Information Sciences 640 (2023) 119065

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

HRL4EC: Hierarchical reinforcement learning for multi-mode epidemic control ☆

Xinqi Du [a,b], Hechang Chen [a,b,d,*], Bo Yang [c,d], Cheng Long [e], Songwei Zhao [a,b]

[a] School of Artificial Intelligence, Jilin University, Changchun 130012, China
[b] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education, China
[c] College of Computer Science and Technology, Jilin University, Changchun 130012, China
[d] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[e] School of Computer Science and Engineering, Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Hierarchical reinforcement learning
Epidemic control
Multi-mode intervention
Deep reinforcement learning

ABSTRACT

Infectious diseases, such as Black Death, Spanish Flu, and COVID-19, have accompanied human history and threatened public health, resulting in enormous infections and even deaths among citizens. Because of their rapid development and huge impact, laying out interventions becomes one of the most critical paths for policymakers to respond to the epidemic. However, the existing studies mainly focus on epidemic control with a single intervention, which makes the epidemic control effectiveness severely compromised. In view of this, we propose a *H*ierarchical *R*einforcement *L*earning decision framework for multi-mode *E*pidemic *C*ontrol with multiple interventions called *HRL4EC*. We devise an epidemiological model, referred to as *MID-SEIR*, to describe multiple interventions' impact on transmission explicitly, and use it as the environment for HRL4EC. Besides, to address the complexity introduced by multiple interventions, this work transforms the multi-mode intervention decision problem into a multi-level control problem, and employs hierarchical reinforcement learning to find the optimal strategies. Finally, extensive experiments are conducted with real and simulated epidemic data to validate the effectiveness of our proposed method. We further analyze the experiment data in-depth, conclude a series of findings on epidemic intervention strategies, and make a visualization accordingly, which can provide heuristic support for policymakers' pandemic response.

1. Introduction

Infectious diseases have always accompanied human history, with outbreaks in different areas of the world at different times [1,2]. For example, the Black Death outbroke in the 14th century and became the fatal pandemic in human history, which caused the death of around 200 million people. The Ebola virus disease (EVD) was discovered in 1976, which is one of the deadliest viral diseases, with frequent outbreaks in Africa [3]. The current Coronavirus Disease 2019 (COVID-19) [4] sweeps worldwide and leads to an unprecedented crisis, and there have been about 760 million confirmed cases so far. Due to the rapid development and huge

☆ This work is partially supported in part by the National Key R&D Program of China (2021ZD0112501 and 2021ZD0112502); the International Cooperation Project of Jilin Province (20220402009GH); the Special Project for Health Research Talents in Jilin Province (2020SCZ39); the National Natural Science Foundation of China (61976102 and U19A2065); and the Graduate Innovation Fund of Jilin University (2022075).

* Corresponding author at: School of Artificial Intelligence, Jilin University, Changchun 130012, China.
E-mail address: chenhc@jlu.edu.cn (H. Chen).

https://doi.org/10.1016/j.ins.2023.119065
Received 7 December 2022; Received in revised form 21 March 2023; Accepted 27 April 2023

Available online 9 May 2023
0020-0255/© 2023 Elsevier Inc. All rights reserved.

impact of infectious diseases, effective epidemic control is significant. The ability to lay out the epidemic intervention, i.e., **HOW** and **WHEN** to lay out **WHICH** epidemic interventions, is an important part of the national response to the epidemic. Thus, there is an urgent need to come up with an intervention strategy to combat the current outbreak effectively [5].

A rich of literature has been proposed to study the problem of epidemic modeling and epidemic control. Due to the scarcity of real-world epidemic data, it is vital to construct the epidemiological model for downstream epidemic control tasks. The epidemic modeling mainly falls into three main categories, i.e., compartmental epidemiological models [6,7], agent-based models [8,9], and machine learning enabled models [10,11]. While epidemic control roughly includes two types of interventions, i.e., pharmacological interventions [12–14], and non-pharmacological interventions [15–17]. Since dynamics and uncertainty accompany epidemic transmission, the rule-of-thumb intervention strategy is not flexible enough for the transient epidemic. Thus, the application of reinforcement learning to epidemic control problems has arisen, thanks to its strong ability in sequential decision-making problems. For example, reinforcement learning has been deployed to cope with the challenge of scarce medical resources allocation in the context of epidemic [18,13,19]. Some researchers propose to employ reinforcement learning for mobility-control strategy [20]. Besides, reinforcement learning is also used to study the city-lockdown strategy with minimized economic loss [16].

Although existing studies have made progress in epidemic modeling and epidemic control, several concerns remain: 1) **Explicitly modeling the external-driven force of epidemic transmission**. Existing models [21,22] only focus on the transmission mode of infectious diseases under internal-driven forces and do not explicitly model the impact of external-driven forces on the transmission. That is, intervention strategies have a direct or indirect impact on the transmission of infectious diseases, which needs to be taken into consideration explicitly. 2) **Multi-mode intervention strategies**. Most existing studies [12,20,15] design strategies with a single intervention, which is biased on real-world problems. In practice, policymakers often deploy multiple interventions simultaneously to reduce the impact of epidemic outbreaks more effectively. In other words, different interventions play different roles in epidemic control. For example, mobility constraints can reduce the infection rate; and increasing temporary medical facilities can accelerate positive tests or improve admission rates, all of which are beneficial to control the further spread of infectious diseases. Motivated by these issues of the existing studies, we aim to take one step further to investigate the multi-mode intervention strategy.

In this work, we propose a Hierarchical Reinforcement Learning decision system for multi-mode Epidemic Control and refer to it as HRL4EC. To enhance the interventions' effectiveness, we employ multiple epidemic interventions and design a novel epidemiological model called *Multilateral-Impact-Driven SEIR* model (MID-SEIR). In contrast to the traditional epidemiological model (describing the internal-driven force of epidemic transmission), it explicitly describes the external-driven force of infectious disease transmission, i.e., the impact of different intervention strategies on the transmission. Meanwhile, hierarchical reinforcement learning is deployed to deal with the complexity of the multi-mode intervention decision problem, which divides the problem into hierarchical subtasks. Extensive experiments are conducted with the real-world epidemic dataset, and the experimental results demonstrate the effectiveness of our proposed method. Besides, the visualization and the analysis of the results are presented to obtain the findings on laying out intervention strategy.

The main contributions of our work are in three-folds:

- A novel decision framework called HRL4EC is devised for epidemic control with a multi-mode intervention strategy, which performs better in epidemic control. To the best of our knowledge, this is the first effort in the literature that proposes to employ multiple interventions to curb epidemic outbreaks in a multi-mode manner.
- An epidemiological model called MID-SEIR is proposed to explicitly describe multiple intervention impacts on epidemic transmission, i.e., external-driven forces. Furthermore, hierarchical reinforcement learning is employed to solve the complex multi-mode epidemic control problem by decomposing the original problem into multi-level control tasks. It makes the decision to lay out the mixture of multiple interventions simultaneously.
- The effectiveness of HRL4EC is validated by conducting comprehensive experiments on real-world and simulation epidemic datasets. Extensive experimental results demonstrate the superiority of HRL4EC in terms of effectiveness in epidemic control, and the result visualizations provide the findings on laying out intervention strategies that can assist policymakers in tackling the epidemic control problem in the future.

The remainder of this paper is structured as follows: we first summarize the existing studies related to our method in Section 2 and then present the notions and basic concepts in Section 3. The details of the proposed model are shown in Section 4. We conduct extensive experiments and display the results in Section 5. Finally, we conclude the paper with discussions in Section 6.

2. Related work

In this section, we briefly discuss the previous studies related to our proposed method from three aspects, i.e., epidemic control, reinforcement learning for epidemics, and hierarchical reinforcement learning.

