REVIEW ARTICLE • OPEN ACCESS

A Reinforcement Learning Based Decision-Support System for Mitigate Strategies During COVID-19: A Systematic Review

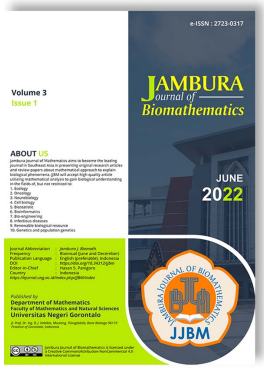
Utti Marina Rifanti *et al.*

Volume 6, Issue 1, Pages 60–70, March 2025

Received 5 February 2025, Revised 18 March 2025, Accepted 9 April 2025, Published Online 13 April 2025

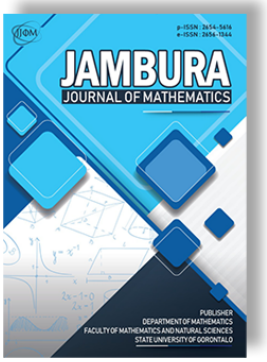
To Cite this Article : U. M. Rifanti *et al.*,"A Reinforcement Learning Based Decision-Support System for Mitigate Strategies During COVID-19: A Systematic Review", *Jambura J. Biomath*, vol. 6, no. 1, pp. 60–70, 2025, *https://doi.org/10.37905/jjbm.v6i1.30513*

JOURNAL INFO • JAMBURA JOURNAL OF BIOMATHEMATICS



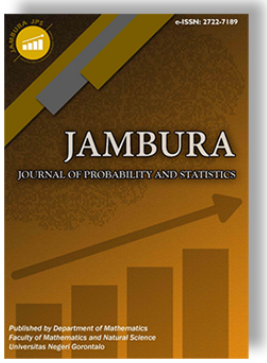
Homepage	:	http://ejurnal.ung.ac.id/index.php/JJBM/index
Journal Abbreviation	:	Jambura J. Biomath.
Frequency	:	Quarterly (March, June, September and December)
Publication Language	:	English
DOI	:	https://doi.org/10.37905/jjbm
Online ISSN	:	2723-0317
Editor-in-Chief	:	Hasan S. Panigoro
Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
Country	:	Indonesia
OAI Address	:	http://ejurnal.ung.ac.id/index.php/jjbm/oai
Google Scholar ID	:	XzYgeKQAAAAJ
Email	:	editorial.jjbm@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS

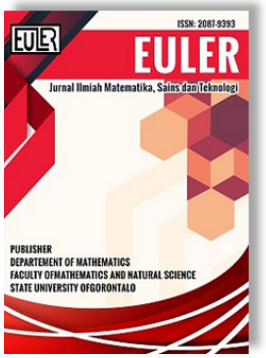


Jambura Journal of Mathematics

Jambura Journal of Mathematics Education



Jambura Journal of Probability and Statistics



EULER : Jurnal Ilmiah Matematika, Sains, dan Teknologi

Jambura Journal of Biomathematics, Volume 6, Issue 1, Pages 60–70, March 2025 https://doi.org/10.37905/jjbm.v6i1.30513

Review Article

Check for updates

A Reinforcement Learning Based Decision-Support System for Mitigate Strategies During COVID-19: A Systematic Review

Utti Marina Rifanti[1,2,*], **Lina Aryati**[1], **Nanang Susyanto**[1], and **Hadi Susanto**[3]

[1]*Department of Mathematics, Universitas Gadjah Mada, Indonesia*
[2]*Department of Telecommunications Engineering, Universitas Telkom, Indonesia*
[3]*Department of Mathematics, Khalifa University, United Arab Emirates*

ARTICLE HISTORY
Received 5 February 2025
Revised 18 March 2025
Accepted 9 April 2025
Published 13 April 2025

KEYWORDS
Reinforcement Learning
Decision Support Systems
COVID-19
Systematic Review
Q-learning
Epidemic Model

ABSTRACT. *The past threat of the COVID-19 pandemic has challenged policymakers to develop effective decision-support systems. Reinforcement learning (RL), a branch of artificial intelligence, has emerged as a promising approach to designing such systems. This systematic review analyzes 20 selected studies published between 2020 and 2024 that apply RL as a decision-making tool for COVID-19 mitigation, focusing on environment models, algorithms, state representation, action design, reward functions, and challenges. Our findings reveal that Q-learning is the most frequently used algorithm, with most implementations relying on SEIR-based models and real-world COVID-19 epidemiological data. Policy interventions, particularly lockdowns, are commonly modeled as actions, while reward functions are health-oriented, economic, or hybrid, with an increasing trend toward multi-objective designs. Despite these advancements, key limitations persist, including data uncertainty, computational complexity, ethical concerns, and the gap between simulated performance and real-world feasibility. This review further identifies a research opportunity to integrate epidemic model formulations with explicit control inputs into RL frameworks, potentially enhancing learning efficiency and bridging the gap between simulation and practice for future pandemic response systems.*

1. Introduction

Coronavirus disease 2019 (COVID-19) is one of the most significant global health crises in recent history, caused by the novel coronavirus SARS-CoV-2. It has had an unprecedented impact on global health, economies, and societies [1–3]. The COVID-19 pandemic requires an urgent response that requires the critical role of decisions by stakeholders. These decisions will shape public health outcomes. Since its emergence in late 2019, countries around the world have implemented various strategies to reduce the spread of the virus. The rapid spread of COVID-19 requires an urgent and comprehensive response. This makes the COVID-19 pandemic an unprecedented challenge. It has reminded constituents around the world that government decision-making can change their lives. Classic epidemiological models and public health strategies have historically been central to infectious disease management. However, the scale and complexity of COVID-19 have highlighted the need for a more dynamic and adaptive approach [4–6]. In response to these challenges, advanced computational approaches such as Reinforcement Learning (RL) have emerged as promising tools for enhancing decision-making processes and optimizing mitigation strategies [7, 8].

RL is a subfield of machine learning that deals with making optimal sequential decisions to control a case. Its algorithms learn over time to choose the best course of action based on feedback from their environment. RL-trained agents can dynamically implement interventions based on current epidemic conditions [9]. This method allows for continuous learning and adaptation based on new data, making it particularly suitable for managing the evolving nature of the COVID-19 pandemic. By leveraging RL, policymakers and health authorities cannot only manage but also optimize interventions. An agent can be trained to take actions based on different available interventions and epidemic situations. This approach can help identify optimal policies that minimize infections, hospitalizations, and deaths while considering the constraints and trade-offs faced by decision-makers.

The emergence of several papers on the application of RL in suppressing the pandemic reflects that RL is quite flexible in handling various aspects of the pandemic response. Regarding optimizing COVID-19 vaccine distribution strategies, research of [10] and [11] show that RL can increase the efficiency and equity of vaccine rollout across the population. In addition, RL methods can also be used to diagnose COVID-19 cases and predict patient outcomes [12]. This systematic review is designed to comprehensively analyze the current research on applying RL-based decision-support systems for COVID-19 mitigation strategies. While previous review papers have explored the application of RL in various domains, there remains a significant gap in understanding how RL has been used as a decision-support mechanism during pandemic scenarios. In particular, the COVID-19 crisis presents a unique context where rapid, data-driven decisions are critical. RL has emerged as a promising tool for dynamically adapting policies in response to evolving public health

*Corresponding Author.

Email : *utti.marina.r@mail.ugm.ac.id* (U. M. Rifanti)
Homepage : http://ejurnal.ung.ac.id/index.php/JJBM/index / E-ISSN : 2723-0317

conditions. This paper aims to fill this gap by offering a focused and systematic review of RL-based decision support systems developed explicitly in COVID-19. We investigate how each study framed its decision-making process through the lens of the environment, algorithm, state representation, action space, and reward design. Although COVID-19 is no longer an active global emergency, the wealth of research produced during this period provides valuable insights for future public health crises. The insights gained from this systematic review equip decision-makers to better handle similar outbreaks in the future.

