Exploring Optimal Control of Epidemic Spread Using Reinforcement Learning

Abu Quwsar Ohi[1,], **M. F. Mridha**[1,*], **Muhammad Mostafa Monowar**[2], **and Md. Abdul Hamid**[2]

[1]Department of Computer Science & Engineering, Bangladesh University of Business & Technology, Dhaka, Bangladesh
[2]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah-21589, Kingdom of Saudi Arabia
[*]firoz@bubt.edu.bd

ABSTRACT

Pandemic defines the global outbreak of a disease that is caused due to some disease containing a high transmission rate. The impact of a pandemic situation can be lessened by restricting the movement of the mass. However, one of its concomitant circumstances is an economic crisis. In this article, we demonstrate what actions an agent (trained using reinforcement learning) may take in different possible scenarios of a pandemic depending on the spread of disease and economic factors. To train the agent, we design a virtual pandemic scenario closely related to the present COVID-19 crisis. Then, we apply reinforcement learning, a branch of artificial intelligence, that deals with how an individual (human/machine) should interact on an environment (real/virtual) to achieve the cherished goal. Finally, we demonstrate what optimal actions the agent perform to reduce the spread of disease while considering the economic factors. In our experiment, we let the agent find an optimal solution without providing any prior knowledge. After training, we observed that the agent places a long length lockdown to reduce the first surge of a disease. Furthermore, the agent places a combination of cyclic lockdowns and short length lockdowns to halt the resurgence of the disease. Analyzing the agent's performed actions, we discover that the agent decides movement restrictions not only based on the number of the infectious population but also considering the reproduction rate of the disease. The estimation and policy of the agent may improve the human-strategy of placing lockdown so that an economic crisis may be avoided while mitigating an infectious disease.

Introduction

Through a pandemic situation, the foremost intention is to produce a vaccine that provides immunity over a particular infectious disease. However, the means of exploring an effective vaccine may take years to develop depending on the disease and some certain criteria. While investigating the vaccine, the loss of a pandemic is to be controlled via proper clinical support, and by reducing the expanse of the disease. Nevertheless, assuring proper clinical care is not possible in a pandemic situation due to a large number of infections over the available limited clinical support. Therefore, lessening the expanse of a disease is the first and foremost effort to overcome the devastation of a pandemic disaster.

Pandemics are often caused by diseases that transmit through person-to-person close contact[1]. At present, pandemics are caused by flu such as Swine flu[2], and Coronavirus[3,4]. Different intervention means are proven to reduce the devastation of a pandemic outbreak[5]. However, these interventions often cause an economic breakdown, and it is not possible to reduce the impact of a pandemic without it[6]. Therefore, a pandemic situation raises challenges to balance the viral spread and a steady economy.

Due to the current COVID-19 pandemic, researchers have been investigating various strategies to reduce the desolation of the pandemic, while striving economical balance. Through several research endeavors, various lockdown strategies have been proposed, such as age-based lockdown[7], n-work-m-lockdown[8], and so on. However, age-based lockdown should not apply for a disease that is critical for all ages. Also, repeated n-work-m-lockdown (n days without lockdown followed by m days of lockdown) strategies may not ameliorate critical pandemic situations. The current challenge of a pandemic situation raises cases such as, (a) is placing a long time lockdown the only way to mitigate a pandemic?, (b) should we place lockdowns while the pandemic situation does not ameliorate?, (c) how should the resurgence of the pandemic be handled?, (d) while mitigating a pandemic, how we could also balance the economical circumstances? In our research endeavor, we attempt to resolve these concerns by combining reinforcement learning and virtual environment based epidemic analyses.

In aspects of mathematics and computer science, the challenge of maximizing a constraint (the economical balance) while minimizing some other factor (reducing the spread of disease) is referred to as an optimization problem. The knowledge of making the best decision of an optimization problem is termed as a policy. The best policy may be found using Reinforcement Learning (RL). In RL, a machine is defined as an actor or agent. The actor performs some actions in an environment and earns a reward for every action. The goal of the actor is to find such a policy that will cause it to acquire the maximum possible reward. Through a proper setup, an RL agent can adapt actions like animals, even like the intelligent ones[9].

Previously, the field of RL was enclosed with implementing dynamic programmings with tabular functions. Q-Learning[10], Double-Q Learning[11] were the fundamental methods of RL. However, the vast improvement of Deep Learning (DL) has enabled it in the usage of RL strategies[12]. In recent times, instead of using tabular functions, Deep Neural Networks (DNNs) are implemented in RL[12]. Deep Reinforcement Learning (DRL) has improved the previous fundamental methods to be implemented using Deep Q-Learning, Double Deep Q-Learning (DDQN), and so on. Also, the current improvement of RL has attracted researchers and therefore, various new implementations are currently available.

The present state of DRL has proven its strength in various platforms such as playing Atari like human[13], chatting like human[14], playing hide and seek[15], and so on. Furthermore, recent improvements in DRL have resulted in beating humans in poker[16], go[17], and even in DOTA-2[18]. DRL is astonishing humans by generating new optimal ideas that were never thought of.

Being inspired by the recent improvements of DRL, in this paper, we search for some optimal ideas on pandemic mitigation. To carry out the exploration, we implement a virtual environment that simulates a pandemic crisis. We consider the disease that causes the pandemic to be transmitted in close contact. A short term memory based DDQN is used as an RL agent. The goal of the agent is to formulate an optimal strategy so that a pandemic crisis may be mitigated while maintaining economical balance. The contribution of our research endeavor includes:

1. We implement a virtual environment that simulates a pandemic situation and also considers economic circumstances.
2. We illustrate the consequences of placing no lockdown, maintaining social distancing, and placing lockdown. The consequences are derived based on the death of population and economic situations.
3. We investigate optimal strategies to reduce the spread of disease using reinforcement learning. Furthermore, we perform extensive analysis and present the reasoning behind the action.

The rest of the paper is organised as follows: In "Methods", the mechanism of the virtual environment is disclosed and the neural network architecture of the agent is defined. In "Results", we explore various control sequences to reduce the spread of the disease and consume our effort to find and analyze the optimal control sequence generated by the agent. Finally, "Discussion" concludes the paper.