2.1. Epidemic control

A substantial amount of literature has studied epidemic control, which mainly consists of: modeling the epidemic dynamics and designing intervention strategies.

For epidemiological model, it can be roughly divided into three categories, i.e., compartmental epidemiological model [23,24], agent-based model [8,25], and machine learning enabled model [6,26]. The compartmental epidemiological models, such as SI [23],

SIR [21], SEIR [22], and so forth, are the most widely used, which describe epidemic transmission among different compartments straightforwardly. However, the existing compartmental epidemiological models lack the ability to describe the impacts of external interventions on infectious diseases.

As for intervention strategies, on the one hand, some researchers propose to control the epidemic by using pharmaceutical interventions (PIs), such as vaccines [12,13,27], and other medical resources [28], which are challenging due to the resources scarcity. On the other hand, non-pharmaceutical interventions (NPIs), such as suppression [15], mobility control [29,30,20] and lockdown [16], have been deployed to hinder the spread of the pandemic, which need to trade off the public health and economic cost.

However, the existing studies mainly focus on making the decision on a single intervention, which does not capture real-world scenarios well. Unlike the existing studies, we mainly focus on incorporating multilateral impacts on epidemic transmission by modeling epidemic dynamics with multi-mode interventions to curb the pandemic outbreaks effectively.

2.2. Reinforcement learning for public health

Reinforcement learning is a powerful method for sequential decision-making problems and has been successfully applied in the fields of games [31], control [32], and healthcare [33,34], and so forth.

Recently, it has also been widely used to assist in the prevention and control of the epidemic, which is roughly in three categories: 1) Reinforcement learning has been applied for epidemic prediction [25], which provides a better approximation of the epidemic dynamics compared to the traditional epidemic model. 2) Reinforcement learning method can be used for epidemic control by searching for the optimal intervention strategies in terms of mobility-control [20], city-lockdown [16], and so forth. 3) Some studies attempt to employ the reinforcement learning methods for medical resource allocations, such as testing resources [18], masks [28], and vaccines [13], which are vital but scarce in the context of the epidemic.

Generally, existing studies mainly focus on single-mode epidemic control while almost unexplored how to deal with multi-mode interventions, which is usually seen in real-world scenarios though. In contrast, this work is the first to model the epidemic control problem with multi-mode interventions.

2.3. Hierarchical reinforcement learning

Hierarchical reinforcement learning attempts to effectively solve complex tasks by dividing the original problem into equivalent sub-problems and solving them in turn. This process may introduce more 'internal reward' to the agent while greatly reducing the complexity of the solution space, which enables an efficient and stable training process. Recently, it has been widely used in many domains.

Specifically, FeUdal Networks [35] is designed for solving complex games by formulating sub-goals as directions in latent state space with meaningful behavioral primitives. HRL-Rec [36] employs hierarchical reinforcement learning in the recommendation system, which jointly recommends heterogeneous items from different sources. Besides, the hierarchical reinforcement learning method is also proposed for resource allocations [28].

However, it is unexplored how to apply hierarchical reinforcement learning to epidemic control problems, which are commonly large-scale and involve complex solution space. It is the first effort to solve the multi-mode intervention decision-making problem with hierarchical reinforcement learning.

3. Preliminaries

This section will present the notations used in our proposed model and review the traditional epidemiological model and deep reinforcement learning, which provide fundamental theoretical support for our proposed method.

3.1. Notations

We attempt to model the underlying problem, i.e., a sequential decision-making problem, as a Markov Decision Process (MDP). Specifically, the MDP can be represented by a tuple $G = \langle S, \mathcal{A}, T, R, \gamma \rangle$, where S is the set of states, and $\mathcal{A}$ is the action space. $T(s'|s,a) := S \times \mathcal{A} \times S \mapsto [0,1]$ is the state transition function used to describe the environment dynamics. $R(s,a) := S \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function capturing the signal from the environment to encourage the agent and $\gamma \in [0,1)$ is the discount factor to balance immediate reward and long-term gain. At each time step t, the agent observes state s_t and executes action a_t decided by policy $\pi(a_t|s_t)$; accordingly, the environment transits to the next state s_{t+1} (or denoted as s') and returns reward r_t to the agent. In reinforcement learning, the objective for the agent is to maximize its expected discounted accumulated return: $\mathbb{E}[R_T] = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t^t\right]$, where T is the final time step (Table 1).

3.2. Epidemiological model

The compartmental model is one of the most widely used models to describe the underlying dynamics of infectious diseases, such as SIR [21], SIRS [24], and so forth. Such models not only provide a concise and compact description of the epidemic dynamics, but are also flexible enough to be extended with additional attributes for specialized situations. In our work, we make the extension

Table 1
Notations.

Symbol	Description
$\mathcal{S}$	The set of the global states
$\mathcal{A}$	The set of the actions for agent
R	Reward function for agent
T	Transition function for agent
π_θ	Policy parameterized by θ for agent
s_t	The state agent observed at time step t
a_t	The action agent executed at time step t
r_t	The reward agent received at time step t
γ	Discount factor

on the susceptible-exposed-infected-recovered (SEIR) compartmental model [22] and assume that birth and death are not taken into consideration. Specifically, it satisfies the following equation:

$$S(t) + E(t) + I(t) + R(t) \equiv N, \tag{1}$$

where at each time step t, $S(t)$ denotes the number of susceptible individuals, $E(t)$ is the number of exposed individuals, $I(t)$ is the number of infected cases, $R(t)$ is the number of recovered cases, and N is the total population. Based on the assumption, SEIR model is defined by the following ordinary differential equations:

$$\begin{aligned}
\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= -\beta S(t)\frac{I(t)}{N}, \\
\frac{\mathrm{d}E(t)}{\mathrm{d}t} &= \beta S(t)\frac{I(t)}{N} - \alpha E(t), \\
\frac{\mathrm{d}I(t)}{\mathrm{d}t} &= \alpha E(t) - \lambda I(t), \\
\frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \lambda I(t),
\end{aligned} \tag{2}$$

where β denotes the rate of exposure to infected cases, α is the epidemic transmission rate, and λ is the recovery rate. Essentially, they describe the $S \to E, E \to I$ and $I \to R$ probabilities at each time step, where $\to$ denotes the epidemiological state transition.

3.3. Reinforcement learning method

Reinforcement learning has achieved success in many domains, such as game playing [31], robotic control [37,32], etc. [38]. Proximal policy optimization (PPO) [39] is one of the most widely applied reinforcement learning methods, and is employed in our work. Two policies are deployed in the PPO, denoted as π_θ and $\pi_{\hat{\theta}}$. Its surrogate objective function is shown as follows:

$$\begin{aligned}
J^{PPO}(\theta) = \mathbb{E}_t\Big[& \min\left(r_t(\theta)A_t^{\pi_{\hat{\theta}}}, \right. \\
& \left. \mathrm{clip}\left(r_t(\theta), 1-\epsilon, 1+\epsilon\right)A_t^{\pi_{\hat{\theta}}}\right)\Big],
\end{aligned} \tag{3}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\hat{\theta}}(a_t|s_t)}$ is the probability ratio for important sampling. clip() is the clipping function to limit $r_t(\theta)$ within $(1-\epsilon, 1+\epsilon)$, and $A_t^{\pi_{\hat{\theta}}}$ is the advantage computed by general advantage estimation method as follows:

$$A_t^{\pi_{\hat{\theta}}} = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots\cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \tag{4}$$

where δ_t is the TD error and defined as follows:

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t), \tag{5}$$

where $V(s)$ is the approximation of the state-value function. Its objective function is shown as follows:

$$J^{Value}(\phi) = \mathbb{E}\left[\left(V_\phi(s_t) - \hat{R}_t\right)^2\right], \tag{6}$$

where $\hat{R}_t$ is the rewards-to-go, i.e., the discounted accumulated reward. In the following section, the on-policy method, PPO, is employed to train the epidemic intervention decision model.

