages of deterioration and a set of maintenance actions. Therefore, condition-based maintenance is a sequential decision-making problem. Reinforcement Learning(RL) is a subfield of Machine Learning proposed for automated decision-making. This article provides an overview of reinforcement learning and deep reinforcement learning methods that have been used so far in condition-based maintenance optimization.

Keyword: Condition-based maintenance; Deep reinforcement learning; Markov decision process; Reinforcement learning.

1. Introduction

Industrial systems are in general subject to degradation because of usage and exposure to environmental factors. This degradation eventually leads to system failure, resulting in safety issues, equipment damage, quality issues, and unexpected machine unavailability [1]. A few decades ago, maintenance was mostly considered something that had to be done after such a failure, but it was also something that was difficult to manage. Maintenance is widely recognized as an essential business function and a critical element of asset management [2]. To keep a system ready for operation over a specified time frame, maintenance actions are required. Traditionally, maintenance actions are classified into corrective maintenance (CM) and preventive maintenance (PM) [3].In CM, a failed system is replaced by a new one, while PM includes specific actions proposed to avoid system failure or reduce the risk of system failure. Recently, another maintenance strategy, the so-called CBM, has received increasing attention thanks to the development of sensor technology. In CBM, the real-time condition of a system is monitored to determine what maintenance needs to be performed [3].

CBM involves making decisions on maintenance based on the actual deterioration conditions of the components [1]. It consists of a chain of states representing various stages of deterioration and a set of maintenance actions [1]. Therefore, CBM is a sequential decision-making problem. Such sequential decision-making problems, often modelled as Markov decision processes (MDPs), could be solved by reinforcement learning (RL) algorithms that have been recently taken attention [4]. Thus, as an optimization tool in the dynamic, uncertain environment, RL could provide an optimal decision strategy (policy) for the CBM problem [5]. For this purpose, the maintenance problem is first converted into an RL framework; then, RL algorithms are applied to obtain an optimal policy[7]. This work aims to review the application of RL as a subfield of Machine Learning (ML) in the maintenance model field. In the following, we first review RL and its algorithms briefly.

RL is a subfield of ML focusing on Artificial Intelligence (AI) which deals with learning from repeated interactions with an environment [6]. A learner (decision maker) is called an agent who interacts with the environment by performing specific actions and receiving feedback from the environment [7]. The feedback is usually termed as a reward. The agent's goal (objective) is to maximize cumulative rewards by learning to perform better [7].An MDP usually describes the environment, consisting of a state space, an action space, a reward function, and state transition probabilities.Therefore, MDP for an RL problem has the following components[11,17,9].

- S is a set of states, and at each time step t, the state is $s_t \in S$.
- $A(S)$ is the set of possible actions, and the action at time t and in state s_t is $a_t \in A(s_t)$.
- $P_{ss'}^{a} = P(s_{t+1} = s' | s_t = s, a_t)$ is the transition probability of beginning in state s' at time $t + 1$, if the system was in state s at time t, and the agent chooses action a_t.
- $r_t$is the reward at time t.
  $\gamma \in (0, 1)$ is a discount factor, and the discount factor essentially determines how much the RL agents care about the rewards in the distant future relative to those in the immediate future.

Figure 1 shows agent-environment interactions in an MDP (for more details, see, e.g., [8]).

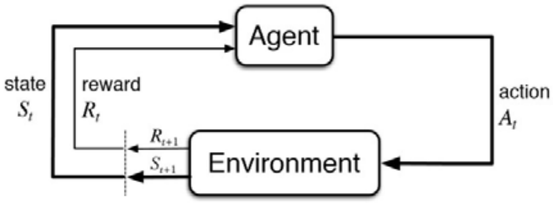

Figure 1. A typical Reinforcement Learning cycle [7]

The maintenance of a system is usually planned based on the system failure mechanism. Generally, a system can fail due to degradation, shock, or both. If the system failure is only because of degradation, then a degradation model (a stochastic process or a deterministic path) is used to model the failure mechanism. In this case, CBM is adjusted based on the information received from the system degradation. If the system failure occurs due to the shocks, a shock model is considered to model the failure mechanism depending on the types of shocks. In this setup, CBM is designed based on information about the shocks, including the number of shocks, their magnitude, and how shock affects the system failure. In a more complex case where the system failure is modeled jointly based on the degradation and arrival shocks, both degradation and shock information are used in the design of the CBM. We will focus on only the first and third cases here.

The paper is organized as follows. Section "Literature review" introduces previous related research on maintenance policies for complex systems with reinforcement learning. The procedure for the CBM approach and taxonomy of Reinforcement learning algorithms are introduced in the section "Research fundamentals." The third section explains two problems of optimal policy in CBM by RL in which the problem is considered a Markov decision-making problem, and the fourth section describes how a semi-Markov decision process formulates a CBM problem to apply an RL approach. The fifth section explains the CBM problem that is modeled as a continuous-state MDP without discretizing the system degradation state, the sixth section illustrates how to find the optimal CBM policy with Deep reinforcement learning (DRL), and the section "Conclusion" presents the conclusion and future work.