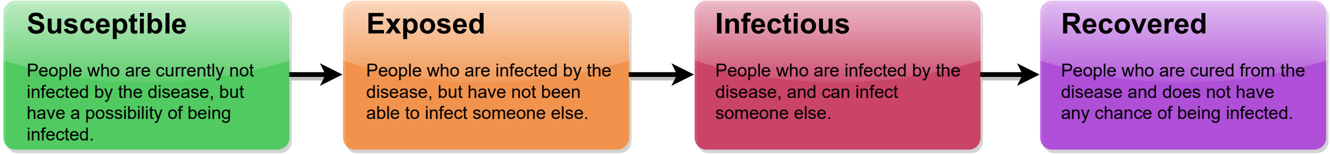
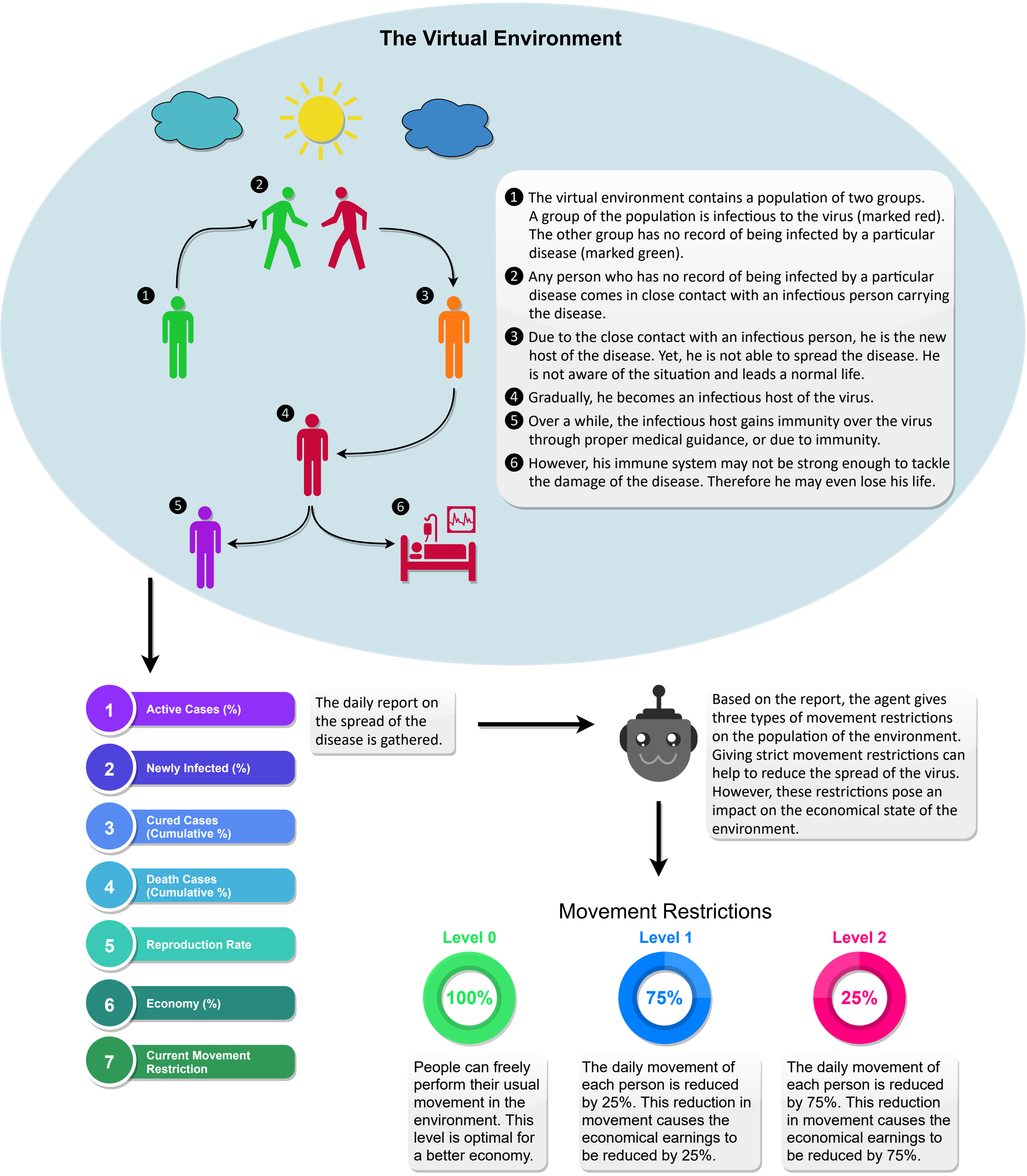

Susceptible
People who are currently not infected by the disease, but have a possibility of being infected.
Exposed
People who are infected by the disease, but have not been able to infect someone else.
Infectious
People who are infected by the disease, and can infect someone else.
Recovered
People who are cured from the disease and does not have any chance of being infected.

Figure 1. The diagram illustrates the different stages of an SEIR compartmental model. Although it can be observed that the infectious population further approaches to recovered state, a portion of the infectious population may not survive the disease and lose their lives.

Methods

To study epidemiology, various compartmental models are being implemented. Compartmental models define a simple mathematical foundation that projects the spread of infectious disease. Currently, various compartmental models are available[19]. Furthermore, different mathematical models are being presented to illustrate the relationship of population heterogeneity and the present crisis of pandemic[20]. These compartmental models are mostly generated using ordinary differential equations (ODE)[21]. Although ODE and other mathematical methods are sufficient in modeling an infectious disease, they lack the randomness of being infected, cured, and death. Also, mathematical models do not include any super-spreaders[22].

Therefore, we implement a virtual environment that solves these issues. The virtual environment is used to generate states and results based on some particular actions. The virtual environment is designed based on the SEIR (Susceptible-Exposed-Infectious-Recovered) compartmental model. Fig. 1 depicts the different stages of SEIR compartmental model. Due to the randomness in various transitions, implementing virtual compartmental models make the problem more challenging. The virtual environment is designed in a 2D grid where the population can randomly move. In each day, the population performs a fixed number of random moves. In Fig. 2 an info-graphic representation of the environment and the training process are illustrated.



1. The virtual environment contains a population of two groups. A group of the population is infectious to the virus (marked red). The other group has no record of being infected by a particular disease (marked green).
2. Any person who has no record of being infected by a particular disease comes in close contact with an infectious person carrying the disease.
3. Due to the close contact with an infectious person, he is the new host of the disease. Yet, he is not able to spread the disease. He is not aware of the situation and leads a normal life.
4. Gradually, he becomes an infectious host of the virus.
5. Over a while, the infectious host gains immunity over the virus through proper medical guidance, or due to immunity.
6. However, his immune system may not be strong enough to tackle the damage of the disease. Therefore he may even lose his life.

The daily report on the spread of the disease is gathered.

Based on the report, the agent gives three types of movement restrictions on the population of the environment. Giving strict movement restrictions can help to reduce the spread of the virus. However, these restrictions pose an impact on the economical state of the environment.

People can freely perform their usual movement in the environment. This level is optimal for a better economy.

The daily movement of each person is reduced by 25%. This reduction in movement causes the economical earnings to be reduced by 25%.

The daily movement of each person is reduced by 75%. This reduction in movement causes the economical earnings to be reduced by 75%.

Figure 2. The infographic illustrates the overall dynamics of the virtual environment, environment features, and the set of possible actions of the agent. Notably, level-1 movement restriction is similar to maintaining social-distancing, and level-2 movement restriction is similar to placing a nationwide lockdown.

Transmission Stages

In the environment, susceptible individuals are infected if they are in close contact with an infectious person. Initially, the infected population is in the exposed stage. After 1-3 days, individuals of the exposed stage is further transmitted to the infectious stage. In this stage, individuals can transmit the disease. The infectious individuals are either recovered after 21-27 days, or they may even lose their lives. The environment is configured so that around 80% of the infected population may survive.

Movement Restrictions

In the virtual environment, the spread of the disease can be mitigated by reducing the movements of the population. There are three levels of movement restrictions in the environmental setup, namely level 0, level 1, and level 2. In level 0, no movement restrictions are enforced. In this state, the population makes the maximum movements. In level 1, the movement of the individuals is restricted by 25%. In general, maintaining social distancing and avoiding unnecessary means is considered to be equivalent to level 1 restriction[23]. In level 2, the movement is reduced by 75% that is similar to a lockdown state[24]. These movement restrictions are provided by the DRL agent. However, although movement restrictions result in reducing the spread of disease, it causes an economic collapse.

Economical Setup

In the virtual environment, each individual contributes to the economy through movement. Therefore, if movement restriction is placed, it has an impact on the economy as well. Each individual contributes a value of 0.8-0.1 by moving. People who did not survive can not make any further contributions to the economy. Therefore, the increasing number of death count has also a negative impact on the economy. Also, the infectious population can not contribute to the economy. Therefore, a high number of active cases has also a negative influence on the economy.

State Genaration

In RL, a state is an observation that passes estimable information to the agent. By analyzing the information, an agent makes an optimal move based on its policy. States can be both finite or infinite. In the virtual environment setup, relevant information about the spread of the disease is passed through a state. Seven parameters are passed as a state of the environment. Fig. 2 illustrates the state parameters as infographic. Active cases represent the number of the population who are in the infectious stage. Newly infected refers to the number of the population who have shifted into the infectious stage on a particular day. Cured cases and death cases illustrate the number of people who have been cured and died from the start of the pandemic, respectively. The reproduction rate represents the average number of people who are being infected by the current infectious population. The economy illustrates the daily economical contribution of the population. Along with the states, the current movement restriction is also presented as a state parameter.