2. Related Works

In recent years, RL has attracted significant attention from researchers. The rise of research on the use of RL can be seen from the many systematic reviews of RL in various application domains. Hamadani et al. [13] presents a comprehensive review of RL techniques for healthcare and robotics, emphasizing algorithmic comparisons and use cases such as cell growth and robotic manipulation. Similarly, Lin et al. [14] focused on the field of Evolutionary Reinforcement Learning (EvoRL) by highlighting its potential to overcome limitations of RL, such as sensitivity to hyperparameters. Tejedor et al. [15] provides a review in a more specific context, namely RL for blood glucose control in diabetic patients. This study explores the role of RL in insulin infusion systems. Zhao et al. [16] examine RL in the prevention and control of noncommunicable diseases with an emphasis on clinical implementation challenges, such as interpretability, training efficiency, and safety. Martins et al. [17] review RL applications in industrial combinatorial optimization, characterizing RL agent designs in terms of state space, action mapping, and reward generation. Tang et al. [18] explore RL-based methods for improving knowledge graph reasoning.

While these reviews provide critical overviews of RL developments, none specifically address the use of RL in decision support systems during public health emergencies, particularly during the COVID-19 pandemic. Furthermore, few reviews adopt a unified analytical framework to compare RL formulations through the lenses of state representation, action modeling, and reward design. In contrast, this review focuses on RL-based decision support systems in response to COVID-19. We systematically classify existing studies according to their state-action-reward structure. This targeted approach provides valuable insights for developing decision-support tools in future public health crises.

3. Reinforcement Learning Based Decision-Support System

Reinforcement Learning (RL) is an approach where an agent interacts with an environment composed of various states. RL involves agents learning through trial and error, receiving feedback in the form of rewards or penalties based on the actions [19]. The agent's role is not to guide but to learn from its actions, accepting punishment for wrong actions and rewards for right actions [20]. The RL method aims to overcome learning and decision-making problems faced in everyday life [21]. At any given time t, the agent observes its current state $s_t \in \mathcal{S}$, where $\mathcal{S}$ represents the set of all possible states of the environment. Based on this observation, the agent selects and performs an action $a_t \in \mathcal{A}$, where $\mathcal{A}$ denotes the set of available actions in state s_t. In response to the agent's action, the environment returns feedback in the form of a reward $r_{t+1} \in \mathcal{R}$, which is received at the subsequent time $t+1$. Here, $\mathcal{R}$ denotes the set of all possible rewards the agent can obtain by performing different actions and visiting various states.

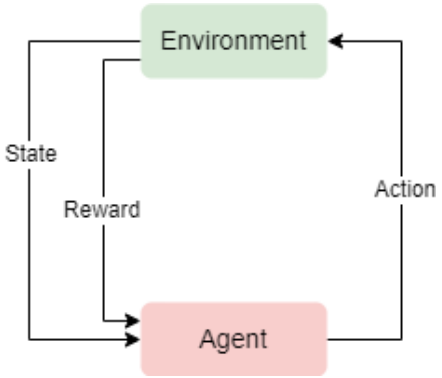


Figure 1. Reinforcement learning framework

As depicted in Figure 1, in state s_t, the agent will choose any action a_t, then the agent receives feedback reward r_{t+1}. Next, in state s_{t+1}, the agent chooses action a_{t+1} and gets new feedback. The agent interacts with its environment to maximize the cumulative rewards it receives over time. The agent understands aspects of the environmental state and chooses appropriate actions. In this case, the agent's main goal is to maximize the cumulative rewards obtained. If G_t is defined as some specific function of the reward sequence, and the sequence of rewards received after time step t is denoted $r_{t+1}, r_{t+2}, \cdots$, then the return G_t is the sum of the rewards:

$$G_t = r_{t+1} + r_{t+2} + r_{t+3} + \cdots + r_T. \tag{1}$$

where T is a final time step. The time of termination, T , is a random variable that normally varies from episode to episode. On the other hand, in many cases the agent–environment interaction does not break naturally into identifiable episodes, but goes on continually without limit. For example, this would be the natural way to formulate an on-going epidemic process-control task. We call these continuing tasks. The return formulation eq. (1) is problematic for continuing tasks because the final time step would be $T = \infty$, and the return, which is what we are trying to maximize, could itself easily be infinite. The additional concept that we need is that of discounting. According to this approach, the agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. In particular, it chooses a_t to maximize the expected discounted return:

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1},$$

where $\gamma \in [0, 1]$ is the discount rate. The primary goal of the agent in RL is to maximize the cumulative rewards received over a certain period. Through repeated interactions with the environment, the collected rewards enable the agent to develop a policy, which guides optimal action selection under different circumstances to maximize rewards. This policy is denoted by π_t, representing the probability of taking a particular action $a_t = a$ given

a specific state $s_t = s$, mathematically expressed as $\pi_t(a|s)$. Formally, a policy is a mapping from states to probabilities of selecting each possible action [9].

In the context of decision support systems (DSS), RL provides a data-driven approach for optimizing sequential decisions under uncertainty. A decision support system powered by RL leverages the agent–environment interaction to improve its decision-making capability over time. The agent selects actions $a_t \in \mathcal{A}$ based on the current state $s_t \in \mathcal{S}$ to maximize the expected return G_t, which is defined as the discounted sum of future rewards. These rewards $r_{t+1}, r_{t+2}, \ldots \in \mathcal{R}$ reflect the effectiveness of actions in achieving the system's goals—such as minimizing infection rates or optimizing resource allocation during a pandemic. The learned decision policy $\pi(a|s)$ maps states to probabilities of actions and serves as the core of the DSS, enabling the system to recommend or automate optimal decisions. The value function $v_\pi(s)$, representing the expected return when starting from state s and following policy π, helps evaluate the long-term benefit of a particular situation. We can define $v_\pi(s)$ formally by [9]:

$$v_\pi(s) = \mathbb{E}_\pi[G_t|s_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s\right], \forall s \in \mathcal{S},$$

where $\mathbb{E}_\pi[\cdot]$ denotes the expected value of a random variable given that the agent follows policy π, and t is any time step. Similarly, the action-value function $q_\pi(s, a)$ estimates the expected return of taking action a in state s, and thereafter following policy π [9]:

$$\begin{aligned} q_\pi(s) &= \mathbb{E}_\pi[G_t|s_t = s, a_t = a], \\ &= \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1}|s_t = s, a_t = a\right]. \end{aligned}$$

Together, these components allow an RL-based DSS to continuously adapt and learn from its environment, making it highly suitable for dynamic domains like healthcare decision-making in response to COVID-19.

4. Methodology

This study adopts the Systematic Literature Review (SLR) methodology to investigate the application of RL in decision support systems for COVID-19 mitigation. A systematic review provides a structured, transparent, reproducible approach to identify, evaluate, and synthesize existing literature to answer a specific research question [22].

4.1. Search strategy

The scope of this review focuses on the use of RL as a decision-making framework in COVID-19 pandemic response scenarios. A literature search was conducted in seven major academic databases: ScienceDirect, Springer Link, Taylor and Francis Online, Nature Research, Sage, ACM, and PLOS One. The publication period was limited to studies published from 2019 to 2024, which aligns with the timeline of the COVID-19 pandemic. The following keywords and their combinations were used to search: "reinforcement learning", "decision-support system", "COVID-19", "pandemic", and "mitigation strategies". Boolean operators (AND, OR) were used to construct the search query, for example ("reinforcement learning" AND "COVID-19") AND ("decision support" OR "policy"). Duplicates were removed manually. The search and screening process used Mendeley Reference Manager to manage references.