2. Literature review

Only a few studies have investigated RL to find an optimal condition-based maintenance schedule to minimize the cost. Adsule et al. [1], modeled the CBM decision-making problem as a continuous semi-Markov decision process (CSMDP), and applied an RL algorithm. Yousefi et al. [9], modeled the CBM decision-making problem as an MDP and also used an RL algorithm. Peng et al. [10], modeled the problem of CBM as a continuous Markov decision-making process without discretizing the degradation states under a Gaussian process (GP) and then applied an RL algorithm. Mahmoodzadeh et al. [11], proposed the CBM optimal policy using an RL algorithm for gas pipelines. Yousefi et al. [6] presented a DRL method to provide a new dynamic maintenance model for a degrading repairable system subject to degradation and random shocks. Zhang et al. [12] proposed a novel and flexible CBM model based on a custom DRL for multi-component systems with dependent competing risks. Table 1providesa summary of the studies mentioned.

Table 1. Summary of existing literature on CBM using the RL algorithm.

Number	Author and Years	Type Problem	Algorithm	Page number	Reference
1	Adsule et al.	continuous semi-Markov decision process (CSMDP)	SMART	5	[1]
2	Yousefi et al.	Markov decision process (MDP)	Q-learning	3	[10]
3	Peng et al.	MDP without discretizing the states	Gaussian Process for reinforcement learning (GPRL)	6	[11]
4	Mahmood	MDP	Q-learning	4	[12]
5	Yousefi et al.	MDP	Deep Q-learning (DQL)	8	[6]
6	Zhang et al.	MDP	Deep Q-learning (DQL)	7	[13]

3. Research fundamentals

3.1 Procedure for CBM approach

The CBM can be done by (1) gathering product status data and monitoring; (2) making a real-time diagnosis of a product status; (3) estimating the deterioration level of the product, and its repairing cost, which depends on the deterioration level, or its replacement cost, and so on; (4) predicting the time of products abnormality; and (5) executing appropriate actions such as repair, replace, left to use as it is, and disposal. Figure 2 shows the generic procedure for implementing CBM.

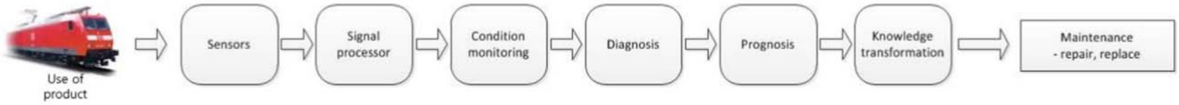

Figure 2. Procedure for CBM approach [14]

3.2 Taxonomy of Reinforcement learning algorithms

The RL algorithms could be classified from different perspectives. Here, we classify the RL algorithms based on whether the environment model is assumed tobe known. A taxonomy of RL algorithms based on such classification is given in Figure 3.