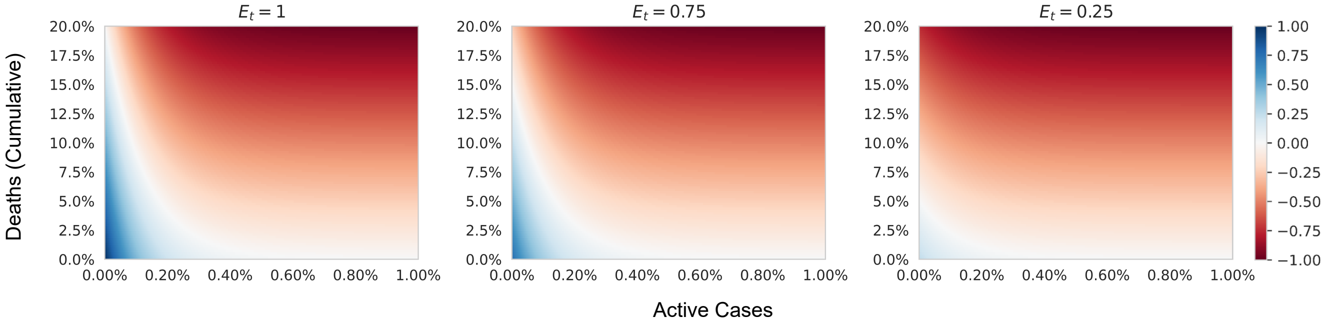


Figure 3. A heatmap representation of the reward function. The horizontal axis represents the percentage of active cases. The vertical axis represents the cumulative death percentage. From left to right, the three heatmaps illustrate the reward distribution in level-0 movement restriction, level-1 movement restriction, and level-2 movement restriction, respectively. In the three restriction levels 0, 1, and 2, the value of E_t is expected to be approximately 1, 0.75, and 0.25, respectively.

Reward Function

In DRL, an action is encouraged and discouraged by a reward function. A reward function encourages an agent to be in a particular state/situation by giving it a high reward for the situation. On the contrary, a particular action or situation is discouraged by giving the agent a low reward. An agent tries to generate such a policy/knowledge so that by following the policy, the agent may avoid the discouraging situation. Through designing a proper reward function, it is possible to generate such an agent that may be able to follow the human desired situation. For the current environment, the reward function is designed as follows,

$$R(s_t) = E_t \times e^{-r \times A_t} - s \times D_t \tag{1}$$

$Where,$

$$\begin{aligned}
E_t &= \frac{CurrentEconomy}{TotalPopulation \times M_t} \\
D_t &= \frac{CumulativeDeath}{TotalPopulation} \\
A_t &= \frac{ActiveCases}{TotalPopulation} \times 100 \\
r &= 8 \\
s &= 5
\end{aligned}$$

The reward function contains three parameters from the environment: the current economy ratio, the current cumulative death ratio, and the current percentage of active cases. Due to the three types of movement restrictions, the economic ratio can be separated into three levels. Due to the direct relationship with movement restriction and economy, level 0, level 1, and level 2 result the value of E_t approximately be close to 1, 0.75, and 0.25, respectively. However, this can be altered due to high death count and randomness. By avoiding the D_t parameter, the correlation of the economical levels and active cases can be utilized. In Fig. 4 a similar situation is illustrated. By utilizing the graph, it can be observed that while the active cases are low, the reward prioritizes higher economical stages. The further increase in active cases lessens the reward of higher economical stages. By setting the value of $r = 8$, the reward of different economical stages is almost the same (the absolute difference is less than 0.001) after crossing 0.82% active cases. This boundary is thought of as a critical point after which, the economy does not matter. After this boundary, the goal becomes to lessen the surge of the disease. Furthermore, including the D_t in the reward function, the agent is also encouraged to reduce the death ratio. Fig. 3 illustrates the relation of reward function relating to the active case percentage and death ratio in three possible economic stages. The impact of the deaths in the reward function is tuned using the parameter s. And $s = 5$ is set to prioritize the negative impact of the deaths.

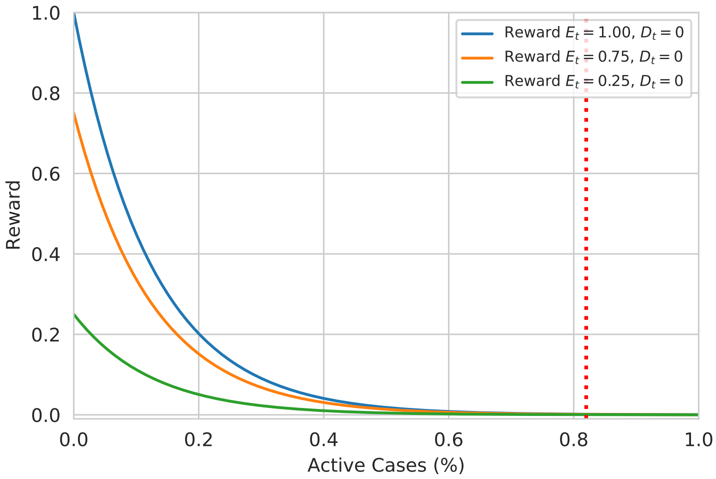


Figure 4. The graph illustrates the decay of reward value concerning the increase in the percentage of active cases (neglecting the cumulative death cases D_t = 0). The value of E_t being 1, 0.75, and 0.25 approximately represents the level-0, level-1, and level-2 movement restrictions. After crossing 0.8% of active cases, the reward of all the different restrictions falls to zero.

Both r and s are the tuning parameters of the reward function. Increasing the value of r causes the reward threshold (described in Fig. 4) to be reduced. Whereas, the value of s defines the significance of death. A higher value of s influences the agent to heavily reduce the death ratio ignoring the economic balance.

The Agent Network

The decision process of the DRL can be considered to be a Markov Decision Process (MDP). In MDP, the environment contains a finite set of states S, with a finite set of actions A. If $s, s^{'} \in S$, and $\alpha \in A$, then the state transition can be represented as,

$$\tau(s^{'}|s,\alpha) \tag{2}$$

The equation states the transition probability of choosing an action α, given an environment state s, and achieving a new state s'. The DRL agent acquires a policy π through bootstrapping. Through this policy, the agent performs an optimal action α_i for

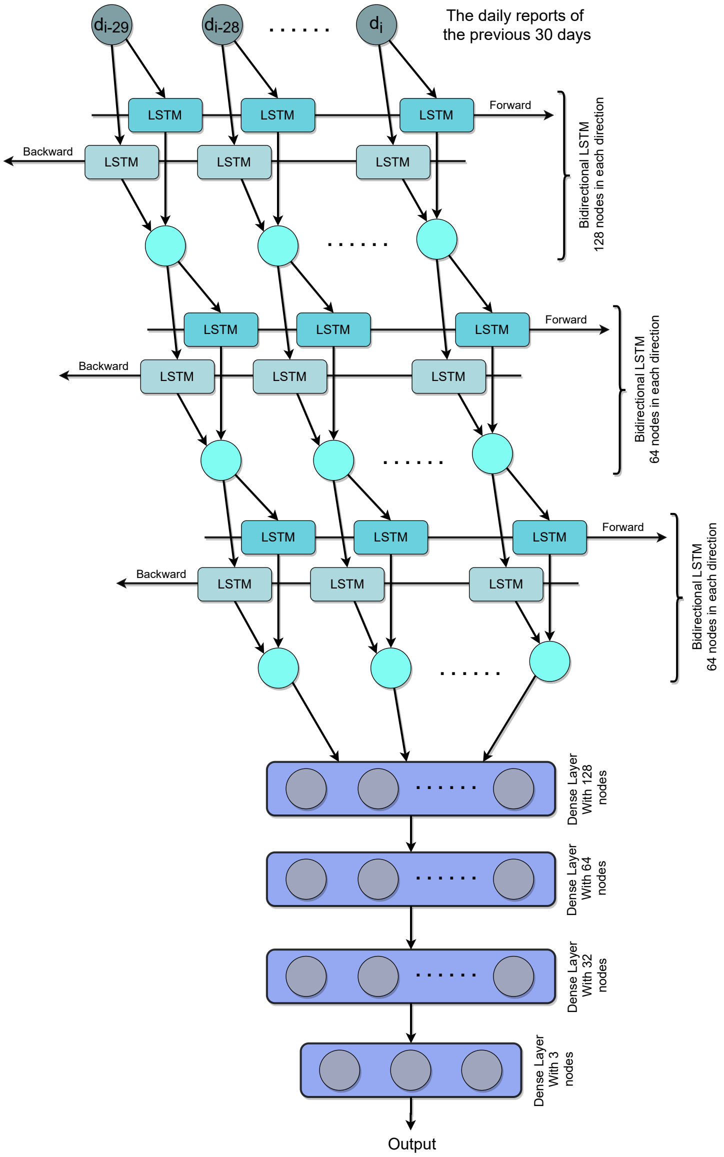

di-29
di-28
di
The daily reports of the previous 30 days
LSTM
Forward
Backward
Bidirectional LSTM 128 nodes in each direction
Bidirectional LSTM 64 nodes in each direction
Bidirectional LSTM 64 nodes in each direction
Dense Layer With 128 nodes
Dense Layer With 64 nodes
Dense Layer With 32 nodes
Dense Layer With 3 nodes
Output

