Optimizing Pandemic Control Strategies: A Deep Reinforcement Learning Approach in Public Health Management

Submitted in partial fulfilment of the requirements of the degree of Doctor of Philosophy
December 2023

Author: Raphael Ibraimoh
Supervisor I : Professor M. Saraee
Supervisor II: Dr Kaveh Kiani

Contents

List of Figures

List of Tables

List of Abbreviations

A2C Advantage Actor Critics

ARDS Acute Respiratory Distress Syndrome

CCM Chronic Care Model

CNN Convolutional Neural Network

ConvNet Convolutional Neural Network

COVID-19 Coronavirus Disease

D3QN Dueling DQN

D4PG Distributed Distribution DDPG

DDPG Deterministic Policy Gradient

DDQN Double DQN

DQN Deep Q-Network

DRL Deep Reinforcement Learning

DRTs Dynamic Treatment regimes

ELU Exponential Linear Unit

HRL High Level Policy

LSTM Long Short Term Memory

MAE Mean Absolute Error

ML Machine Learning

MLSTM Modified Long Short Term Memory

PPO Proximal Policy Optimization

ReLU Rectified Linear Unit

RMSE Root Mean Square Error

RL Reinforcement Learning

RNN Recurrent Neural Network

SVG Stochastic Value Gradient

TRPO Trust Region Policy Optimization

UK United Kingdom

WHO World Health Organisation

Declaration of authorship

I, Raphael Ibraimoh, declare that this thesis titled Innovative Strategies for Pandemic Control: Unleashing Deep Reinforcement Learning in Public Health Management, COVID-19 Case Study and the work presented within are my own and no part of the work contained in this thesis has been given in support of any application for any other degree or qualification at the University of Salford or any other university or institution of learning.

I have maintained professional integrity during all aspects of my research degree and, I have followed the Institutional Code of Practice and the Regulations for Postgraduate Research Degrees.

Where I have consulted the published work of others, this is always clearly attributed. Where I have quoted from the work of others, the source is always given. This research received no external funding, and I declare no conflict of interest.

Acknowledgement

By virtue of His boundless benevolence, the Almighty God has bestowed this opportunity upon me. For this, I am profoundly, genuinely, and unconditionally appreciative and grateful. Incorporating the support of my family, friends, and supervisors was crucial in ensuring the successful completion of my thesis. Sincerely, I wish to convey my appreciation to my primary advisor, Professor Mohamad Saraee, for his extraordinary expertise, in addition to his unwavering support and guidance throughout the entire process of developing my thesis, until its completion. I am profoundly appreciative of Dr. Kaveh Kiani's ongoing guidance, unwavering support, and stern advice, which he provided throughout the course of my PhD. Furthermore, I would like to extend my utmost gratitude to all those who contributed their time voluntarily to this endeavour in order to aid others; without them, this undertaking would have been rendered unattainable. Their unwavering affection and support throughout my life, which has served as a source of inspiration and motivation, for which I am extremely grateful to my family. The magnitude of my affection and gratitude towards them, to whom I owe everything, cannot be adequately conveyed. My profound appreciation goes to my late spouse, Dr. Patricia Onyemaechi Ibraimoh, for her unwavering support in my pursuit of a PhD while she was alive. Additionally, I am indebted to my sons, Raphael Ibraimoh Jr., Victor Ibraimoh, and Samuel Ibraimoh, for their motivation, guidance, and support throughout my entire academic career. This undertaking would have been unattainable without the support, compassion, solace, and altruism supplied by them. In conclusion, I wish to extend my sincere appreciation to the University of Salford for their generous support.

Abstract

COVID-19, also known as the SARS-CoV-2 coronavirus, has paralysed the world and forced people to change their lifestyles. Since COVID-19 deaths are increasing daily, the disease has become a global public health issue. Different countries used different public health guidelines to avoid human-to-human transmission. Personal hygiene, hand washing and sanitization, face masks for social distance, comprehensive testing, and, in the worst case, a lockdown and travel restriction are rules.

This research seeks the optimal lockdowns and border control approach for timely lockdown and travel limitations. This thesis attempts to use UK data from the global pandemic dataset. The data was trained using DRL algorithm to determine lockout and travel limitation timing. This is the first study to use deep reinforcement learning to determine the best UK lockdown and border control method. A unique base model, Duelling Double Deep Q-Network (D3QN), a variation of the Deep Q-Network algorithm (DQN), was used to train COVID-19 epidemic dataset and evaluated on test data. Public health and government will be able to execute prompt and appropriate lockdown and border control policies to minimise the disease's spread, improving people's quality of life and lowering costs.

Initial lockdown and travel restrictions reduced COVID-19 load. However, our agency advised the UK to lock down or restrict travel before or on the index case (the first deceased recoded). Moreover, the agent frequently called for a full lockdown, border closures, travel restrictions, and more harsh security measures than public health. This study assesses the positive effects of preventing COVID-19's spread on population health while considering its negative economic and social effects. Finally, average moving reward was used to compare baselines.

Chapter 1

Introduction

Any nation's public health system must work cooperatively with the rest of society to prevent illnesses, extend life, and avoid disabilities. As everyone in a social unit is exposed to the same level of influences that affect their health and wellbeing, it is a collaborative effort. How well-developed the public health system is will determine how high the quality of life is. Many health advantages as well as social advantages come with good mental health. As both equitable social advantages and balanced, healthy lives are supported by positive psychological performance. A good social objective in and of itself is social wellbeing, which includes mental wellbeing. As a result, it has become a definite and stated objective of the government in many nations.

There is a new wave of sickness that is more deadly in terms of transmissions and mortality rates, which is of great worry to public health due to the quick expansion of human-to-human transmission, the lack of immunisations, and the virus's challenging mutational behaviour. The spread of this virus has turned into a pandemic, destroying not only national economy but also the mental and physical health of the populace. On December 31, 2019, Wuhan, China, reported the index case of COVID-19.

Within a month after the initial announcement, there were around 75,000 confirmed cases of the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) on the Chinese mainland. This disease quickly spreads with an outbreak into Korea and Italy. The overwhelming number of individuals who are infected with this disease within a short period of time has made it a global problem. There has not yet been a potential vaccination to guarantee that people are immune to this terrible sickness as of the 28th of February 2021, when this proposition was made. Lack of personal protective equipment, hospitals, self-quarantine facilities, critical care capacity, and appropriate treatment lead to a surge in further health crises in some nations and territories. COVID-19 is thought to be responsible for 0.5 million fatalities worldwide, while this number may be conservative due to insufficient testing in developing and underdeveloped nations.

An infectious respiratory and vascular disease, COVID-19, affects people. The illness' typical signs and symptoms include a fever, cough, exhaustion, shortness of breath, and a loss of taste and smell. The time it takes a human to contract a virus and start showing symptoms is known as the incubation period, and it can be anywhere from one day and 14 days. The majority of those affected only experience minor symptoms, but some go on to suffer acute respiratory distress syndrome (ARDS). Unfortunately, those who overcame the acute stage would sustain long-term organ damage, mainly to the heart and lungs. Additionally, those who transiently entered the acute phase would endure one or more of these symptoms, such as fever, weakness in the muscles, exhaustion, memory loss, and other symptoms, for an extended period.

Despite the possibility of vaccines that would require people to wait longer to receive them, there is still a need for People should practise social distance, good personal hygiene including washing their hands, and keeping their hands away from their eyes, mouth, nose, and ears, according to the WHO and public health departments of many nations. To reduce the danger of transmission, face masks are recommended for usage in public settings. The disease's potential impact on individuals who may not manifest any outward signs is more alarming. Testing is useless in such asymptomatic patients who lack symptoms. A global economic and financial catastrophe has also been caused by the COVID-19's propagation and the response actions that were taken. Because of varying cultures, climates, and demographics, it can be difficult to identify the right kind and degree of public health policies for each nation and territory. International organisations are working nonstop to ease this suffering and safeguard people's priceless lives. NHS corona dataset, UK (https://www.gov.uk/coronavirus),UNICEF, (https://www.cdc.gov/coronavirus/2019-ncov /prevent-getting-sick/prevention.htm)CDC, WHO (https://www.who.int), and numerous other organisations are just a few of them. To slow the coronavirus's rapid spread, these institutions are recommending individuals to follow certain guidelines through the public health system.

Throughout the past few decades, machine learning (ML) techniques have advanced remarkably in a variety of fields, most notably in the field of public health (Vlado et al., 2019). Virtually every element of modern life is supported by machine learning, including social network functionality, e-commerce recommendation systems, and smart devices like smartphones and cameras. Supervised and Unsupervised learning, two categories under which machine learning is typically categorised, have been employed to some extent in the prediction of clinical risk of diabetes from electronic hospital records (EHR) (Huang et al., 2007). The application of machine learning (ML) techniques is concentrated on predictive analyses, such as future forecasting or, more precisely, disease diagnostics, which identifies the route to disease therapy. Finding the best policy regime for a patient's therapy, meanwhile, is difficult. As a result, ML algorithms only recommend a generalised treatment that is not tailored to a specific patient. Since reinforcement learning (RL) has lately acquired popularity in video games, it can now fill the gaps left by ML (Mohammad and Mahmoud ,2020).

The reward mechanism used in RL is based on trial and error. This sequential decision-making method has a delayed reward (Nemati et al., 2016). Because of the non-optimized training dataset, the results are typically optimal techniques (Raghu et al., 2017). In their most recent research, Phuong et al. (2018) applied RL to identify the best diabetes type 1 control strategy. The minimum and HAVORKA models were combined in the study to assess and maintain blood glucose levels. Mahsa et al (2019).'s use of the conventional Q-learning reward system to identify the best type 1 diabetes care.

Deep learning (DL) is "Shallow" in comparison to common machine learning methods like decision trees, logistic regression, support vector machines (SVMs), linear regression, and more. To learn representations of data with a variety of abstractions, DL uses computer models consisting of multiple processing layers. There have been numerous DL studies on diabetes, but they have all relied on the Boltzmann machine approach to categorise and identify different forms of diabetes (Zeki et al., 2012). The combination of deep learning with reinforcement learning is known as deep reinforcement learning. The reinforcement learning uses deep neural networks as its approximation components. A novel method for simulating basic or complicated real-life settings is DRL. Mnih et al. (2015) reported progress in learning control strategies for the Atari game directly from high-dimensional sensory inputs.

In this report, I will discuss how DRL algorithms can be used to determine the appropriate course of action for public health to safeguard individuals from the spread of COVID-19. The agent, which is the model software, can identify the appropriate timing for lockdown and travel restriction can be applied as policy, that will drastically decrease the spread of the disease, maintain quality of life, and overall keep the economy going.

1.1 Research Background

Since the World Health Organisation was notified about the SARS-CoV-2 virus at the end of December 2019, at least 143 activities and interventions have taken place in the UK to restrict and mitigate the infection's spread. Considering the classification approach described by (Moy et al., 2020) and the Oxford stringency index created by (Hale et al. 2020) to investigate these treatments. The policy classification proposes numerous COVID-19 intervention classes that grow in severity and subsequently de-escalate when governments wind down response actions. These classifications pertain to COVID-19 containment and mitigation policies, with the goal of reducing the severity of the impact on health and increasing care. Economic and health technology interventions are also classified.

Some venues like the pubs and restaurants were allowed to open in England on July 4th, but not in Scotland until July 15th. Indoor restaurants and pubs in Wales could open on August 3rd, and indoor restaurants and pubs selling food in Northern Ireland could open on July 3rd, during the early phases of the pandemic, the government focused on viral prevention and mitigation, most of the measures were related to containment policies. Following the introduction of social distancing measures and fines, the number of measures rose, as did the stringency of containment measures. The major limitation was the suspension of non-essential services on March 16th, followed by a lockdown on March 23rd. This lockdown required everyone to stay at home and work from home as much as possible, with just one hour of exercise, food shopping and prescription trips allowed every day, and a social distancing measure of 2 metres. Nonetheless, cases in the UK continued to grow until levelling as the number of daily cases slowly decreased, as seen by the decreasing confirmation rate. The daily peak in simulated symptomatic and lab-confirmed cases occurred on April 1st and May 1st, respectively. The severity of policies was lessened or deescalated after the peak of instances (and the number of confirmed cases continued to fall).

Zhou and Khan(2021) look at how different social groups' wages, time utilisation, and subjective well-being altered during the pandemic in the UK. They examine within-individual changes in labour income, paid work time, housework time, childcare time, and distress level during the three lockdown periods and the easing period between them (from April 2020 to late March 2021) using longitudinal data from the latest UK Household Longitudinal Survey (UKHLS) COVID study and earlier waves of the UKHLS. (Zhou and Khan, 2021) discovered that as the epidemic progressed, COVID-19 and its linked lockdown measures had unequal and variable effects on people's income, time consumption, and subjective well-being based on their gender, ethnicity, and educational level in the UK. Finally, the magnitude of the effects of COVID-19 and COVID-induced measures, as well as the rate at which these effects manifested, varied across social groups with different types of vulnerabilities.

Recent study indicates that the COVID-19 epidemic and accompanying social and economic interventions, such as physical separation and business closure, have varied effects on different social groupings. In the UK, for example, women and parents are shown to have experienced a greater decrease in subjective wellbeing (David and Jones, 2020). Pierce et al., (2021) affirm that, Black, Asian, and minority ethnic (BAME) immigrants were more likely to face economic difficulties in the immediate aftermath of the first national lockdown (Hu, 2020) Furthermore, among those with COVID-19, people of BAME origin in the UK had a greater death rate than white people (Petel et al., 2020). These earlier findings showed the existence of acute disparities in repercussions for different social groups, but our understanding of the long-term effects of COVID-19 and related measures remains restricted. The COVID-19 epidemic has already lasted more than a year, with the UK experiencing three national lockdowns. Early research was limited by data that encompassed only two-time intervals, such as before and immediately after the first lockdown notification. Little is known about how unequal societal repercussions manifest themselves at various stages of the COVID-19 epidemic, particularly with recurrent lockdowns. This omission limits our knowledge of how COVID-19 and COVID-induced social policies, such as physical distancing measures, working from home, and the closure of enterprises, which have been changing on a weekly or even daily basis, affect people's lives. Documenting the evolution of the impacts of COVID-19 and COVID-induced measures is critical for understanding the ramifications of this quickly developing pandemic and assisting policymakers in planning for future waves and pandemics.

A more detailed and up-to-date studies on how inequalities have evolved as the COVID-19 pandemic spreads in waves and various strategies to contain it have been implemented over the last year. We conducted analyses on nationally representative population data from the most recent UK Household Longitudinal Survey (UKHLS), which was conducted prior to the first lockdown in March 2020, during the first lockdown from April to June 2020, during the ease of the first lockdown (June to September 2020), and during the latter two lockdowns (November 2020 and January 2021 to March 2021). In this work, we contribute to COVID-19 research by giving a dynamic picture of how people's labour earnings, time use, and wellbeing varied across different stages of the pandemic. We also investigated whether and how much the disparities in these outcomes by gender, ethnicity, and educational level have changed in the last year. In what follows, (Petel et al., 2020) first evaluate the most recent research on the impact of COVID-19 and COVID induced measurements on people's life, concentrating on three dimensions of social inequality: gender, race/ethnicity, and education. Their research discusses the evolution of the COVID-19 pandemic, and the UK lockdown measures from March 2020 to April 2021. Following that, (Petel et al., 2020) introduce the data and its longitudinal architecture, which allows us to compare the information of the same individuals before the start of the epidemic and at different time periods during the past year. Finally, they provided the results of fixed-effect regression analysis and discuss their findings.

Although different data mining approaches have been employed in the past, previous research difficulties can be grouped into two categories: data gathering and prediction strategies. With the introduction of COVID-19 in the United Kingdom, data collecting is limited. The low number of Biobank participants with COVID-19 and serology data leads to model overfitting or misprediction (Willette et al., 2022).

1.2 Research Motivation

The necessity to help public health determine the best strategic policy to advise in preventing the spread of pandemics, as COVID-19 in this instance, is what drives this research. To slow down or stop the spread of the COVID-19 virus. Coronavirus is one of several public health challenges that urgently calls for government intervention in terms of policy and management. Otherwise, if an appropriate and effective policy is not in place, it will spread and cause an increasing number of deaths. As of the time this report was being written, 17.9 million people worldwide were infected by the coronavirus by December,2020, which also caused 0.6 million deaths worldwide. As more people are tested, more cases are coming to light. The above-mentioned need for the DRL algorithm further motivates this research, which aims to investigate the effectiveness of the impact of the best policy that the public health department will advise the populace on, to stop the spread of the virus while also maintaining positive economy.

1.3 Research Problems

COVID-19 is an infectious disease, because of several acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The people affected with COVID-19 normally show symptoms that include persistence dry cough, fever, fatigue, shortness of breath, and loss of taste and smell. In older for public health management and government to curb the spread of this disease, there is need to formulate an optimal policy that will not only flatten the curve of spread but also bring it down to a manageable number.

A review of the relevant literature uncovered many key themes about COVID-19 prevalence. Existing research has greatly aided our understanding of the importance of ML in controlling the spread of COVID-19 in the UK. There are, however, a few critical research gaps that must be addressed.

First, most research from the United Kingdom and other nations have simply looked for trends to predict COVID-19 progression (Xin et al., 2022, Willette et al., 2022, Petel et al., 2020, David and Jones, 2020, Hu, 2020). Overfitting may occur due to the researcher's insufficient data set. Given the importance of determining the optimal timing for lockdown and travel restrictions, this is a critical research gap.

Second, there is not a single study that was conducted in the United Kingdom that looked into the optimal timing for travel restrictions and lockdowns that was based on a deep reinforcement learning technique. The few studies that did employ analytics and trustworthy data never examined the usage of neural networks and reinforcement learning techniques. Since the findings of their investigation might be misleading, this is another crucial void that needs to be filled as soon as possible.

Third, DRL has been the only method for determining the optimal period for lockdown and travel restriction (Harshad et al., 2020, Gloria et al., 2021, Kailiang, 2022,). In a prior study, DRL was utilised to forecast cases of confirmation, mortality, and recovery, and the results were compared to the time when lockdown and social isolation were imposed (Dong et al., 2022).