4. Methodology

4.1. System overview

To solve the multi-mode interventions decision-making problem in epidemic control, we propose a Hierarchical Reinforcement Learning Decision Framework called HRL4EC for Epidemic Control, with a multilateral-impact-driven epidemiological model. The overview of the HRL4EC architecture is shown in Fig. 1.

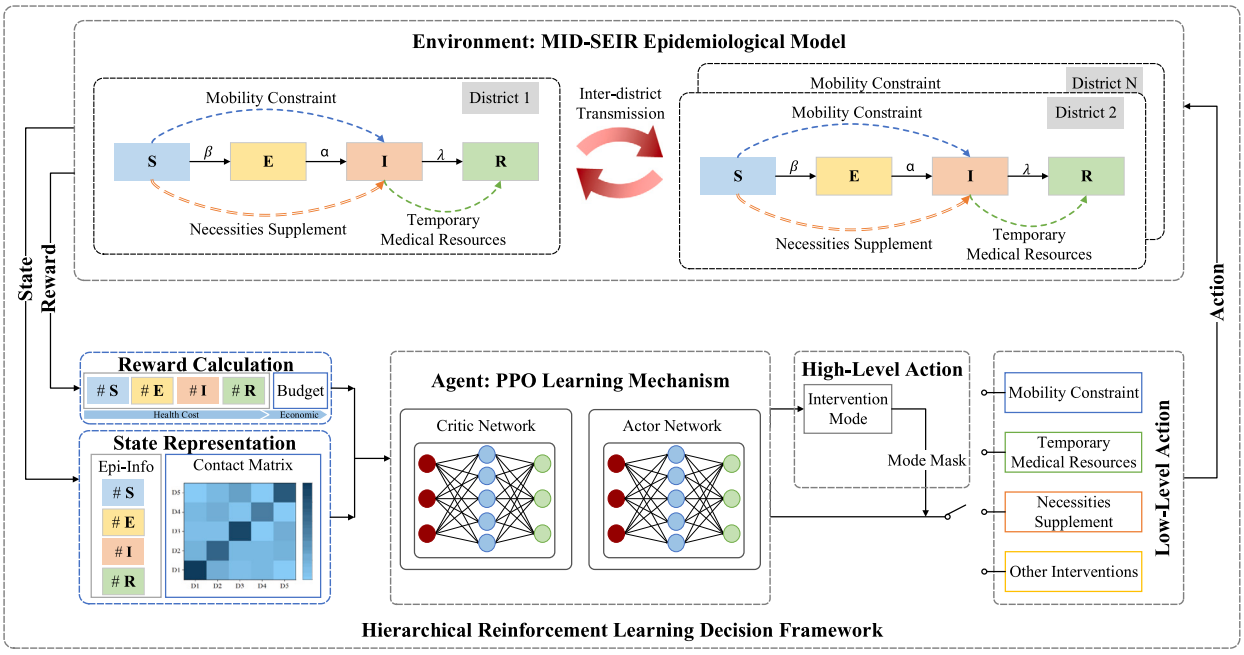


Fig. 1. The overall architecture of the proposed HRL4EC. 1) MID-SEIR Epidemiological Model: it explicitly describes the multiple interventions' impacts on epidemic transmission and is employed as the environment of the decision framework. 2) Hierarchical Reinforcement Learning Decision Framework: it models the epidemic control problem as a reinforcement learning problem and decomposes the multi-mode control problem into a multi-level control problem by abstracting high-level and low-level actions.

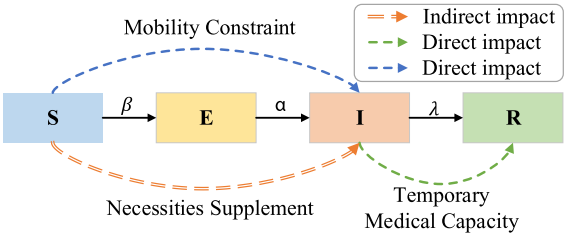


Fig. 2. The structure of the epidemiological model, i.e., Multilateral-Impact-Driven SEIR model (MID-SEIR). The colored blocks indicate the groups belonging to different epidemiological states. The colored arrows denote the impacts of the different interventions on epidemic transmission.

First, we devise a novel epidemiological model, called Multilateral-Impact-Driven SEIR (MID-SEIR), to simulate the epidemic transmission under various intervention modes (Section 4.2). It is used to describe the epidemic transmission and collect the interaction data for decision algorithm training. Second, hierarchical reinforcement learning is employed to solve the complex multi-mode intervention decision problem (Section 4.3), which alleviates the challenge of the large-scale solution space. The details about the proposed model can be found in the following section.

4.2. *MID-SEIR epidemiological model*

A major challenge to epidemic control is the scarcity of infectious disease transmission data. It is known that it cannot directly verify the effectiveness of intervention strategy in real-world scenarios to collect infectious disease data, which poses a challenge for its evaluation. To tackle this challenge, substantial epidemiological models [24,8] have been proposed to model epidemic transmission. However, most of the existing epidemiological models [21,22] focus on describing the underlying infectious disease transmission patterns and do not take into account the impact of external interventions on transmission.

A novel epidemiological model is proposed, called *Multilateral-Impact-Driven SEIR* model (MID-SEIR). It extends the traditional SEIR model and incorporates the impact of various interventions on epidemic transmission. The state transition diagram of MID-SEIR is illustrated in Fig. 2, where the colored dashed arrows indicate the direct or indirect influence on epidemic transmission carried out by the different intervention modalities.

For clarity, three types of intervention are considered, including pharmacological intervention (PI) and non-pharmacological intervention (NPI), such as mobility constraint (NPI), temporary medical resources (PI), and necessities supply (NPI), in MID-SEIR model:

- **Mobility Constraint** reminds the residents to keep social distance or limits the range of their activity, which is conducive to reducing the probability of exposure to the virus environment. It has a direct impact on state transmission as shown by the blue arrows in Fig. 2.
- **Temporary Medical Resources**, such as testing reagents, drugs and hospital beds, have a direct impact on epidemic control by quickly identifying positive cases and speeding up the recovery rate of infected cases, as shown by the green arrows in Fig. 2. In this way, the risk of the secondary spread of the epidemic can be evitable.
- **Necessities Supply** affects the emotions of the residents, which in turn has a bearing on how well they cooperate with the demands of epidemic control, as shown by the orange arrow in Fig. 2. Scarcity of supplies during the epidemic outbreaks increases the anxiety and panic among residents [40], which indirectly hinders the progress of epidemic control.

To reflect the impact of the mobility constraint, we first divide the blocks according to the administrative areas and obtain the contact frequency among the blocks. $\mathcal{M}$ denotes the contact frequency matrix, which is defined as follows:

$$\mathcal{M} = \begin{bmatrix} m_{11} & \cdots & m_{1|D|} \\ \vdots & \ddots & \vdots \\ m_{|D|1} & \cdots & m_{|D||D|} \end{bmatrix}, \mathcal{M} \in \mathbb{R}^{|D|\times|D|}. \tag{7}$$

The human contact frequency between block i and block j is denoted as m_{ij} in matrix $\mathcal{M}$, and its value is proportional to the risk of infection. In other words, when the policymaker lays out the mobility constraint, the human contact frequency m_{ij} will be changed accordingly, and thus they can master the process of epidemic transmission. For intra-block, MID-SEIR is used to model the spread of infectious diseases.