Figure 5. The memory-based agent neural network architecture of the agent. The agent uses three bidirectional LSTM layers with 128, 64, and 64 nodes, respectively. It is further followed by four dense layers of 128, 64, 32, and 3 nodes, respectively.

a given state s, represented as, $\pi(\alpha_i|s)$. The optimal action is chosen based on the state-value function $V^{\pi}(s)$ that defines the chained reward value. The reward value is a chain multiplication of discount value γ and state rewards R. This can be presented as,

$$V^{\pi}(s) = \mathbb{E}_{\pi} \sum_{k=0}^{n} \left[\gamma^{t+k} R_{t+k} | s_0 = s\right] \tag{3}$$

An optimal policy π^* finds the best possible state-value function, that can be defined as,

$$V^*(s) = max_{\pi} V^{\pi}(S) \quad \forall s \in S \tag{4}$$

As the transition of an MDP ($\tau(s^{'}|s,\alpha)$) is unknown, a state-action function $Q^{\pi}(s,a)$ is generated. The state action function mimics the value state-value function $V^{\pi}(s)$ and also tries to identify best action α. The state-action function greedily chooses the actions for which, it gains the maximum state-value.

The $Q^{\pi}(s,a)$ function is defined as the DRL agent. In the experiment, we study with memory-based DRL agents since the memory-based agent perceives further possibilities and takes optimal decisions and acquires better rewards[25]. Among different memory sizes, we found that the DRL agent makes better actions with a minimal memory of 30 days. The agent is implemented using three bidirectional Long Short Term Memory (LSTM). Bidirectional LSTM performs optimally when there exist both forward and backward relationships in a portion of data[26]. In the case of this epidemic data, using bidirectional LSTMs provides the following benefits: (a) Select an optimal action based on previous data, and (b) Estimate the influence of selecting a particular action. The agent uses three bidirectional LSTM layers, followed by four dense layers. In Fig. 5 the memory-based DRL agent architecture is depicted.

DDQN method is used to train the agent. The DDQN architecture uses an actual agent and a target agent. Traditionally, in DDQN, both agents contain the same network structure. Furthermore, the traditional DDQN training process is implemented to train the architectures[27]. The agent is trained over 7000 episodes and without any pre-knowledge and human interpolation. To explore the environment properly, random movements were made in the training episodes. The training is started with a random movement ratio of $\varepsilon = 1$, and it is continuously decayed as $\varepsilon = max(\varepsilon - \varepsilon/(6000), 0.1)$. To propagate the future rewards to any particular state, the discount value (γ) is set to be 0.9.

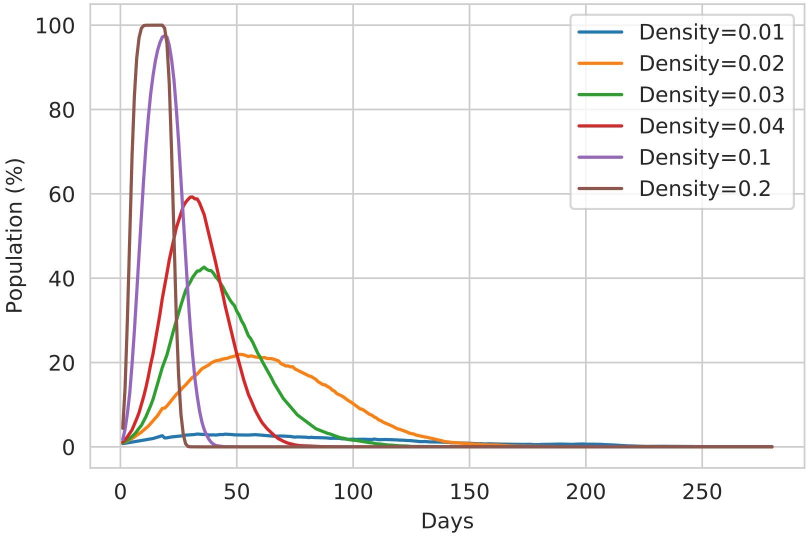


Figure 6. Simulation of active cases of the environment, based on different population density. Increasing the density of the population also increases the probability of contact between people. Therefore, the spread of disease also increases.

Results

The experiments are conducted in a virtual environment that is implemented on a quadratic time complexity based algorithm, presented in Algorithm 1. Therefore, we experiment with a limited number of 10,000 population and a default daily movement of 15 steps. Through our investigation, we found that the spread of the disease in the environment acts differently based on the density of the population. In Fig. 6, we illustrate distinguishable waves of active cases over different rates of population density. Due to the high density of the population, the probability of contact between two different person increases. Therefore, the rate of spread of a disease depends on the density of the population. On the contrary, in the environment, the reproduction rate of a disease is not dependent on the population density. In Table 1, the mean and median reproduction rate is reported, tested over different population densities.

The increase in density does not alter the reproduction rate of the environment. Furthermore, efforts have been made to reasoning the cause[28]. The mean and median of the reproduction rate of the virtual environment closely simulates the estimated reproduction rate evaluated in Wuhan[29]. Nevertheless, the population density of 0.01 does not spread the disease properly. On the contrary, the population density of 0.04, 0.1, 0.2 excessively spreads the disease. Therefore we conduct our experiment on the population density of 0.02 and 0.03. The overall implementation is conducted using Python[30], Keras[31], and TensorFlow[32]. Matplotlib[33] is used for graphical representations.

Evaluation of Different Control Sequences

Fig. 8 presents a datasheet of the virtual environment simulation and Fig. 7 represents the initial positioning of the infectious population over the environment. The datasheet is separated into four individual graphs. In the current simulation, no lockdown is placed (level-0 restriction). The graph indicates a raise in active cases by simultaneously infecting 20% of the population.

Table 1. A comparison of the reproduction rate in different population density. The comparison is represented in mean±std format of the data collected in ten individual runs.

Area	Population	Density	R_0 mean	R_0 median
1000x1000	10000	0.01	2.87±0.19	2.84±0.11
708x708	10000	0.02	3.2±0.30	2.84±0.02
577x577	10000	0.03	3.4±0.23	2.94±0.08
500x500	10000	0.04	3.4±0.18	2.76±0.11
316x316	10000	0.1	3.3±0.40	2.73±0.05
224x224	10000	0.2	3.4±0.12	2.9±0.05

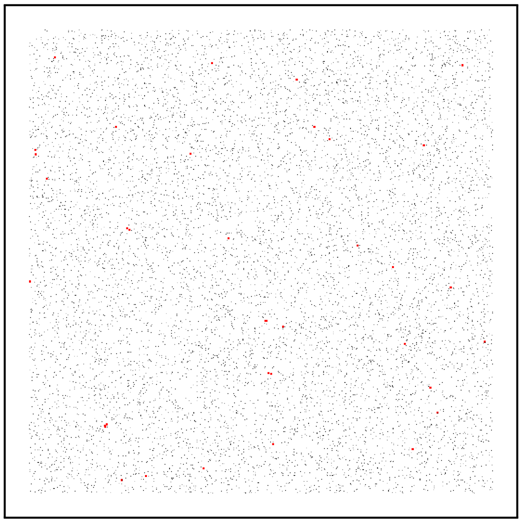


Figure 7. This graph represents the initial state of the pandemic. The black dots denote the position of the susceptible population. The red dots denote the position of the infectious population. The virtual environment contains 10,000 population, in which, 70 (0.7%) of them are infectious. This is a challenging scenario because the infectious is heavily spread all over the regions. The density of the environment is set to be 0.02.