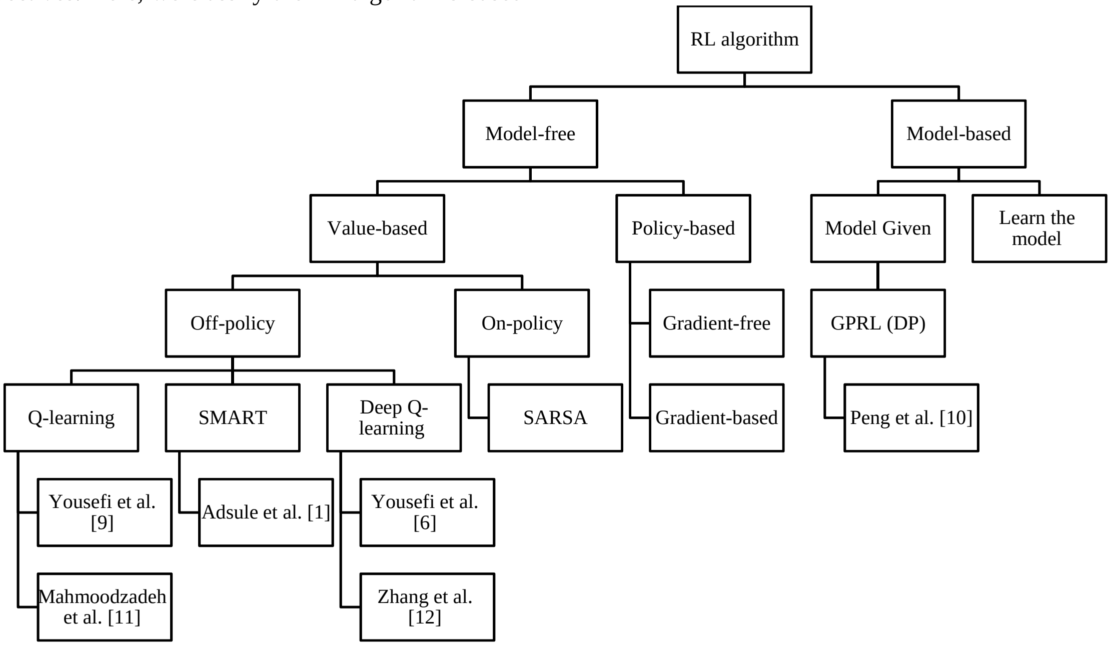

Figure 3. Taxonomy of Reinforcement learning algorithms [8]

Note that a "model" means an ensemble of acquired environmental knowledge. Whether the environment model is used or not, RL algorithms can be classified into model-free and model-based classes [7]. In model-based RL, all elements of the environment MPD are known, and the RL algorithms will use them in learning the optimal policy [7]. The model-based methods can be split into two categories: given model and learning the model [7]. In the given model methods, the reward function and the transition process can be accessed directly by the agent (e.g., Gaussian Process for reinforcement learning [GPRL]) [10].

In contrast, in learning the model methods, the agent can learn the model from interactions with the environment first and then apply the learned model to find the optimal policy [7].Model-based approaches can become impractical in many realistic applications (Huang [18]).Alternatively, the optimal policy can be obtained directly without knowing the environment model. This class is called model-free RL. The model-free methods fall into two main categories: value-based and policy-based. The value-based methods usually imply that first learning the action-value function (Q(s, a): cumulative discounted reward by starting from state s and taking action a), and then obtaining the optimal action corresponds to the highest cumulative discounted reward based on the learned Q (s, a) [8]. Another approach is optimizing the policy directly (without learning Q(s, a)), which is called the policy-based method. The value-based methods are divided into the on-policy and off-policy methods. The on-policy methods learn or improve the policy that the agent is acting upon in its interactions with the environment, such as SARSA [7], whereas the off-policy methods can learn or improve a policy that is different from the one that the agent is using to take action in the environment such as Q-learning [10,12], Deep Q-learning [6, 13] and SMART [1]. With off-policy methods, the experience of other agents interacting with the environment can also be used to find the optimal policy. The policy-based methods are classified into two categories: gradient-based and non-gradient-based. The gradient-based methods can be used to improve parameterized policies, and the non-gradient-based method is applied to optimize less complicated policies. More details about the classification of RL algorithms can be found in [6].

4. Finding the optimal policy with the RL approach to solve the CBM problem as an MDP

Yousefi et al. [9], considered an RL approach to develop a new dynamic CBM policy for multi-component systems with individually repairable components. The following assumptions concerning to failure model have been made in their work.

1. Each component is subject to two competing failures: the process of degradation and random shock.
2. A gamma process is used to model the degradation path of each component.
3. Shock arrivals occur as a homogeneous Poisson process.
4. Each incoming shock may cause the system to fail immediately due to its magnitude, and it also affects the degradation path of the components.

Let X_i(t) be the i^{th} component degradation level.To apply the RL approach, they converted the optimal maintenance problem to an MDP problem based on the following assumptions:

1. State space is $S = \{0,1,2,3,4\}$ where
$$S = \begin{cases} 0 & X_i(t) = 0 \\ 1 & 0 < X_i(t) \le H_i^3 \\ 2 & H_i^3 < X_i(t) \le H_i^2 \\ 3 & H_i^3 < X_i(t) \le H_i^1 \\ 4 & H_i^1 < X_i(t) \end{cases}$$
and H_i^k, $k = 1,2,3$ are some prefixed known degradation thresholds.
2. Actions space is $A = \{a_1, a_2, a_3\}$ where a_1,a_2, anda_3 are “do nothing”,“repair a component”, and “replace a component”, respectively.