The overall epidemiological model with four states to describe the transmission of the epidemic is expressed as the following ordinary differential equations:

$$\begin{aligned} \frac{\mathrm{d}S_i}{\mathrm{d}t} &= -\sum_{j\in D} \frac{\beta m_{ij}\left((1+\delta_{anx})\epsilon(\cdot)E_j + \mu(\cdot)I_j\right)}{N_j} S_i, \\ \frac{\mathrm{d}E_i}{\mathrm{d}t} &= \sum_{j\in D} \frac{\beta m_{ij}\left((1+\delta_{anx})\epsilon(\cdot)E_j + \mu(\cdot)I_j\right)}{N_j} S_i - \alpha E_i, \\ \frac{\mathrm{d}I_i}{\mathrm{d}t} &= \alpha E_i - (1+\delta_{acc})\lambda I_i, \\ \frac{\mathrm{d}R_i}{\mathrm{d}t} &= (1+\delta_{acc})\lambda I_i, \end{aligned} \tag{8}$$

where β denotes the probability of transmission per contact with infectious virus carriers. α is the rate of transfer from exposed to infectious. λ is the recovery rate of infectious individuals. We denote the remaining number of susceptible people, the number of exposed cases, the number of infected cases, and the number of recovered individuals in each block i as S_i, E_i, I_i and R_i, and the sum of these four groups as N_i. The $\epsilon(\cdot)$ denotes the proportion of exposed cases that are infectious. For example, COVID-19 has an incubation period that is also infectious. The $\mu(\cdot)$ indicates the proportion of the infected cases who are negligently protected. Besides, we newly incorporate the multilateral intervention impacts driving the pandemic transmission into the epidemiological model. δ_{anx} denotes the degree of residents' anxiety, and its value is determined by the degree of necessities supply $a_L^{\mathcal{NS}}$, i.e., $\delta_{anx} \propto 1/a_L^{\mathcal{NS}}$. δ_{acc} indicates the advantages that temporary medical resources bring to curing infected cases.

Algorithm 1: Hierarchical reinforcement learning for multi-mode epidemic control (HRL4EC).

Input: The epidemiological parameters, policy parameters and value function parameters.
Output: The parameters of the policies.

Initialize the epidemiological model with epidemiological parameters;
Initialize the contact matrix $\mathcal{M}$;
Initialize the policy networks $\pi_{\phi_H}, \pi_{\phi_L}$, critic networks V_θ and replay buffer $\mathcal{RB}$.
for $k = 0, 1, 2, \cdots, max_episodes$ **do**
  for $step = 1, 2, \cdots, max_length$ **do**
    Observe state s and choose high-level intervention mode: $a_H \leftarrow \pi_{\phi_H}(s)$;
    Choose the granularity of each interventions $a_L = [a_L^{\mathcal{MC}}, a_L^{\mathcal{TMR}}, a_L^{\mathcal{NS}}]$ according to π_{ϕ_L};
    Execute actions $a = mask(a_H, a_L)$, receive s' and r from the environment;
    Store transitions in $\mathcal{RB}$, and set $s \leftarrow s'$;
  Compute rewards-to-go $\hat{R}_t$ in the replay buffer $\mathcal{RB}$;
  Compute advantage estimates $\hat{A}_t$;
  Update the policies parameters by maximizing the objective function, i.e., Eq. (3), via stochastic gradient ascent with Adam;
  Fit value function, i.e., Eq. (6), by regression on mean-squared error via gradient descent method;
Return the parameters for policy networks and critic networks.

4.3. Hierarchical reinforcement learning decision framework

To solve the complex problem of epidemic multi-mode intervention decisions, where the enormous solution space and the high dynamics within the pandemic pose a challenge for the policymaker, we devise a hierarchical decision framework with reinforcement learning, referred to as hierarchical reinforcement learning for epidemic control (HRL4EC).

Generally speaking, the multi-mode intervention decision problem is hierarchically decomposed into multiple sub-problems, which alleviates the challenge of complexity. Then, we deploy deep reinforcement learning with the advantage of time-series decision-making to learn the multiple interventions layout strategies, i.e., **WHEN** and **HOW** to lay out **WHICH** interventions. In the following, the intervention decision problem is modeled as a reinforcement learning task by specifying the key components, i.e., environment, agent, state, action, reward, and transition function:

Environment: Considering that directly applying the intervention policy to realistic scenarios poses a threat to public health and property, we utilize the idea of the sim-to-real and design an epidemiological model, MID-SEIR, as the environment for agent interaction. The detail of MID-SEIR can be referred to in Section 4.2.

Agent: The agent takes the role of a policymaker in the real-world, making decisions on appropriate intervention by judging the current situation of the pandemic.

State: The temporal integration of the current status of the pandemic ($\#S, \#E, \#I, \#R$) and the degree of human contact ($\mathcal{M}$) at time step t is used as the state vector $\mathbf{s}_t$ in reinforcement learning.

Action: To cope with the problem of great complexity in epidemic control, hierarchical reinforcement learning is adopted to divide the problem hierarchically and seek multi-level control solutions. Specifically, the different control levels (also known as options) in epidemic control problems are abstracted, and the upper level masters the lower level. The detailed definitions used in epidemic control problem are as follows:

1) *High-Level action* a_H: it is defined as the mode of implementing the epidemic intervention, i.e., single-mode intervention or multi-mode intervention, where in the single-mode setting: the agent simply adopts one of the interventions for epidemic control (in this work, mobility constraint is used in the single-mode setting, which can effectively control the epidemic [41]). In contrast, in the multi-mode setting: the agent simultaneously lays out multiple interventions to control the further development of the pandemic collaboratively. The high-level action a_H is determined by the policy π_{ϕ_H}. It addresses the problem of **WHICH** interventions to deploy.

2) *Low-Level action* a_L: it is defined as the fine-grained implementation strength of interventions. In this paper, three interventions are considered, i.e., mobility constraint, temporary medical resources, and necessities supply, all of which have a direct or indirect impact on curbing the spread of infectious diseases. It addresses the problem of **WHEN** and **HOW** to lay out the interventions. The details are as follows:

- For mobility constraint, the low-level action is denoted as $a_L^{\mathcal{MC}} \in [0, 100\%]$, which straightforwardly affects the contact matrix $\mathcal{M}$ (refer to Section 4.2). Generally, $a_L^{\mathcal{MC}} = 0\%$ corresponds to no constraint on residents' mobility; $a_L^{\mathcal{MC}} = 100\%$ indicates that the most strict constraint is imposed, and only the essential activities (such as healthcare and supplies delivery) are going on. Accordingly, the value of m_{ij} in contact matrix $\mathcal{M}$ decreases w.r.t. the initial value in matrix $\mathcal{M}$ but is still greater than zero, i.e., $m_{ij} > 0$.
- For temporary medical resources, the low-level action is denoted as $a_L^{\mathcal{TMR}} \in [L, M, H]$, which has an impact on the recovery rate λ. When $a_L^{\mathcal{TMR}} = L$, it indicates that healthcare is provided by the medical facilities already available in the city; otherwise, the government makes the effort to provide more drugs or hospital beds to improve patient recovery rates λ. Specifically, it is assumed that the improvement δ_{acc} is approximately 3% ($a_L^{\mathcal{TMR}} = M$) and 6% ($a_L^{\mathcal{TMR}} = H$), respectively.
- For necessities supply, the low-level action is denoted as $a_L^{\mathcal{NS}} \in [Non, L, M]$, which primarily impacts human emotions and further influences the effectiveness of the epidemic control. The policymaker makes no consideration to ensure the supply of daily necessities when $a_L^{\mathcal{NS}} = Non$. In contrast, when $a_L^{\mathcal{NS}} = L$ (or M), more measures are taken to ensure daily necessities supply, which helps to mobilize the residents' initiative. Specifically, it is assumed that δ_{anx} is approximately -5% ($a_L^{\mathcal{NS}} = L$) and -10% ($a_L^{\mathcal{NS}} = M$), respectively.