This novel DRL method is based on Duelling Double Q-Network (D3QN) and uses available data to learn from the environment the ideal policy for implementing lockdown and travel restrictions. This technique will assist the government and public health in the United Kingdom in determining the best strategy for controlling disease transmission over time. The application of MLSTM aids in forecasting the future development of this disease, paving the path for contingency measures.

This research would help to enhance knowledge and discover practical applications in the field of DRL to combat the spread of COVID-19 and other pandemics by solving these large gaps in research.

1.4 Research Aim and Objectives

Unlike past studies, this research will analyse the potential impact on the quality of life and the economy by DRL in determining the best time to implement the lockdown and travel restrictions. COVID-19, like many other diseases, has a greater impact on vulnerable individuals, such as the elderly, people with disabilities, ethnic minorities, and those living in disadvantaged areas. Allowing the virus to spread exponentially would result in societally unacceptable consequences in terms of loss of life and illness.

Given the scarcity of research on the detailed use of DRL to find the appropriate timing for both lockdown and travel restriction in the UK, as well as an analysis of the potential impact of disease on the economy and the adverse quality of life it caused, the purpose of this report will be to factor in the economy and quality of life when using deep reinforcement learning on COVID-19 data in the UK.

Based on the motivation, the aim of this Report is to use a state-of-the-art DRL algorithm to identify the best timing for lockdown, and border control restrictions that would significantly slow the spread of the disease while having little to no negative economic effects.

The research main objectives:

1. To conduct an extensive exploratory analysis of Epidemic COVID-19 data. Investigate, to reveal patterns in death, recovery and confirmed cases of COVID-19 in the UK, pre-processing, and feature engineering as necessary.

2. To train DRL algorithm to modify the LSTM model for improved prediction of the confirmed, recovered and death cases of COVID-19 in the UK.

3. To train a Deep Reinforcement Learning (DRL) algorithm, the D3QN to finding the right timing of optimal policy for lockdown and travel restriction in the UK.

4. To develop a framework that will clearly spelt out the research approach and results.

1.5 Research Contributions

This research is about DRL, which evaluates the possible effect of combining COVID-19 Epidemic data with Lockdown and local/international travel ban data to determine the ideal timing for lockdown and travel restriction. The current report makes significant contributions by attempting to fill several gaps.

First, this study adds to the minimal research on deep reinforcement learning that has been done to curb the spread of COVID-19 in the UK. This is one of the first studies to look at determining the best approach for when the UK's public health and government suggest a lockdown and travel ban.

Second, by utilising CRISP-DM standard Data Science methodology to determine the optimal policy for lockdown and travel ban timing to curtail the spread of COVID-19 in the UK. This study adds to the existing literature on DRL as an agent to finding the appropriate timing for introducing lockdown and travel ban in the UK. This is also one of the first studies to investigate DRL in relation to COVID-19 dissemination using normal data mining approaches.

Third, our work adds to the small amount of research on using untrustworthy data from the Epidemic COVID-19 dataset. The information is primarily aggregated global totals, therefore it is a true representation of the COVID-19 spread in the UK. DRL requires a huge amount of data to be trained to near precision, as opposed to a little amount of data, which could lead to overfitting.

Fourth, no previous research has looked at the unique D3QN, which is a version of the Deep Q-Network algorithm, according to the literature review conducted for this study. Because the off-policy learning required to seek an optimal policy from data emanated from other behaviour policies generates overestimation and replay buffer (Li,2019), the Double Deep components help to control the overestimation of the policy (Van et al., 2016) and (Wang et al., 2003).

Fifth, previous research has concentrated on using DRL with little amounts of data or with simulated data. One of the first research to employ a novel discrete state and action representation space-based reinforcement learning paradigm. Because (Kailiang,2022)'s continuous action space could lead to incorrect prediction and learning because the pandemic could come to a halt at any point, D3QN beats DDPG (Julious and Deshendran, 2019) in terms of limited training.

This work would contribute to the new field of DRL by assisting in the management of COVID-19 spread in the United Kingdom. It will also increase public faith in public health and government decision-making, as well as public respect and compliance with both public health and government directives.

1.6 Scope of Research/Research focus

RL itself is a broad area of artificial intelligence (AI), as well as the deep learning. All aspects of both reinforcement and deep learning that do not have any direct link with use to finding the optimal policy would not be explained in depth or totally disregarded.
The obvious methods used to control the spread of COVID-19 in the United Kingdom is by personal hygiene, face mask, quarantine, track, trace, and testing. Those infected with this virus are made to isolate for 14days so as not to reinfect others in case they are asymptomatic. It is feasible to test everybody and very expensive especially when one

does not show any symptoms. Another approach is by disinfection; all the streets, every home and office will be sanitized. This approach is also very expensive and may crumble the economy. Another common approach advised by the public health is one or combination of the followings: washing of hands, social distancing, locking down activities, travel restrictions and quarantine facilities to isolate those that will show symptoms.

However, this research work will focus on the appropriate timing and intensity for optimal policy or strategy on lockdown and travel restriction that will slow down the spread of the virus to a controllable level that will have less burden on the economy.

1.7 Ethical and legal considerations

Ethical considerations were thoroughly addressed throughout the experiment, including obtaining informed consent, safeguarding participant confidentiality, and minimizing potential harm. Legal considerations included rigorous adherence to data protection rules and intellectual property rights, which protected the interests of all stakeholders. The comprehensive approach to these issues indicates a thorough awareness of the professional norms and rules required for doing responsible research like this one.

1.8 Research Structure

The study's context was provided in chapter one. The research problems, aims, and objectives have been delineated, and the significance of this study has been substantiated. Chapter two of this study will look at the existing literature on the prevalence of COVID-19, with the aim of identifying research directions and gaps within the broader context of determining optimal policies.

Chapter three will undertake a comprehensive examination of the research methodology and the technical framework used in this research work and strategies for enhancing skill development within the realm of deep and reinforcement learning and shows several assessment metrics employed in the domain of Deep Reinforcement Learning.

Chapter four of the study will encompass the presentation of the data preparation process and the subsequent creation of the model. This work aims to provide a justification for the utilisation of a qualitative, inductive research methodology. Additionally, it will analyse the overall study framework, including its inherent limitations.

Chapter five provides a comprehensive examination of the research findings and analyses. Chapter six adequately addresses the research study's conclusion and offers recommendations for future research endeavours.

Chapter 2

Literature Review

2.1 Background and Related work

The COVID-19 pandemic, caused by the new coronavirus SARS-CoV-2, has had a dramatic influence on communities all over the world. The UK was no exception, with waves of infections, healthcare issues, and societal changes. This review of the literature looks at significant studies and research on the spread of COVID-19 in the UK, highlighting factors impacting transmission, government interventions, healthcare systems, and lessons learned. Thus background study will be examining those factors influencing the transmission, government responses, healthcare systems, and lessons learned of the COVID-19 in the United Kingdom.

2.1.1 Factor Influaecing Trasmission

1. Population Density and Urbanization:

The COVID-19 pandemic, driven by the new coronavirus SARS-CoV-2, has put worldwide public health systems under unprecedented strain. The virus has had a substantial impact on the United Kingdom (UK), causing researchers to investigate the elements that contribute to the infection's spread within the country. Population density and urbanisation are important factors that have been extensively researched for their possible impact on the transmission of infectious diseases such as Covid-19.Because of increased interpersonal connections and decreased physical distancing chances, population density, defined as the number of people living per unit area, plays an important role in disease transmission. The population density in the United Kingdom, particularly in places like London, varies, which could affect the virus's transmission. In densely populated urban areas, public transportation, shared living spaces, and workplaces are frequently overcrowded, which can enhance the fast spread of respiratory viruses. Souch and Cossins (2020) discovered a link between population density and the initial COVID-19 transmission rate in different UK regions.

This shows that higher population density locations saw faster virus propagation, maybe due to the increased possibility of intimate contact between individuals. The expansion of cities and urban areas, which is characterised by the growth of cities and urban regions, is closely tied to population density and can magnify the spread of infectious illnesses. Increased migration, foreign travel, and human mobility result from urbanisation, all of which contribute to the rapid spread of diseases. The largest cities in the United Kingdom, such as London, Birmingham, and Manchester, are economic, social, and cultural centres that attract people from all over the world. These interactions may increase the likelihood of virus introduction and transmission. According to Lucas et al. (2021), higher degrees of urbanisation are associated with an increase in COVID-19 cases in various UK regions, showing that urban areas are more sensitive to disease outbreaks due to increased connectivity and contact opportunities.

Socioeconomic gaps exacerbate the influence of population density and urbanisation on

Covid-19 transmission. The wealth distribution in cities is generally uneven, with densely populated areas facing higher poverty rates and limited access to healthcare resources. Such discrepancies can worsen the spread of COVID-19 by interfering with persons' ability to follow preventive measures such as self-isolation and appropriate medical care. Aldridge et al. (2020) discovered that neighbourhoods in the United Kingdom with lower socioeconomic level had greater rates of COVID-19 cases and deaths, underscoring the impact of social variables in shaping illness outcomes. Government initiatives and regulations have been critical in reducing the impact of population density and urbanisation on COVID-19 spread in the United Kingdom. Lockdowns, travel restrictions, and social distancing laws were implemented to lower contact rates and limit transmission within densely populated areas. According to Flaxman et al. (2020), severe actions in the UK, including a nationwide lockdown, significantly decreased the transmission of the virus and avoided a considerable number of infections.

Finally, the combination of population density, urbanisation, and socioeconomic conditions has influenced the spread of COVID-19 in the United Kingdom. Highly crowded urban regions with higher levels of urbanisation have shown an enhanced vulnerability to the virus's quick transmission. Socioeconomic gaps within these urban areas amplify the disease's impact. Government efforts, on the other hand, have been helpful in slowing the spread of the virus in densely populated areas. Understanding the complex connections between these components as the pandemic evolves is critical for developing successful measures to contain and manage future outbreaks.

2. International Travel International travel has been a major driver of COVID-19's global spread, particularly its impact on the United Kingdom (UK). As a highly networked nation with a high volume of foreign travel, the UK has a significant risk of importing and spreading the virus. Air travel, in particular, plays an important role in easing the movement of people across borders, opening up paths for the virus to enter and spread within the UK. COVID-19 was introduced to the UK mostly through international travel, when persons harbouring the virus arrived from several worldwide hotspots. Russell et al. (2020) discovered that the time and amount of international arrivals were substantially associated to the initial rise in COVID-19 cases in the United Kingdom. Airport hubs such as London Heathrow, one of the busiest airports in the world, aided in the virus's rapid spread across the country. Travellers travelling from areas where epidemics were active inadvertently carried the virus, resulting in localised clusters and eventual community transmission. Travel restrictions, quarantine procedures, and border controls were among the government actions aimed at controlling the spread of COVID-19. According to Chinazzi et al. (2020), adopting travel restrictions dramatically reduced the importation of COVID-19 cases and thereby delayed the virus's transmission in the UK. Although effective, these procedures posed economic and logistical obstacles by disrupting international trade and tourism.

Finally, international travel has been critical in the spread of COVID-19 throughout the United Kingdom. The UK's connection with the rest of the world, helped by its busy airports and transit networks, contributed to the virus's rapid introduction and early spread. Travel restrictions and quarantine precautions were implemented on time, which helped to prevent further transmission. Balancing the benefits of foreign travel with the need to prevent disease importation will remain a crucial challenge as the UK navigates future phases of the pandemic.

3. Age and Vulnerability

In the UK, age and susceptibility have emerged as significant factors impacting the transmission and effect of COVID-19. Individuals over the age of 65 and those with underlying health issues are at a higher risk of severe illness and death, which contributes to disparities in disease outcomes. The severity of COVID-19 instances has been determined by age. The elderly, particularly those over 65, are more vulnerable to serious respiratory problems and have a higher mortality rate. Docherty et al. (2020) discovered that age was a major predictor of mortality in COVID-19 patients, with the risk increasing exponentially with age. This age-related susceptibility has raised the strain on healthcare systems and demanded preventative measures for the elderly, such as shielding and prioritisation of vaccination.

The impact of COVID-19 in the UK has been exacerbated by vulnerability caused by underlying health illnesses such as cardiovascular disease, diabetes, and respiratory disorders. Patients who have prior diseases are more prone to become really unwell and require hospitalisation. Comorbidities were highly related with death in COVID-19 cases, according to Williamson et al. (2020), emphasising the necessity of identifying and safeguarding vulnerable people. Government policies have prioritised the protection of the most vulnerable citizens. Vaccination initiatives have prioritised the elderly and those with pre-existing health issues, with the goal of reducing the burden of severe cases and mortality. Lockdown measures, such as school closures and employment restrictions, were also put in place to protect vulnerable populations from the infection.

Finally, age and vulnerability variables have had a considerable impact on the transmission and impact of COVID-19 in the United Kingdom. The elderly and people with underlying health issues are at a higher risk of serious illness and death. Because of these differences, public health interventions such as vaccine prioritisation and preventive measures to safeguard vulnerable communities have been implemented. Addressing the special requirements of these groups is a critical aspect of good pandemic management as the pandemic evolves.

2.1.2 Government Responses and Policies

1. Lockdown Measures

Lockdown procedures have been critical in limiting the spread of COVID-19 in the United Kingdom (UK). These measures, which included stay-at-home orders, the closure of non-essential businesses, and gathering restrictions, were put in place to decrease human interaction and disrupt transmission chains. According to research, lockdowns have been beneficial in reducing the transmission of the infection. According to Flaxman et al. (2020), the initial statewide lockdown in the UK significantly reduced the transmission rate of Covid-19, resulting in a considerable decrease in new cases. The study emphasised the need of early and strict actions in slowing the spread of the pandemic.

Lockdowns have not only helped to reduce infection rates, but they have also given healthcare systems much-needed breathing room. Lockdown procedures have saved hospitals from becoming overburdened by lowering the number of severe cases occurring at the same time, allowing healthcare resources to be allocated more efficiently. According to Imai et al. (2020), lockdowns and other non-pharmaceutical measures reduced COVID-19-related mortality and relieved pressure on the healthcare infrastructure. The impact of lockdown measures, however, is not without obstacles. Lockdowns have economic, social, and mental health ramifications, such as job losses, economic downturns, and increased rates of anxiety and despair. Balancing the health advantages of lockdowns against their societal costs is a difficult task for politicians. Furthermore, the success of lockdowns can be determined by public compliance, enforcement measures, and individuals' ability to follow constraints.

Finally, lockdown measures have been critical in slowing the spread of COVID-19 in the United Kingdom. Early and strict interventions have been shown to be successful in lowering transmission rates and limiting overburdening of the healthcare system. The broader impact of lockdowns on economic, social, and emotional well-being, on the other hand, must be carefully monitored. As the pandemic progresses, a comprehensive approach that takes into account both the health and socioeconomic components will be critical for effective pandemic containment.

2. Testing and Contact Tracing

COVID-19 testing and contact tracing have emerged as essential measures for curbing its spread in the United Kingdom (UK). These methods are intended to swiftly identify and isolate sick persons, interrupting the transmission chains and limiting further spread. Effective testing has been critical in detecting symptomatic and asymptomatic cases, allowing for prompt isolation and treatment. The United Kingdom initially encountered difficulties in increasing testing capacity, however attempts were made to increase testing availability and accessibility. Larremore et al. (2020) found that frequent and extensive testing, combined with timely reporting of results, can greatly reduce the reproduction number of COVID-19, hence limiting its spread.

Contact tracing supplements testing by identifying and informing people who have had close contact with a confirmed case. In the United Kingdom, digital methods and manual contact tracing have been used to identify and isolate probable cases. According to Hinch et al. (2020), prompt and efficient contact tracing combined with isolation measures is useful in reducing outbreaks, especially when accompanied with testing.The success of testing and contact tracing, on the other hand, is dependent on public engagement, prompt reporting of test findings, and cooperation with isolation measures. The effectiveness of these tactics can be hampered by a lack of public trust, privacy concerns, and delays in result announcement.

Finally, testing and contact tracing have been critical in controlling the spread of COVID-19 in the United Kingdom. Rapid case identification, isolation, and interruption of the transmission chain are key components of pandemic management. Effective implementation necessitates striking a compromise between increasing testing capacity, ensuring timely result reporting, and encouraging public confidence and engagement. As the epidemic progresses, it will be critical to refine these evidence-based tactics and adapt to new difficulties.

3. Vaccination Campaigns

Vaccination initiatives were critical in changing the course of the COVID-19 pandemic in the United Kingdom (UK). As vaccinations became accessible, they provided a strong tool for reducing virus propagation, controlling severe disease, and eventually achieving population-level immunity. Vaccination efforts in the United Kingdom have been phased, with high-risk groups like as the elderly, healthcare professionals, and people with underlying health issues being prioritised. Dagan et al. (2021) found that vaccination of elderly people resulted in a significant reduction in COVID-19 cases, hospitalisations, and severe outcomes, highlighting the importance of prioritisation measures in preventing transmission.

The success of immunisation campaigns in reducing COVID-19 spread is dependent on getting adequate vaccine coverage in the population. Vaccination not only protects individuals directly, but also helps to herd immunity, in which a substantial proportion of the population becomes immune, limiting the virus's capacity to propagate. According to Bubar et al. (2021), achieving high vaccine coverage in the UK significantly reduced transmission and the overall impact of Covid-19. Vaccine distribution issues, vaccine reluctance, and the appearance of novel variations continue to be problems for the effectiveness of immunisation campaigns. Maintaining public trust in vaccine safety and efficacy while resolving logistical challenges is critical for achieving broad immunisation and reducing transmission.

Finally, vaccination initiatives have proven to be a game changer in the control of COVID-19 in the United Kingdom. Prioritising high-risk groups and obtaining high immunisation coverage have resulted in fewer cases and less severe outcomes. Addressing problems and modifying tactics will be critical to ensuring long-term control of the pandemic as the UK continues its vaccination efforts.