Without placing any lockdown, the disease affects more than 80%, among which, around 20% of the population loses their lives. Due to the huge decrease in the population, an impact is also measured in the economical state of the environment. As the non-survivals could not contribute to the economy, the economic ratio of the environment falls around 0.20 due to the loss of the population. Therefore, considering the economy, it can be determined that placing no lockdowns in a pandemic situation may not be a good solution. The reproduction rate of the disease is mostly in a close interval of 2 to 5. However, a surge in the reproduction rate is reported after passing 160 days of the pandemic, due to the superspreaders.

The effect of social-distancing (level-1 restriction) is presented in Fig. 9. By maintaining social-distancing, around 20% spread of the disease can be reduced, along with 10% fewer deaths. Also, the surge of active cases is reduced by around 10%. However, due to social-distancing, the economic ratio is decreased by around 0.2. The impact of lockdown (level-2 restriction) is presented in Fig. 10. From the illustration, it can be stated that placing lockdown heavily decreases the spread of disease. On the contrary, placing lockdown also causes the economy to collapse. The simulation also points out that the spread of disease can be fully halted by placing a 63 days lockdown. However, in the real world scenario, complete elimination of a disease through lockdown is near impossible.

Fig. 11 illustrates the restrictions that the agent placed in the virtual environment of population density 0.02. The initial state of the environment starts with a devastating pandemic situation, in which, the disease infects almost 1% of the population. Therefore, the agent places multiple 30-40 days of lockdown segments to reduce the spread of the disease. Then the agent removes the restrictions and stables the economy. However, multiple smaller peaks of active cases are reported in an approximately 100 days cycle. The agent reduces the spread of the disease by performing two types of actions. At first, the agent activates a cyclic lockdown to level the spread of the virus by keeping the economy steady as much as possible. Finally, the cyclic lockdown is followed by a 10-20 days long lockdown. By further analyzing the reproduction rates of the environment, it can be concluded that this combination optimally reduces the reproduction rate below 1. Reducing the reproduction rate

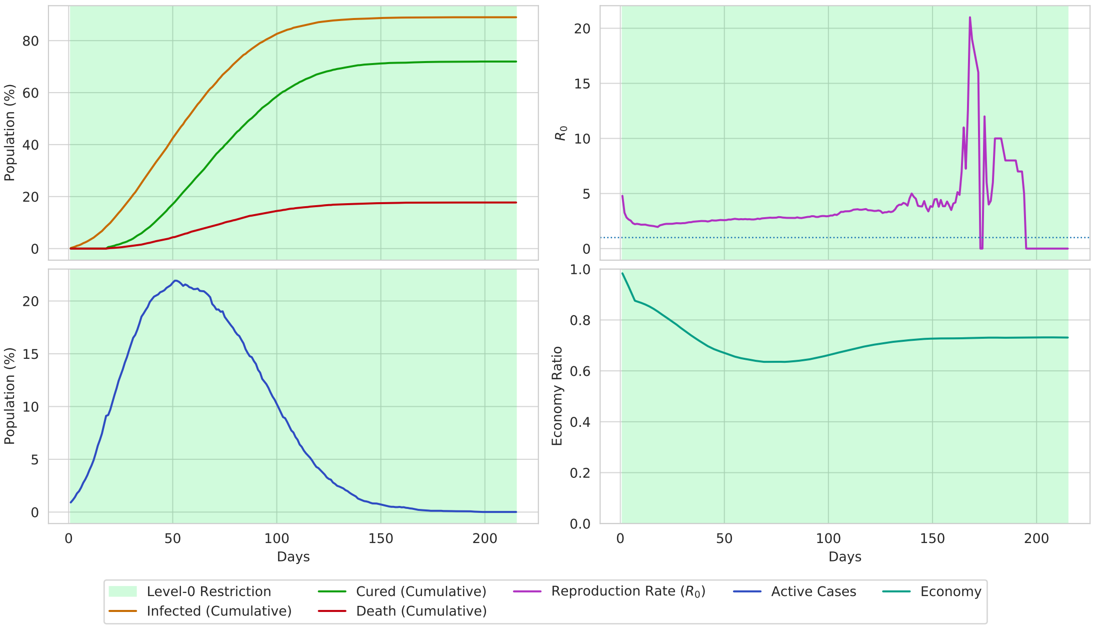


Figure 8. A simulation of the virtual environment (0.02 population density) by placing level-0 restriction. The upper-left portion illustrates the cumulative sum (in percentage) of infected, cured, and dead of the overall population. The upper-right portion illustrates the reproduction rate of the disease. The lower-left portion indicates the percentage of the active cases of the population. The lower-right portion determines the economical state through the spread of the disease. A massive surge of active cases is reported on reaching the 50th day of the pandemic. Around 20% of the population dies due to the disease if no lockdown is placed and no social-distancing is maintained.

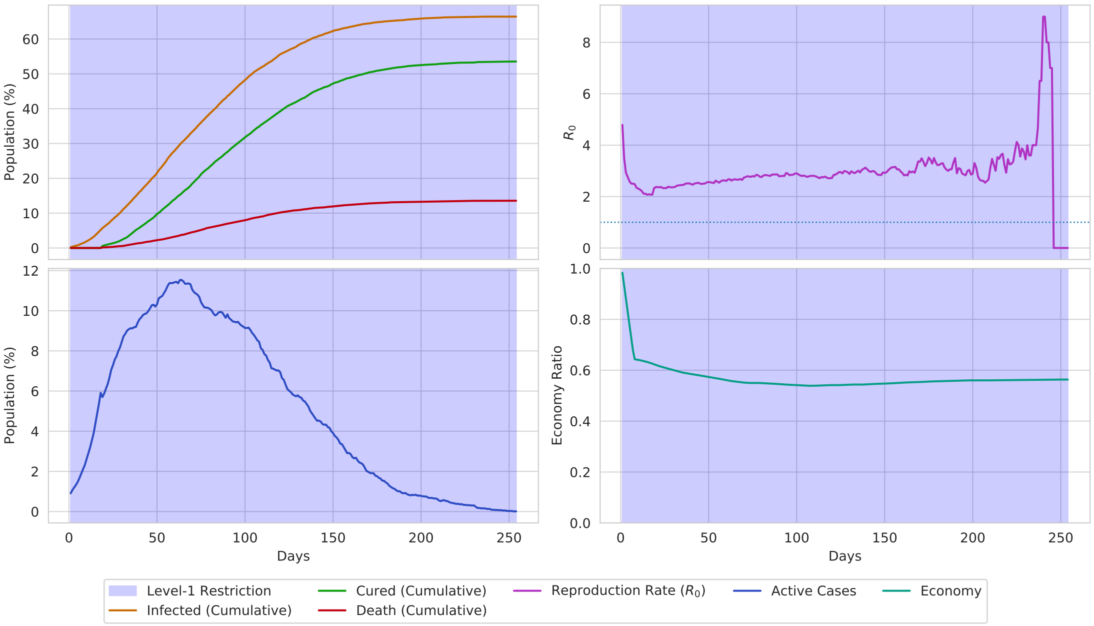


Figure 9. A simulation of the virtual environment (0.02 population density) only if social-distancing is maintained. Due to social-distancing, the spread of the disease is reduced. Hence, the total number of infections is reduced by 20%, along with a reduction in deaths by 10%. Although the economical ratio is reduced by 0.2, it is considerate, relating to the reduced spread of the virus.