Using the Q-learning method, they obtained the optimal maintenance actions for all the system degradation states. As an advantage, this method provides a dynamic maintenance policy for each specific degradation state of the system, which is more beneficial than the fixed maintenance plan. In another study, Mahmoodzadeh et al. [11], proposed a CBM policy via the RL method for gas pipelines. Gas pipeline systems are one of the largest energy infrastructures in the world and are known to be very efficient and reliable. However, this does not mean they are prone to no risk. Corrosion is a significant problem in gas pipelines that imposes large risks, such as ruptures and leakage to the environment and the pipeline system. Therefore, various maintenance actions are performed routinely to ensure the integrity of the pipelines. The costs of corrosion-related maintenance actions are a significant portion of the pipeline's operation and maintenance costs. Minimizing this high cost is a highly compelling subject that many studies have addressed. Mahmoodzadeh et al. [11], investigated the benefits of applied RL techniques to the corrosion-related maintenance management of dry gas pipelines. In the mentioned work, as the first step, the pipeline's corrosion maintenance problem has been converted to a sequential decision-making problem by defining the problem in an MDP format. Because the scope of the research is the corrosion of the pipeline, the state definition should include all the essential information to predict the next corrosion status given the action. Therefore, they initially designed the state definition to include the depth and length of the corrosion. However, instead of directly taking the value of the depth and length, the max-normalized version of them has been considered and removed the agent's dependency on the pipeline's parameters. Equations (1) and (2) define the corrosion depth and length where the maximum corrosion depth is the wall thickness, and the maximum corrosion length has been estimated by running the model for 40 years without maintenance.

$$CDP = \frac{corrosion\ depth}{maximum\ corrosion\ depth} \tag{1}$$

$$CLP = \frac{corrosion\ length}{maximum\ corrosion\ length} \tag{2}$$

Representing the corrosion state with only the depth and length is inaccurate because the next stage of the corrosion is not predictable without knowing the rate of corrosion degradation. Therefore, the corrosion rate has been added to the state variables. They assumed the agent's access to the state variables is feasible only through monthly inspections of the corrosion depth and length. Therefore, the corrosion rate has been derived by comparing the current month’s corrosion with the previous month’s corrosion .Since corrosion is a slow and gradual process, the agent does not need high precision in state representation. The corrosion rate is represented (CRP) as a binary variable with a value of 0 when there is no corrosion aggravation and 1 when the corrosion exacerbates. The following equation formulates the corrosion rate presence as the third state variable.

$$CRP = \begin{cases} 0 \ \ if \ \ CDP_{t-1} = CDP_t \ \ and \ \ CLP_{t-1} = CLP_t \\ 1 \ \ if \ \ CDP_{t-1} \neq CDP_t \ \ and \ \ CLP_{t-1} \neq CLP_t \end{cases} \tag{3}$$

Thus, they have discretized the state variables into 24 bins as shown in Table 2.

Table 2. Discretized representation of the state space [11]

CRP	$\frac{\text{CDP}}{\text{CLP}}$	0-20%	20-40%	40-60%	60-100%
0	0-33%	0	1	2	3
0	33-66%	4	5	6	7
0	66-100+%	8	9	10	11
1	0-33%	12	13	14	15
1	33-66%	16	17	18	19
1	66-100+%	20	21	22	23

A discrete action space of size 5 is considered for the agent as follows, {Do nothing, Batch corrosion inhibitor, Internal coating, Cleaning pigging, Replacement}.The details of the considered maintenance actions are shown in Table 3.

Table 3. The maintenance scheduler set of maintenance actions [11]

Actions	Descriptions	Comments
Do nothing	• No mitigation is done	• The corrosion proceeds • Corrosion inhibitor is added from the inlet of the pipeline
Batch corrosion inhibitor	• A chemical that adsorbs onto the metal surface and reacts with it to form a protective film	• Corrosion rate drop is based on the inhibitor efficiency • Effective only within its lifetime
Internal coating	• An artificial coating that isolates the pipe from the corrosive environment and prevents water from reaching the pipe surface	• No corrosion propagation during its lifetime
Cleaning pigging	• A gadget that effectively cleans up liquids, corrosive solids and debris	• No corrosion propagation during its lifetime
Replacement	• Replace the corroded segment with a new one	• Renew corrosive environment • No more corrosion defects

The total reward after each month has been defined as the algebraic summation of the cost of failure, life extension reward, and cost of maintenance**,** Mahmoodzadeh et al. [11].The approach used in this research is entirely data-driven and model-free. The agent treats the model as a black box that mimics a real pipeline and emits the required data for the learning process. The Q-learning algorithm for the problem of pipeline optimal corrosion maintenance management has been applied. The results show that applying the proposed condition-based maintenance management technique can reduce up to 58% of the maintenance costs compared to a periodic maintenance policy while securing pipeline reliability.

5. Finding the optimal policy by RL approach to solve the CBM problem as a continuous semi-Markov decision process

Adsule et al. [1] modeled the CBM decision problem as a continuous semi-Markov decision process (CSMDP). SMDPs generalize MDPs by allowing the state transitions to occur continuously and irregularly. They employed an RL algorithm to learn optimal maintenance decisions and inspection schedules based on the current health status of a component by maximizing the average reward of a CSMDP for their CBM problem. The following assumptions are made for the model:

1. The health of a component is assessed at different time intervals.
2. A stochastic model is used to capture the deterioration progress as a function of time.
3. A hypothetical component is considered with a hard-facing layer, and it is assumed that the layer thickness decreases over time due to wear.
4. The acceptable minimum layer thickness threshold is known, deterministic, and fixed.
5. The component is failed if the layer thickness is less than the threshold value.