2.1.3 Healthcare Systems and Capacity

1. Healthcare Overload

Healthcare overload, defined as overburdening healthcare systems as a result of an increase in COVID-19 cases, has had a significant impact on the virus's transmission in the United Kingdom (UK). As the pandemic spread, hospitals saw difficulties in managing patient influx, affecting both COVID-19 and non-COVID-19 care. High levels of COVID-19 patients might put a pressure on hospital resources, resulting in a shortage of ICU beds, ventilators, and medical professionals. According to Remuzzi and Remuzzi (2020), healthcare overload can lead to higher mortality due to decreased availability to critical care for severe COVID-19 cases. During moments of healthcare congestion, the United Kingdom, like many other countries, was forced to make difficult judgements about resource allocation and triage methods.

COVID-19 patients are not only affected by healthcare overload, but ordinary medical care is also disrupted. Patients who did not have COVID-19 experienced delays in elective procedures, treatments, and screenings, potentially resulting in unfavourable health outcomes. Thornton (2020) conducted a study on the impact of healthcare overload on cancer care in the United Kingdom, emphasising the importance of a balanced strategy to addressing both COVID-19 and non-COVID-19 healthcare demands.To alleviate healthcare overburden, the government responded with lockdowns and public health measures. These actions were designed to flatten the epidemic curve, minimise case counts, and keep hospitals from being overburdened. Davies et al. (2020) found that early treatments in the United Kingdom were beneficial in delaying and lowering healthcare overload, saving lives and maintaining healthcare capacity.

Finally, healthcare overburdening has had a considerable impact on the spread of COVID-19 in the United Kingdom. Overburdened healthcare systems struggle to provide proper treatment for COVID-19 and non-COVID-19 patients, resulting in higher mortality and delayed medical services. Government initiatives to decrease transmission and flatten the epidemic curve have been crucial in preventing healthcare overload. As the pandemic progresses, it is critical to maintain a balance between COVID-19 treatment and ordinary healthcare services in order to protect public health.

2. Telemedicine and Digital Health

The COVID-19 pandemic has accelerated the implementation of telemedicine and digital health solutions in the United Kingdom (UK) as a means of mitigating virus propagation while assuring continuous healthcare delivery. These technology improvements have been critical in the provision of medical care, the maintenance of health services, and the reduction of the strain on healthcare facilities.

Telemedicine, which includes virtual consultations and remote patient monitoring, has become a key tool in the management of COVID-19 patients and the maintenance of ordinary medical care. Telemedicine lowers the need for in-person visits by allowing patients to obtain healthcare services from the comfort of their own homes, hence reducing potential virus exposure. Greenhalgh et al. (2020) conducted research that emphasised the benefits of telemedicine during the pandemic, citing its potential to give timely care, facilitate physical separation, and conserve healthcare capacity. Mobile apps and wearable gadgets, for example, have been utilised for symptom tracking, contact tracing, and public health surveillance. The National Health Service (NHS) of the United Kingdom (UK) launched the NHS COVID-19 app for contact tracing, alerting anyone who have been in close vicinity to a confirmed case. Ferretti et al. (2020) shown that digital contact tracing, when paired with other techniques, can help reduce virus spread. While telemedicine and digital health solutions provide tremendous benefits, there are still limitations. Not everyone has equal access to technology, which may exacerbate health inequities. Furthermore, preserving public trust and protecting sensitive health information requires assuring data privacy and security.

Finally, telemedicine and digital health solutions have been critical in limiting the spread of COVID-19 in the United Kingdom. Remote healthcare delivery, symptom monitoring, and contact tracking are all possible with these technologies, decreasing the need for physical encounters. As the epidemic progresses, incorporating telemedicine and digital health into healthcare systems can help to improve the efficiency and resilience of healthcare delivery while also assisting in pandemic management.

3. Health Inequities

Inequities in health have had a substantial impact on the spread and impact of COVID-19 in the United Kingdom (UK). Differential outcomes among diverse population groups have been attributed to socioeconomic inequality and unequal access to healthcare services. Lower-income individuals, racial and ethnic minorities, and people with underlying health issues have faced a disproportionate burden of COVID-19 instances and severe outcomes. Pan et al. (2020) found that locations with higher levels of deprivation in the UK had higher incidence of COVID-19-related mortality. Factors such as crowded living circumstances, restricted access to preventive services, and unequal distribution of healthcare facilities have worsened health disparities.

Occupational variables have also contributed to health disparities and the spread of COVID-19. Individuals in lower-paying jobs with few alternatives for remote work have been more vulnerable to the virus. Lewer et al. (2020) discovered that workers in low-income jobs were more likely to contract COVID-19, demonstrating how socioeconomic disparities overlap with occupational characteristics. To address health inequality and the development of COVID-19, the government has implemented targeted initiatives, public health campaigns, and immunisation campaigns that prioritise disadvantaged people. Initiatives to address structural inequalities, such as improving access to healthcare, affordable housing, and job opportunities, are critical in addressing the underlying causes of health disparities.

Finally, health disparities have had a considerable impact on the transmission of COVID-19 in the United Kingdom. Vulnerable communities with inadequate resources and access to quality healthcare have had greater incidence of infection and poor outcomes. Addressing health imbalances through targeted treatments and structural changes is critical not only for pandemic preparedness, but also for constructing a more equitable and resilient healthcare system.

2.1.4 Lessons Learned and Future Preparedness

1. Early Interventions

Early efforts were critical in defining the course of the COVID-19 epidemic in the United Kingdom (UK). Rapid deployment of preventative measures, testing, and public health campaigns has been critical in limiting the spread of the virus and mitigating its impact. Among the first strategies intended at reducing COVID-19 transmission were government-enforced lockdowns, travel restrictions, and social distancing measures. Flaxman et al. (2020) discovered that the timing of these treatments had a considerable impact on the pace of virus propagation in the UK. Early and strong actions were linked to fewer COVID-19-related deaths and less healthcare overload. When applied early, testing and contact tracing have also helped to slow the virus's spread. The UK's Test and Trace programme, which sought to detect and isolate patients, was critical in containing outbreaks. Walker et al. (2021) emphasised the importance of timely contact tracing in lowering transmission rates in their investigation.

Public health programmes, such as those promoting hand hygiene, mask use, and immunisation, have aided in raising awareness and promoting adherence to preventive behaviours. Early communication and precise guidance from health officials were critical in changing public behaviour and lowering the danger of transmission. Early intervention, on the other hand, is dependent on public compliance, communication techniques, and effective enforcement. Balancing the economic and social costs of interventions against the health benefits has also presented difficulties.

Finally, early treatments have been critical in slowing the development of COVID-19 in the United Kingdom. Lockdowns, testing, contact tracing, and public health campaigns were implemented quickly, which helped to reduce transmission rates and minimise hospital overburden. As the epidemic progresses, a proactive approach that incorporates evidence-based therapies and adapts to new difficulties will be critical for effective pandemic containment.

2. Data Sharing and Survillance

Data exchange and surveillance have been critical in understanding and managing COVID-19's growth in the United Kingdom (UK). Timely epidemiology and health data collection, analysis, and distribution have enabled informed decision-making, resource allocation, and successful public health responses. Real-time data sharing has enabled health officials to trace the virus's spread, identify potential hotspots, and track changes in transmission patterns. The COVID-19 Dashboard in the United Kingdom, as well as other data systems, provides up-to-date information on cases, hospitalisations, and mortality rates. Kass-Hout et al. (2020) found that transparent data exchange was critical for guiding public health interventions during the epidemic.

monitoring systems such as syndromic monitoring and genomic sequencing have shed light on the virus's evolution and transmission. Genomic sequencing, in particular, has aided in the identification and tracking of variations, allowing for targeted therapies. The COG-UK consortium has been instrumental in sequencing SARS-CoV-2 genomes in the UK, which has aided in understanding transmission patterns. Data sharing and surveillance also aided modelling efforts to estimate the pandemic's trajectory and assess the effectiveness of countermeasures. Davies et al. (2020) found that combining epidemiological data with modelling can help inform decision-making and public health policies. However, issues such as data privacy, standardisation, and data quality continue to be issues. Maintaining public trust and making informed judgements requires ensuring that collected data is accurate, representative, and secure.

Finally, data exchange and surveillance have been critical in understanding and controlling the spread of COVID-19 in the UK. These practises provide important insights into transmission dynamics, direct public health interventions, and allow for more informed decision-making. Maintaining comprehensive data collection, exchange, and surveillance procedures will be critical for efficient pandemic management as the epidemic evolves.

3. Collaboration and Communication

In the United Kingdom (UK), effective teamwork and communication have been critical in overcoming the problems posed by the COVID-19 epidemic. The timely exchange of information, expertise, and resources among government agencies, healthcare institutions, researchers, and the general public has been critical in controlling the virus's spread and mitigating its damage. Cross-sector coordination has facilitated coordinated pandemic responses. Government agencies, healthcare providers, and public health organisations have collaborated to devise strategies, distribute resources, and put interventions in place. Greenhalgh et al. (2020) found that collaborative efforts are critical in creating and implementing public health interventions such as lockdowns and testing programmes. Global collaboration has encouraged knowledge exchange and learning from other countries' experiences. The dissemination of best practises, research findings, and lessons learned has sped up the creation of effective methods and solutions. International alliances such as COVAX have worked to provide fair access to vaccines around the world.

Communication has been crucial in getting accurate information out to the people. Transparent and straightforward information from health officials has aided people in understanding the virus, preventive measures, and vaccine efforts. Allington et al. (2020) discovered the importance of effective communication in increasing public compliance with protective behaviours. However, communication issues include dealing with misinformation and preserving public trust. Misinformation and rumours can stymie efforts to contain the virus's spread. Health communication initiatives should be adapted to specific communities and take cultural and language diversity into account.

Finally, collaboration and communication have been critical in controlling the spread of COVID-19 in the United Kingdom. Coordination of reactions, global collaborations, and open communication have enabled the implementation of effective initiatives and resource mobilisation. As the pandemic progresses, open lines of collaboration and good communication will be critical for mitigating the virus's impact and guaranteeing public safety.

2.2 Additional DRL Related Work

2.2.1 Healthcare

The surge of enormous multimodality data availability has increasingly driven and improves computational models and algorithms, the capabilities of Artificial intelligence (AI) in public healthcare in the last decade (Dilsizian et al., 2014). The proliferation in trend has led to increase in participation in the use of advanced data analytics and machine learning approaches in some healthcare applications (Johnson et al., 2016). RL being a subfield in machine learning does well in theoretical and technical accomplishments in generalization, thus increasing the applications of RL in real-life problems, computer vision, in playing games, business management, robotics control, autonomous driving, financial and natural language (Mnih et al.,2015).

RL techniques have received substantial research in developing effective cancer chemotherapy treatment plans. For the first time, Zhao et al. (2009) used the model-free, temporal difference (TD) technique known as Q-learning to decide how much chemotherapeutic drug to use. Virtual clinical trial data from in vivo tumour growth patterns was quantitatively constructed using the mathematical chemotherapeutic model described by numerous Ordinary Difference Equations (ODE). Support vector regression (SVG) (Vapnik et al., 1997) and extremely randomised trees (ERT) (Ernst et al., 2005) are two explicit machine learning techniques that were used to fit the approximated Q-functions to the given trial data. It was shown that using these batch learning techniques, it was possible to simulate optimal strategies directly from clinical trial data.

Based on an ODE-based tumour growth model given by de (Pillis and Radinsky,2003) and the Natural AC (NAC) technique (Kober and Peters, 2009) for the medication scheduling of cancer chemotherapy, (Ahn and Park, 2011) investigated the applicability of the NAC approach. By constantly injecting drug from the start until the right time, the NAC technique might find an efficient drug scheduling policy by minimising the tumour cell population and the drug amount while maximising the populations of normal and immune cells. This strategy outperformed the conventional pulsed chemotherapy procedure, which delivers the medication periodically over a period of many hours.

The work Hassani et al. (2010), in which naive discrete Q-learning was used, likewise confirmed the superiority of continuous dosage treatment over a burst of dosing treatment. More recently, to create efficient drug dosing regimens for patient groups with various characteristics, Padmanabhan et al. (2017) developed various formulations of the reward function in Q-learning. To estimate Q values in a simulation of an advanced general cancer trial, Humphrey (2017) investigated several supervised learning algorithms, including Classification and Regression Trees (CART), random forests, and modified versions of Multivariate Adaptive Regression Splines (MARS). Another important cancer treatment option is radiotherapy, and several research have used RL methods to create automated radiation adaption procedures (Feng et al., 2018) by changing the portion size during treatment, Jalalimanesh et al. (2017) introduced an agent-based simulation model and Q-learning algorithm to optimise dose calculation in radiotherapy.

Although, most research in Covid-19 are relatively new in machine learning and artificial intelligence, however, there have been several works on reinforcement learning, deep learning, and deep reinforcement learning. Ying et al, (2017) used deep reinforcement learning for Dynamic Treatment Regimes on medical registry data. This deep reinforcement framework involved the use of supervised learning steps to predicting the best action that could be made by an expert by estimating the value function that is long time for the Dynamic Treatment Regimes. The research framework (DRL) was motivated on the Centre for international Bone Marrow Transplant Research registry, the implementations focused on prevention and the treatments of acute and chronic disease. The initial implementation study showed results that accurately predict the diagnosis or decision that could have been made by human experts.

In the intensive care unit (ICU), Sepsis is a leading cause of death and making cost on hospitalization overwhelming. Raghu et al, (2017) were able to use deep reinforcement learning to treat Sepsis in the ICU. The fact that individual response differently to the same treatment, thus there is no specific agreed treatment for Sepsis, However, in Raghu et al study, they use the deep reinforcement learning to finding the optimal strategies from training samples data that are not portray optimal behaviour, thus increasing the survivor rate of Sepsis in the ICU department. Ngo et al. (2018) proposed a reinforcement learning based algorithm to optimize glucose in the blood for patients with type-1 diabetes. The algorithm proposed helps in form of policy to regulate the optimal insulin to be injected into Type-1 diabetes patient. This study used historical 10 years of clinical data of patients that were treated in a general hospital while the technique uses reinforcement learning agent which runs through different states of the patient and explores the response of patient when the patient is provided with insulin of different doses. The reward is a function of the difference between the actual level of glucose in blood in response to the insulin intake and targeted level of glucose.

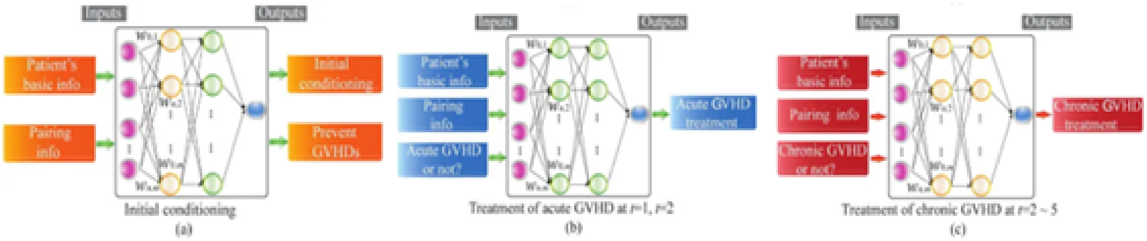


Figure 2.1: Treatment of chronic GVHD-Dynamic Treatment Regimes (DTR): A review of recent reinforcement learning applications to healthcare by Isaac Godfried Towards Data Science

Lin et al. (2020), in their recent research made use of chest computed tomography (CT) scan images to differentiate between COVID-19 and Pneumonia chest images. In Lin et al research work, they extracted different features in the CT images to detect the any chest x-rays effected with COVID-19 by using deep learning (COVNet). The result showed a robust model, haven able to decipher between pneumonia and non-pneumonia chest x-ray images. Five years CT scan images from six different hospital between August 2016 and February 2020 were collected for this study. Sensitivity, specificity, and the area under curve were used to check the performance of the model.

Through the validation, the model achieved an accuracy of 89.5% having specificity of 0.88 and sensitivity of 0.87. Early dynamics of COVID-19 pandemic infections were analysis by Machine language (ML) based on the existing US data from 20 January 2020 when the index case was confirmed (Malik, 2020). Public health insights like the rate of mild infection becoming critical, infectious force, asymptomatic infections estimation, and forecasting new confirmed cases over time by analysing number of infection growth over time. Malik work shows that the proposed technique is efficient and robust making its application to any virus. Yan et al. (2018) explored the audio-based cough assessment. Their work assessed cough frequency, intensity of coughing and properties of coughing sound. Monitoring coughing tool uses machine language cough recognition algorithm for detecting the cough.

These classifier algorithms include support vector machine (SVM), naive Bayesian classifier (Bayesian), neural network (NN), hidden Markov model (HMM), and dynamic time warping (DTW) (Liu and Du, 2009). Hemdan et al., (2020) and Wan et al., (2020) applied COVIDX-Net, a deep learning framework to assist radiologists to automatically detect COVID-19 x-rays images. The COVIDX-Net contains seven different architectures of deep convolutional neural network models, which are the modified Visual Geometry Group Network (VGG19) and the second version of Google MobileNet. Each deep neural network model can analyse the normalized intensities of the X-ray image to classify the patient status either negative or positive COVID-19 case. Marco et al., (2020) used mobile phone tracking data to measure mobility restriction and compare the data to number of new SAR-CoV-2 positive cases daily in three most affected area in Italy.

Another remarkable study was carried out in China where the first index case was established, Chinazzi et al. (2020) explored data from 200 countries and territories using the global epidemic and mobility model (GLEAM) to model travel restriction. The model showed 77% reduction in cases imported from other countries due to travel restriction.

Early lockdown was shown to be more effective than response delayed in China (Tian et al., 2020). Chinazzi, M et al, (2020) research work assessed the efficacy of travel restrictions for different transmission scenarios; travel ban was only meaningful if combined with a 50% or higher transmission reduction.

Taking early decision on locking down and pre-emptive travel restriction was shown in (Tian, H et al.,2020) article were considerably effective than response delayed. Previous studies show that machine learning and other artificial intelligence algorithms are used in identifying diseases based on existing data and making prediction with simulation. The proposed approach in this research work is not the same with the previous literature studied. There would be no supervised learning involved, the algorithm learns by interacting with the environment. Based on the dynamism of the virus, any change in the environment is captured by the RL agent with appropriate reward function.