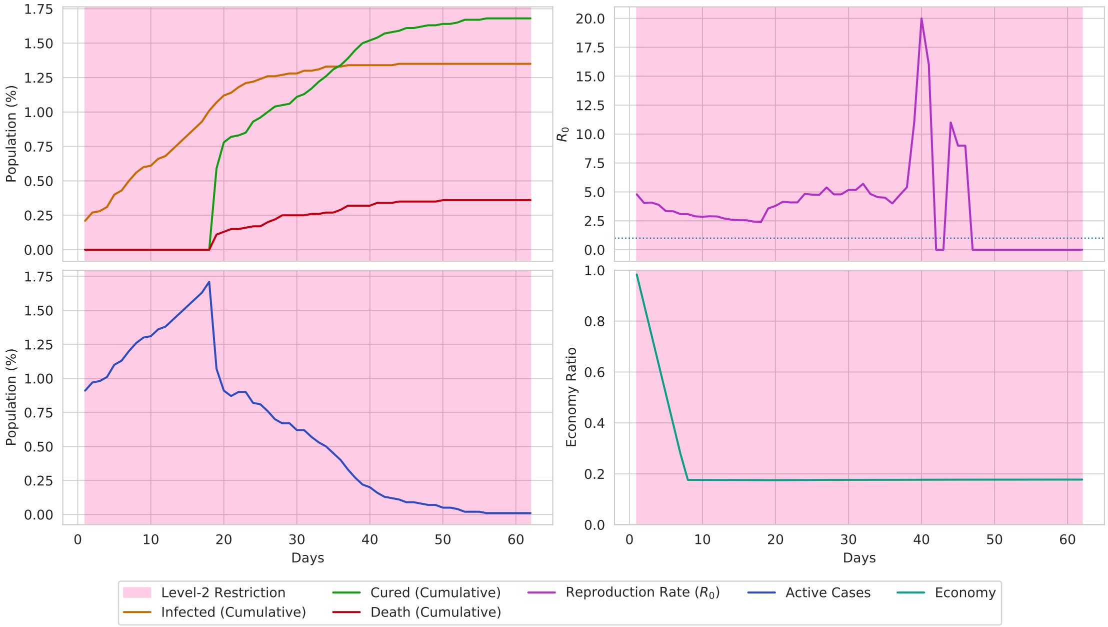


Figure 10. A simulation of the virtual environment if full lockdown is ordered. Due to strict lockdown, the spread of the virus fully stops after 60 days. However, this is almost impossible to occur in a real-world scenario. Furthermore, lockdown causes the economical ratio to be decreased to less than 0.2.

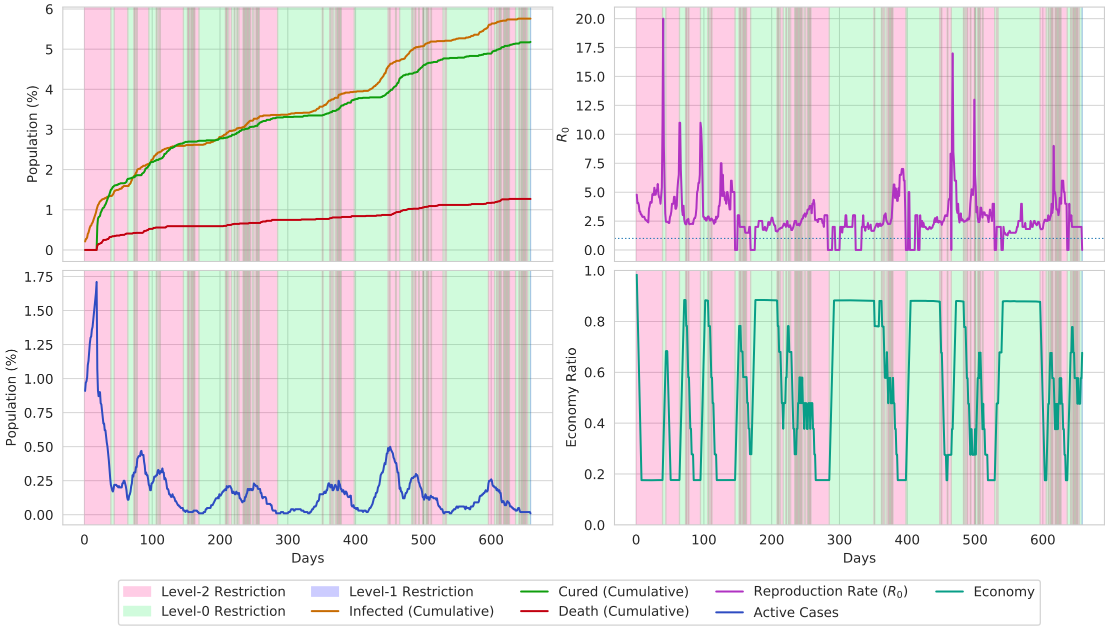


Figure 11. The graphs represent the movement restrictions provided by the agent. The red region of the graph denotes the days when a lockdown is placed. The green region of the graph denotes the days when no lockdown is placed. In the early stage of the environment, the agent places multiple 20-40 days lockdown to reduce the spread of the disease. In the later stage, to control the resurgence of the disease, the agent performs a cyclic lockdown (1-3 days cycle) followed by a 10-15 days lockdown to reduce the future spread of the virus. It can be also analyzed that the agent mostly follows this pattern when both the active cases percentage and the reproduction rate is high.

causes the spread of the disease to be halted. In Figure 12 and 13, the action sequences of the agent are illustrated for an environment of population density 0.01 and 0.03, respectively. In both cases, the agent follows a cyclic lockdown if the situation

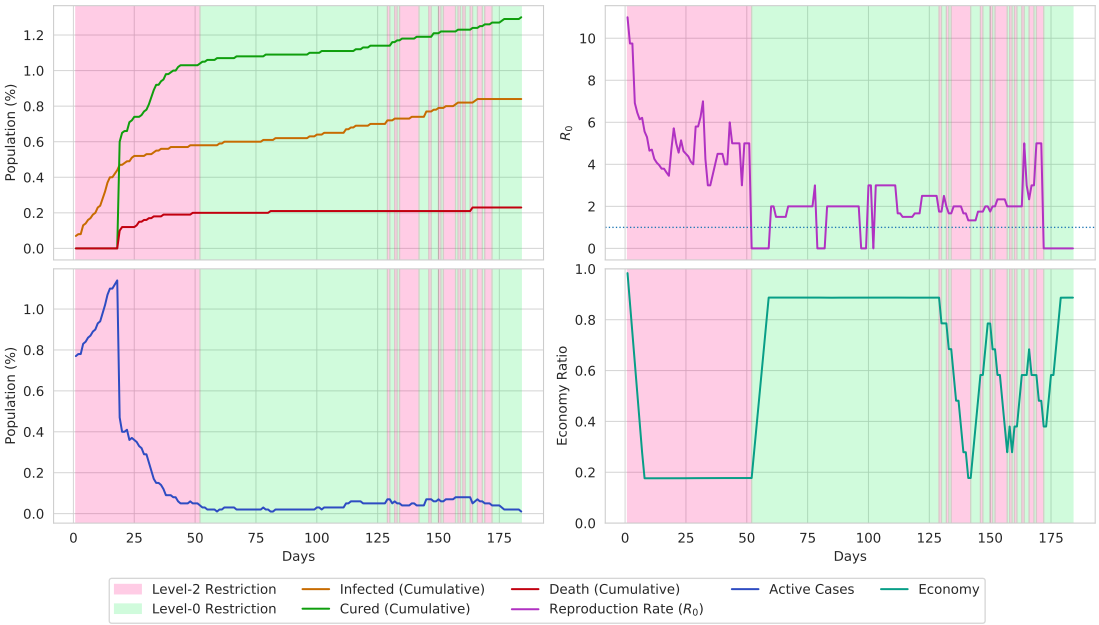


Figure 12. This graph illustrates the actions performed by the agent in a 0.01 population density environment. The other environmental parameters are kept unchanged. The graph resembles a similar action pattern of the agent observed in a 0.02 population density environment. However, less population is infected due to the lesser spread of the disease.