4.2. Selection and review process

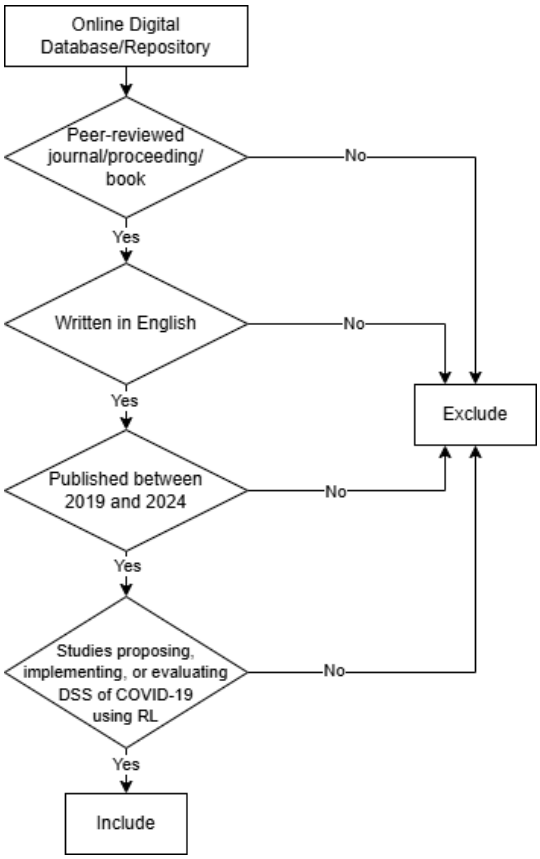


Figure 2. Inclusion and exclusion criteria

The selection process followed three main stages: (1) initial screening based on title and abstract; (2) full-text evaluation for relevance; and (3) data extraction and classification. To ensure objectivity and consistency, the inclusion criteria were applied systematically as illustrated in Figure 2.

1. Peer-reviewed journal articles, conference proceedings, or book chapters.
2. Written in English.
3. Published between 2019 and 2024.
4. Studies applying RL to COVID-19 decision support systems (DSS), including proposals, implementations, or evaluations of such systems.

As shown in Figure 2, studies were first screened based on source type, language, and publication year. The final and most critical criterion was the explicit focus on applying RL in developing or evaluating decision support systems related to COVID-19. Studies that failed to meet these four criteria were excluded from the review. In other words, studies published outside the specified time frame, not peer-reviewed, not written in English, theoretical studies without implementation or evaluation, and studies without a decision support component will be excluded from the review process.

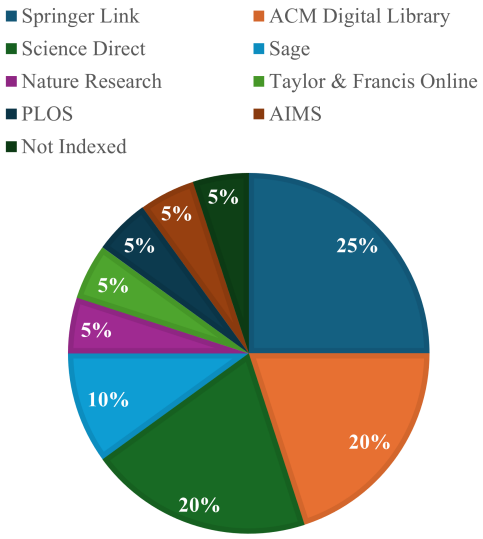


Figure 3. Distribution of selected articles

A total of 20 papers published between 2020 and 2024 were included in the final review. The number of studies included in the final review was limited, reflecting the topic's specificity. Most of the selected studies came from reputable academic publishers to ensure the credibility and quality of the reviewed papers. The distribution of publication sources is as follows:

1. Springer Link – 5 papers
2. ScienceDirect – 4 papers
3. ACM Digital Library – 4 papers
4. Taylor & Francis Online – 1 paper
5. Nature Research – 1 paper
6. Other Academic Publishers (SAGE, AIMS, PLOS) – 4 papers
7. Not Indexed – 1 paper

As shown in Figure 3, the selected studies were distributed across various reputable digital repositories. The majority of articles were sourced from Springer Link (25%), Science Direct (20%), and the ACM Digital Library (20%), followed closely by Sage (10%). Other databases, including Nature Research, Taylor & Francis Online, PLOS, AIMS, and non-indexed sources, each contributed 5% to the total. This distribution reflects the wide range of scholarly platforms that have published studies on reinforcement learning in COVID-19-related decision support systems.

4.3. Data categorization

The reviewed studies are analyzed based on five key aspects that define the structure of RL-based decision support systems for COVID-19 mitigation, namely (1) environment model, (2) RL algorithm, (3) state representation, (4) action space, and (5) reward function. In addition to these five aspects, challenges and limitations reported in the literature are discussed to provide a more comprehensive picture.

1. The environment model is a simulation framework in which the RL agent operates. It simulates the complexity of the real world and provides the basis for state transitions and reward signals.
2. The RL algorithms referred to in this section are the approaches used in the studies. The algorithms used in one study and another may differ, usually depending on the action space (discrete or continuous) and the complexity of the simulation environment.
3. State representations describe how agents perceive the environment. Common state variables include epidemiological indicators (e.g., infected, cured, deceased), health care system capacity, reproduction rate (R_0), and economic factors such as GDP. These features often come from the underlying environmental model.
4. The action space defines the set of interventions available to an agent. Commonly modeled actions include non-pharmaceutical interventions (e.g., lockdowns, social distancing, testing) and pharmaceutical interventions (e.g., vaccination and treatment). These actions can be discrete or continuous.
5. Reward functions measure the objective to be optimized while assessing how good actions perform at a given state. Rewards during a pandemic are typically designed to punish high infection or death rates, healthcare overload, or economic losses while incentivizing low-cost interventions.

5. Results and Discussion

Research on RL for pandemic control has seen rapid growth, driven by the urgent need for adaptive and data-driven strategies to support timely and effective policymaking. The challenges posed by the COVID-19 pandemic have highlighted the importance of intelligent systems that can respond dynamically to evolving conditions and uncertainties. This study presents a systematic review of RL-based frameworks to support decision-making during the COVID-19 pandemic. In the following section, we summarize the findings from the reviewed literature by analyzing the fundamental components of RL-based decision support systems. In addition, we present the challenges and limitations of the reviewed studies. The results and discussion of the review are organized into six core aspects: environment models, RL algorithms, state representations, action spaces, reward functions, and challenges and limitations. Each subsection will discuss its respective components and synthesize the findings from selected studies to reveal trends, strategies, and research opportunities in the COVID-19 pandemic DSS literature with RL.

5.1. Environment models

This section will summarize the environmental models used by the studies reviewed in this paper. Table 1 presents not only the environmental models used but also the software or framework and data used in each study.

Based on Table 1, most of the studies used continuous epidemiological models, specifically variations of the SIR (Susceptible Infectious Recovered) and SEIR (Susceptible Exposed Infectious Recovered) frameworks. Some modified these standard epidemiological models, such as SEIRD by adding a mortality (D) compartment [30, 38, 42], SQEIR by adding a quarantine (Q) compartment [40], and specialized variants such as SIDARTHE (Susceptible Infected Diagnosed Ailing Recognized Threatened Healed Extinct) [25]. Some studies used additional compartments to define the infected state into presymptomatic and