The epidemiologists and scientists could not explain the behaviour of COVID-19 numerous symptoms with regards to its mutation and how it affected by demographic changes. Thus, this research would consider the affected of the pandemic on the economy, pattern of spread, its effect on quality of life and other parameters into the reword function to enable the RL agent learned the optimal or minimal policy from the data available. In this research work, we would be focusing on finding optimal policy for controlling the spread of the virus, timing and intensity of the policy would be considered, however, the evolving symptoms of COVID-19 are ignored.

2.2.2 Robotics

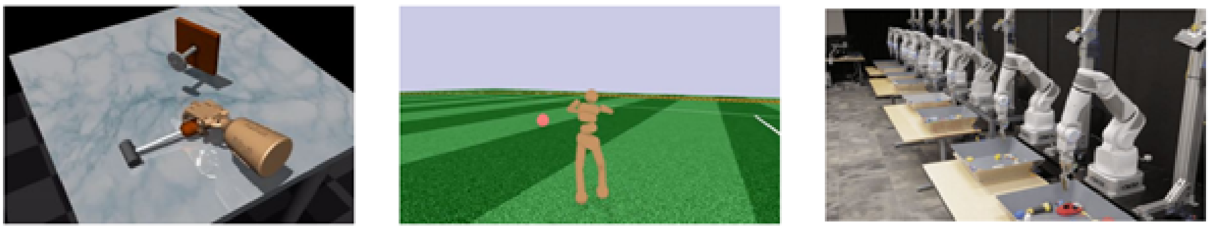


Figure 2.2: TRobotics in RL (source: RL in Robotics.pdf (ieor8100.github.io)

A traditional application of reinforcement learning is in robotics. The end-to-end training of the perception and control systems was suggested by Levine et al. (2016a) to directly transfer raw picture observations to torques at the robot's motors. To overcome the problem that supervised learning typically does not achieve strong, long-horizon performance, GPS alternates between trajectory-centric RL and supervised learning, obtaining the training data from the policy's own state distribution.

Pre-training is used by GPS to cut down on the quantity of experience data needed to train visuomotor policies. On a variety of real-world manipulation tasks requiring localization, visual tracking, and managing complex contact dynamics, as well as in simulated comparisons with earlier policy search techniques, good performance was attained. This is the first technique that can teach profound visuomotor strategies for complicated, high-dimensional manipulation abilities with direct torque control, according to the scientists (Levine et al., 2016a). By maximising cumulative reward while solving an RL problem and considering un/self-supervised tasks to increase data efficiency and task performance, Mirowski et al. (2017) was able to develop the navigation skill. Unlike traditional methods like Simultaneous Localization and Mapping (SLAM), which use explicit location inference and mapping for navigation, with this method, navigation is a by-product of the goal-directed RL optimisation problem. This could displace the well-liked SLAM, which typically requires manual processing.

2.2.3 Computer Vision

Computer vision is the study of how computers interpret digital photos or videos. The following sections cover recognition, motion analysis, scene understanding, integration with Natural Processing Language (NLP), and visual control after providing background information on computer vision. In tasks like object segmentation, articulation model estimation, object dynamics learning and haptic property estimation, object recognition or categorization, multimodal object model learning, object pose estimation, grasp planning, and manipulation skill learning, reinforcement learning would be a crucial component of interactive perception (Bohg et al., 2017). By concentrating just on the prominent components, RL can increase classification efficiency for images. RL can be more effective for visual object localization and detection than methods involving exhaustive spatial hyporeport searching and sliding windows because it strikes a balance between sampling more areas for greater accuracy and ceasing the search once the target's location is known with sufficient certainty.

2.2.4 Robotics

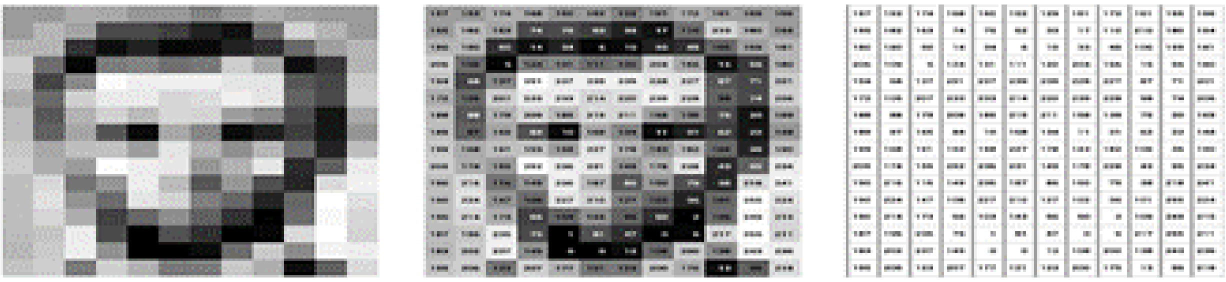


Figure 2.3: Computer Vision (source: Towards Data Science

To focus on a chosen series of regions or places from an image or video for image classification and object recognition, Mnih et al. (2014) presented the recurrent attention model (RAM). The model was trained using reinforce to get over the non-differentiability problem, and experiments were conducted on an image classification job and a dynamic visual control problem. By transforming a bounding box using transformation actions to find the target objects' most precise location, Caicedo and Lazebnik (2015) suggested an active detection model for object localization with DQN.

By maximising the long-term reward associated with localization accuracy over all objects with DQN, Jie et al. (2016) suggested a tree-structure RL strategy to search for objects sequentially while considering both the current observation and prior search pathways. Mathe et al. (2016) suggested using policy search to recognise visual objects. For collaborative object search, Kong et al. (2017) used cooperative multi-agent RL with inter-agent communication. For the categorization of multi-label images, Welleck et al. (2017) suggested a hierarchical visual architecture with an attention mechanism. For video face recognition, Rao et al. (2017) suggested an attention-aware deep RL algorithm.

2.3 Systematic Literature Review

Based on the review of previous studies, this section performed a systematic review of only those studies that seemed to apply deep reinforcement learning technique to finding the optimal timing for both lockdown and travel restriction in the United Kingdom.

2.3.1 Methods

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) reporting guidelines (Page et al., 2021) guided the selection of this research. PRISMA recommends a checklist of 27 components for a systematic literature review's sections and content, as well as a four-phase flowchart for article selection. In order to meet the requirements of this research project, the PRISMA guideline was modified. The flowchart for paper selection consists of three stages: identification, screening, and eligibility.

- Identification: This stage comprises identifying information sources and a search strategy that yields a list of possible publications. The number of articles selected during the identification phase is based on the keywords' deep reinforcement learning AND optimal policy , AND lockdown AND travel restrictions, AND optimal strategy AND appropriate timing. Screening means choosing articles from the first phase and getting rid of the ones that don't fit with the study's goal.
- Eligibility: The third phase comprises a more in-depth examination of publications and the selection of those that are relevant to the research themes., displays the number of papers at each level and their subsequent progression.

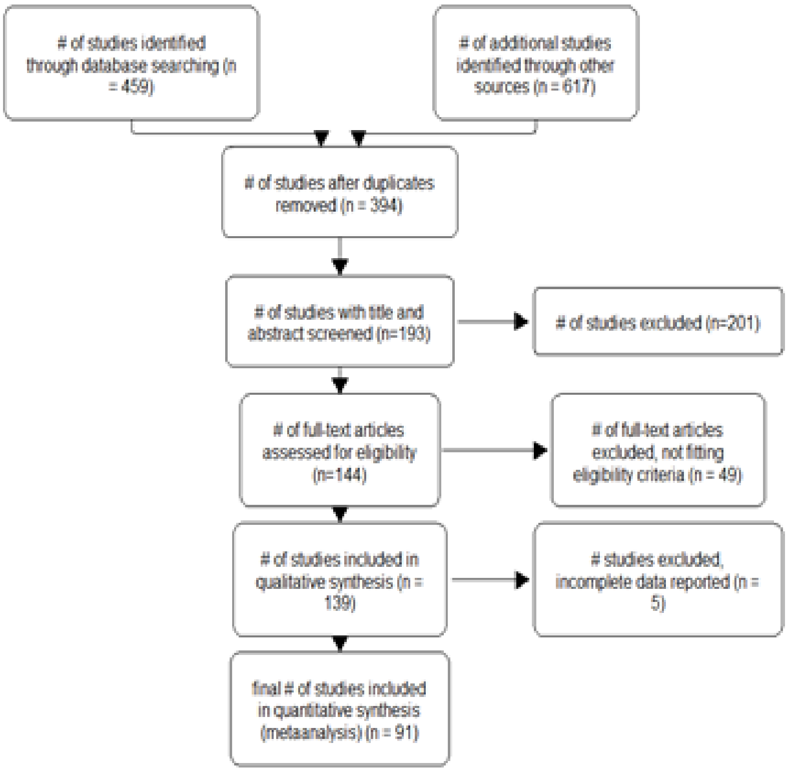

# of studies identified through database searching (n = 459)
# of additional studies identified through other sources (n = 617)
# of studies after duplicates removed (n = 394)
# of studies with title and abstract screened (n=193)
# of studies excluded (n=201)
# of full-text articles assessed for eligibility (n=144)
# of full-text articles excluded, not fitting eligibility criteria (n = 49)
# of studies included in qualitative synthesis (n = 139)
# studies excluded, incomplete data reported (n = 5)
final # of studies included in quantitative synthesis (metaanalysis) (n = 91)

Figure 2.4: The three stages of the study selection process: identification, screening, and eligibility

Selection: The next stage is to choose information sources for conducting literature searches. Despite the fact that there are several search engines and academic databases, the research focused on scholarly and research databases, including places where data mining methods for crime prediction and social media usage can be acquired.

This research studied an optimal search process that integrates several academic search databases, with searches done at the greatest attainable level of specificity. According to Bramer et al. (2017), if the research topic is more interdisciplinary, a bigger scientific database like Web of Science is likely to be beneficial. Nevertheless, according to this research, Scopus is a bigger database than the Web of Science. Scholarly opinion is divided regarding the usefulness of Google Scholar for multidisciplinary research. According to

Sirotkin (2013), some search engines, particularly Google Scholar, have a predisposition to selectively reveal information by using algorithms that personalise content for users.

This phenomenon is known as the filter bubble effect. Haddaway et al. (2015) revealed that while looking for certain papers in Web of Science, the majority of the results were also found in GS. When similar search terms were used in Web of Science and GS (10-67 percent overlap), GS missed some crucial information in five out of six case studies, according to their results. Consequently, publications that are unavailable through Google Scholar or Scopus were collected through the library of the University of Salford. The identified papers from each database were imported straight into the free citation manager Endnote. Finally, duplicates were removed from each database, resulting in 394 papers for the screening step.

As a result, items older than ten years old were excluded during the screening process. Second, duplicates in the selected datasets were eliminated. Finally, each article was scrutinised to eliminate those judged "irrelevant." This was accomplished by defining "relevant" articles based on the three criteria listed below. The first need is that the record cover lockdown and border control in the UK that uses DRL as an agent to finding the appropriate timing. Papers dealing with DRL ,lockdown,travel restrictions, optimal policy or strategy. The second need for research to be regarded "relevant" is that it use an exploratory or cluster analysis data mining technique. The third component is the use of COVID-19 epidemic data.

The final step in the screening process was to exclude extraneous articles that the authors couldn't read owing to subscription constraints. The screening process yielded 193 items that were appropriate for the third and final rounds. The abstracts and main body of each of the 193 papers were inspected and analysed during this final step, the eligibility phase (e.g., study field, data, methodology, and findings). The goal was to collect information for the paper's qualification requirements. These are divided into three categories: Procedures; significance: the data item's relevance and purpose; and research characteristics: data item study area, data source, data mining methodologies, and evaluation metrics. Following that, each category and its associated data pieces were scrutinised. The first data point indicates the type of publication. Conference papers are occasionally excluded from literature reviews because their quality is not assessed in the same way as International Scientific Indexing (ISI) publications.

However, in particular subjects, such as computer science, several conferences are acknowledged as extremely credible publishing channels. During the screening procedure, a large number of papers prepared by professionals in computer or information science were discovered; so, conference papers were not rejected at this time. In total, seven novels or book parts were eliminated. The next two "relevance" criteria (i.e., relevance and purpose) address the papers' content's conformity to the subject of this research. The document's relevance was double-checked at this stage. Several articles that appeared relevant throughout the screening process . Despite the fact that lockdwon and travel restrictions were mentioned in the abstract, the authors acknowledged to the fact that they were providing a framework or survey for future research. This research effort now includes the data item "data sources" in order to combine methodologies for modelling and assessing correlations between dependent and independent variables (e.g., optimal policy). Because of these criteria, 49 papers were excluded.

Finally, four additional quality and consistency criteria for the selected articles were offered. Nonetheless, as detailed in the Results section, there are significant discrepancies amongst the research. The final two conditions are the limitation to empirical data analysis (e.g., no proof-of-concept or purely methodological study) and the use of measurements to evaluate model performance (e.g., accuracy, precision). The latter two criteria ensure that we only review publications that are relevant to this investigation. There were 53 manuscripts that were rejected due to criteria relating to their research relevance.

2.3.2 Study Quality

Before each phase, the chief supervisor and co-supervisor of this research assessed the publications that composed this study. This research was further analysed and argued until all parties reached an agreement on the next step. To ensure methodological consistency, the study's results were subjected to multiple cross-checks. Throughout the final step (eligibility), the research student assessed the papers several times to ensure that all eligible publications were included. Concerning the results subsections of the four study stages (”Study characteristics,” ”Overview of selected publications on the DRL on COVID-19 ” ”using DRL to finding appropriate timing for both lock down and travel restrictions in the UK and epidemic data,” and ”factors to consider when analysing optimal policy”).

To extract information organised as data items, a three-step approach was used: extract, discuss, and analyse. First, the authors manually extract the data components and values from the articles by reading them (1-extract). The researcher and supervisors then discussed and evaluated the data items and their values (2-discussion/consensus). If the information remains ambiguous, it is compared to freely available information for clarity (3-analysis). This data was formatted as a matrix, with rows representing papers and columns indicating various processing data (for example, a data item is the year of publication). The study scale was used to assess the risk of bias in each study. As stated throughout the eligibility method, regional and temporal constraints were established to ensure that we analyse medium-to-large-scale research and that the results are not biassed by location or season. Furthermore, we found neither duplicate publications (two or more articles with identical samples and procedures) nor any study characteristics, such as unique and unusual attributes or research topics.

2.3.3 Results

2.3.3.1 Study Characteristics

This inquiry provided 81 publications in the end. Approximately 89% of the articles chosen provides United Kingdom trends using COVID-19 epidemic data, whereas approximately 19 of the total publications (11%) only look at the deep reinforcement learning based on simulated data.

2.3.3.2 Deep Reinforcement Task

Out of the total of 69 publications analysed, about 32 of them exclusively utilised COVID-19 data and implemented categorization algorithms in order to allocate resources. Twelve distinct Deep Reinforcement Learning (DRL) techniques were employed in the study, however using simulated data. Conversely, the remaining approaches employed neural network applications to predict patterns in the variables of recovery, death, and confirmed cases.

2.3.3.3 Analysis and Discussion

As previously mentioned, the selected papers encompass the authors' proposed baseline models for machine learning, reinforcement learning models, or deep reinforcement learning models. The evaluation metrics for these models are widely recognised in the field of deep reinforcement learning, as evidenced by previous studies on reinforcement learning and deep reinforcement learning. However, in addition to optimising policy, few writers underscore the importance of integrating or implementing many evaluation measures.

Comparing the evaluation findings of the 69 articles proves to be a tough task, mostly due to several elements that contribute to the complexity of the task. These factors include the presence of varied numbers of variables, diverse study fields, and different methodologies employed in data mining. The examination of their commonalities may be deemed necessary. In the context of classification and forecasting tasks, accuracy (n = 18) and precision (n = 8) are the evaluation measures that are commonly employed. The evaluation measures that are commonly employed include Mean Squared Error (MSE, n = 2) and Root Mean Squared Error (RMSE, n = 2). The investigation focused on evaluating model performance by utilising the top two assessment criteria, namely accuracy and precision. Publications that utilise reinforcement and deep reinforcement learning employ various performance metrics, such as learning curve, average reward, discount return, and comparisons between baseline and policy evaluation.

2.4 Summary of the Findings

The primary objective of this study is to employ Deep Reinforcement Learning (DRL) as an agent in order to determine the most effective policy for implementing lockdown measures and travel restrictions based on data pertaining to the COVID-19 epidemic. In order to gain a thorough understanding and evaluate the current status of empirical research on DRL with a specific emphasis on lockdown, a complete literature review was undertaken using the reporting requirements known as ”PRISMA” (Liberati et al., 2009). When considering data mining methods, authors commonly presented conventional machine learning techniques and, to a lesser extent, deep learning approaches. The comprehensive review revealed that the performance of the majority of deep reinforcement learning (DRL) models exhibited inconsistency.

Furthermore, it was shown that the performance of these models occasionally differed based on the specific dataset and pre-processing techniques employed. A range of performance evaluation metrics were employed, with prediction accuracy, precision, and recall emerging as the three most prominent metrics. Ultimately, the practise of partitioning data into train and test sets emerged as the prevailing method for evaluating models. However, researchers also employed cross-validation and bootstrapping techniques for this purpose.

Furthermore, essential aspects of research inquiries were provided in a confusing manner or omitted entirely. With regards to the final point, we propose providing the subsequent details: study area, dimensions, duration of sampling, specific months, classification type, characteristics of the samples, techniques employed for feature engineering and selection, data pertaining to class imbalance problem, exploratory analysis conducted, and temporal unit. This encompasses a total of 10 elements. Additionally, it is necessary to include a proposed methodology, a recommended methodology, a baseline methodology, evaluation metrics, and a validation methodology.

Chapter 3

Research Methodology & Technical Framework

3.1 Introduction

This chapter of the research introduces the reader to the models that will be studied in greater depth in later sections, with a focus on their prevalence and utility in data mining, as well as their potential for application in DRL as an agent to determine optimal policy. They were chosen because they were the most common DRL learning models discovered in earlier studies employing a data mining method to combat COVID-19 spread.