The maintenance action choices considered to be available to the decision-maker are

1. No maintenance action (NA).
2. Minor maintenance (MM): minor maintenance means that a failed system is restored just back to a functioning state. After minor maintenance, the system continues as if nothing had happened. The likelihood of system failure is the same immediately before and after a failure. A minimal repair thus restores the system to an “as bad as old” condition.
3. Replacement through PM.
4. Replacement through CM.

The choice of “no action” means no maintenance action is required and the component is allowed to work in its current state.

In this case, the maintenance action "minor maintenance" (MM) refers to the re-lubrication of the component surface, which will reduce its wear rate. A PM action results in the planned replacement of the components, which means we stop the machine with proper scheduling. The CM happens when the component fails. A reward based on the component's health is inversely proportional to the health index (HI). If the HI value of the component is high, the agent will receive a less negative reward and vice versa. This will motivate the agent to keep the component in a healthy state. In this research, an application of the SMART algorithm is demonstrated for CBM using a case of wear deterioration of a component. The SMART algorithm is a model-free, average-reward algorithm for continuous-time SMDPs. It is a generic algorithm that can be applied to any component that deteriorates with time and usage in which an RL agent prescribes an optimal or near-optimal maintenance action along with the time for the next inspection to minimize the cost. The uniqueness of the approach proposed in the article by Adsule et al. [1] lies in the fact that it attempts to optimize the maintenance action choice and the inspection schedule (with non-constant inspection intervals) simultaneously.

6. Finding the optimal policy by RL approach to solve the CBM problem as a continuous-state MDP

Peng et al. [10], modeled the problem of CBM as a continuous Markov decision-making process without discretizing the degradation states of the system through the RL method and Gaussian process (GP). Gaussian process regression (GPR) has been used as a function approximation to model the state transfer and state value functions in RL setup. The additional assumptions for the system have been listed as follows:

1. A non-repairable system is continuously monitored or periodically inspected before each decision epoch.
2. The condition of the system is a continuous random variable, denoted as $X(t)$, which satisfies the Markovian property.
3. The system fails when $X(t)$ reaches a pre-determined end-of-life threshold, H.
4. Maintenance decisions are made at equally spaced decision epochs based on the observed condition of the system as follows:
    - If the system condition exceeds the end-of-life threshold, $X(t) > H$, then CM is performed, incurring a combined cost of the replacement cost, C_R, and a penalty cost due to downtime, C_D.
    - If the system condition exceeds a threshold for replacement, H_p, where H_p<$X(t) < H$, then PM is implemented even though the system is still functioning, and only the replacement cost, C_R, is incurred.
    - If the system condition is less thanH_p, no action (N) is needed.
5. After a PM or CM action the state of the system becomes as-good-as-new state X (0) = 0.

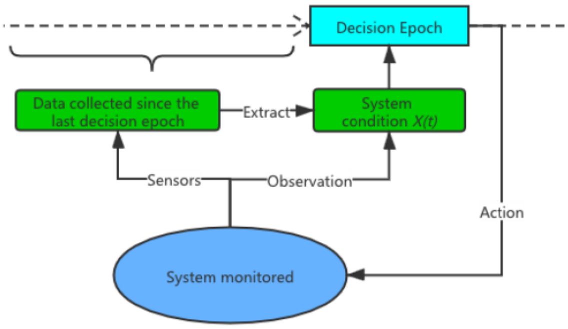

Figure4. The Framework of MDP for CBM [1]

The overall framework of the MDP for the condition-based maintenance proposed in this work is shown in Figure4.The system under maintenance is monitored to obtain its state at each decision epoch. The state of the system can be directly represented by its sensor data, which commonly leads to large state space. To avoid the computational burden, the system states can be extracted from the sensor data or determined by field experts' observations. Based on the system state at each decision epoch, an action is taken on the system maintained.

The MDP model can be fully described by the quintuple $\{\tau, S, A, p(s_0 |s, a), r(s, a)\}$, where τ stands for the countable set of decision epochs. The MDP model can be fully described by the quintuple $\{ \tau, S, A, p(s_0 \ |s, a), r(s, a)\}$, where$\tau$stands for the countable set of decision epochs. In their MDP model for CBM, S is the set of all possible values of the continuous system condition.Although the system is constantly monitored, only its conditions at the decision epochs have been taken for decision-making with the Markovian property assumption. At each decision epoch, if the system is in state $s \in S$, an action $a \in A$ is made that incurs an expectedreward of $r(s,a)$. In CBM model, $A = \{N, PM, CM\}$ denoting three different actions. The reward function r is evaluated based on the total costs related to maintenance: r = 0, -C_R, or -(C_R+C_D) for a = N, PM, or CM, respectively. The state transition probability distribution has been represented by $p(s_0 |s, a)$. When a = PM or a = CM, $p(0|s, a)$ = 1 for all s. When a = N, the state transition probability, $p(s_0 |s, a)$, has been estimated from existing degradation paths. When S is a countable set, $p(s_0 |s, a)$ provides a value for each s, $s_0 \in S$, $a \in A$. Otherwise, $p(s_0 |s, a)$ has been assumed to be a probability density function.