Table 1. Environment Models

Ref.	Environment Models	Software/Framework	Data Used
[23]	Modified SEIR	Matlab	COVID-19 data (Qatar)
[24]	SEIR	Python	COVID-19 data (New York)
[25]	Modified SIR	Python	COVID-19 data (Moroccan)
[26]	Modified SEIR	Python	COVID-19 data (New York)
[27]	SEIR	Python	Not explicitly mentioned
[28]	SIR	Python	Not explicitly mentioned
[29]	Modified SEIR	Not explicitly mentioned	COVID-19 data (New Jersey)
[30]	SEIRD	Python	Not explicitly mentioned
[31]	SIRD	Python	COVID-19 data (global)
[32]	Modified SEIR	Not explicitly mentioned	COVID-19 data (Changchun, Shanghai)
[33]	SIR	Not explicitly mentioned	COVID-19 data (China)
[34]	SIR	Not explicitly mentioned	COVID-19 data (New York)
[35]	SEIR	Python	COVID-19 data (United States, Beijing)
[36]	Modified SIR	Python	Not explicitly mentioned
[37]	Modified SEIR	Not explicitly mentioned	COVID-19 data (Austin, Texas)
[38]	Modified SEIRD	Not explicitly mentioned	COVID-19 data (United States)
[39]	Agent-based model	Not explicitly mentioned	Not explicitly mentioned
[40]	Modified SEIR	Not explicitly mentioned	Not explicitly mentioned
[41]	Modified SEIR	Not explicitly mentioned	Not explicitly mentioned
[42]	SEIRD	Python	COVID-19 data (Utah)

asymptomatic, then divide presymptomatic into mild, severe, hospital, and ICU categories [29, 41]. These models were mainly chosen due to their effectiveness in representing disease transmission dynamics at the population level and their computational efficiency. Specifically, modified versions of the SEIR model were employed in 40% of the reviewed studies, while the standard SEIR model appeared in 15%. SIR-based models, including standard and modified variants, accounted for 25% of the studies, followed by SEIRD and SIRD models. A small portion (5%) adopted a modified SEIRD model, and only one study (5%) explored an agent-based modeling approach. This distribution indicates that researchers primarily rely on epidemiological compartment models due to their mathematical tractability and ability to capture key disease dynamics.

Regarding the software or computational framework used to implement these environmental models, Python emerged as the most frequently used programming language. It appeared in 50% of the studies. Its popularity is due to its reasonably diverse scientific libraries and its ability to integrate with RL, such as TensorFlow, PyTorch, and OpenAI Gym. Python is the most frequently used software or framework due to its accessibility, extensive libraries, and flexibility in simulation tasks. Some construct systems related to COVID-19 pandemic simulations, such as PandemicSimulator [37]. The remaining studies either did not explicitly specify the software used (45%) or mentioned tools such as Matlab (5%). These findings highlight the trend towards using open source and flexible platforms, particularly Python, to build and train RL-based decision support systems in pandemic DSS.

Regarding the data sources used in the simulations, most studies used COVID-19 epidemiological data from various locations, such as New York, Qatar, Morocco, Austin (Texas), and China, and others. Some studies did not explicitly mention their data sources that may indicate the use of synthetic or hypothetical datasets. In addition, one study uses an agent-based model, which emphasizes detailed individual-level interactions but does not explicitly specify its data sources.

These findings highlight a strong preference among researchers for continuous SEIR-based models and real-world epidemiological data despite increasing exploration of alternative modeling approaches and data scenarios to effectively address the complex nature of the COVID-19 pandemic. In addition, Python is the most commonly used programming language due to its flexibility, availability of scientific libraries, and ease of integration with RL frameworks.

5.2. *Reinforcement learning algorithms*

The choice of RL algorithm usually depends on the environment model's complexity or the state-action space's dimensionality. This section discusses the various types of RL algorithms adopted in the reviewed studies as summarized in Table 2. The reviewed studies employed a range of RL algorithms that can be broadly grouped into four classes, each reflecting different capabilities and complexities in addressing epidemic control problems. This classification underlines the evolution from simple tabular methods to more complex multi-objective frameworks. This evolution indicates increased adaptability, scalability, and realism in simulated epidemic environments.

Table 2. Reinforcement Learning Algorithms

Ref.	RL Algorithm
[23],[24],[28],[30],[37]	Q-Learning
[29],[33],[34],[35],[42]	DQN
[26],[27]	DDQN
[25],[32]	PPO
[40]	DDPG
[31]	D3QN
[38]	MARAAC
[41]	PCN
[36]	PPO and SAC
[39]	Q-learning, SARSA, DQN, and DDPG

a) Tabular Methods

Q-learning is the most commonly used algorithm, appearing in six studies as the primary approach or in combination with others (30% of the total). This includes classic tabular applications as well as comparative baselines. This algorithm was employed by studies such as [23], [24], [28], [30], and [37], which typically

Table 3. State Space Parameters

Ref.	State Space Parameters	Category
[23]	Active cases	Epidemiological states
[24]	Susceptible, exposed, infected, recovered, death	Epidemiological states
[25]	Transmission rate, identification rate, death rate, reinfection rate	Latent or derived variables
[26]	Active cases	Epidemiological states
[27]	Active cases, infected, recovered, deaths, reproduction rate, GDP	Economic and epidemiological
[28]	Active cases	Epidemiological states
[29]	Susceptible, vaccinated, infected, hospitalized, recovered, death	Epidemiological states
[30]	Susceptible, exposed, infected, recovered, death	Epidemiological states
[31]	Active cases, recovery, deaths, acceleration rate, GDP	Economic and epidemiological
[32]	Susceptible, exposed, infected, recovered	Epidemiological states
[33]	Active cases, annual GDP	Economic and epidemiological
[34]	Active cases, economic score, social rate	Economic and epidemiological
[35]	Susceptible, exposed, infected, ascertained, removed	Epidemiological states
[36]	Susceptible, infected, recovered, vaccinated	Epidemiological states
[37]	Test results, hospitalizations	Epidemiological states
[38]	Susceptible, exposed, asymptomatic, infected, recovered, death	Epidemiological states
[39]	Discrete time steps (months)	Temporal abstraction
[40]	Susceptible, quarantined, exposed, infected, recovered	Epidemiological states
[41]	Susceptible, exposed, infected, recovered, death	Epidemiological states
[42]	Susceptible, exposed, infected, recovered, death	Epidemiological states

model epidemic environments using compartmental approaches with a manageable number of states. [39] also incorporated Q-learning in their comparative experiments alongside other algorithms.

b) Deep Reinforcement Learning

The second group consists of value-based deep RL approaches, particularly those based on the Deep Q-Network (DQN) family. DQN and its variants also appear prominently, with DQN used in five studies (25%), Double DQN (DDQN) in two studies (10%), and Dueling Double DQN (D3QN) in one study (5%). These methods represent the transition from tabular to deep value-based approaches, enabling RL to scale to more complex epidemic environments. These methods use neural networks to approximate value functions and can handle larger state spaces. DQN and its variants, including DDQN and D3QN, were utilized by studies such as [29], [33], [34], [35], and [42]. Enhancements like DDQN and D3QN were specifically chosen to address overestimation bias and improve learning stability, as seen in [26], [27], and [31]. In addition, [39] also explored DQN-based methods in their frameworks.

c) Policy Gradient Methods

The policy gradient methods group directly optimizes the policy instead of estimating the value function. These methods are used in studies with continuous action spaces. Policy gradient methods such as Proximal Policy Optimization (PPO) and Soft Actor-Critic (SAC) are employed in three studies (15%), demonstrating the growing interest in algorithms that support continuous action spaces. One study uses a Deep Deterministic Policy Gradient (DDPG), which combines actor-critic architecture with deterministic policies. PPO was employed by [25], [32], and [36], with the latter also integrating the SAC to enhance exploration. Meanwhile, DDPG, known for its suitability in deterministic continuous control tasks, was used by [40] and also considered in the comparative analysis in [39].

d) Multi-Objective and Hybrid RL Methods

The final category emerged in the form of multi-objective and hybrid RL methods, reflecting a shift towards more variable decision-making. In addition, two studies (10%) using methods in this category explored more specialized frameworks. The Multi-Agent Regularized Actor-Critic (MARAAC) applied by [38] to handle decentralized decision-making, while [41] leverages Pareto Conditioned Networks (PCN) to address multi-objective optimization in epidemic control. These methods balance multiple, potentially conflicting, objectives such as health outcomes and socio-economic impacts.