3.1.1 Deep Learning

"Shallow" learning contrasts with deep learning. There is an input layer and an output layer for many machine learning techniques, such as logistic regression, support vector machines (SVMs), decision trees, and boosting predictive model. The inputs may be manually changed prior to training via feature engineering. Between the input and output layers in deep learning, there may be one or more hidden layers. We compute the input to each unit at every layer except the input layer as the weighted sum of units from the previous layer. Then, we typically apply a nonlinear transformation or activation function, such as logistic, tanh, or more recently, rectified linear unit (ReLU), to the input of a unit to obtain a new representation of the input from the previous layer. We have weights on the connections between the units at each tier. We can compute error derivatives backward and backpropagate gradients towards the input layer after calculations flow forward from input to output, at output layer and each hidden layer, so that weights can be adjusted to optimise some loss function.

A feedforward deep neural network, also known as a multilayer perceptron (MLP), uses a mathematical function made up of numerous smaller functions at each layer to transfer a collection of input values to output values. A feedforward deep neural network having convolutional, pooling, and fully connected layers is called a convolutional neural network (CNN). CNNs are modelled after simple cells and complex cells in visual neuroscience and are designed to process data with multiple arrays, such as colour image, language, audio spectrogram, and video. They take advantage of the properties of such signals, including local connections, shared weights, pooling, and the use of many layers (LeCun et al., 2015). By including shortcut connections to learn residual functions with reference to the layer inputs, ResNets (He et al., 2016d) are intended to make the training of very deep neural networks easier. With hidden units to preserve the history of previous components, a recurrent neural network (RNN) is frequently used to process sequential inputs like speech and language, element by element. When unfolded during forward computing, an RNN can be thought of as a multilayer neural network with all layers sharing the same weights. The gradient may disappear, and it is difficult for RNN to maintain information for a very long time.

LSTM and gated recurrent units (GRU) with gating techniques to modify information through recurrent cells were proposed to overcome such problems (Hochreiter and Schmidhuber, 1997; Chung et al., 2014). All the deep neural networks can be trained using gradient backpropagation or one of its variations. Dropout is a regularisation technique that trains an ensemble of sub-networks by randomly eliminating non-output units from the parent network (Srivastava et al., 2014). The purpose of batch normalisation (Ioffe and Szegedy, 2015) is to expedite training by minimising internal covariate shift, i.e., the change in parameters of previous layers will modify the distribution of inputs for each layer. To recover the compositional hierarchies in various natural signals, deep neural networks automatically train representations from raw inputs. Higher-level features are made up of lower-level ones, for example, the hierarchy of objects, parts, themes, and local combinations of edges in images.

A key concept in deep learning is distributed representation, which states that several features may represent one input and that each feature may represent a variety of inputs. The exponential difficulties posed by the curse of dimensionality are offset by the exponential benefits of deep, dispersed representations. End-to-end training, as used by AlexNet (Krizhevsky et al., 2012) with raw pixels for image classification, Seq2Seq (Sutskever et al., 2014) with raw sentences for machine translation, and DQN (Mnih et al., 2015) with raw pixels and score to play games, refers to a learning model that uses raw inputs without manually creating features to produce outputs.

3.1.2 LSTM Network

To train deep learning models for multi-step-ahead prediction for time series prediction, the original time series must be converted into a state-space vector. According to Taken's Theorem (1980), the reconstruction can accurately capture key characteristics of the original time series. As a result, given an observed time series x(t), one can create an embedded phase space $Y(t) = [(x(t), x(tT), \ldots, x(t(D1)T)$ where T is the time delay, D is the embedding dimension and N is the length of the original time series. To effectively employ Taken's theorem for reconstruction, suitable values for D and T must be chosen. Taken's demonstrated that $D = 2d + 1$ would be sufficient if the initial attractor has dimension d (Takens, 1980).

Recurrent neural networks (RNNs) can be classified as a variant of long short-term memory (LSTM) networks. The output generated in the preceding phase is thereafter utilised as the input for the ongoing step of a RNN. The LSTM model was developed by Hochreiter and Schmidhuber in 1997. The paper discussed the problem of long-term reliance in RNNs, wherein the RNN is capable of predicting words based on current input but struggles to predict words that are stored in its long-term memory. The effectiveness of RNNs decreases as the length of the gap rises. By default, LSTM has the capability to retain information for extended periods. The tool is employed for the processing, forecasting, and classification of time-series data.

RNNs equipped with LSTM architecture have been specifically designed to effectively process sequential data, encompassing time series, speech, and textual information. LSTM networks are well-suited for many applications such as language translation, speech recognition, and time series forecasting due to their ability to effectively capture and learn long-term dependencies within sequential data. The network faces difficulties in learning long-term dependencies due to the presence of only one hidden state in a conventional RNN. The aforementioned problem is effectively addressed by LSTM models by the incorporation of memory cells, which serve as reservoirs capable of retaining information over extended periods. The memory cell is regulated by three gates, namely the input gate, the forget gate, and the output gate. The function of these gates is to ascertain the appropriate data to be inputted into the memory cell, extracted from it, and afterwards outputted.

The input gate is responsible for managing the data that is added to the memory cell. The forget gate is responsible for controlling the information that is wiped from the memory cell. In addition, the output gate serves to regulate the data that is emitted from the memory cell. LSTM networks provide the capability to acquire knowledge of long-term dependencies by making informed decisions regarding the retention or elimination of information as it traverses the network. By employing a stacking technique, it is possible to construct deep LSTM networks that possess the capability to identify very complex patterns within sequential data.

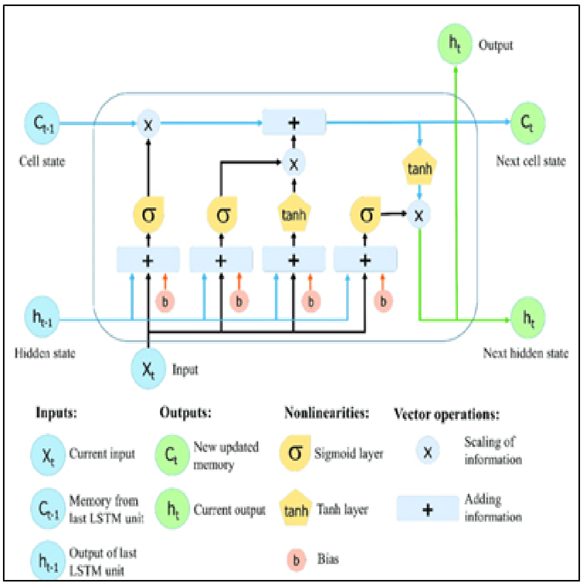


Figure 3.1: The structure of the LSTM neural network: reproduced from (Yan et al., 2019).The LSTM cell processes information in sequential time steps. At each time step, x_t produces output h_t while retaining and updating the cell state C_t.

A hidden state output is calculated by the LSTM network model and h_t by:

$$i_t = \sigma(x_t U^i + h_{t-1} W^i) \tag{3.1}$$

$$f_t = \sigma(x_t U^f + h_{t-1} W^f) \tag{3.2}$$

$$\tilde{C}_t = \tanh(x_t U^c + h_{t-1} W^c) \tag{3.3}$$

$$C_t = \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t) \tag{3.4}$$

$$h_t = \tanh(C_t) * o_t \tag{3.5}$$

where the terms input, forget, and output gates, respectively, at time t, are denoted by

i_t, f_t, and o_t. Memory cell c is referred to here. The terms x_t and h_t, respectively, stand for the quantity of input features and quantity of hidden units. W and U are the weight matrices that are modified together with b, which is the bias, during learning. Keep in mind that each gate has the same proportions. dh is determined by the hidden state's size. $\tilde{C}_t$ is the present cell memory and is the intermediate cell state. $C_0 = 0$ and $h_0 = 0$ provide the initial values at $t = 0$. Note that we indicate element-wise multiplication with a star $(*)$.

A memory cell is added to the hidden layer of an RNN deformation structure called an LSTM to constrain the timeline data's memory data. Data is transferred between various cells of the hidden layer by a few adjustable gates, allowing control of the memory while ignoring the volume of the prior and current data. The status of the memory cell is managed by two LSTM gates. The first is the forget gate, which demonstrates how much "memory" from the cell's final instant can be saved; the second is the input gate, which establishes how much current-time information can be kept in the cell's status and controls the fusion ratio of "old" data and "present" incentive. Finally, the output gate of the LSTM is intended to regulate the volume of outputs of cell status information.

Here, the linear and non-linear activation functions of the LSTM are utilised. The non-linear tanh function is employed by the conventional LSTM. Choosing the most effective activation function for the best results

$$f(x_t) = \frac{exp(x_i)}{\sum_j exp(x_j)} \tag{3.6}$$

Deep learning encourages RL improvement, and the deep reinforcement learning (DRL) domain is described by DL methods inside RL. Deep learning enables RL to scale up previously intractable problems, such as high-dimensional state settings, interruption zones, and decision-making. Every reinforcement learning function, including the value function, Q function, transformation system, and reward function, is approximated by deep reinforcement training using a deep neural network. An RL system called Q-Learning determines which action an agent should do based on an action-value role. Depending on an action-value role, this establishes the relevance of continuing in a specific state and completing a specific task there. By creating an algorithm to reduce temporarily deviating from policy, reinforcement learning has made one of its most significant advances. Using Q-Learning, a target system's state-action value function is measured to determine the highest value in selecting the action. When given a current state (S) and an action (A), function Q provides an estimated reward for that action in that state. As a result, Q functions start off by giving arbitrary fixed values before looking into the situation.

States are regarded to be collections of high, moderate, and low, and decisions are made in response to the states. Here, the decision is made in accordance with the growth in the number of confirmed and fatal cases. When the predictions of verified cases and death cases are accurate, a reward is given. The action-value function defined as:

$$q(s, a; \theta) \approx Q^*(s, a) \tag{3.7}$$

where is a reference variable that represents the edge weights of the network and q (s, a) denotes the neural network approximation. The neural network receives a state as its input and produces approximation q values for unrelated activities Q = q(s, a, θ)| $a \in A$s. By depreciating Q (s, a; θ) anticipated faults, the system is trained. The DRL agent acting at time t is a_t = argmax_aq(s_t, a; θ), where q (s_t, a; θ) for different activities is provided by the network's results. If, for instance, the state changes to s_{t+1} and the compensation is r_{t+1}, then different actions are offered by the quantum neural networks (QNN) outputs. If the state changes to s$_{t+1} andtherewardis$r$_{t+1}$, then the equation (s_t,a_t,t_{t+1},s_{t+1}) establishes an "experience sample" that might be used to train the network. The prediction error of the network for the specific experience sample (s_t,a_t,r_{t+1},s_{t+1}) is specified for training as:

$$L_{st}, a_t, r_{t+1}, s_{t+1}(\theta) = (y_{rt+1}, s_{t+1} - q(s_t, a_t; \theta))^2 \tag{3.8}$$

Where θ = weights; and ($y_{t+1,s_{t+1}}$)= targetoutput, which is defined as:

$$(y_{rt+1}, s_{t+1}) = r_{t+1} + \gamma max'_a q(s_{t+1}, a'; \theta) \tag{3.9}$$

MLSTM Workflow

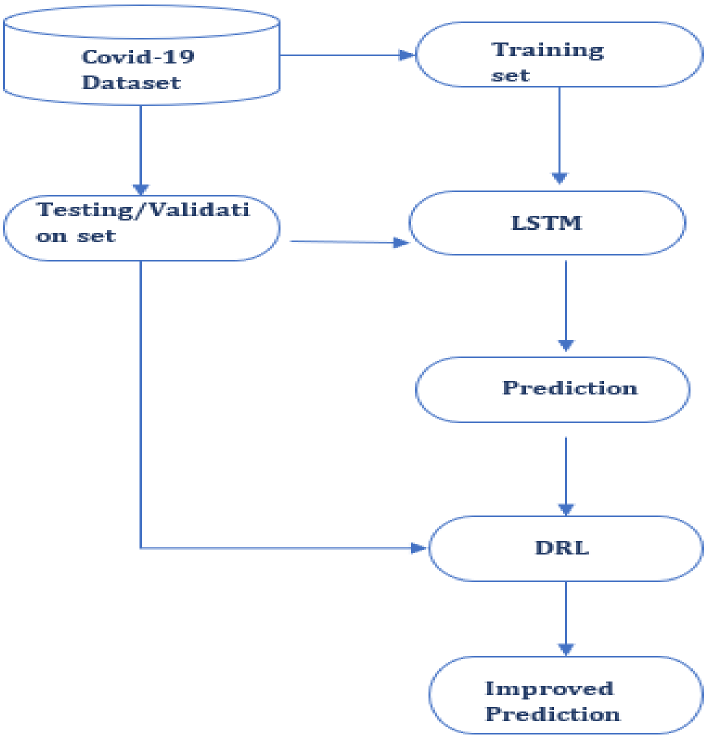


Figure 3.2: MLSTM Workflow

Figure 3.2 illustrates the procedural steps undertaken to execute the predictive component of the research study. Deep reinforcement learning was employed as an extension of the

Long Short-Term Memory (LSTM) model in order to enhance the accuracy of predicting recovery, confirmation, and mortality situations.

3.1.3 Differentiable Computational Graphs

A group of techniques known as deep learning are used to build parametric differentiable computation graphs and train them using gradient-based optimisation. In this section, we go through a few of the revolutionary aspects of deep learning. An artificial neural network (ANN), also known as a deep learning model, is a network of parameterized functional modules that are typically arranged in layers to enable computation to be carried out from the input levels to the output layers. Units or neurons are the names for the values kept in each layer. Theorem of Universal Approximation: According to the universal approximation theorem, a neural network with at least one hidden layer can arbitrarily well estimate any continuous function by changing the number of units. Furthermore, more hidden layers enable networks to do more intricate hierarchical calculations. Deeper networks are typically more effective but more challenging to train. From this potent depth dimension originates the word "deep" in deep learning.

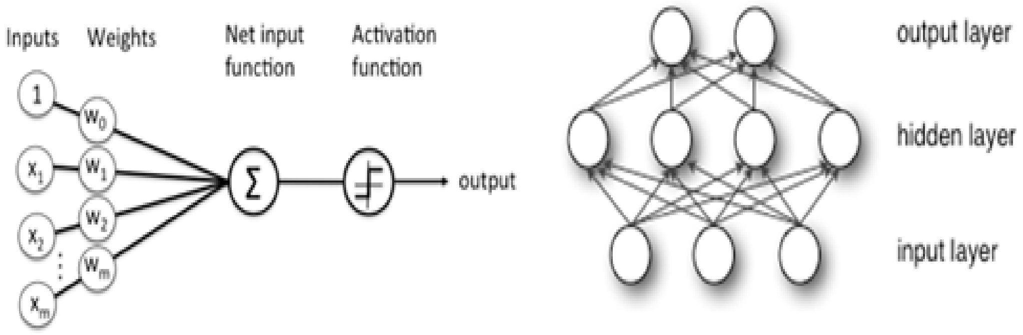


Figure 3.3: Neural network

3.1.4 Optimisation

Backpropagation

Gradient-based optimisation, particularly gradient descent, and its variants, is the foundation of deep learning. An initial forward pass through the network is carried out with a batch of inputs to carry out gradient descent. The outputs are produced, and the loss is calculated using them. The list of partial derivatives, or the gradient of the loss, must then be computed with respect to each network parameter. Using the backpropagation algorithm, also known as backprop, which uses the chain rule and computes the gradient one layer at a time, iterating backward from the last one, this is efficiently accomplished in a single backward run.

Automatic Difference

The derivative function of each layer must be known to employ the chain rule; older implementations of the backpropagation technique required the user to provide each of these. Fortunately, automatic differentiation, also known as autograd, is a feature of all contemporary deep learning frameworks that automatically determines the derivatives by identifying the main computation units employed, such as sums, products, exponentials, powers, or sine functions, and using their known derivatives.

Optimizers

To hasten convergence, simple gradient descent can be improved. For instance, momentum produces updates by linearly adding the most recent update to the gradient of the current update. The inertia of the generated updates enables them to escape flat areas of the search space and overcome oscillations of noisy gradients. Updates can also be made using additional techniques, such as those based on second order derivatives. There is a trade-off between sample efficiency and clock time when using some optimisation approaches because they use a lot of memory and compute. AdaGrad (Duchi et al., 2011), RMSProp (Tieleman Hinton, 2012), and Adam (Kingma Ba, 2014) are notable instances of optimizers.

Batches

Gradient descent reduces the value of a loss function when the gradient is calculated over the entire dataset, in theory, provided a sufficiently low learning rate. A dataset is rarely fully utilised for each training stage. An The simple fact that this is rarely feasible due to the excessive memory requirement is the obvious cause. It would simply require more memory than is typically available on computers to feed a whole dataset with potentially millions of instances, compute the neural activations for each example in parallel, and execute backpropagation. Stochastic gradient descent is typically used in practise. It uses repeated gradient descent steps with smaller data batches to approximate the gradient descent of the entire batch. It is also important to note that due to the parallelization possibilities of hardware and software, batches of inputs typically enable faster inference or backprop per input than for individual calls.

3.1.5 Layers

The fundamental units of neural networks are layers. By stacking them, networks become deeper and more expressive.

Skip connections

Deep neural network becomes vanishing gradient problem, in which gradients get smaller with depth, frequently affects deep neural networks. Some activation functions emphasise this by squashing the values and lowering gradients. There are certain methods to get around this problem, like skip connections, where the output of earlier layers is used at different points. The residual neural network, or ResNet, is the most prevalent type of network using skip connections. ResNet adds the output of a block of layers to the input of the same block to keep information and enable shorter channels for gradients to pass through.

Activation Functions

Most currently used layers convert their inputs along basic linear paths. After such layers, nonlinearities must be exploited to construct strong functions. Deep neural networks are more expressive than linear transformations due to the alternation of linear parametric operations with nonlinear (typically nonparametric) transformations. The ReLU (rectified linear unit), ELU (exponential linear unit), sigmoig (logistic), tanh (hyperbolic tangent), and softplus are the most widely used activation functions.