To learn from existing samples, deterministic policiesπhave been considered that are functions,π: $S \rightarrow A$, which assign a single action to each range of state:

$$\pi(s) = \begin{cases} N & s \in [0, H_p) \\ PM & s \in [H_p, H) \\ CM & s \in [H, \infty) \end{cases}$$

Tabular solving methods have been used to solve the maintenance problems mentioned in the previous sections. To handle a large or continuous state space that cannot be addressed by the tabular method, one can turn to the function approximation to model the state transitions of the system and the value functions (both state-value functions and state-action value functions). A general approximator is preferred when there is not enough information on the possible function to approximate value functions. Although neural networks can model various relationships, they usually require a large amount of data. The GPR can fit small datasets without loss of generality. As an application, they have demonstrated their proposed method to model the battery maintenance decision-making problem by an MDP, where the GPR describes the system dynamics and value functions. Using NASA battery randomized usage data, the Gaussian Process for reinforcement learning (GPRL) algorithm has been applied over the state value iteration. Compared with discrete MDPs, the GPRL algorithm appeared to return a similar optimal policy while being computationally more efficient. They showed that GPRL could save up to 11.9% (varies by different values of H) of the average cost compared to the MDP results.It is worth mentioning that the GPs have been widely adopted for stochastic modeling processes in reliability and maintenance studies. Also, as a general nonparametric model, GPR gains a reputation for its universality and good utilization of data, which is also easy to implement [15].

7. Finding the optimal CBM policy with Deep reinforcement learning (DRL)

Most existing research on CBM assumes that preventive maintenance should be conducted when the degradations of system components reach specific threshold levels upon inspection. However, searching for optimal maintenance threshold levels is often efficient for low-dimensional CBM. Still, it becomes challenging if the number of components gets larger, especially when those components are subject to complex dependencies. Another limitation of most existing CBM models is that they often ignore competing for failure risks when incorporating various types of dependencies, which are common in many real-world systems [16, 17].In this context, competing risk refers to a system failure due to the failure of any of its components. For instance, a modern computer could fail due to the failure of its CPU, storage unit, or operating system, whichever occurs first. The competing risks also impose an economic dependency among components since the system's downtime after one component fails is shared by all the components. Such economic dependency should be considered, which further makes the CBM challenging. Therefore, establishing a general CBM model that jointly incorporates component-wise dependencies and competing risks is necessary. Otherwise, the CBM planning could be inefficient and suboptimal, incurring higher operational and maintenance costs.

Most applications of the traditional RL have been limited to domains where the features can be handcrafted or represented in low-dimensional state spaces. Therefore, directly applying the traditional RL to maintenance planning of K-component systems with complex component-wise interactions would be computationally inefficient and challenging. To overcome this challenge, Zhang et al. [12] proposed a novel and flexible CBM model based on a custom DRL for multi-component systems with dependent competing risks.

DRL is an approach in machine learning that blends reinforcement learning techniques with strategies for deep learning. This type of learning requires computers to use sophisticated learning models and look at large amounts of input in order to determine an optimized path or action. Their proposed CBM model for a K-component system is different from the existing models in two ways:

1. It jointly incorporates stochastic dependency, economic dependency, and competing for failure risks among components.
2. It completely excludes the concept of maintenance thresholds, which are key decision variables in conventional CBM policies.

Specifically, the proposed model directly maps the multi-component degradation measurements at each inspection epoch to the maintenance decision space with a cost minimization objective, and the leverage of DRL enables high computational efficiency and thus makes the proposed model suitable for both low and high dimensional CBM problems.

They have shown that the system deterioration and maintenance process can be formulated as an MDP, and a Deep Q-learning (DQL) algorithm has been selected for the maintenance decisions making. The DQL is a value-based algorithm combining Q-learning and deep learning to approximate the Q-value function. In other words, the DQL is an alternative for Q-learning to solve RL problems with huge state and action spaces or when the state or action spaces are continuous. Specifically, the DQL algorithm aims to recognize patterns instead of mapping every state to its best action. The difference between Q-learning and DQL is illustrated in Figure5.