The Low-Level action a_L is determined by the policy π_{ϕ_L}. Finally, $\mathrm{mask}(\cdot,\cdot)$ is the masking function to retain the availed intervention action in a_L to execute.

Reward: The reward function mainly consists of two terms, one of which is related to public health (refers to r_h) and the other to socio-economic overheads (refers to r_e). In this way, it encourages the agent to search for an effective intervention policy, which minimizes the economic damage while ensuring the containment of the pandemic outbreaks. Here, the reward function is denoted as R, and its definition is as follows:

$$R = (1-\omega) \times r_h + \omega \times r_e, \tag{9}$$

where ω is the weight hyper-parameter to balance the significance of the public health and economic losses. The public health related term r_h is defined as the following equation:

$$r_h = \mathrm{Normalize}(\Delta S) = \frac{\#S_t - \#S_{t+1}}{N} \tag{10}$$

Table 2
The attributes of the areas where COVID-19 local outbreaks. Part of them are related to the risk of epidemic transmission, which are marked with ↑ or ↓.

	Changchun	Shanghai
↓ Size (Km^2)	24,744	6,340
↑ Population (K)	9,087	24,871
↑ Density (K/Km^2)	0.36	3.92
– Start Date	04 March 2022	02 March 2022
– End Date	02 April 2022	31 March 2022

where ΔS denotes the difference in the number of people who are susceptible on t and $t+1$. Besides, the definition of the economic loss is:

$$r_e = B\big(\mathbf{MI}, b, \mathbb{1}[Flag_m = 1]\big), \tag{11}$$

where b is the rest of the budget, $\mathbf{MI} = [a_L^{\mathcal{MC}}, a_L^{\mathcal{TMR}}, a_L^{\mathcal{NS}}]^T$ is the vector of the intervention decisions (which are with different costs). $\mathbb{1}[x = y]$ is equal to 1 if x equals to y, otherwise it is equal to 0. Therefore, $\mathbb{1}[Flag_m = 1]$ is an indicator function to determine whether a multi-mode intervention setting is used. Here, $B(\cdot)$ is the function to calculate the rest of the budget for deploying epidemic interventions.

Transition Function: The epidemic transmission dynamics can be described by the ordinary differential equation in Equation (8). Besides, random noise is added to simulate the uncertainty which accompanies the epidemic development.

During the training phase, the widely used deep reinforcement learning, PPO [39], is employed to search for the optimal intervention policy which can effectively alleviate the influence caused by the pandemic outbreaks. The detailed learning process is described in Algorithm 1. In general, the advantages of our method, HRL4EC, are two-fold:

- Firstly, we simultaneously consider the external and internal impact of the epidemiological model and refer to our model as the MID-SEIR model, which is more effective in describing the epidemic dynamics than the traditional ones.
- Secondly, we decompose the complex multi-mode epidemic-intervention decision problem and employ hierarchical reinforcement learning to search for the optimal policy, curbing the epidemic outbreaks with less economic cost.

5. Experiments

To empirically evaluate our approach, extensive simulation experiments are carried out and compared with real-world data collected on COVID-19 local outbreaks all over the world. Specifically, four research questions are investigated as follows:

- **RQ1**: Is the proposed model effective, i.e., whether the learned intervention strategy can reduce the number of cumulative infected cases?
- **RQ2**: What lessons can be drawn from the intervention strategies learned?
- **RQ3**: What about the generalizability of HRL4EC?
- **RQ4**: What influence do the various hyperparameter settings have on the proposed model?

5.1. Experimental setup

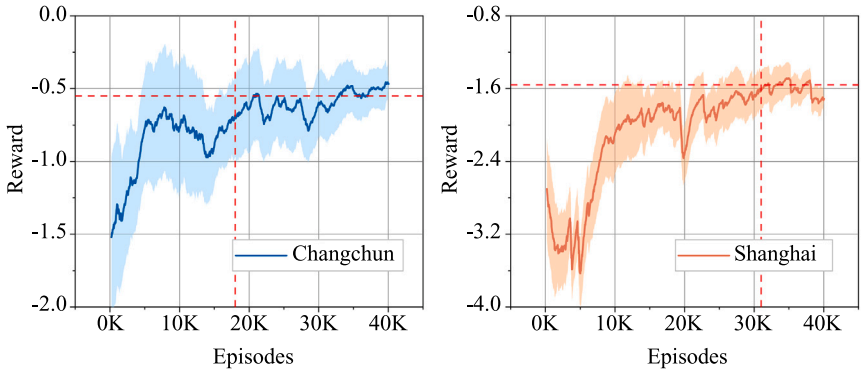
5.1.1. Dataset

To evaluate the effectiveness of our method, HRL4EC, in tackling epidemic control, extensive simulation experiments are conducted in two cities in China with real-world pandemic data. Table 2 presents information about these two cities, i.e., Changchun and Shanghai. They have different urban attributes: Shanghai is an international city with a high population density, leading to higher transmission risk, whereas Changchun has a low population density. Besides, both of them suffered from COVID-19 outbreaks in 2022. We collect the epidemic dataset about reported infected cases in these two cities for simulation experiments.

5.1.2. Baselines

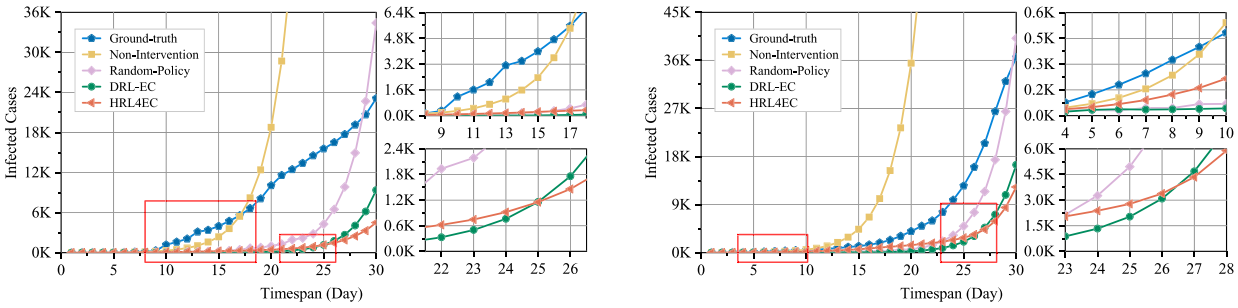
Several baseline methods are compared with our method, including manual methods and reinforcement learning based methods, to demonstrate the superiority of HRL4EC, which are detailed as follows:

- Ground-truth: It represents the baseline designed by policymakers in real-world, and its infected cases are collected from official site.
- Non-intervention: It does not lay out any interventions during epidemic outbreaks, which is used to measure the advance of other methods.
- Random-policy: During the training, it randomly selects the intervention to lay out for curbing the epidemic outbreaks, which is used for comparison.



(a) Changchun (b) Shanghai

Fig. 3. The training results in Changchun and Shanghai scenarios. The X-axis denotes the episode number and Y-axis indicates the per-step average reward during training.



(a) Cumulative Infected Cases in Changchun (b) Cumulative Infected Cases in Shanghai

Fig. 4. The cumulative infected cases (averaged over ten runs) comparisons with different intervention strategies. The Y-axis indicates the number of cumulative infected cases, and the X-axis denotes the duration of pandemic outbreaks. The left and right panels present the results in Changchun (04 March to 02 April 2022) and Shanghai (02 March to 31 March 2022), respectively. Local areas full of details are zoom in for clarity.

- DRL-EC [42]: It proposes a deep reinforcement learning framework to learn the intervention strategy to prevent COVID-19 from spreading further. Only a single intervention, i.e., lockdown-city, is considered for the epidemic control problem.