Overall, the reviewed studies show a diverse application of RL algorithms. This diversity of algorithm choices indicates the adaptability of RL in addressing the diverse challenges of COVID-19 decision support systems. Among all algorithms, Q-learning is the most frequently adopted method, appearing in about 30% of the reviewed studies.

5.3. *State Space Representation*

The design of the state space directly affects the agent's ability to perceive and respond to the environment. Various parameters representing the pandemic conditions have been used to define the state space, as shown in Table 3. Table 3 presents the state space parameters used across the reviewed studies and provides a comprehensive view of how the environment is structured in RL-based decision support systems for epidemic control. The state space defines a set of observable variables the agent perceives from the environment at each decision point. These variables are important because they serve as the information base from which the agent selects actions to optimize long-term outcomes.

Table 3 shows that most reviewed studies (14 out of 20, or 70%) represented the environment using epidemiological states. This reflects the reliance on compartmental models to capture disease dynamics in RL-based decision-making frameworks. This category is the most commonly used among the reviewed studies. It includes classical compartmental model variables such as susceptible (S), exposed (E), infected (I), recovered (R), and de-

Table 4. Action Space Parameters

Ref.	Action Description	Intervention(s)
[23]	Discrete action set	Lockdown, travel restriction, hygiene habits, health-care treatment
[24]	Multi-level lockdown actions	Lockdown
[25]	Multiple interventions	Lockdown, travel restriction, social distancing, mask-wearing, testing, treatment, vaccination
[26]	Binary policy actions	Lockdown
[27]	Multi-level lockdown actions	Lockdown
[28]	Percentage-based action levels	School closure, hybrid learning
[29]	Multiple interventions	Testing, Contact Tracing, Quarantine, school closure
[30]	Binary policy actions	Lockdown
[31]	Multi-level lockdown actions	Lockdown
[32]	Percentage-based action levels	Lockdown, temporary medical resources
[33]	Multi-level lockdown actions	Lockdown
[34]	Binary policy actions	Lockdown
[35]	Multiple interventions	Mask-wearing, isolation, school closures, work from home, lockdown
[36]	Percentage-based action levels	Mask-wearing, vaccination, school closure, work from home
[37]	Discrete action set	Not explicitly mentioned
[38]	Multi-level lockdown actions	Lockdown
[39]	Multiple interventions	Testing, sanitization, lockdown
[40]	Multiple interventions	Quarantine, vaccination, and treatment
[41]	Multiple interventions	Social contact reduction
[42]	Percentage-based action levels	Lockdown

ceased (D), as seen in the SEIR or SEIRD formulations. Studies by [23], [24], [29], [35], and [42] used these states to represent epidemic dynamics. In addition to these classical variables, some studies incorporated epidemiological observations such as hospitalizations and testing outcomes as part of the state space. For example, [37] models the state using hospitalizations and test results, allowing agents to respond directly to indicators of disease severity. This representation allows RL agents to make decisions based on the internal evolution and impact of the disease observed in the population.

There are 15% studies (4 of 20) adopted mixed economic and epidemiological indicators. These studies aimed to balance health-related and socioeconomic outcomes. Some studies expanded the country's representation by incorporating socioeconomic indicators variables. For example, [31], [33], and [34] included data on GDP, economic impact, or public policy indicators such as lockdown stringency or quality of life metrics. These features allow RL models to optimize trade-offs between public health interventions and their societal consequences.

Only one study (5%) used latent or derived variables, such as transmission and identification rates, which are often inferred rather than directly observed [25]. Another study (5%) used temporal abstraction, defining states based on discrete time intervals rather than epidemiological indicators [39]. A subset of studies includes latent parameters or estimations derived from the underlying epidemic model, such as transmission rates, identification rates, or detection probabilities. For example, [25] and [27] incorporate these inferred parameters to capture dynamic aspects of disease spread, which may not be directly observable.

This distribution highlights that most RL-based epidemic models rely on classical epidemiological constructs. Meanwhile, growing interest is in integrating economic and social considerations to support more holistic decision-making. Some studies adopt classical epidemiological variables derived from compartmental models, such as SEIR or SEIRD, while others include observational indicators, such as active cases, hospitalizations, or test results. Some studies incorporate economic variables, latent parameters (e.g., transmission or identification rates), or abstract temporal representations such as discrete time steps. This diversity in state design reflects the different modeling priorities adopted by researchers.

In the context of COVID-19 mitigation, state space encompasses many factors that describe the current pandemic status and relevant societal conditions. These factors include the number of active cases, recovered cases, deaths, hospital bed availability, health service capacity, vaccination status, reproduction rate, human mobility, economic activity, and more. Each state in the state space incorporates these factors at a given time, providing a holistic picture of the pandemic.

5.4. Action Space

The action space in RL consists of all possible actions that the agent (policymakers) can take to influence the system. In the first step, researchers investigate a range of realistic possible actions. For COVID-19 mitigation, the action space includes various interventions to control the spread of the virus. Each action in the action space represents a decision or policy that can alter the course of the pandemic by affecting the state space variables. RL does not judge which action is the best but determines which action is appropriate for a particular state. RL helps determine which action is optimal to use in a state in order to reduce COVID-19 cases. Various action parameters taken by policymakers to manage COVID-19 cases were considered from several reviewed study results, as shown in Table 4.

Table 4 summarizes the action space used in the reviewed studies. Based on the review, actions consist of implementing lockdowns, mandating masks in public spaces, enforcing social distancing measures, setting up quarantine zones, allocating additional resources to hospitals, recruiting and deploying healthcare workers, expanding testing and contact tracing efforts, accelerating vaccination, school and workplace closure, and others. Each study defines its own action set, reflecting the available interventions to control the spread of COVID-19. These actions may vary in form and complexity, ranging from simple binary de-

cisions to multi-dimensional intervention combinations.

The reviewed studies show a variety of approaches in defining the action space for RL-based decision support systems in COVID-19 mitigation. As summarized in Table 4, the most common category is multiple interventions, which account for 30% (6 of 20) of the studies. This category reflects that agents consider multiple control measures simultaneously, such as lockdowns, mask-wearing, testing, vaccination, and healthcare. Following this, multilevel lockdown measures appear in 25% (5 of 20) of the studies, indicating a focus on varying levels of mobility restrictions. Percentage-based action levels account for 20% (4 of 20) of the studies, offering a more granular representation of intervention intensity. Binary policy actions, which represent simple yes/no decisions (e.g., implementing or not implementing a lockdown), account for 15% (3 of 20) of the reviewed papers. Finally, discrete action sets, which typically include a predefined set of intervention options, are observed in 10% (2 of 20) of the studies.

This variety in action representations highlights the adaptability of RL frameworks in capturing real-world decision complexities. Even when the same intervention type (such as lockdown) is applied, it can be encoded as binary, multi-level, or percentage-based actions, depending on the design choices and policy goals of the study. In particular, the multiple intervention category emerged as the most frequently adopted approach. This reflects the growing recognition of the complex nature of pandemic control, whereby a combination of measures, such as lockdowns, mask use, testing, vaccination, and strengthening of healthcare systems, is needed to effectively mitigate the spread of COVID-19. Overall, the diverse action space underscores the importance of flexibility and comprehensiveness in designing RL-based decision-support systems for pandemic response.

5.5. Reward Functions

The reward function is the heart of RL systems, shaping how agents evaluate and improve their actions over time. The reward function is an essential element for successful RL design. Therefore, experts are free to define the reward function based on the goal they want to achieve. For the same reason, several reviewed studies have produced various types of reward functions based on the goal they want to achieve, as shown in Table 5.