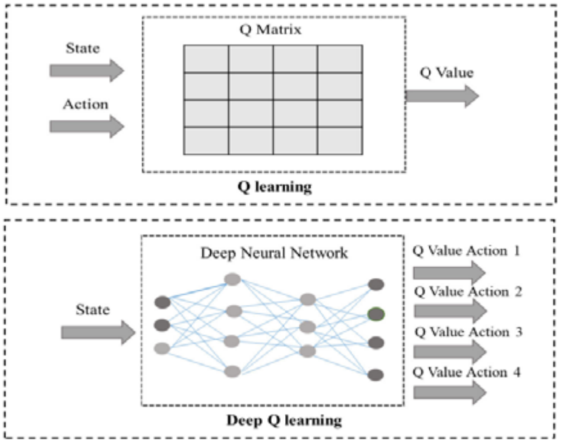

Figure5. Q-learning vs. Deep Q-learning [6]

In another study, Yousefi et al. [6] proposed a DRL method to provide a new dynamic maintenance model for a degrading repairable system subject to degradation and random shock. The following assumptions concern to the

A gamma process is used to model the degradation path of each component.

1. The total degradation of each component is computed by the summation of the internal degradation process and the damages from arrived shocks.
2. The system is inspected periodically at specific intervals, and at each inspection, maintenance actions can be implemented on all the components based on their degradation level.
3. The exact level of system degradation instead of discretizing the state space is considered, which creates an infinite number of states for the maintenance problem.
4. Five different actions of "nothing", "imperfect repair", "repair", "imperfect replace", and "replace" can be performed on all the components. Table 4shows the description of each action.
5. The components of the system are degrading separately with completely different degradation behavior, and all of them are subject to random environmental shocks.

Table 4. The maintenance scheduler set of maintenance actions [6]

Action	Description	Description of action effects
0	Do nothing	The system degrades more based on gamma process
1	Imperfect repair	The system is required but the repair was not perfect
2	Repair	The system is repaired
3	Imperfect replacement	The system becomes very close to new , but the level is not zero
4	replacement	The system is good as new and degradation level goes to zero

The problem has been formulated as an MDP with an infinite number of states, and the DQL algorithm was used to solve the problem and find the best maintenance action dynamically.

8. Conclusion

In this paper, we reviewed CBM-developed models recently with different setups by using RL and DRL methods. Since using RL methods in modeling maintenance problems is fairly new, the existing work in this area is limited. Even such limited literature has shown that the novel RL and DRL methods can provide more accurate and efficient optimal policy for CBM problems than traditional methods. Specifically, among different RL algorithms, it seems that more recently developed RL model-free algorithms such as DQL [12] and their extensions (see, for example, chapter 8 of [18] and [7])outperform the traditional approaches. Such algorithms required assumptions on the true model structure and can more flexibly mimic environmental trends. Alternatively, the RL model-based algorithms can offer promising results for more complex CBM settings where the model-free algorithms may not be as efficient. Although such model-based algorithms make strong assumptions about the environment mechanism (that may or may not be correct), they can be employed for a broader range of CBM problems (e.g., GPRL for continuous-time CBM problems [10]).

Systematic reviews, like ours, are crucial to illustrate the potentials of RL based on the recent developments of CBM problems, to provide a thorough source of the existing work, and ultimately to reflect the gaps and opportunities in the literature as future work for the researchers in this field. In the ML and AI era, RL algorithms are frequently being developed or improved. More studies are required to employ the more recent RL algorithms for CBM problems. Additionally, different RL algorithms are developed for different purposes. Focusing on CBM application and comparing the performance of different RL algorithms for different problem settings is another crucial gap in the current literature.

9. References

[1] A.Adsule, M.Kulkarni, &A.Tewari, "Reinforcement learning for optimal policy learning in condition-based maintenance, "*IET Collaborative Intelligent Manufacturing*, vol.2, no.4, pp.182-188,2020.

[2] C. Andriotis, K. Papakonstantinou, "Managing engineering systems with large state and actions paces through deep reinforcement learning," *Reliab. Eng. Syst. Saf.*vol.191, 106483,2019.

[3] M. Rausand, & A. Hoyland, *System reliability theory: models, statistical methods, and applications*, John Wiley & Sons.(2003).

[4] R. Bellman, "A Markovian decision process. Journal of mathematics and mechanics," 679-684, 1957.

[5] J. Huang, Q. Chang, & J. Arinez, "Deep reinforcement learning based preventive maintenance policy for serial production lines," *Expert Systems with Applications*, vol.160, 113701, 2020.

[6] N. Yousefi, S. Tsianikas, &D. W. Coit, "Dynamic maintenance model for a repairable multi-component system using deep reinforcement learning," Quality Engineering, vol.34, no.1, pp.16-35, 2022.

[7] R. Sutton, A. Barto, "Reinforcement Learning: An Introduction, "*MIT Press*: Cambridge, MA, USA, 2018.

[8] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, Bellemare, M. G., ... & Hassabis, D. "Human-level control through deep reinforcement learning". *nature*, *vol.518*(7540), pp.529-533, 2015.