*5.2. Results and analysis (**RQ1**)*

We conduct the experiments in two cities, i.e. Changchun and Shanghai, and obtain the training results. Considering the randomness in the epidemic transmission simulation, the results are averaged over ten runs. Fig. 3 shows the obtained reward during model training, which indicates that our method can converge on both cities and demonstrates that our method is effectiveness.

To demonstrate the superiority of our method, we present the infected cases to make a comparison of our method and other baselines quantitatively. As shown in Fig. 4, it presents the changes in the cumulative infected cases under different intervention strategies and zooms in on the local areas to make a detailed observation. In Fig. 4a, it is evident that: 1) Random-Policy with multi-mode interventions alleviates the early stage of a pandemic outbreak; however, conversely, in the later stage, the infected cases proliferate rapidly compared to the Ground-truth, which is because of the intervention strategies that can not maximize the utility of the budget for epidemic control. 2) DRL-EC reduces the infected cases sharply by laying out a single intervention, i.e., mobility constraint. 3) HRL4EC performs better than DRL-EC and Random-Policy in the long term, i.e., its cumulative infected cases are lower than DRL-EC. 4) It is worth noting that the difference in cumulative infected cases between these three methods increases significantly over time.

Without loss of generality, the results in Shanghai have the same trend in Fig. 4b. In general, HRL4EC slows the increase of cumulative infected cases in all scenarios and outperforms other methods in the long term, further demonstrating the effectiveness of HRL4EC. The possible reasons behind the above phenomenon are as follows: 1) A multi-mode, fine-grained infectious disease intervention strategy is more conducive to the regular control of the pandemic. 2) HRL4EC has the ability to assist policymakers in seek to a better intervention decision policy, especially when the policymakers confront complex situations and countless choices for decision-making.

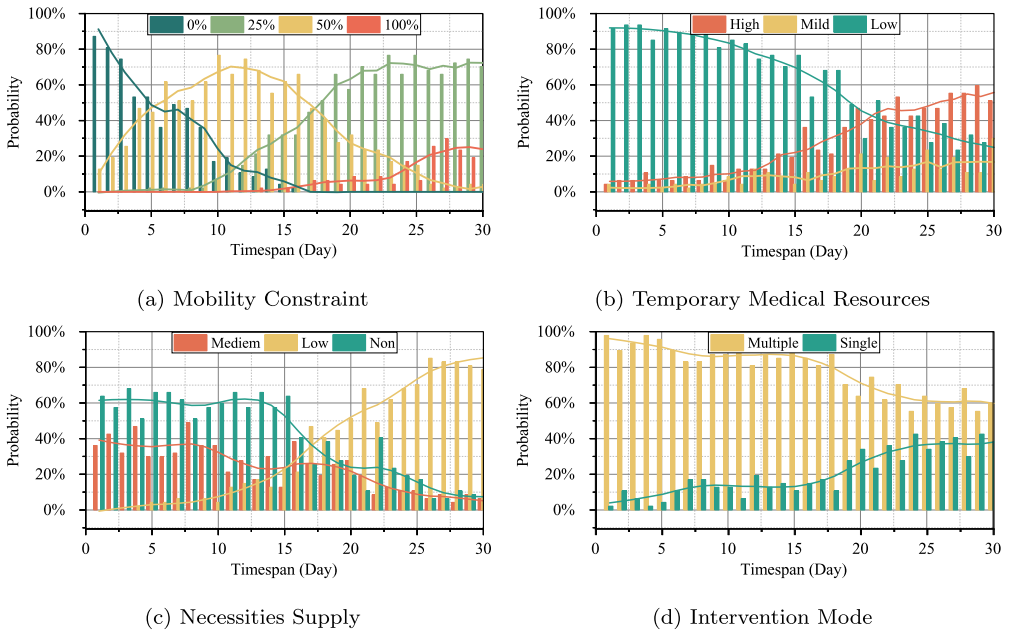


Fig. 5. Visualization of learned intervention strategies in Changchun, including (a) mobility constraint, (b) temporary medical resources, (c) necessities supply, and (d) intervention mode. The Y-axis denotes the probability of laying out the intervention, and the X-axis denotes the duration of the epidemic outbreak.

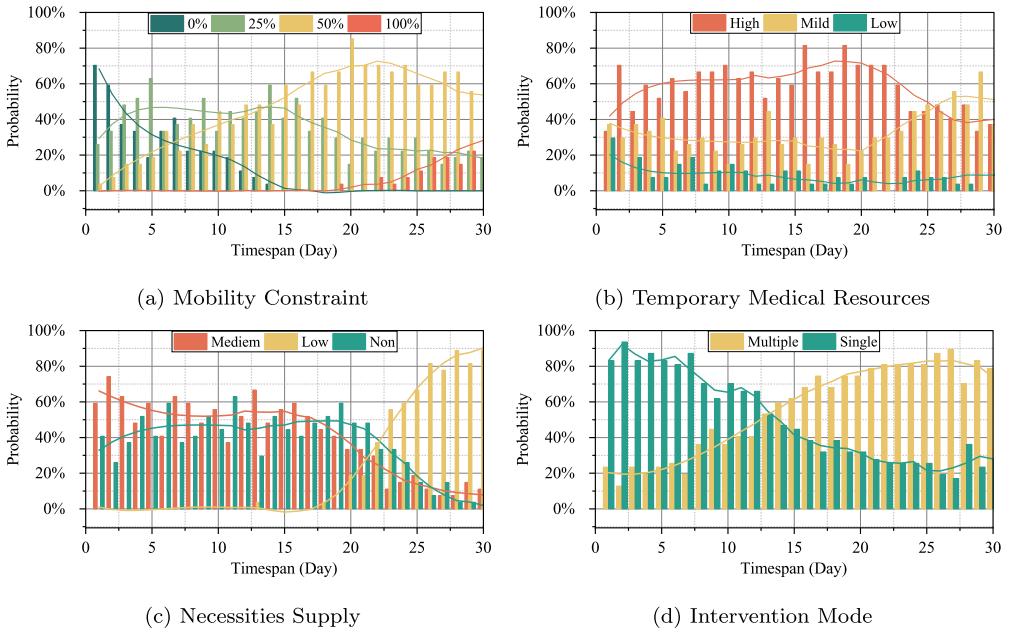


Fig. 6. Visualization of learned intervention strategies in Shanghai, including (a) mobility constraint, (b) temporary medical resources, (c) necessities supply, and (d) intervention mode. The Y-axis denotes the probability of laying out the intervention, and the X-axis denotes the duration of the epidemic outbreak.

*5.3. Findings of the experiments (**RQ2**)*

The visualization of the intervention strategies learned by HRL4EC is shown in Fig. 5 and Fig. 6, demonstrating how the frequency of laying out interventions changes with the duration of the infectious disease outbreak. Besides, we conduct an in-depth analysis of them and come up with the following findings.

For mobility constraint, HRL4EC prefers a moderate intensity of mobility constraint, i.e., 25% and 50% intensity, probably due to the consideration that strict mobility constraint can be seriously detrimental to socio-economics. In other words, an appropriate mobility constraint can effectively control the development of infectious diseases while maintaining socio-economic stability. Com-



Fig. 7. Phase visualization of the epidemic disease intervention strategy in Changchun and Shanghai. For clarity, we temporally aggregate every 10 steps of the daily intervention strategies into one phase and take the $\log(f(a))$ value to show. The inner circle indicates the type of interventions, and the outer circle corresponds to the intensity of each intervention. Comparing the top and bottom panels shows similar trends in both Changchun and Shanghai.

paring the intervention strategies in Changchun (Fig. 5a) and Shanghai (Fig. 6a), it is obvious that as population density increases, the demand for high-intensity mobility constraints increases.

For temporary medical resources, they are conducive to enhancing the recovery rate of the infectious and positive-testing rate, respectively, which have a direct impact on epidemic control by reducing the chance of virus transmission. Fig. 5b and Fig. 6b present the intervention strategies of temporary medical resources laid out within different cities, i.e., Changchun and Shanghai, indicating that its demand significantly increases (the orange-colored bar) with the outbreak duration and population density.