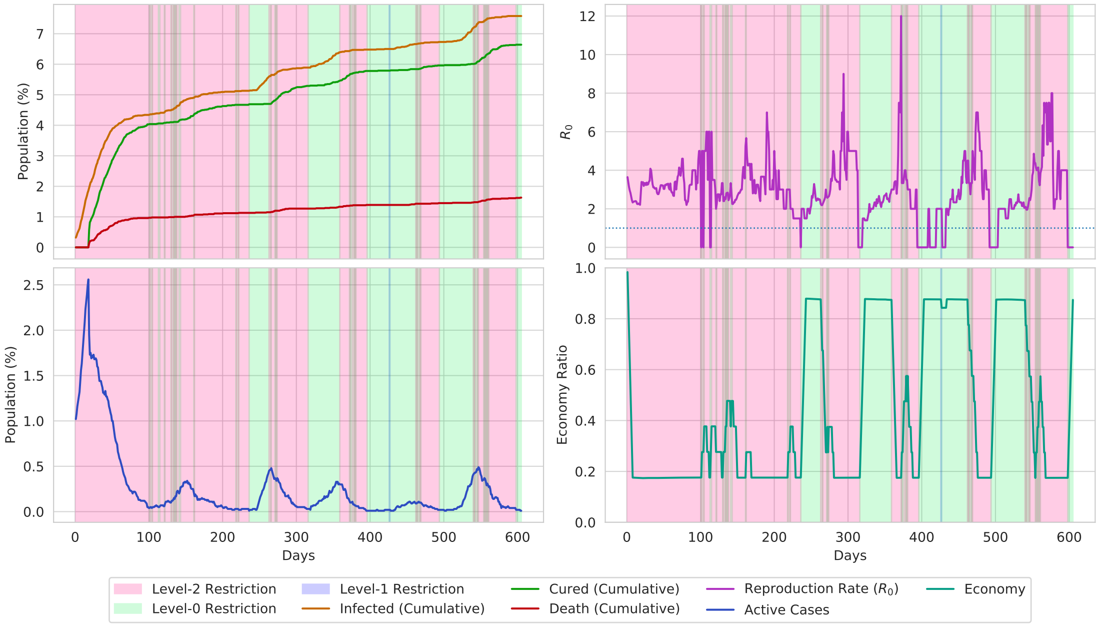


Figure 13. This graph illustrates the actions performed by the agent in a 0.03 population density environment. The other environmental parameters are kept unchanged. The graph resembles a similar action pattern of the agent observed in a 0.02 population density environment. However, due to increased population density, the spread of disease is also increased. Therefore, the agent mostly places strict lockdown instead of cyclic lockdown.

is less severe; otherwise, it places a full lockdown. Furthermore, by closely evaluating the reproduction rate and active cases of the environment, a pattern of the lockdown placement can be observed.

The agent places lockdown based on the active cases and the reproduction rate. However, it can be observed that the agent sometimes avoids placing lockdown when the reproduction rate is high. The agent only places lockdown when the value of active cases and reproduction rates are high. It further removes the lockdown when the reproduction rate is less than 1. To discover the reason for the action, let us consider the following formula,

$$\delta Inrease_{disease} = ActiveCases \times R_0 \tag{5}$$

The equation formulates the possible number of people who may get infected in the next day. The reproduction rate R_0 represents the average number of newly infected cases caused by an infectious person, and the value of $ActiveCases$ indirectly represents the number of infected persons in a single day. Therefore, the increase in infectious cases can generally be formulated using Equation 5. The agent places strict lockdown actions when the value of the equation 5 becomes too high. On the contrary, for minor cases, the agent follows a cyclic lockdown phase. This causes optimally controlling the spread of the disease below a particular percentage.

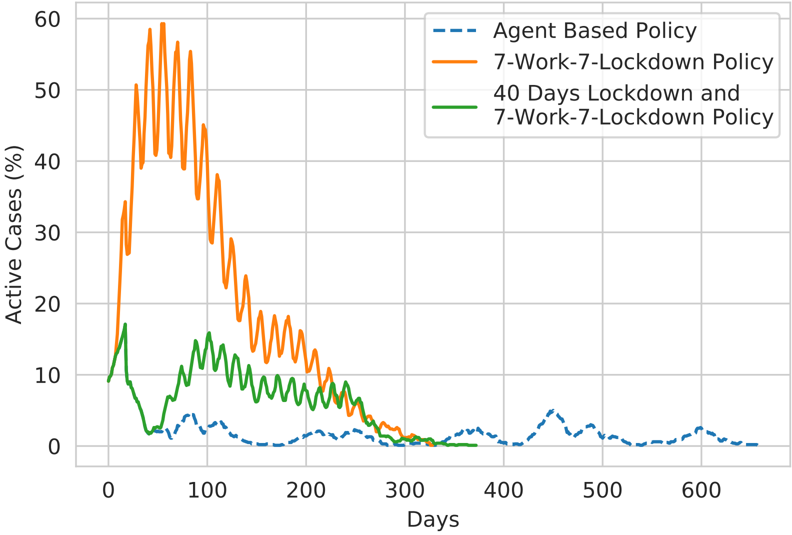


Figure 14. The graph presents a comparison of the agent's policy with the traditional n-work-m-lockdown policy. The comparison is formed on a 10000 population with a density of 0.02. By only maintaining a 7-work-7-lockdown policy, a rapid spread of the virus can not be halted, and therefore, a total of 34.5% of the population gets infected. Furthermore, if the 7-work-7-lockdown policy is applied after a full lockdown of 40 days, the overall infection is decreased to 11.5%. However, the agent generated policy mostly flattens the curve.

In Figure 14, we further compare the agent's policy with the traditional n-work-m-lockdown policy. From the comparison, it can be justified that only maintaining the n-work-m-lockdown policy is not an optimal solution to mitigate a pandemic. Furthermore, adding 40 days of full lockdown before following the n-work-m-lockdown policy reduces the first surge of the disease. However, the n-work-m-lockdown policy does not control the spread of the disease properly. Therefore, a resurgence of the disease is observed. From the general comparison, it can be validated that an agent can optimally control a pandemic crisis if proper training method is implemented.

Discussion

The paper motivates the readers towards the achievements and advancements of reinforcement learning through its application for controlling the pandemic crisis. We introduce a virtual environment that mostly relates to a pandemic situation, and sedulously investigate new tactics to mitigate disease by applying reinforcement learning. In what follows, we perform a pensive analysis of the impact of lockdown, social-distancing, and using agent-based solutions to prevent the mitigation of disease. We find our proposed scheme to be convincing in achieving optimal decision balancing the overweening pandemic and economic situation. We strongly believe that the contribution of this research endeavor will unite the epidemic study with reinforcement learning, and may help the human race to defend against the pandemic crisis.

References

1. Earn, D. J., Dushoff, J. & Levin, S. A. Ecology and evolution of the flu. *Trends ecology & evolution* **17**, 334–340 (2002).

2. Butler, D. Swine flu goes global: New influenza virus tests pandemic emergency preparedness. *Nature* **458**, 1082–1084 (2009).

3. De Wit, E., Van Doremalen, N., Falzarano, D. & Munster, V. J. Sars and mers: recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523 (2016).

4. Yang, Y. *et al.* The deadly coronaviruses: The 2003 sars pandemic and the 2020 novel coronavirus epidemic in china. *J. autoimmunity* 102434 (2020).

5. Qualls, N. *et al.* Community mitigation guidelines to prevent pandemic influenzaunited states, 2017. *MMWR Recomm. Reports* **66**, 1 (2017).

6. Anderson, R. M., Heesterbeek, H., Klinkenberg, D. & Hollingsworth, T. D. How will country-based mitigation measures influence the course of the covid-19 epidemic? *The Lancet* **395**, 931–934 (2020).

7. Acemoglu, D., Chernozhukov, V., Werning, I. & Whinston, M. D. A multi-risk sir model with optimally targeted lockdown. Tech. Rep., National Bureau of Economic Research (2020).