[9] N. Yousefi, S. Tsianikas, &D. W. Coit, "Reinforcement learning for dynamic condition-based maintenance of a system with individually repairable components, "*Quality Engineering*,vol.32,no.3,pp. 388-408, 2020.

[10] Peng, S. (2021). Reinforcement learning with Gaussian processes for condition-based maintenance. Computers & Industrial Engineering, 158, 107321.

[11] Z. Mahmoodzadeh, K. Y. Wu, E. L. Droguett, & A. Mosleh, "Condition-based maintenance with reinforcement learning for dry gas pipeline subject to internal corrosion," *Sensors*, vol.20,no.19,5708,2020

[12] N. Zhang, & W. Si, "Deep reinforcement learning for condition-based maintenance planning of multi-component systems under dependent competing risks, "*Reliability Engineering & System Safety*, vol.203, 107094, 2020.

[13] M. Knowles, D. Baglee, & S. Wermter, "Reinforcement learning for scheduling of maintenance". In International Conference on Innovative Techniques and Applications of Artificial Intelligence (pp. 409-422). Springer, London.2010, December.

[14] J. H. Shin, &H. B. Jun, "On condition-based maintenance policy," *Journal of Computational Design and Engineering*, *vol.2*, no.2, pp.119-127,2015.

[15] M. Ebden "Gaussian processes for regression: A quick introduction". Available: https://arxiv.org/pdf/1505. 02965. pdf

[16] R. Grande, Th. Walsh, and J. How, "Sample efficient reinforcement learning with Gaussian processes," in proceeding of 31th conference on Machine Learning, china, 2014.

[17] C. Sammut, &G. I. Webb, (Eds.). "Encyclopedia of machine learning," Springer Science & Business Media.2011.

[18] S. Huang, "Introduction to Various Reinforcement Learning Algorithms. Part I (Q-Learning, SARSA, DQN, DDPG),"*Towards Data Science*, vol.12.2018.

[19] M. Sewak, "*Deep reinforcement learning*," Springer Singapore, 2019.

[20] T.Cheng, M.D.Pandey, JA. V.D. Weide,"The probability distribution of maintenance cost of a system affected by the gamma process of degradation: finite time solution, "*Reliab Eng. Syst Saf.*, vol.108,pp.65–76,2012

[21] B.D. Jonge, & P. A. Scarf, "A review on maintenance optimization, "*European journal of operational research*, vol.285, no.3, pp.805-824,2020.

[22] H. Dong, Zh. Ding, &sh. Zhangde, "Deep Reinforcement Learning Fundamentals, Research and Applications", Princeton University,china2020.

[23] C. Henk ,A Tijms. "First Course in Stochastic Models," John Wiley & Sons, Ltd, 2004.

[24] W. Li, H. Pham, "An inspection-maintenance model for systems with multiple competing processes, "*IEEE Trans Reliability*, vol. 54,pp.318–27.2005

[25] Y-HLin, Y-FLi, E. Zio, "Fuzzy reliability assessment of systems with multiple-dependent competing degradation processes, "*IEEE Trans Fuzzy Syst*, vol.23,pp.1428–38,2014.

[26] T. Matiisen, ”Demystifying Deep Reinforcement Learning” Available: neuro.cs.ut.ee.(December 19, 2015)

[27] F. S. Melo, "Convergence of Q-learning: A simple proof," Institute of Systems and Robotics, Tech. Rep, 1-4.,2001.

[28] Y. Nevmyvaka, Y. Feng, & M.Kearns, "Reinforcement learning for optimized trade execution," In Proceedings of the 23rd international conference on Machine learning (pp. 673-680), 2006.

[29] Z. Tian, H. Liao "Condition-based maintenance optimization for multi-component systems using proportional hazards model, "*Reliab Eng. Sys. Saf.*, vol. 96, pp.581–9,2011.

[30] H.V. Hasselt, A. Guez, & D. Silver, "Deep reinforcement learning with double q-learning," In Proceedings of the AAAI conference on artificial intelligence (Vol. 30, No. 1), 2016.

[31] C.J. Watkins, P. Dayan, "Q-learning Machine, *Learning* 1992, 8, 279–292.

[32] E. Yaakov, ShieMannor, and R. Meir. "Bayes meets Bellman: The Gaussian process approach to temporal difference learning," In Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 154–161, 2003.

[33] N. Zhang, Q. Yang, "Optimal maintenance planning for repairable multi-component systems subject to dependent competing risks, "*IIE Trans*, vol. 47, pp.521-532,2015.

[34] Y.Zhao, M. R.Kosorok,&D.Zeng, "Reinforcement learning design for cancer clinical trials, "*Statistics in medicine*, vol.28, no.26, pp.3294-3315,2009.