Table 5 summarizes the reward functions used in 20 studies that applied RL as a decision support system in COVID-19. The reward function is crucial for the algorithm's optimal performance. It is a critical component that guides the learning process by providing feedback to the agent about the desirability of actions. The success of an RL method as a decision support system depends on how well the reward function represents the goals of the application designer and how well it assesses progress in achieving those goals [9, 15]. The design of the reward function plays a critical role in guiding the RL agent to learn the optimal policy that balances public health objectives and socioeconomic constraints. For the case of COVID-19 mitigation, the reward function needs to encapsulate the objectives of minimizing the negative impacts of the pandemic while balancing another important economic factor. Researchers freely define the reward function, so in this review study, we found various reward functions that vary greatly, as shown in Table 5.

In the context of COVID-19 decision support systems, reward design reflects the balance between epidemic control and socioeconomics. Table 5 presents the various reward functions used in the reviewed studies. They are categorized into three thematic groups: health-oriented, economy (or cost)-oriented, and hybrid. This classification highlights the different priorities across RL-based DSS models, underlining the importance of reward design in aligning agent behavior with public health goals and policy objectives. Furthermore, it shows that the choice of evaluation criteria is closely related to how the reward function is formulated, i.e., whether it focuses on infection suppression, economic sustainability, or intervention efficiency. Furthermore, Table 6 will briefly present this discussion.

a) Health-Oriented Rewards

These reward functions prioritize public health goals by minimizing infection rates, reducing mortality, and maintaining healthcare system stability. For example, [26] imposes a penalty when ICU occupancy exceeds a critical threshold, incentivizing agents to keep hospitalization rates manageable. Similarly, [37] and [38] impose penalties on scenarios where the number of critically ill patients exceeds hospital capacity. Infection-related metrics are central to studies such as [24] and [41], which incorporate hospital congestion probabilities and explicitly target reductions in infection and hospitalization rates. Across these studies, evaluation criteria typically include total infections, ICU burden, and mortality, focusing on health system outcomes.

b) Economy (or Cost)-Oriented Rewards

This category focuses on minimizing the economic impact of interventions or maintaining economic activity during the pandemic. For example, [30] penalizes resource use and policy enforcement costs. In addition, [33] formulates rewards on GDP outcomes that link policy choices to economic performance. Meanwhile, [36] integrates economic sustainability by penalizing productivity losses or excessive costs associated with interventions. The evaluation metrics in these studies focus on minimizing intervention costs and maximizing utility from an economic perspective.

c) Hybrid Rewards

Hybrid reward functions represent the majority approach in the reviewed studies, reflecting an attempt to balance health outcomes with economic considerations. For example, [23] integrates hospital capacity management, infection control, and intervention costs into a unified reward structure. A comprehensive formulation that considers quality of life, virus transmission, economic indicators, and resource consumption is considered in [39]. Studies such as [25], [27], [28], [29], [32], [34], [35], and [42] also use multifactorial reward designs to capture the complex interactions between public health goals and socioeconomic impacts. Evaluation criteria in this category often include composite indicators, such as cost-effectiveness per infection averted, healthcare burden indices, or policy efficiency metrics, which align with real-world decision-making's complexity.

Analyzing reward functions in RL-based decision support systems for COVID-19 highlights a strong trend toward hybrid approaches that balance health and economic objectives. In partic-

Table 5. Reward Functions

Ref.	Reward Function	Notation and Description
[23]	$r = r_1 + r_2 + \beta_w r_3$	r_1 : hospital capacity; r_2 : the difference of $I(t)$ and the desired value; r_3 : action cost; β_w : cost of interventions;
[24]	$r = \frac{k}{max(\mathbb{K})e^{-p_{queue}}}$	$k \in \mathbb{K}$: contact index; p_{queue}: the probability of hospital queue.
[25]	$r = h + e$, if $h, e > 0$ $r = 0$, otherwise	h (health score): public health performance; e (economic score): the budget to invest in an action
[26]	$r = 0$, if $ICU_{error} < margin$ $r = -\alpha \lvert ICU_{error} \rvert$, if $ICU_{error} \geq margin$	$ICU_{error} = ICU_{actual} - ICU_{threshold}$
[27]	$r = e^{mA_t} E_t - nD_t$	E_t: current economy; A_t: active cases; D_t: death.
[28]	$r = \alpha \sum_{i=1}^{C} A_i - (1-\alpha) \sum_{i=1}^{C} I_i$	A: allowed students; I: infected students; $\alpha \in [0, 1]$.
[29]	$r = \lambda E_t - \mu I - \rho D + \pi(S + V_1 + V_2)$	S: susceptible; I: infected; D: death; V_i: people with i-th vaccinated; E_t: economic score.
[30]	$r = -$cost	The negative of the costs associated with implementing the interventions in action a (lower the cost, higher the reward).
[31]	$r = r^{crc} + 0.5r^{crd} + 0.5r^{crr}$	crc: infection cases; crd: death cases; crr: recovery cases.
[32]	$r = (1-\omega)r_h + \omega r_e$	r_h: health score; r_e: economic score; ω: parameter weights to balance the significance of public health and economic losses.
[33]	$r = \sum_{j=1}^{3} C_j \mathbb{I}(a = j) \frac{G_l}{365}$	$C_1 = 0$; G_l: gross domestic product (GDP); $C_j \sim N\left(\mu_j, \sigma_j^2\right)$ for $j \in \{2, 3\}$.
[34]	$r = \alpha_0 \frac{E_t}{E_0} - \alpha_1 \frac{X_t^I}{M} - \alpha_2 \phi(a)$ $E_0 = 0$	E_t: economic, freedom, and happiness level; X_t^I: infected person; M: total population.
[35]	$r = \sum_{t=1}^{T} (\epsilon r_h - r_e)$	r_h: health score; r_e: economic score; ϵ - the trade-off weight to coordinate r_h and r_e.
[36]	$r = -$cost	The negative of the costs associated with implementing the interventions in the action a (lower the cost, higher the reward).
[37]	$r = a \max\left(\frac{n^c - C^{\max}}{C^{\max}}, 0\right) + b \frac{stage^p}{\max_j stage_j^p}$	n^c: persons in critical condition; $C^{\max}$: maximum hospital capacity; $stage \in [0, 4]$: restriction stages; a, b, p: weighting coefficients
[38]	$r = \alpha \max\left(\frac{p^C - M}{M}, 0\right) + \beta \sum_i w_i Loc_i$	p^C: persons in critical condition; M: hospital capacity; Loc_i: weight for area i; w_i: lockdown level at area i; α, β: weighting coefficients.
[39]	$r = w_1 LQ + w_2 Ec - w_3 Sp - w_4 Rs$	LQ (Living Quality): quality of life metric; Ec (Economy): economic condition; Sp (Spread): virus spread indicator; Rs (Resources): resource usage; w_1, w_2, w_3, w_4 - weighting coefficients
[40]	$r = -(aS + b\,I + c\,u_1 + d\,u_2 + e\,u_3)$	S: susceptible; I: infected; $u_i(t)$: control inputs; a, b, c, d, e: coefficients to balance infection and control.
[41]	$R_{ARI} = -\left(\sum_{k=1}^{K} S_k - \sum_{k=1}^{K} S'_k\right)$; $R_{ARH} = -\sum_{k=1}^{K} H_k^{new}$; $R_{SB} = \sum_{i=1}^{K} \sum_{j=1}^{K} (C - C')_{ij} [S_j S_i + R_j R_i]$	R_{ARI}: attack rate (infections); R_{ARH}: new hospitalizations; R_{SB}: social burden based on contact reduction; $S_k(s)$: susceptible individuals; $R_k(s)$: recovered individuals; C, C': social contact before and after interventions; $H_k^{new}(s)$: new hospitalizations.
[42]	$R_\alpha(x) = f_\alpha(a) - f_h(I_{x,A}) - \eta\theta I_{x,A} C_I$	$f_\alpha(a)$: monetary cost; $f_h(I_{x,A})$: hospitalization cost; η: objective term; θ: mortality rate; $I_{x,A}$: number of infected; C_I: cost per mortality.

ular, hybrid reward functions were the most frequently adopted, appearing in 50% of the studies reviewed. This design combines public health priorities with economic considerations, often through weighted sums or composite metrics that combine epidemiological indicators (such as infection rates and hospital capacity utilization) with economic measures (such as intervention costs and quality of life). Health-oriented rewards accounted for 35% of the studies, primarily focusing on minimizing infections and mortality. Meanwhile, economic (or cost) in 15% of the studies, oriented rewards prioritize minimizing policy costs and maintaining economic activity. This distribution underscores the recognition that practical policy recommendations require careful consideration of public health outcomes and socioeconomic impacts, ensuring that RL agents can support decision-making processes in a balanced and contextually sensitive manner.