Fully Connected Layers

The fully connected layer, sometimes referred to as dense or perceptron, is the most basic sort of layer. One hyperparameter is necessary: the quantity of output units. It has two parameters: a bias vector b and a weight matrix W. The output tensor Y is produced as follows: Y = f (W.X + b) given an input tensor X and an activation function f . A stack of entirely connected layers is known as a multilayer perceptron (MLP). Neurons in the visual brain serve as a loose inspiration for a 2D convolutional layer. Over a 2D feature map, it applies a kernel with a receptive field. A tensor with width, height, and depth is a 2D feature map. The number of channels is another name for the depth number. In contrast to colour RGB images, which have three channels for red, green, and blue, grayscale images only have one channel. A kernel with width, height, and depth defines a convolution when it is moved over an input feature map to produce an output feature map. By applying the same kernel in different locations across the input feature map, the resulting pattern matching is translation invariant.

Recurrent Layers

To allow neural networks to process sequences of input data and retain information, recurrent layers have been introduced. They all rely on some internal state being carried over between unroll steps, and mostly have one hyperparameter specifying the size of the generated output. The simplest form of recurrent neural network (RNN) takes in input both the current input and the previous output, concatenates them, and generates a new output using an internal dense layer. This output is then used itself with the next input to generate the next output, and so on. The sequence of inputs and outputs can be used in various ways depending if the task at hand is many to one (only the last output is sent to the next layer), one to many (the same input is repeated multiple times in input, and the many outputs are sent to the next layer), or many to many (at every unroll step, a different input is provided, and the generated output is sent to the next layer).

In any case, the last output generated by a simple RNN is the result of several passes through the same dense layer. This makes the retention of information increasingly harder for the network, and the unroll of a recurrent layer is effectively the same as a deep network with a repeated layer, creating issues with vanishing gradients once again. To overcome this issue, long short-term memory (LSTM) (Hochreiter Schmidhuber, 1997) and gated recurrent unit (GRU) (Cho et al., 2014) layers have been proposed. They rely on gating mechanisms involving multiple internal dense layers controlling the amount of forgetting and addition performed at every unroll step, drastically increasing memory capacities and limiting vanishing gradient issues.

Attention Layers

The padding, which indicates how many pixels were added to the input feature map's borders, and the dilatation. The term "attention" has been used to describe techniques that allow certain elements of a huge amount of information to be highlighted in the deep learning literature. Currently, queries, keys, and values are used as the primary types of attention by the well-known Transformer network (Vaswani et al., 2017). A possible distinct collection of unordered vectors is projected into query vectors on one side while an unordered set of vectors is projected into key and value vectors on the other. After exponentiation and normalisation, scalars produced by the dot product of the keys and queries yield coefficient weights.

To create a new set of values, these weights are applied to the relevant value vectors in a weighted sum. The new vectors then aggregate data from various areas of the input sets. Transformers were first used in translation jobs, where the source language's text was represented by the first input set and the target language's text was represented by the second input set. An very efficient method of executing the translation without the need for memorising is to pay attention to both input sequences. Without requiring recurrent networks to iteratively encode certain sequences, attention attends directly to portions of the inputs. Being able to manage dynamic set sizes and producing "soft weights" that fluctuate with the inputs, these techniques offer a major paradigm breakthrough in deep learning.

Probability Distribution

Many neural networks use probability distributions as a final step. For instance, classifiers, text or image generators, and policies all make use of them. logical operation to establish the most basic type of probability distribution,

Logical Function

A single scalar x, commonly known as "logit," is used to represent the probability of one result and is passed via the sigmoid function 1/ (1 + exp(-x)). A big negative logit represents a chance of 0, whereas a large positive logit denotes a probability of 1.

Softmax

A softmax is a logistic function that generalises to many classes. A logit vector x is used to model the probability of classes after being put through a normalised exponential function. A class i likelihood is given by exp(xi)/ $\sum jexp(xj)$.

Gaussian

A multivariate Gaussian distribution, also known as a multivariate normal distribution, can be utilised for continuous spaces. It is a higher-dimensional generalisation of the one-dimensional (univariate) normal distribution. The covariance matrix of its probability density function, which depicts how the distribution spreads across its dimensions, has a distinctive bell-shaped shape that is centred on its mean. The covariance matrix must be positive definite, which makes it challenging for a neural network to represent. Because of this, most multivariate normal distributions are constructed as diagonal distributions, where the covariance matrix is zero everywhere but the main diagonal (scale). Each dimension becomes independent as a result. So, the mean and scale of a diagonal Gaussian are each represented by a separate vector. The term ”standard normal distribution” refers to a specific instance of this distribution when the scale vector is filled with ones.

Reparameterization Trick

The generated stochastic samples do not directly allow gradients to flow through a neural network when a probability distribution is applied somewhere in the centre of the network. The reparameterization approach can be used to fix this problem. It is based on rewriting the sample generation process as a differentiable function of the network's deterministic values and a random sample supplied by an external distribution of probabilities. For instance, a sample from a diagonal Gaussian distribution can be produced by combining a mean vector µ and a scale vector produced by a neural network with a random noise vector , produced by a multivariate standard normal distribution. This sample has the formula y = µ + σ ϵ and permits gradients to pass across the points µ and σ.

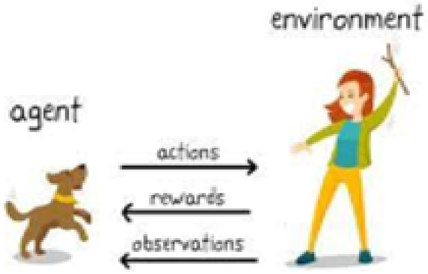


Figure 3.4: Reinforcing a dog:The dog is traduced to different actions , while the dog observes and does the actions as commanded, the dog is rewarded with treat.

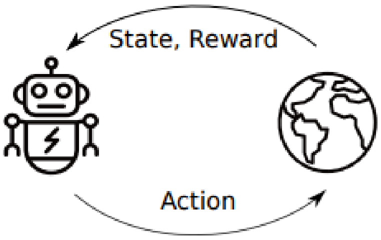


Figure 3.5: Action-Perception loop. The agent acts in a way that has an impact on the environment. The environment releases a reward, which the agent then observes along with the environment's present condition. www.Flaticon.com icons created by Good Ware and Turkkub.

3.2 Reinforcement Learning

The branch of machine learning that deals with sequential decision-making is called reinforcement learning (RL). An agent is given a notion of cumulative rewards and is required to make decisions in each environment to optimise it. It will become evident that this formalisation may be used for a wide range of tasks and that it captures many crucial aspects of artificial intelligence, including a sense of cause and effect as well as a sense of uncertainty and nondeterminism. An important feature of RL is that an agent picks up positive behaviours. This indicates that it gradually adjusts existing behaviours and skills or learns new ones.

One aspect of RL is that, tasks are defined in terms of scalar rewards that an agent can get by interacting with the environment. In most cases, the objective is to maximise a return, which is an accumulation of these benefits. In a non-deterministic setting, the objective is often to maximise return expectations. A policy, which is a function that maps a state input to an action, or a conditional distribution of actions given a state, is typically how a solution is presented. This is the principal object that describes the agent's behaviours. A policy's fundamental form is one that is reactive and chooses an action based on its present observations. A policy might, for instance, be a look-up table with a cell for a response action for each state and a size equal to the number of states. Finding the appropriate actions to enter the table's cells in this scenario is part of the learning process. The table's activities ought to provide a policy that offers agents the highest return.

The core of RL is creating the learning process. When and how to gather information from an environment should be thought out during the learning process. An agent can use its activities to investigate its surroundings and gather the information required to enhance its policy. It must also determine which acts will result in a reward in the far future using the data it has collected. When the state space is big, a representation like a look-up table may present issues. One issue is that it would require a significant amount of memory. Another issue is that it is challenging to travel to enough states to fill up every table cell.

The concept of a function approximator was developed to address this issue. A parametric function called a function approximator is used to calculate an interest value. The data is stored more compactly in the parametric function's parameters by using function approximators. A learning algorithm can also naturally handle interpolation by selecting the function that best fits the data. The function approximator can then "guess" the results of data that have not yet been seen. One of the key elements that makes it possible for RL to be used in issues with a wide state space is its interpolation capability. Learning in RL needs propagation of associative information from an observed outcome back to the action that took place in the distant past, in contrast to supervised learning. When the crucial acts are temporally removed from the rewards, learning becomes more difficult. This gives rise to the concept of temporal abstraction, which aims to break down lengthy action sequences into more streamlined shorter sequences.

We investigate a class of policies with hierarchical representations to temporally abstract an agent's activity. A policy that combines sub policies and is activated by higher level policies, which in turn activate lower-level policies, is an example of a hierarchical policy. The specifics of the decisions made when taking primitive actions are abstracted away by information propagation at the high-level policy level. RL methods such as hierarchical reinforcement learning (HRL) techniques use this type of policy representation (Barlo and Mahadevan, 2003).

3.2.1 Exploration Problem in Reinforcement Learning

The exploration approach of a DRL system is one of its most troublesome features. The challenge is finding the information that the learning algorithms need. The most basic version of exploration techniques lacks an understanding of ”what” to explore. Instead, it makes a random decision about what to do at each time-step. Surprisingly, this type of exploration, called dithering exploration (Osband and Vana, 2017), is the most widely used tactic in DRL since it is so easy to use and performs reasonably well. Greedy exploration is the best-known application of such a tactic, which functions by generally adhering to an estimated best policy and sporadically uniformly selecting a random action.

The goal of random action is to study and enhance the currently considered best policy. Searching for a specific target is a better exploring method. The guided exploration approach is used in this situation (Osband and Vana, 2016). Due to its goal-oriented nature, guided exploration got its moniker. With orthogonal objectives (or other goals) that may be distinct from the primary purpose, these exploration tactics delve further. In HRL, random exploration strategies at lower temporal scales (lower frequency) can lead to guided exploration at higher scales.

The HRL policy structure is to blame for this. A high-level policy (option) in HRL(hierarchical reinforcement learning) activates lower-level policies (policies) in a hierarchical manner, which then carry out basic operations. When a high-level policy employs a random exploration, such as greedy, it will choose a random option and activate it until it is terminated. This adherence to an option's behaviour over several time steps is a directed behaviour. As a result, at this scale of time, the agent might be considered as engaging in directed exploration. As an alternative, a few exploration tactics look for 'valuable' information directly. However, it is still unclear exactly what constitutes "useful" information. The setting of intrinsically driven agents has been addressed in this field of study, where the main question is how to measure an agent's curiosity is up for debate (Schmidhuber, 2010). For instance, the uncertainty of an agent's present understanding of the world can serve as the foundation for inquiry. The agent may have an innate desire to improve the model of the knowledge it has regarding the environment. In low dimensional contexts, this idea of leveraging uncertainty for exploration has been well investigated (Strehi et al, 2006).

Another crucial feature of RL is that it relies on trial-and-error learning, as opposed to, say, dynamic programming, which presupposes complete environment information. The RL agent just must be able to interact with the environment and gather information; it does not need to have comprehensive knowledge of or control over the environment. In an offline context, experience is gained beforehand and used as a batch for learning (hence the term batch RL for the offline setup).

3.2.2 Framework

Markov decision process: An agent engages in successive interactions with its environment in the traditional reinforcement learning framework. Markov decision process (MDP), a mathematical framework used to represent decision-making, aims to describe the three factors necessary for decision-making: sensation, action, and objective. MDPs are an extension of Markov chains that include actions for goal-oriented goals and rewards. A state space S, an action space A, a reward function r (s, a, s), and a state-transition distribution p $(s|s, a)$ often make up the description of an MDP. The agent is given a state at each time step t and creates an action at. In response, the environment offers a new state (s$_{t+1}$) and a reward (r_{t+1}). There are some states that can be terminal, meaning that once they are attained, no more interaction is possible.

A decision-maker maps states to probabilities over actions using a policy, symbolised by the symbol π. States in MDPs are referred to as Markovian because they include all the data needed to simulate state transitions and rewards and, thus, to reflect ideal policies. There is no need for further material, such as a history of past states. An environment's internal states do not necessarily have to match those offered to an agent. For instance, a physics simulator's complete state might include variables like friction coefficients or the magnitude of the gravitational acceleration if they remain intact, they do not need to be communicated to the agent. Like this, numerous approaches and different points of reference can be used to represent an object's state, including its coordinates, rotation, and velocity.

3.2.3 Episodes

Typically, the interaction between the agent and the environment occurs in a limited number of stages known as episodes. The agent first senses the initial state s_0 and outputs the first action a_0 after sampling it from a distribution $p_0(s_0)$. Following that, the environment changes to the following state, s_1, based on s_0 and a_0, and produces a reward, r_1. The agent generates a new action, a_2, in response to this new state and reward. The interaction loop continues in this manner until the episode reaches a terminal state and ends. A terminal state can be the conclusion of a maze for a path-finding job, a lethal hit from an enemy in a video game or losing balance and falling for a robot.

3.2.4 Objective

Sum of Rewards

The decision-maker's goal is to create a policy πthat maximises a return R, which is a cumulative function of the future rewards. But there are several problems that can arise when using the rewards total directly. For instance, the sum would be boundless if episodes could endure forever, and an infinite number of awards could be earned. In other cases, even when there are fewer incentives available, there is no motivation to claim them sooner rather than later because the amount is time indifferent.

Discounted Sum of Rewards

The most typical alteration used to reward instead of using a simple sum is a geometric progression utilising a discount factor of $\gamma \in [0, 1)$ .The two difficulties mentioned above are resolved by this transformation: it makes short-term incentives more attractive and limits the sum of reduced rewards than those that are long-term. It is important to note that a modest discount factor can cause policies to become overly myopic and eagerly choose little immediate rewards over larger delayed ones. Except when otherwise noted, the remainder of this argument will assume that returns are discounted.

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{3.10}$$

3.2.5 Policy Evaluation

Explains what a state is worth when it abides by a policy. It represents the anticipated return under the assumptions that we follow our policies and begin in the state, s. Here is the equation for the state value function:

$$V^{\pi}(s) = \underset{\pi}{\mathbb{E}}[G_t | s_t = s] \tag{3.11}$$

This equation 3.2 is called a Bellman equation as it connects the value of a state to the values of the successor states. Intuitively, the Bellman optimality equation should output a value for a state under an optimal policy that must equal the expected return for the best action from that state.

$$V*(s) = max_{a \in A}\, q_{\pi *}(s, a) \tag{3.12}$$

$$V*(s) = max_{a \in A}\, \underset{\pi *}{\mathbb{E}}[G_t | S_t = s, A_t = a] \tag{3.13}$$

$$V * (s) = max_{a \in A} {}_{\pi} *^{\mathbb{E}} [R_{t+1} + \gamma G_t | S_t = s, A_t = a] \tag{3.14}$$

$$V * (s) = maxa \in A {}_{\pi} *^{\mathbb{E}} [R_{t+1} + \gamma V * (S_{t+1}) | S_t = s, A_t = a] \tag{3.15}$$

$$V * (s) = max_a {}_{\pi} *^{\mathbb{E}} \Sigma(s', r | s, a) + [r + \gamma V * (s')] \tag{3.16}$$

The Bellman equation for optimality $q \in (s)$ state-value function is:

$$q \in (s, a) = \mathbb{E}[R_{t+1} + \gamma max_a \; q * (S_{t+1}, a') | S_t = s, A_t = a] \tag{3.17}$$

$$q * (s, a) = \sum s', rp(s', r | s, a)[r + \gamma max_a q * (s', a') \tag{3.18}$$

$$\mathbf{Q}^{\pi}(s, a) = \underset{\pi}{\mathbb{E}}[G_t | s_t = s, a_t = a] \tag{3.19}$$

Action value function: Indicates the value of acting in accordance with a policy in a particular state. It is the expected return given the state and an activity taken in accordance with a policy:

3.3 The Bellman Equation

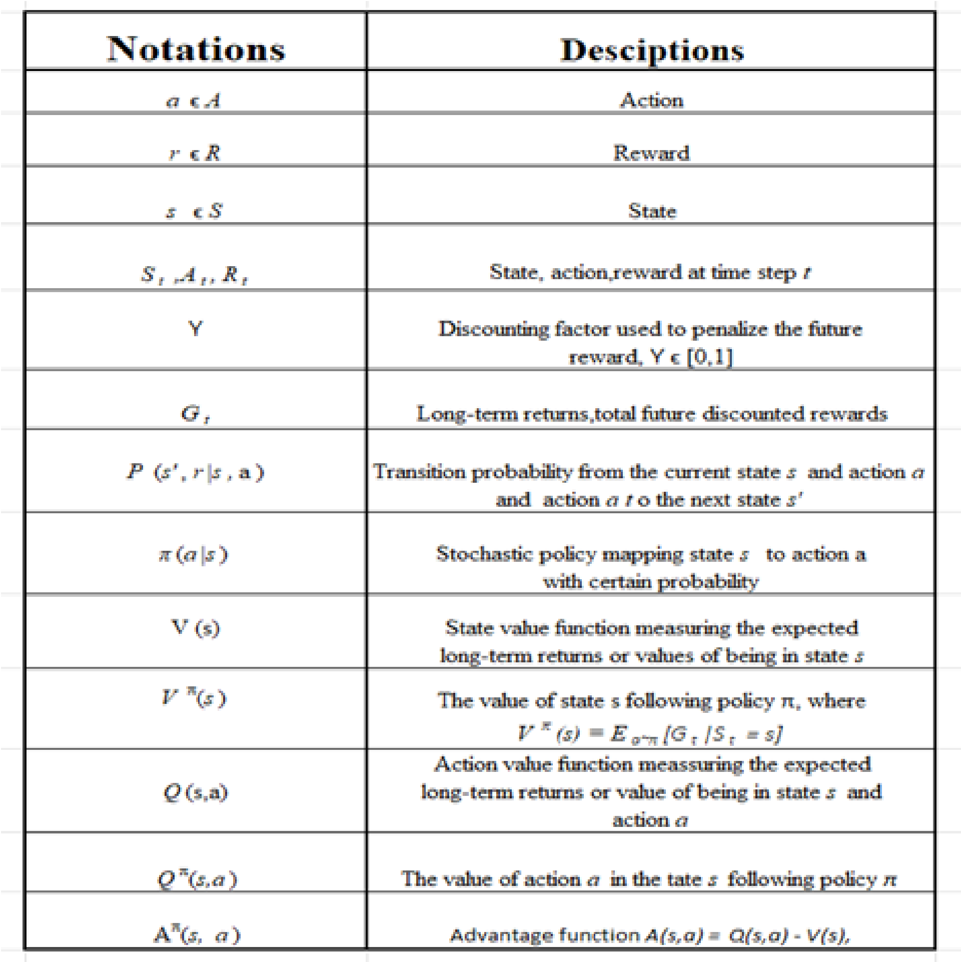
An approximation of the Q-value can be found using the Bellman equation below

$$Q(s_t, a_t) = Rt + 1 + (s_{t+1}, a\prime) \tag{3.20}$$

If an agent is in state st and then performs an action at, the environment will respond by giving the agent a reward of level R_{t+1} and will place the agent in a new state s_{t+1}. Because the agent does not know the rewards of future time steps while he is interacting with the environment (R_{t+2},R_{t+3}, etc.), the Bellman equation approximates the Q-value by assuming that the agent would pick the action $a\prime$ at state S_{t+1} that yields the highest return when following his policy. This is done because the agent does not know the rewards of future time steps when he is dealing with the environment. The table 3.1 provides basic mathematical notations used in this report.