8. Karin, O. *et al.* Adaptive cyclic exit strategies from lockdown to suppress covid-19 and allow economic activity. *medRxiv* (2020).

9. Cully, A., Clune, J., Tarapore, D. & Mouret, J.-B. Robots that can adapt like animals. *Nature* **521**, 503–507 (2015).

10. Watkins, C. J. & Dayan, P. Q-learning. *Mach. learning* **8**, 279–292 (1992).

11. Hasselt, H. V. Double q-learning. In *Advances in neural information processing systems*, 2613–2621 (2010).

12. Arulkumaran, K., Deisenroth, M. P., Brundage, M. & Bharath, A. A. A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866* (2017).

13. Mnih, V. *et al.* Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

14. Serban, I. V. *et al.* A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349* (2017).

15. Baker, B. *et al.* Emergent tool use from multi-agent autocurricula. *arXiv preprint arXiv:1909.07528* (2019).

16. Brown, N. & Sandholm, T. Superhuman ai for multiplayer poker. *Science* **365**, 885–890 (2019).

17. Silver, D. *et al.* Mastering the game of go without human knowledge. *nature* **550**, 354–359 (2017).

18. Berner, C. *et al.* Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680* (2019).

19. Brauer, F. Compartmental models in epidemiology. In *Mathematical epidemiology*, 19–79 (Springer, 2008).

20. Britton, T., Ball, F. & Trapman, P. A mathematical model reveals the influence of population heterogeneity on herd immunity to sars-cov-2. *Science* (2020).

21. Yong, B. & Owen, L. Dynamical transmission model of mers-cov in two areas. In *AIP Conference Proceedings*, vol. 1716, 020010 (AIP Publishing LLC, 2016).

22. Galvani, A. P. & May, R. M. Dimensions of superspreading. *Nature* **438**, 293–295 (2005).

23. Gollwitzer, A., Martel, C., Marshall, J., Höhs, J. M. & Bargh, J. A. Connecting self-reported social distancing to real-world behavior at the individual and us state level. *PsyArXiv preprint* (2020).

24. Aloi, A. *et al.* Effects of the covid-19 lockdown on urban mobility: Empirical evidence from the city of santander (spain). *Sustainability* **12**, 3870 (2020).

25. Williams, J. D. & Zweig, G. End-to-end lstm-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269* (2016).

26. Ding, Z., Xia, R., Yu, J., Li, X. & Yang, J. Densely connected bidirectional lstm with applications to sentence classification. In *CCF International Conference on Natural Language Processing and Chinese Computing*, 278–287 (Springer, 2018).

27. Van Hasselt, H., Guez, A. & Silver, D. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence* (2016).

28. Hu, H., Nigmatulina, K. & Eckhoff, P. The scaling of contact rates with population density for the infectious disease models. *Math. biosciences* **244**, 125–134 (2013).

29. Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklöv, J. The reproductive number of covid-19 is higher compared to sars coronavirus. *J. travel medicine* (2020).

30. Oliphant, T. E. Python for scientific computing. *Comput. Sci. & Eng.* **9**, 10–20 (2007).

31. Gulli, A. & Pal, S. *Deep learning with Keras* (Packt Publishing Ltd, 2017).

32. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, 265–283 (2016).

33. Hunter, J. D. Matplotlib: A 2d graphics environment. *Comput. science & engineering* **9**, 90–95 (2007).

Author contributions statement

All authors conceived the project idea and designed the data analysis. Abu Quwsar Ohi and M. F. Mridha collected the data, undertook the data analysis, wrote the manuscript and prepared the figures and tables. Muhammad Mostafa Monowar and Md. Abdul Hamid reviewed the manuscript, figures and tables and also provided advice and guidance throughout the study.

Competing interests

The authors declare no competing interests.

Additional information

A pseudocode of the environment is represented in Algorithm 1.

Algorithm 1: A pseudocode of the execution process of the virtual environment.

```
Input: The size of the grid $s$,
       The size of population $N$,
       Number of infectious population $M$,
       Number of days staying exposed $E_t$,
       Number of days staying infectious $I_t$,
       Number of daily movements $M_t$,
       Number of days $D$
$S \leftarrow \{(x,y,d) \in \mathbb{N} | 0 \leq x,y \leq s \; and \; d = 0\} \quad and \quad |S| = N - M$ ;     /* susceptible population */
$E \leftarrow \{\}$ ;                                                         /* exposed population */
$I \leftarrow \{(x,y,d) \in \mathbb{N} | 0 \leq x,y \leq s \; and \; d = 0\} \quad and \quad |I| = M$ ;         /* infectious population */
$R \leftarrow \{\}$ ;                                                       /* recovered population */
$P \leftarrow S \cup E \cup I \cup R$ ;                                               /* total population */
$Economy \leftarrow 0$ ;                                             /* total economic transaction */
/* loop for each day */
for $day \leftarrow 1$ to $D$ do
    /* loop for each step */
    for $m_t \leftarrow 1$ to $M_t$ do
        /* loop for each person */
        for $p \in P$ do
            /* $\mathbb{Z} \cap [-1,1]$ defines picking a random integer from -1, 0, and 1 */
            $x_t \leftarrow max(min(p(x) + \mathbb{Z} \cap [-1,1], s), 0)$ ;        /* making valid movements in the grid */
            $y_t \leftarrow max(min(p(y) + \mathbb{Z} \cap [-1,1], s), 0)$;
            $z_t \leftarrow p(z) + 1$ ;                                          /* updating the day counter */
            /* if the person is in recovered state */
            if $p \in R$ then
                $P \leftarrow (P - p) \cup \{(x_t, y_t, 0)\}$ ;           /* no state upates for recovered population */
            /* if the person is in infectious state */
            else if $p \in I$ then
                /* $\mathbb{N} \cap [0,7]$ defines picking a random integer in range [0, 7] */
                if $z_n - \mathbb{N} \cap [0,7] \geq I_t$ then
                    /* randomly choose if a person survives, and the probability distribution of choosing 1 over
                       0 is 1:5 */
                    if $\mathbb{N} \cap [0,1] = 1$ then
                        $P \leftarrow P - p$ ;                          /* dead person are removed from the states */
                    else
                        $R \leftarrow R \cup \{(x_t, y_t, 0)\}$ ;              /* recovered person is moved to recovered state */
                        $P \leftarrow (P - p) \cup \{(x_t, y_t, 0)\}$;
                    $I \leftarrow I - p$;
                else
                    $P \leftarrow (P - p) \cup \{(x_t, y_t, z_t)\}$;
                    $I \leftarrow (I - p) \cup \{(x_t, y_t, z_t)\}$;
            /* if the person is in exposed state */
            else if $p \in E$ then
                /* $\mathbb{N} \cap [1,2]$ defines picking a random integer 1 or 2 */
                if $z_n - \mathbb{N} \cap [1,2] \geq E_t$ then
                    $E \leftarrow E - p$;
                    $I \leftarrow I \cap \{(x_t, y_t, 0)\}$;
                    $P \leftarrow (P - p) \cup \{(x_t, y_t, 0)\}$;
                else
                    $E \leftarrow (E - p) \cup \{(x_t, y_t, z_t)\}$;
                    $P \leftarrow (P - p) \cup \{(x_t, y_t, z_t)\}$;
            /* if the person is in susceptible state */
            else
                /* check if the person is in close contact with any of the infectious person */
                if $(x_n + [-1,1], y_n + [-1,1], \mathbb{N}) \in I$ then
                    $I \leftarrow I \cup \{x_t, y_y, 0\}$ ;                          /* move the person in exposed state */
                    $P \leftarrow (P - p) \cup \{(x_t, y_t, 0)\}$;
                else
                    $P \leftarrow (P - p) \cup \{(x_t, y_t, z_t)\}$
            /* except of the person is not infectous (and not dead) he/she contributes to the economy */
            if $p \notin I$ then
                $Economy \leftarrow Economy + \mathbb{R} \cap [0.8, 1]$;
```