5.6. *Challenges and Limitations*

Although RL has shown potential in supporting decision-making for pandemic mitigation, this review identifies several challenges and limitations identified by the reviewed studies. These limitations include data constraints, computational demands, ethical issues, and practical gaps between simulation and

Table 6. Thematic Classification of Reward Functions

Ref.	Category
[24], [26], [31], [37], [38], [40], [41]	Health-oriented
[30], [33], [36]	Economy (or cost)-oriented
[23], [25], [27], [28], [29], [32], [34], [35], [39], [42]	Hybrid

real-world implementation.

a) Data Limitations

Many studies rely on uncertain, incomplete, or delayed epidemiological data. For example, [26] and [24] emphasize the difficulty in obtaining accurate and real-time data, which is critical for effective RL training. [23] and [35] also highlight that RL models often assume ideal data availability, which is rarely the case in real-world scenarios. This raises concerns about the robustness and generalizability of RL policies trained on such data.

b) Computational Challenges

Several studies report the computational burden associated with complex epidemiological models and the iterative nature of RL. [23] and [30] note that high model complexity can hinder scalability and responsiveness, especially in emergency contexts where timely decisions are critical. Even models designed for efficiency, such as PaCAR [35], may still require significant resources when scaled.

c) Ethical and Policy Considerations

Ethical implications are often underexplored in RL-based systems. While some studies, such as [24], address the trade-off between health outcomes and economic impacts, there is limited attention to issues such as fairness, equity, and public acceptance of automated decisions. Additionally, [30] and [35] caution that decisions derived from RL models should be carefully evaluated by human experts before implementation, especially when involving restrictive measures such as lockdowns.

d) Effectiveness vs. Real-World Implementation

Most studies have demonstrated the usefulness of RL as a decision-support for COVID-19 systems in reducing infections and minimizing costs in simulated environments. However, significant differences between model assumptions and the complexity of real-world problems mean that its practical implementation remains uncertain. For example, the policies developed in the studies of [26] and [23] were successful in controlled simulations but lacked validation in real public health settings. Similarly, [35] and [30] highlighted that translating RL policies into actionable strategies is constrained by unpredictable population behavior and political constraints. These factors require monitoring and testing before RL-informed policies can be adopted. Bridging this gap between theoretical performance and applied feasibility remains a significant hurdle.

In summary, while RL offers a promising avenue for data-driven policy design during the pandemic, addressing its limitations is critical. Future research should focus on improving model robustness under real-world data conditions, improving computational efficiency, incorporating ethical frameworks, and developing implementation strategies that bridge the gap between simulation and practice.

6. Conclusion

This systematic review provides a comprehensive overview of RL applications in decision support systems for COVID-19 mitigation. Our findings reveal that most studies rely on SEIR-based epidemiological models and real-world COVID-19 data, with Python and widely adopted libraries such as TensorFlow and OpenAI Gym as the dominant development platforms. Regarding algorithm selection, Q-learning emerged as the most frequently used method due to its simplicity and interpretability. The state space generally incorporates infection-related indicators, while action design largely centers on policy interventions such as lockdowns or mobility restrictions. Reward functions exhibit a variety of objectives, spanning health-oriented, economy-oriented, and hybrid formulations. In particular, there is increasing attention to multi-objective reward design to balance public health priorities with socio-economic considerations.

In addition to summarizing methodological trends, the review highlights the critical limitations of current RL-based approaches. Data quality and availability, computational demands, and unexplored ethical dimensions remain significant challenges. Furthermore, while RL has shown promising performance in simulation environments, its application in real-world decision-making is still limited, partly due to the discrepancy between model assumptions and actual policy contexts.

These insights underscore the potential of RL to improve pandemic response strategies while also pointing out areas where further progress is needed. In light of this, our analysis draws attention to a neglected opportunity. While most existing studies use continuous Ordinary Differential Equations (ODE)-based epidemic models primarily as simulation environments, they rarely integrate these dynamics directly into state representations with explicit control inputs. Future research should explore using discrete ODE formulations and embedding control variables in the state space to better align with the RL structure and improve learning efficiency. Such integration could drive a more robust learning process for more effective and accountable decision-support tools.

Author Contributions. **Rifanti, U. M.**: Conceptualization, methodology, formal analysis, writing—original draft, software, resources, visualization, and project administration. **Aryati, L.**: Conceptualization, supervision validation, formal analysis, and writing—review. **Susyanto, N.**: Conceptualization, supervision, validation, investigation, and writing—review and editing. **Susanto, H.**: Conceptualization, validation, writing—review, supervision, and software.

Acknowledgement. The authors would like to thank Universitas Gadjah Mada and Universitas Telkom for the excellent academic environment and access to necessary resources, as well as the anonymous reviewers for their valuable suggestions and comments.

Funding. The authors would like to thank the Ministry of Higher Ed-

ucation, Science, and Technology (Kemdiktisaintek) for funding this research through the Indonesian Education Scholarship (BPI) decree number 00827/BPPT/BPI.06/9/2023.

Conflict of interest. The authors declare no competing interests.

Data availability. Not applicable.

References

[1] B. Sawicka *et al.*, *Chapter 14 - The coronavirus global pandemic and its impacts on society*. Elsevier, 2022, pp. 267–311. ISBN:9780323851565. DOI:10.1016/B978-0-323-85156-5.00037-7

[2] C. L. Atzrodt *et al.*, "A guide to covid-19: a global pandemic caused by the novel coronavirus sars-cov-2," *FEBS Journal*, vol. 287, no. 17, pp. 3633–3650, 2020. DOI:10.1111/febs.15375

[3] W. H. O. (WHO), "Weekly epidemiological update on covid-19 - 1 february 2022," *URL: https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19—1-february-2022*, Accessed on 1 February 2022.

[4] M. Rayungsari, M. Aufin, and N. Imamah, "Parameters estimation of generalized richards model for covid-19 cases in indonesia using genetic algorithm," *Jambura Journal of Biomathematics (JJBM)*, vol. 1, no. 1, pp. 25–30, 2020. DOI:10.34312/jjbm.v1i1.6910

[5] M. C. Schippers and D. C. Rus, "Optimizing decision-making processes in times of covid-19: Using reflexivity to counteract information-processing failures," *Frontiers in Psychology*, vol. 12, 2021. DOI:10.3389/fpsyg.2021.650525

[6] N. Susyanto and J. P. Arcede, "Unveiling sir model parameters: Empirical parameter approach for explicit estimation and confidence interval construction," *Jambura Journal of Biomathematics (JJBM)*, vol. 5, no. 1, pp. 54–62, 2024. DOI:10.37905/jjbm.v5i1.26287

[7] P. Libin *et al.*, *Deep reinforcement learning for large-scale epidemic control*. Springer, 2021, pp. 155–170. DOI:10.1007/978-3-030-67670-4_10

[8] B. Lin, G. Cecchi, and D. Bouneffouf, *Psychotherapy ai companion with reinforcement learning recommendations and interpretable policy dynamics*. ACM, 2023, pp. 932–939, ISBN:9781450394192. DOI:10.1145/3543873.3587623

[9] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press Cambridge, Massachusetts, 2018.