3.4 Sarsa Algorithm

The state-action value function Q of the policy is iteratively approximated in the Sarsa algorithm. This policy, which interacts with the environment, is created by combining exploratory data with the learnt Q-values. activities that are greedy. For instance, the -greedy exploration technique chooses the uniformly from A with probability and the greedy action arg $max_a Q(s, a)$ with probability $(1 - \epsilon)$. Another exploration technique is called Boltzmann exploration, which creates a Boltzmann distribution from the Q-values and assigns a higher probability to actions with higher values. The loss function is given for a sample s, a, r, s', and a', which gives the method its name:



Notations	Desciptions
$a \in A$	Action
$r \in R$	Reward
$s \in S$	State
S_t ,A_t, R_t	State, action,reward at time step t
Y	Discounting factor used to penalize the future reward, Y ϵ [0,1]
G_t	Long-term returns,total future discounted rewards
$P(s', r \| s, \mathrm{a})$	Transition probability from the current state s and action a and action a t o the next state s'
$\pi(a \| s)$	Stochastic policy mapping state s to action a with certain probability
V (s)	State value function measuring the expected long-term returns or values of being in state s
$V^{\pi}(s)$	The value of state s following policy π, where $V^{\pi}(s) = E_{a \sim \pi}[G_t \| S_t = s]$
Q (s,a)	Action value function meassuring the expected long-term returns or value of being in state s and action a
$Q^{\pi}(s,a)$	The value of action a in the tate s following policy π
$\mathrm{A}^{\pi}(s, a)$	Advantage function $A(s,a) = Q(s,a) - V(s)$,

Table 3.1: Useful RL notations

$$L\theta(s,a,r,s',a') = \frac{1}{2}(Q_{\theta}(s,a)(r+\gamma Q(s\prime,a')))^2 \tag{3.21}$$

taking the gradient of this loss with respect to θ, we get:

$$\nabla_{\vartheta} L_{\vartheta}(s,a,r,s',a') = (Q_{\theta}(s,a)(r+(s',a')))\nabla_{\vartheta} Q_{\vartheta}(s,a) \tag{3.22}$$

Using this gradient to update θ enhances the estimation of Q^{π}. It's interesting to note that while Q changed, π also altered, but eventually the policy and Q-function would converge. In addition, as the policy improves, the quantity of exploration can be decreased over time, allowing the Q-function to monitor changes in. Because the data generated by a policy is only valid for the policy that generated it and needs to be deleted following an update, Sarsa is known as an on-policy algorithm.

3.5 Q-Learning

Q-Learning, which stands for Quality Learning, is a reinforcement learning algorithm that teaches an agent to take the appropriate action given a state. Q-Learning's purpose is to learn a policy that maximises overall reward. In other words, the Q-Learning algorithm is responsible for learning the Q-values in a Q-table. When the Q-table converges, we have arrived at an optimal policy, in which the agent understands the consequences of performing a certain action in a specific state. The agent begins with a Q-table that is initialised with zeros, then through contact with the environment, he updates this table based on the Bellman Equation, and eventually (after convergence), this table becomes his reference to choose the best feasible action given a state. Q-Learning is known with a strong proof of convergence established by (Tommi et al.,1993). It directly computes the optimal action policy without any intermediary cost evaluation step and without using any model. Q-Learning is model free and off-policy algorithm established on the bases of the Bellman Equation given in bellman equation above.

3.6 Deep Reinforcement Learning

In this section, we will discuss background survey of existing DRL work done previously. We also give an overview of supporting concepts that are considered fundamentally related to the realization of the proposed research.

- State: Actions are the Agent's methods which allow it to interact and change its environment, and thus transfer between states. Every action performed by the Agent yields a reward from the environment. The decision of which action to choose is made by the policy

- Action: Actions are things that an agent can do in the environment. Actions can be moves allowed by the rules of play. In RL, we distinguish between two types of actions: discrete or continuous. Discrete actions form the finite set of mutually exclusive things an agent could do, such as move left or right. Continuous actions have some value attached to the action.

- Reward: Rewards provide evaluative feedbacks for a RL agent to make decisions. It can be positive or negative, large, or small. The purpose of the reward is to tell the agent how well they have behaved.

- Environment: The environment is external to an agent, and its communication with the environment is limited by rewards (obtained from the environment), actions (executed by the agent and given to the environment), and observations (some information besides the rewards that the agent receives from the environment).

- Agent: An agent is somebody or something who/which interacts with the environment by executing certain actions, taking observations, and receiving eventual rewards for this. In most practical RL scenarios, it's our piece of software that is supposed to solve some problem in a more-or-less efficient way.

- Value Function: A value function is a prediction of the expected, accumulative, discounted, future reward, measuring how good each state, or state-action pair, would be.

- Policy: A policy maps state to action, and policy optimization is to find an optimal mapping.

3.6.1 Model-Based and Model Free

To explore the taxonomy in reinforcement learning, we first discuss model-based and model-free methods. A model, such as a deep neural network, is a specific function with initialised parameters (pre-trained model) or learned parameters (well-trained model). In contrast, a "model" in model-based reinforcement learning refers to a collection of learned environmental knowledge. Remember that there are five factors in the Markov decision process (MDP) labelled as S, A ,P ,R ,γ, S and A indicate the environment's status space and action space; $p(s \mid s, a)$ represents a transition function, which provides the likelihood of the environment changing from s to s when an agent acts a; and signifies the discount factor R(s, a) is a reward function that returns a reward when an agent does action an at state s. The agent is typically unaware of the reward function R and the transition function ($p(s'|s, a)$. To make use of the incentive feedback, the agent must make a few mistakes and learn through monitoring the surroundings.

Predicting environmental components is one approach. Although the reward function R and P are unknown, the agent can nevertheless collect some samples by acting in the environment. The values of $p(s'|s, a)$ and r can be predicted by supervised learning if the samples (s, a, s', r) are large enough. Once all the components are known, planning techniques can be applied immediately. This approach is known as model-based. Another approach is to actively search for the best policy rather than modelling the environment. For instance, the policy gradient method directly looks for the best policy in the policy space while the Q-learning algorithm selects the actions with the highest Q-values and converges to an optimal Q-value function. These two algorithms directly seek the largest reward rather than concentrating on the model. Model-free is the term for this method. Whether the agent will acquire or learn the model (or dynamics) of the environment, such as the transition function and the reward function, distinguishes model-based from model-free approaches.

Model-based approaches can be divided into two groups: those that use pre-existing models and those that build new models. The reward function and transition process models can be accessible directly by the agent for the methods that function with a specific model. For instance, the AlphaGo algorithm (Silver et al. 2016, 2016) specifies the Go game's rules, which are easily comprehensible in computer language. For the agent to assess and enhance its policy, the transition function and reward function in Go are both well-known. The complexity or opaqueness of the environment prevents the methods used in the other category from learning the model directly, but an agent can learn the model through interactions with the environment and then use it to develop policies. The World Models algorithm (Ha and Schmidhuber 2018), the I2A algorithm (Racanière et al., 2017), and others are typical instances for the second category. Like the World Models technique, the agent gathers some input (S_t, A_t, S_{t+1}) from a random strategy and uses a variational auto-encoder (VAE) to encode it into a low dimensional latent vector z t (Baldi, 2012). Then, a prediction model of the future latent vector z is learned using these data $(Z_t, A_t,$ and $Z_{t+1})$. Following that, the agent can use the learnt model to refine its policy.

The main benefit of model-based approaches is the ability to predict future states and rewards using the environment model, which enables the agent to plan more effectively. Pure planning and expert iteration are two typical approaches (Sutton and Barto, 2018). For instance, the AlphaGo algorithm (Silver et al. 2016) uses expert iteration while the MBMF algorithm (Nagabandi et al., 2018) uses pure planning techniques. Model-based approaches have the drawback that the model is typically unavailable and that the dynamics of the environment might be complicated, and sometimes not even explicitly represented. Additionally, the practised learned models are typically erroneous, which introduces bias into estimation. When implemented in the real world, a policy that was estimated and improved using a biased model typically fails.

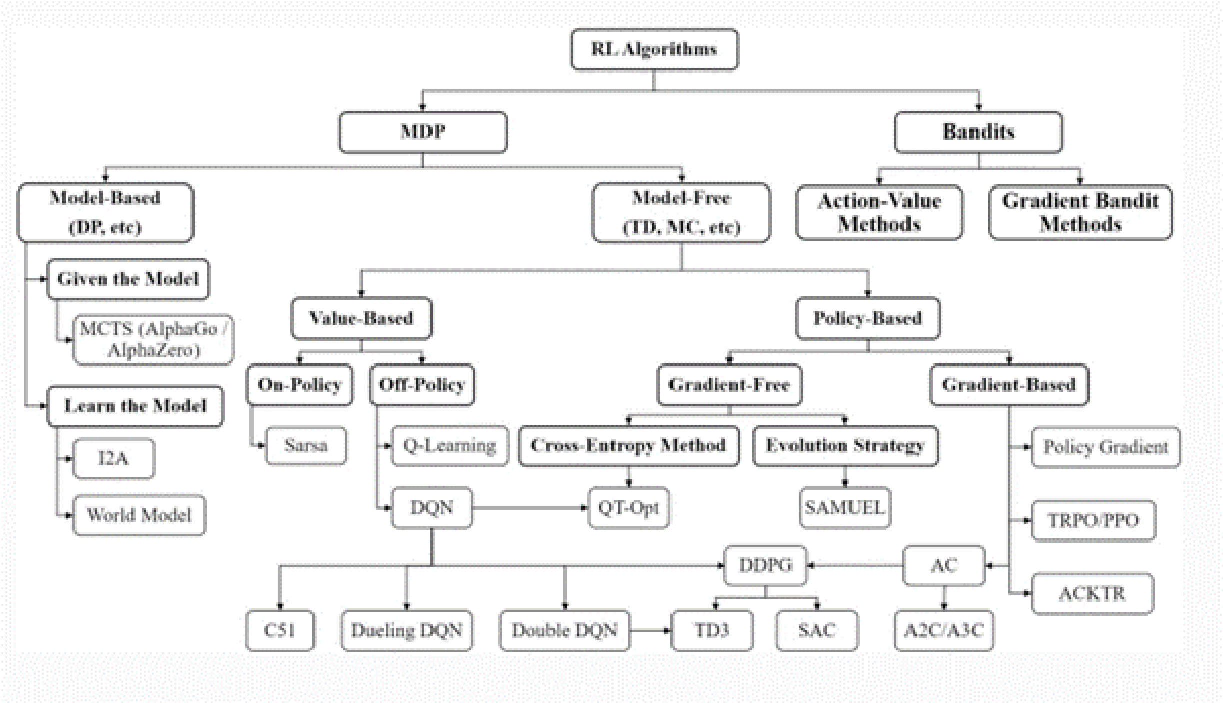


Figure 3.6: Taxonomy of Reinforcement learning, Source: Hangming and Tianyang(2020)

3.6.2 Model-Free Method

methodologies avoid attempting to create an environment model. The agent immediately engages with the environment, and as it explores more samples, it becomes more effective. Model-free methods are easier to implement than model-based methods since they don't depend on the model, which might be challenging to understand if it isn't provided. Model-free approaches, however, nevertheless frequently experience issues of their own. Sometimes the time commitment, equipment wears and tears, and safety concerns associated with exploration in the actual world can be very significant. In the case of an automatic pilot, for instance, we cannot train an agent to explore in the real world using a model-free method without taking any further safety measures because any traffic accident will be too expensive to sustain. Model-free algorithm includes the deep Q-networks (DQN) algorithm (Mnih et al. 2015), policy gradient (PG) approaches (Sutton et al. 2000), the deep deterministic policy gradient (DDPG) algorithm (Lillicrap et al. 2015), and others. However, model-based methods are becoming more and more crucial because model-free methods have low sample efficiency. Model-based algorithms include, for instance, AlphaGo (Silver et al. 2016) and AlphaZero (Silver et al. 2018, 2017).

3.6.3 Value-Based and Policy-Based

Value-based approaches and policy-based methods are the two basic categories for policy optimisation in deep reinforcement learning. The actor-critic class of algorithms and other algorithms, like QT-Opt (Kalashnikov et al., 2018), which use the value function for updating the policy, are created by combining the two. The action-value function $Q\pi(s, a)$ must typically be optimised when using value-based approaches. The optimal policy can be calculated by $\pi*$ $\approx$rgmax$\pi Q\pi$("$\approx$ " due to the approximation error), where$Q\pi*(s, a) = max_a Q\pi(s, a)$ is the optimal value function after optimisation. The high sample efficiency, low estimation variance, and difficulty of falling into a local optimum are the benefits of the value-based approach. The continuous action space problem is typically too complex for it to solve, and the *epsilon* greedy strategy and the max operator, like in DQN, can easily lead to overestimation.

Q-learning (Watkins and Dayan, 1992), DQN (Mnih et al., 2015), and its derivatives are two popular value-based algorithms: Prioritised Experience Replay (Schaul et al., 2015) increases learning efficiency by weighting the data based on TD error; Dueling DQN (Wang et al., 2016) enhances network structure. To increase approximation capacity, it divides the action-value function Q into the state-value function V and the advantage function A; Double DQN (Van Hasselt et al. 2016) selects and assesses actions with various parameters; Retrace (Munos et al. 2016) updates the calculation method for the Q value and lowers the variance of value estimation; and Noisy DQN (Fortunato et al.,2017). The policy-based approach directly optimises the policy by iteratively updating it until the cumulative return is maximised. A policy-based method provides the advantages of simplified policy parameterization, better convergence, and suitability for continuous or high dimensional action space when compared to the value-based method. PG (Sutton et al., 2000), TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017; Heess et al., 2017), and others are examples of standard policy-based algorithms.

To avoid policy collapse and increase algorithm stability, TRPO and PPO limit the update step based on PG. The more common approaches, in addition to the straightforward value-based and policy-based approaches, are those that combine the two and produce an actor-critic framework. The actor-critic method combines the advantages of the value-based method and the policy-based method by learning a Q function or value function to increase sample efficiency using the value-based methods and a policy function that is appropriate for discrete or continuous action space using the policy-based methods. This kind of strategy can be seen as an upgrade to the policy-based method for lowering sampling variance or as an extension of value-based methods in continuous action space. The benefits of the two ways are combined in this strategy, but the drawbacks are also carried over. For instance, the actor struggles with inadequate exploration, and the reviewer likewise struggles with overestimation. Actor-critic (AC) method (Sutton and Barto, 2018), among others, is a popular deep reinforcement learning actor-critic algorithm. It has undergone a number of changes, including: (1) DDPG (Lillicrap et al. 2015) inherits DQN's target network, and the actor is a deterministic policy; (2) A3C (Mnih et al. 2016) extends AC to asynchronous and parallel learning, disturbs the correlation between data, and improves the speed of data collection and training; (3) TD3 (Fujimoto et al. 2018) introduces clipped double Q-learning mode and delayed policy update strategy.

3.6.4 Monte Carlo and Temporal Difference

Some algorithms had already examined the distinctions between the Monte Carlo (MC) and temporal-difference (TD) approaches. For the sake of wrapping up, I will once more summarise their distinctions here. TD is a middle ground between MC techniques and dynamic programming (DP). Both TD and DP use bootstrapping for estimate, and neither TD nor MC require complete environmental knowledge. The method used for the learning update is where MC and TD diverge the most. DP can update at every time step, but MC must wait until the end of an episode to do so. Due to this distinction, TD methods will be able to have larger biases while MC methods will have larger variances.