For necessities supply, it has an indirect impact on epidemic control. Specifically, the lack of necessities supply will increase citizens' anxiety and affect their level of cooperation with the epidemic, while increasing the necessities supply will with more additional overhead. As shown in Fig. 5c and Fig. 6c, it is preferable to provide low necessities supply for long-term epidemic control. At the early stage, either the provision of supplies, or budget savings to deploy other interventions (i.e., non-necessities supply) can be helpful for epidemic control. If policymakers attempt to provide supplies, the execution probability will increase with the number of people, as the orange-colored bar indicates in Fig. 5c and Fig. 6c.

For intervention mode, two modes are available in HRL4EC, namely single-mode, and multi-mode. As Fig. 5d and Fig. 6d show, multi-mode intervention is preferred during epidemic control. It is worth noting that there are some exceptional cases where single-mode (i.e., strict mobility constraint) can also be chosen: 1) cities with low population densities, where a single, strict mobility constraint can be used for more prolonged outbreaks of infectious disease; 2) cities with high population densities can be controlled early with single-mode intervention to avoid causing large-scale transmission.

In summary, the above findings, related to different interventions, can provide some heuristic suggestions to help policymakers develop sound intervention strategies, i.e., **HOW** and **WHEN** to lay out **WHICH** epidemic interventions, to curb the further spread of the pandemic.

*5.4. Generalizability analysis (**RQ3**)*

To validate the generalizability of HRL4EC, we conduct experiments in Shanghai and Changchun. Specifically, Shanghai is a crowded cosmopolitan city with more than ten times the population density of Changchun, which leads to different pandemic situations. Previous experimental results in Section 5.2 have already shown that our method works well in both scenarios.

Besides, the total epidemic outbreak is divided into three phases and take the $\log(f(a))$ value to make a comparison, where $f(a)$ is the number of times each intervention was selected during epidemic control. Fig. 7 presents the average $\log(f(a))$ of all mentioned interventions, where the inner circle indicates the type of interventions, and the outer circle corresponds to the intensity of the interventions to be deployed. By comparing the top and bottom panels in Fig. 7, we find that the proportion of the 0% mobility

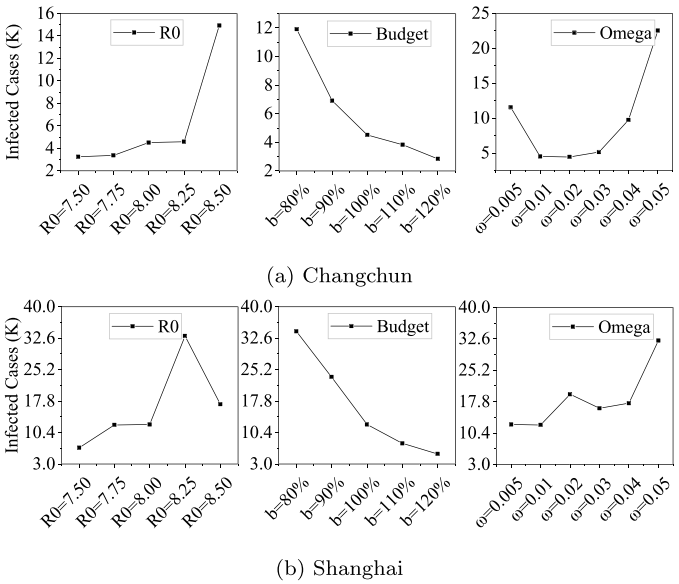


Fig. 8. Performance comparisons with different hyperparameter settings. The Y-axis shows the cumulative infected cases. The X-axis denotes the settings of hyperparameter.

constraint decreases gradually from Phase 1 to Phase 3. In contrast, the 100% mobility constraint increases both in Shanghai and Changchun. For other types of interventions, the learned strategies also have similar trends of changes from Phase 1 to Phase 3 in both cities, proving that HRL4EC can provide general intervention suggestions for policymakers to tackle pandemic outbreaks in different cities. It further demonstrates that HRL4EC has good generalizability to various situations.

*5.5. Parameter sensitivity (**RQ4**)*

We investigate the performances of HRL4EC with different hyperparameter settings, i.e., the basic reproductive number $R0$, initial budget b, and loss weight ω. The basic reproductive number $R0$ is used to indicate the intensity of infectious disease transmission, the initial budget b denotes the total cost used for interventions, and the loss weight balance ω balances the significance of economic loss and health loss. Single-parameter sensitivity analysis is employed to study uncertainty in the output of our proposed model, by varying one parameter while fixing the others each time. As shown in Fig. 8, the range of $R0$ is set in $[7.5, 8.5]$, and the model performs effectively. In the main experiments, $R0 = 8.0$ is used to simulate epidemic transmission. For initial budget b, we set $80\%, 90\%, 100\%, 110\%, 120\%$ of the budget in the main experiments. It is obvious that the infected cases degradation from 80% to 100% is more obvious than from 100% to 120%, and $b = 100\%$ is chosen in experiments. As for ω, the smaller ω pays more attention to epidemic control but damages the social economy. Finally, $\omega = 0.01$ is adopted in experiments, which balances the economy while maintaining epidemic control.

6. Conclusion

In this work, we investigate the problem of multi-mode intervention decision-making for epidemic control. We propose a hierarchical reinforcement learning decision system for multi-mode epidemic control by transferring the complex intervention decision-making problem into a multi-level control problem, referred to as HRL4EC. Besides, a novel epidemiological model, called MID-SEIR, is devised to serve as the environment for HRL4EC, which has the ability to capture external interventions' impacts on epidemic dynamics. Extensive experiments with real-world and simulation epidemic data demonstrate the effectiveness of HRL4EC. Furthermore, the visualizations of the experimental results uncover the regulars in laying out intervention strategies, which provide insights for policymakers to cope with pandemic outbreaks in the future. How to deploy HRL4EC into multiple cities and collaborate among them is interesting yet not explored so far, which we will investigate in the future. Meanwhile, we will also improve HRL4EC to provide adaptive epidemic control for multiple epidemics by providing different granularity of interventions.

CRediT authorship contribution statement

Xinqi Du: Conceptualization, Methodology, Software, Writing – original draft. **Hechang Chen:** Conceptualization, Formal analysis, Writing – review & editing. **Bo Yang:** Conceptualization, Supervision. **Cheng Long:** Writing – review & editing. **Songwei Zhao:** Investigation, Software.

Declaration of competing interest

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

Data availability

Data will be made available on request.