[10] R. Damaševičius, R. Maskeli□nas, and S. Misra, *Using Reinforcement Learning for Optimizing COVID-19 Vaccine Distribution Strategies*. Springer, Cham, 2023, pp. 169–196. DOI:10.1007/978-3-031-33183-1_10

[11] R. Awasthi *et al.*, "Vacsim: Learning effective strategies for covid-19 vaccine distribution using reinforcement learning," *Intelligence-Based Medicine*, vol. 6, p. 100060, 2022. DOI:10.1016/j.ibmed.2022.100060

[12] S. Chen *et al.*, "Reinforcement learning based diagnosis and prediction for covid-19 by optimizing a mixed cost function from ct images," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 11, pp. 5344–5354, 2022. DOI:10.1109/JBHI.2022.3197666

[13] M. N. A. Al-Hamadani *et al.*, "Reinforcement learning algorithms and applications in healthcare and robotics: A comprehensive and systematic review," *Sensors*, vol. 24, no. 8, p. 2461, 2024. DOI:10.3390/s24082461

[14] Y. Lin *et al.*, "Evolutionary reinforcement learning: A systematic review and future directions," *Mathematics*, vol. 13, no. 5, p. 833, 2025. DOI:10.3390/math13050833

[15] M. Tejedor *et al.*, "Reinforcement learning application in diabetes blood glucose control: A systematic review," *Artificial Intelligence in Medicine*, vol. 104, p. 101836, 2020. DOI:10.1016/j.artmed.2020.101836

[16] Y. Zhao *et al.*, "Systematic literature review on reinforcement learning in non-communicable disease interventions," *Artificial Intelligence in Medicine*, vol. 154, p. 102901, 2024. DOI:10.1016/j.artmed.2024.102901

[17] M. S. E. Martins, J. M. C. Sousa, and S. Vieira, "A systematic review on reinforcement learning for industrial combinatorial optimization problems," *Applied Sciences*, vol. 15, no. 3, p. 1211, 2025. DOI:10.3390/app15031211

[18] Z. Tang *et al.*, "A systematic literature review of reinforcement learning-based knowledge graph research," *Expert Systems with Applications*, vol. 238, p. 121880, 2024. DOI:10.1016/j.eswa.2023.121880

[19] J. Escobar-Naranjo *et al.*, "Autonomous navigation of robots: Optimization with dqn," *Applied Sciences*, vol. 13, no. 12, p. 7202, 2023. DOI:10.3390/app13127202

[20] R. Vaish *et al.*, "Machine learning applications in power system fault diagnosis: Research advancements and perspectives," *Eng. Appl. Artif*, vol. 106, p. 104504, 2021. DOI:10.1016/j.engappai.2021.104504

[21] D. Qu *et al.*, "A two-stage decomposition-reinforcement learning optimal combined short-time traffic flow prediction model considering multiple factors," *Applied Sciences (Switzerland)*, vol. 12, no. 16, 2022. DOI:10.3390/app12167978

[22] L. Uttley *et al.*, "The problems with systematic reviews: a living systematic review," *Journal of Clinical Epidemiology*, vol. 156, pp. 30–41, 2023. DOI:10.1016/j.jclinepi.2023.01.011

[23] R. Padmanabhan *et al.*, "Reinforcement learning-based decision support system for covid-19," *BSPC*, vol. 68, p. 102676, 2021. DOI:10.1016/j.bspc.2021.102676

[24] S. Roy, R. Dutta, and P. Ghosh, "Towards dynamic lockdown strategies controlling pandemic spread under healthcare resource budget," *Applied Network Science*, vol. 6, 2021. DOI:10.1007/s41109-020-00349-0

[25] M. A. Chadi and H. Mousannif, "A reinforcement learning based decision support tool for epidemic control: Validation study for covid-19," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022. DOI:10.1080/08839514.2022.2031821

[26] M. Arango and L. Pelov, "Covid-19 pandemic cyclic lockdown optimization using reinforcement learning," *ArXiv*, vol. abs/2009.04647, 2020. DOI:10.48550/arXiv.2009.04647

[27] A. Q. Ohi *et al.*, "Exploring optimal control of epidemic spread using reinforcement learning," *Scientific Reports*, vol. 10, no. 1, p. 22106, 2020. DOI:10.1038/s41598-020-79147-8

[28] E. A. Ondula and B. Krishnamachari, "Using reinforcement learning for operating educational campuses safely during a pandemic (student abstract)," *AAAI Conference*, vol. 36, no. 11, pp. 13025–13026, 2022. DOI:10.1609/aaai.v36i11.21649

[29] S. Bushaj *et al.*, "A simulation-deep reinforcement learning (sirl) approach for epidemic control optimization," *Annals of Operations Research*, vol. 328, pp. 245–277, 2023. DOI:10.1007/s10479-022-04926-7

[30] H. Khadilkar *et al.*, "Optimising lockdown policies for epidemic control using reinforcement learning," *INAE*, vol. 5, pp. 129–132, 2020. DOI:10.1007/s41403-020-00129-3

[31] G. H. Kwak, L. Ling, and P. Hui, "Deep reinforcement learning approaches for global public health strategies for covid-19 pandemic," *PLoS ONE*, vol. 16, no. 5, p. e0251550, 2021. DOI:10.1371/journal.pone.0251550

[32] X. Du *et al.*, "Hrl4ec: Hierarchical reinforcement learning for multi-mode epidemic control," *Information Sciences*, vol. 640, p. 119065, 2023. DOI:10.1016/j.ins.2023.119065

[33] R. Wan, X. Zhang, and R. Song, "Multi-objective model-based reinforcement learning for infectious disease control," in *ACM SIGKDD Conference*, pp. 1634–1644, 2021. DOI:10.1145/3447548.3467303.

[34] A. Vereshchaka and N. Kulkarni, "Optimization of mitigation strategies during epidemics using offline reinforcement learning," in *SBP-BRiMS*, vol. 12720, pp. 35–45, 2021. DOI:10.1007/978-3-030-80387-2_4.

[35] X. Guo *et al.*, "Pacar: Covid-19 pandemic control decision making via large-scale agent-based modeling and deep reinforcement learning," *Medical Decision Making*, vol. 42, no. 8, pp. 1064–1077, 2022. DOI:10.1177/0272989X221107902

[36] A. Mai *et al.*, "Planning multiple epidemic interventions with reinforcement learning," in *IJCAI Proceedings*, pp. 6147–6155, 2023. DOI:10.24963/ijcai.2023/682

[37] V. Kompella *et al.*, "Reinforcement learning for optimization of covid-19 mitigation policies," *CEUR Workshop Proceedings*, 2020, pp. 1–8. DOI:10.48550/arXiv.2010.10560

[38] K. Zong and C. Luo, "Reinforcement learning based framework for covid-19 resource allocation," *Computers & Industrial Engineering*, vol. 167, p. 107960, 2022. DOI:10.1016/j.cie.2022.107960

[39] M. I. Uddin *et al.*, "Optimal policy learning for covid-19 prevention using reinforcement learning," *Journal of Information Science*, vol. 48, pp. 336–348, 2022. DOI:10.1177/0165551520959

[40] N. Ghazizadeh *et al.*, "Modeling and control of covid-19 disease using deep reinforcement learning method," *MBEC*, vol. 62, no. 12, pp. 3653–3670, 2024. DOI:10.1007/s11517-024-03153-5

[41] M. Reymond *et al.*, "Exploring the pareto front of multi-objective covid-19 mitigation policies using reinforcement learning," *Expert Systems with Applications*, vol. 249, p. 123686, 2024. DOI:10.1016/j.eswa.2024.123686

[42] S. N. Khatami and C. Gopalappa, "Deep reinforcement learning framework for controlling infectious disease outbreaks in the context of multi-jurisdictions," *Mathematical Biosciences and Engineering*, vol. 20, no. 8, pp. 14306–14326, 2023. DOI:10.3934/mbe.2023640