References

[1] A. Trilla, G. Trilla, C. Daer, The 1918 “Spanish flu” in Spain, Clin. Infect. Dis. 47 (5) (2008) 668–673.
[2] H. Chen, B. Yang, J. Liu, Partially observable reinforcement learning for sustainable active surveillance, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2018, pp. 425–437.
[3] E. Nicastri, G. Kobinger, F. Vairo, C. Montaldo, L.E. Mboera, R. Ansunama, A. Zumla, G. Ippolito, Ebola virus disease: epidemiology, clinical features, management, and prevention, Infect. Dis. Clin. 33 (4) (2019) 953–976.
[4] World Health Organization, World health organization coronavirus (Covid-19) dashboard, 2021. (Accessed 15 March 2023).
[5] S. Iketani, L. Liu, Y. Guo, L. Liu, J.F.-W. Chan, Y. Huang, M. Wang, Y. Luo, J. Yu, H. Chu, et al., Antibody evasion properties of Sars-cov-2 omicron sublineages, Nature 604 (7906) (2022) 553–556.
[6] S. Ghamizi, R. Rwemalika, M. Cordy, L. Veiber, T.F. Bissyandé, M. Papadakis, J. Klein, Y. Le Traon, Data-driven simulation and optimization for Covid-19 exit strategies, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3434–3442.
[7] C. Zhan, L. Shao, X. Zhang, Z. Yin, Y. Gao, K.T. Chi, D. Yang, D. Wu, H. Zhang, Estimating unconfirmed Covid-19 infection cases and multiple waves of pandemic progression with consideration of testing capacity and non-pharmaceutical interventions: a dynamic spreading model, Inf. Sci. 607 (2022) 418–439.
[8] A. Chandak, D. Dey, B. Mukhoty, P. Kar, Epidemiologically and socio-economically optimal policies via Bayesian optimization, Trans. Indian Nat. Acad. Eng. 5 (2) (2020) 117–127.
[9] S.B. Shuvo, B.C. Molokwu, Z. Kobti, Simulating the impact of hospital capacity and social isolation to minimize the propagation of infectious diseases, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3451–3457.
[10] Y. Chen, N. Crespi, A.M. Ortiz, L. Shu, Reality mining: a prediction algorithm for disease dynamics based on mobile big data, Inf. Sci. 379 (2017) 82–93.
[11] L. Zou, L. Xia, L. Hou, X. Zhao, D. Yin, Data-efficient reinforcement learning for malaria control, arXiv preprint, arXiv:2105.01620, 2021.
[12] R. Awasthi, K.K. Guliani, S.A. Khan, A. Vashishtha, M.S. Gill, A. Bhatt, A. Nagori, A. Gupta, P. Kumaraguru, T. Sethi, Vacsim: learning effective strategies for Covid-19 vaccine distribution using reinforcement learning, Intell.-Based Med. 6 (2022) 100060.
[13] A. Beigi, A. Yousefpour, A. Yasami, J. Gómez-Aguilar, S. Bekiros, H. Jahanshahi, Application of reinforcement learning for effective vaccination strategies of coronavirus disease 2019 (Covid-19), Eur. Phys. J. Plus 136 (5) (2021) 1–22.
[14] M. Liao, H. Liu, X. Wang, X. Hu, Y. Huang, X. Liu, K. Brenan, J. Mecha, M. Nirmalan, J.R. Lu, A technical review of face mask wearing in preventing respiratory Covid-19 transmission, Curr. Opin. Colloid Interface Sci. 52 (2021) 101417.
[15] P. Yang, G. Yang, J. Qi, B. Sheng, Y. Yang, S. Zhang, G. Bi, X. Mao, The effect of multiple interventions to balance healthcare demand for controlling Covid-19 outbreaks: a modelling study, Sci. Rep. 11 (1) (2021) 1–13.
[16] C. Colas, B. Hejblum, S. Rouillon, R. Thiébaut, P.-Y. Oudeyer, C. Moulin-Frier, M. Prague, Epidemioptim: a toolbox for the optimization of control policies in epidemiological models, J. Artif. Intell. Res. 71 (2021) 479–519.
[17] C. Sun, Z. Zhai, The efficacy of social distance and ventilation effectiveness in preventing Covid-19 transmission, Sustain. Cities Soc. 62 (2020) 102390.
[18] H. Bastani, K. Drakopoulos, V. Gupta, I. Vlachogiannis, C. Hadjichristodoulou, P. Lagiou, G. Magiorkinis, D. Paraskevis, S. Tsiodras, Efficient and targeted Covid-19 border testing via reinforcement learning, Nature 599 (7883) (2021) 108–113.
[19] S. Chen, X. Qiu, X. Tan, Z. Fang, Y. Jin, A model-based hybrid soft actor-critic deep reinforcement learning algorithm for optimal ventilator settings, Inf. Sci. 611 (2022) 47–64.
[20] S. Song, Z. Zong, Y. Li, X. Liu, Y. Yu, Reinforced epidemic control: saving both lives and economy, arXiv preprint, arXiv:2008.01257, 2020.
[21] W.O. Kermack, A.G. McKendrick, A contribution to the mathematical theory of epidemics, Proc. R. Soc. Lond. Ser. A, Contain. Pap. Math. Phys. Character 115 (772) (1927) 700–721.
[22] M.Y. Li, J.S. Muldowney, Global stability for the seir model in epidemiology, Math. Biosci. 125 (2) (1995) 155–164.
[23] H.W. Hethcote, The mathematics of infectious diseases, SIAM Rev. 42 (4) (2000) 599–653.
[24] P.R. Odell, Book review: Bailey, N.T.J. 1975: The mathematical theory of infectious diseases and its application. London: Griffin: Thom, R. 1975: Structural stability and morphogenesis. Reading, Massachusetts: Benjamin, Prog. Hum. Geogr. 7 (3) (1983) 442–444.
[25] Z. Tang, K. Yan, L. Sun, W. Zhan, C. Liu, A microscopic pandemic simulator for pandemic prediction using scalable million-agent reinforcement learning, arXiv preprint, arXiv:2108.06589, 2021.
[26] J. Kaminsky, L.T. Keegan, C.J.E. Metcalf, J. Lessler, Perfect counterfactuals for epidemic simulations, Philos. Trans. R. Soc. B 374 (1776) (2019) 20180279.
[27] J. Ren, M. Liu, Y. Liu, J. Liu, Optimal resource allocation with spatiotemporal transmission discovery for effective disease control, Infect. Dis. Poverty 11 (1) (2022) 34.
[28] Q. Hao, F. Xu, L. Chen, P. Hui, Y. Li, Hierarchical reinforcement learning for scarce medical resource allocation with imperfect information, in: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 2955–2963.
[29] P.J. Libin, A. Moonens, T. Verstraeten, F. Perez-Sanjines, N. Hens, P. Lemey, A. Nowé, Deep reinforcement learning for large-scale epidemic control, in: Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part V, Springer, 2021, pp. 155–170.
[30] Q. Hao, L. Chen, F. Xu, Y. Li, Understanding the urban pandemic spreading of Covid-19 with real world mobility data, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3485–3492.
[31] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.
[32] T.P. Lillicrap, J.J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv preprint, arXiv:1509.02971, 2015.
[33] A. Lakhan, M.A. Mohammed, S. Kozlov, J.J. Rodrigues, Mobile-fog-cloud assisted deep reinforcement learning and blockchain-enable iomt system for healthcare workflows, Trans. Emerg. Telecommun. Technol. (2021) e4363.
[34] S. Tang, J. Wiens, Model selection for offline reinforcement learning: practical considerations for healthcare settings, in: Proceedings of the 6th Machine Learning for Healthcare Conference, vol. 149, PMLR, 2021, pp. 2–35.

[35] A.S. Vezhnevets, S. Osindero, T. Schaul, N. Heess, M. Jaderberg, D. Silver, K. Kavukcuoglu, FeUdal networks for hierarchical reinforcement learning, in: Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017, pp. 3540–3549.
[36] R. Xie, S. Zhang, R. Wang, F. Xia, L. Lin, Hierarchical reinforcement learning for integrated recommendation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 4521–4528.
[37] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al., Mastering the game of go without human knowledge, Nature 550 (7676) (2017) 354–359.
[38] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, 2018.
[39] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint, arXiv:1707.06347, 2017.
[40] J. Mackolil, J. Mackolil, Addressing psychosocial problems associated with the Covid-19 lockdown, Asian J. Psychiatry 51 (2020) 102156.
[41] K. Chen, C.S. Pun, H.Y. Wong, Efficient social distancing during the Covid-19 pandemic: integrating economic and public health considerations, Eur. J. Oper. Res. 304 (1) (2023) 84–98.
[42] X. Du, T. Liu, S. Zhao, J. Song, H. Chen, District-coupled epidemic control via deep reinforcement learning, in: International Conference on Knowledge Science, Engineering and Management, Springer, 2022, pp. 417–428.