3.6.5 On-Policy and Off-Policy

From a policy standpoint, there is a distinction between on-policy and off-policy .Off-policy methods evaluate or improve a separate policy from the one used to generate the data, in contrast to on-policy approaches, which aim to evaluate or enhance the policy that is used to make choices. The on-policy method necessitates that the agent itself engage in environmental interaction; hence, the policy that engages in environmental interaction and the policy that is being improved must be the same. It is not necessary for the off-policy technique to follow it; instead, the policy can be improved by learning from how other agents interact with the environment. Sarsa is a popular on-policy technique that chooses an action based on the existing policy, carries it out, and then utilises the results to update the policy. Therefore, the updated policy and the policy that interacts with the environment are both the same. The Q function is updated as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_t + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

(3.23)

A classic off-policy method is q-learning. The policy that interacts with the environment and the updated policy are not the same since it adopts the max operation and a $\in$-greedy policy when choosing actions. The Q function is updated as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_t + \gamma_a^{max} Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]. \tag{3.24}$$

3.7 Deep Q-Network (DQN)

Deep-Q-Learning is a variant of Q-Learning in which a Neural Network (NN) is used in place of a table to learn the Q-value function for a variety of states and actions. Deep-Q-Learning is also known as Q-Learning 2.0. This neural network is referred to as a Deep-Q-Network (DQN) by us. This DQN takes a state as its input and computes the corresponding Q-values for each action as its output after receiving the state. A straightforward explanation of the DQN algorithm may be found in the form of a regression problem, in which the NN makes predictions for the Q-values of all possible actions based on a state. with contrast to the traditional method of learning through supervision, known as supervised learning, with reinforcement learning (RL), we do not have access to the accurate target values. Because of this, we train the network in a supervised online method and use the Bellman Equation to come up with an approximation of the goal Q-values is defined as follow.

$$Q_{target}(S_t, a_t = R_{t+1} + \gamma max Q_{pred}(s_{t+1}, a') \tag{3.25}$$

3.8 Double-Q-Learning

The problem with DQN, which is otherwise one of the most powerful algorithms, is that not only the predicted Q-values rely on the NN, but also the target Q-values that we estimate using the Bellman equation. This is the reason why, when we update the weights of the network to give us a prediction that is close to the target value, the target value will also change because of the update to the weights of the network. Because of this association, the double DQN was developed in (Van et al., 2015) to break it. Instead of using a single NN, the double DQN uses two NNs. One of them, known as the learning network, is the one that is routinely updated at each timestep, while the other, known as the target network, is the one that is utilised to compute the target values. When the learning network is updated in this manner, the target variables won't be affected in any way. The weights of the learning network are periodically incorporated into the target network, which is periodically updated. Van et al.(2015) gives more specific details regarding the operation of the algorithm. However, this report tends to use Dueling double-Q-learning, and this will be explained extensively next.

3.9 Dueling DQN (D3QN)

Up until this point, the Deep Learning Models (here the term models refer to their usage as in supervised learning models as opposed to the MDP model) that we discussed were 'Sequential' architectures (sequential architectures and sequential models may have had different meanings in different contexts) a distinct meaning in the context of deep learning). In these models, every neuron in a certain layer could only be connected to the neurons in the layer immediately before and following their own layer. This restriction applied to all the neurons in that layer. To put it another way, these model architectures did not have any branches or loops of any kind.

Even though DQN and Double DQN each featured two Q networks, there was only one deep learning model, and the values of the other (target) network were just periodic copies of the values of the active (online) network. In Dueling DQN, we have a non-sequential architecture of deep learning in which, after the convolutional layers, the model layers divides into two independent streams (sub-networks), with each having their own fully-connected layer and output layers. This allows the model to train in a more efficient manner than traditional sequential architectures. The first of these two branches or networks corresponds to the Value function, which can be utilised to 'guess' the value of a certain state and only contains one node in its output layer. This branch or network is corresponding to the Value function. The second branch or network is known as the "Advantage" network, and it is responsible for calculating the value of the "advantage" of performing a certain action in comparison to the base value of existing in the present condition.

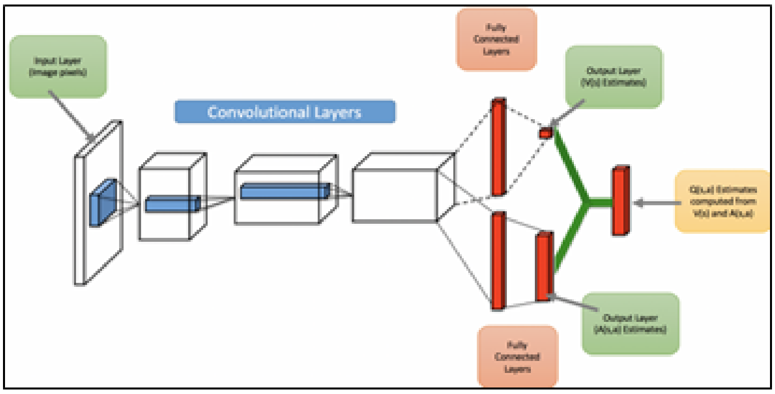


Figure 3.7: Schematic structure of Dueling Q Network.

However, the Q Function in Dueling DQN still reflects the Q Function in any a typical Q Learning algorithm, and so the Dueling DQN algorithm should work conceptually in the same manner that a typical Q Learning algorithm works by estimating the absolute action values or Q estimates. As a result, we must estimate the action-value/Q estimates as well. Remember that action-value is the absolute value of performing a specific action in a specific condition. So, if we combine (add) the output of the state's base value (first network/ branch) and the incremental 'advantage' values of the actions from the second ('advantage') network/ branch, we can effectively estimate the action-value or are Q Values as required in Q Learning. This can be expressed mathematically as follows:

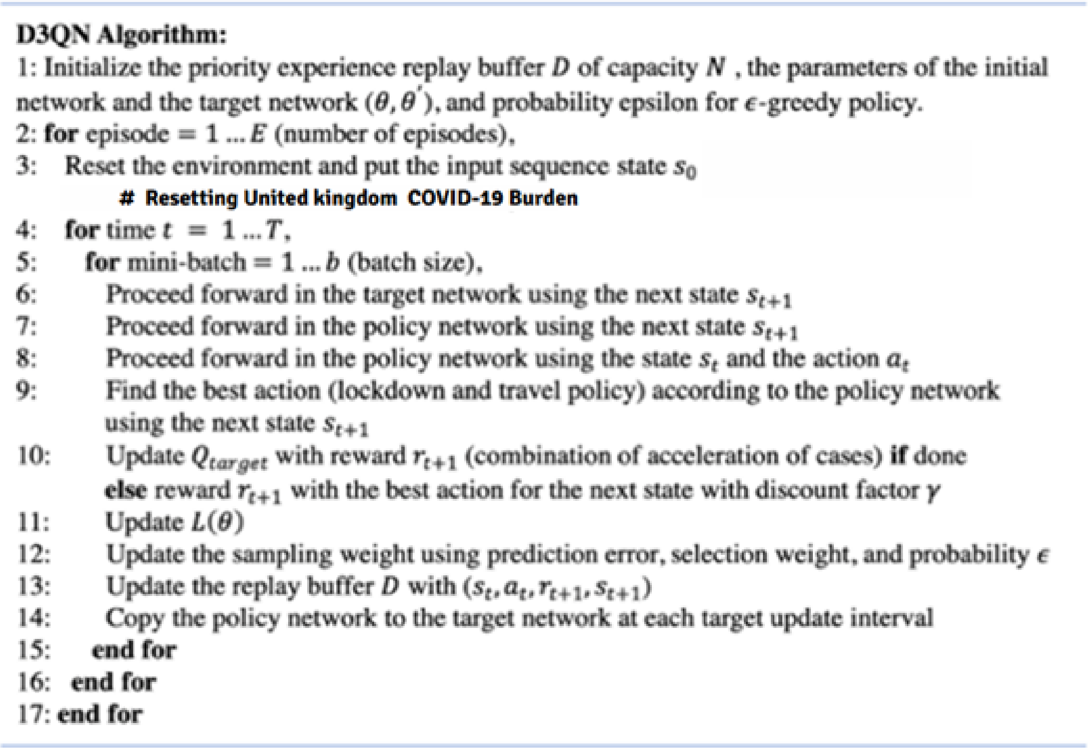
$$Q_{(s,a;,\alpha,\beta)} = V_{(s;\theta,\beta)} + (A_{(s,\alpha;\theta,\alpha)} - max_{a' \in |A|} A_{(s,a';\theta,\alpha)}) \tag{3.26}$$

The words Q, V, s, a, and a' in equation above have the same consistent meaning as in the table of mathematical notations above . Furthermore, the letter 'A' represents the advantage value. $'\theta'$ denotes the parameter vector of the convolutional layer that is shared by both the 'value' and 'Advantange' networks. $'\alpha'$ represents the $'Advantange'$ network's parameter vector, and $'\beta'$ represents the 'State-Value' function's parameter vector. Since we've reached the domain of function approximators, the values of any network are indicated regarding the parameters of the 'estimating' network to differentiate between the values/estimates of the same variable inferred via many different estimating functions. The equation in simple terms means that the Q value (the subscripts θ, α, β to the Q indicate that the Q estimates here are as computed from the estimating model which has three series of parameters or is a function of θ, α, β) for a given state-action combination is equal to the value of that state or absolute utility of being in that state as estimated from the state-value (V) network (the subscripts θ, β of V in the equation denote that the state value is estimated from Corrections for "identifiability" complete the equation. Examining the "identifiability" further, From the basic, obvious explanation above equation may have been as easy as below:

$$Q_{(s,a;\theta,\beta)} = V(s;\theta,\beta) + A(s,a;\theta,\alpha) \tag{3.27}$$

However, the issue with this straightforward construction is that, while we might obtain the value of Q (action values), if the values of S and A are supplied, the inverse is not true. This is illustrated in equation 3.22 above. This means that we were unable to "uniquely" deduce the values of S and A from a given value of Q. This is referred to as 'unidentifiability'. Equation 3.23, which is a superior variation of equation 3.21, is shown below. The final component of equation 3.21 is slightly altered in equation 3.23. Although the numbers are somewhat off-targeted when a constant is subtracted, this has little impact on learning because the value comparison is still there. Additionally, the stability of the optimisation is increased by the equation in this form.

$$Q_{(s,a;\theta,\alpha,\beta)} = V(s;\theta,\beta) + (A_{(s,a;\theta.\alpha)} - \frac{1}{|A|}\sum_{a} A_(s,a;\theta,\alpha)) \tag{3.28}$$



```
D3QN Algorithm:
Initialize the priority experience replay buffer $D$ of capacity $N$ , the parameters of the initial network and the target network $(\theta, \theta')$, and probability epsilon for $\epsilon$-greedy policy.
for episode = 1 ... $E$ (number of episodes),
  Reset the environment and put the input sequence state $s_0$
      # Resetting United kingdom COVID-19 Burden
  for time $t = 1 ... T$,
    for mini-batch = 1 ... $b$ (batch size),
      Proceed forward in the target network using the next state $s_{t+1}$
      Proceed forward in the policy network using the next state $s_{t+1}$
      Proceed forward in the policy network using the state $s_t$ and the action $a_t$
      Find the best action (lockdown and travel policy) according to the policy network
      using the next state $s_{t+1}$
      Update $Q_{target}$ with reward $r_{t+1}$ (combination of acceleration of cases) if done
      else reward $r_{t+1}$ with the best action for the next state with discount factor $\gamma$
      Update $L(\theta)$
      Update the sampling weight using prediction error, selection weight, and probability $\epsilon$
      Update the replay buffer $D$ with $(s_t, a_t, r_{t+1}, s_{t+1})$
      Copy the policy network to the target network at each target update interval
    end for
  end for
end for
```

Figure 3.8: D3QN algorithm for optimal policy implementation of COVID-19 ,Source: (Huang et al., 2018)

The figure 3.7 shown above is the step by step implementation of the D3QN algorithm for the derivation of optimal policy for appropriate timing of lockdown and travel ban of both local and international travels.

3.10 Framework for DRL on COVID-19 in the UK

The development of a comprehensive framework for deep reinforcement learning (DRL) on COVID-19 to finding the optimal policy for lockdown and travel ban in the UK entails the integration of methodologies derived from deep learning and reinforcement learning to effectively tackle the unique obstacles associated with the ongoing pandemic. Presented below is a comprehensive framework that might serve as an initial reference point.

3.10.1 Problem Definition

The main problem is to designing an effective policy for appropriate timing for lockdown and travel restrictions to combat the spread of COVID-19 in the UK. Before the policy design, there is need to find the severity of the state of COVID-19 with regards to death, recovery and confirmed cases.

3.10.2 Data Collection

Collecting relevant data for training and evaluation. This include; COVID-19 case data(confirmed, recovered and deaths), Policy interventions data(lockdown measures and travel ban).

3.10.3 State Representation

The task involves establishing the state representation that will be employed by the Deep Reinforcement Learning (DRL) model. The process entails determining the necessary information that the model need in order to make informed decisions.

3.10.4 Action Space

This defines the actions that our DRL(D3QN) agent can take. Adjusting the severity of lockdown and travel restriction(Low, medium or high).

3.10.5 Reward Function

In order to assess the desirability of various situations, it is necessary to establish a reward function. The reward function plays a crucial role in shaping the learning process of the deep DRL agent. See section 5.10.

3.10.6 Model Architecture

This has to do with choosing the appropriate deep neural network, in our case D3QN.

3.10.7 Training

Training our DRL(D3QN) model using historical data.

3.10.8 Evaluation

Assess the efficacy of the trained model by conducting a performance evaluation on an independent test dataset. Utilise appropriate metrics that align with the objectives of your framework, such as the accuracy of predictions, efficiency in resource utilisation, or the implications of policy suggestions.

3.11 Evaluation Metrics

The evaluation metrics employed in deep reinforcement learning (DRL) can exhibit variability contingent upon the particular situation at hand. Nevertheless, there exist various prevalent metrics and methodologies that are utilised to evaluate the efficacy of DRL algorithms. The following are several significant evaluation measures and methodologies:

3.12 Reward Function

The fundamental evaluation metric commonly utilised in the field of reinforcement learning is the cumulative reward or return acquired by the agent over its temporal trajectory. The reward function establishes the objective particular to the task, and the agent's efficacy is commonly evaluated based on the cumulative reward it acquires over the training or testing process.

3.13 Average Reward

The average reward is a popular assessment metric in Deep Reinforcement Learning (DRL) for assessing an agent's performance during training or testing. It indicates how effectively the agent performs on average across episodes or time steps. In DRL, you can calculate and analyse the average reward metric as follows:

3.13.1 Calculation of Average Reward

1. During training or testing, the agent interacts with the environment, taking actions and receiving rewards at each time step.
2. Calculate the cumulative reward obtained by the agent in each episode (the sum of rewards obtained in that episode).
3. After a set number of episodes or time steps, calculate the average of these cumulative rewards to get the average reward.

Mathematically, the average reward can be expressed as:

$$\text{Average Reward} = \frac{1}{N}\sum_{i=1}^{N} R_i \tag{3.29}$$

Where:

- N is the total number of episodes or time steps.
- R_i is the cumulative reward obtained in episode i

The average reward serves as a measure of the agent's total task performance. A larger average reward often signifies superior performance on the part of the agent, whereas a lower average reward implies inferior performance. Nevertheless, the understanding and analysis of the mean reward are contingent upon the particular job at hand and the explicit formulation of the reward function.

It is imperative to acknowledge that although the average payout serves as a valuable metric, it may not comprehensively encompass all facets of the agent's performance. For instance, the provided information does not encompass details regarding the agent's approach to exploration, and it may not accurately depict the agent's proficiency in accomplishing a particular task objective.

3.14 Discounted Return

The discounted return, akin to the average reward, is frequently employed in order to incorporate the concept of temporal discounting of rewards. The phenomenon prioritises present gains above future rewards to a greater extent.

3.15 Learning Curve

Learning curves illustrate the progression of an agent's performance as it undergoes training or over a period of time. The aforementioned curves offer valuable insights into the learning process, stability, and convergence of the agent.

3.16 Policy Evaluation

Metrics pertaining to the efficacy of the acquired policy, such as the extent of state exploration, the accuracy of action selection, and the precision of value estimation, can offer valuable insights into the behavioural patterns exhibited by the agent.

3.17 Comparison to Baselines

It is customary to evaluate the performance of the Deep Reinforcement Learning (DRL) agent by comparing it to baseline methods, which may include conventional reinforcement learning algorithms or, if applicable, human performance.

Other evaluation metrics are the Success Rate,Exploration vs Exploitation Trade-off,Entropy, Safety and Ethical Metrics and Generalization which valuate the agent's ability to generalize its learned policy to unseen or slightly modified environments, which is crucial in real-world applications.

Chapter 4

Data Preparation & Model Development

The development of data mining approaches aimed to enhance the comprehensiveness of the information finding process, moving beyond the sole utilisation of statistical or machine learning techniques. The primary objective of these workflows is to initiate a systematic process that commences with the identification of relevant inquiries and proceeds with the targeted analysis of unprocessed, often disorganised data with the aim of uncovering novel insights. While the building of models is an essential component of the knowledge discovery process, it is important to note that the selection and deployment of these models constitute only a small portion of the overall time invested. On the other hand, the process of data gathering, manual cleaning, and preparation can necessitate a significant allocation of time.

The predominant emphasis in data science research has been on the technical proficiencies essential for the field, with limited attention given to the challenges associated with project management in data science. Furthermore, only a small number of scholars have undertaken an analysis of the diverse methodologies employed by data science teams (Saltz and

Hotz, 2020). This study utilised the cross-industry standard procedure for data mining (CRISP-DM) to examine the potential for enhancing crime prediction by integrating Twitter sentiment polarity with historical crime records. The aim was to explore the feasibility of leveraging social media data from Twitter to improve crime prediction. The CRISP-DM framework, as described by Wirth and Hipp (2000), serves as a systematic approach to converting business difficulties into data mining tasks and effectively implementing data mining projects, regardless of the specific application area or technology utilised. The implementation being discussed is a widespread adoption of the Knowledge Discovery (KD) approach as described by Brachman and Anand in their work published in 1994.

Figure 4-1 depicts the interconnections among the six stages of the CRISP-DM process model. The initiation of a data mining project entails the articulation of the project's objectives, a process that is encompassed under the "Business Understanding" phase. The application of predictive analytics to enhance the operational performance and effectiveness of machinery is a common goal within the realm of crime prediction research in business settings. The aforementioned objective is subsequently transformed into a distinct data mining task, which involves determining the suitable machine learning technique and assessing the variables' impact on the model. During the phase known as "Data Understanding," hypotheses are formulated to uncover latent information that may be relevant to the project aim. These hypotheses are derived from prior experience and are supported by qualifying assumptions. In the context of crime prediction, it is advisable to retrain the model by employing diverse sampling approaches. This approach aims to determine the ideal hyperparameter that would yield the most effective models, utilising all of the major characteristics or variables present in the dataset. The picture illustrates a concise summary of the sequential stages involved in the conventional approach to data science utilising the Cross-Industry Standard Process for Data Mining (CRISP-DM).

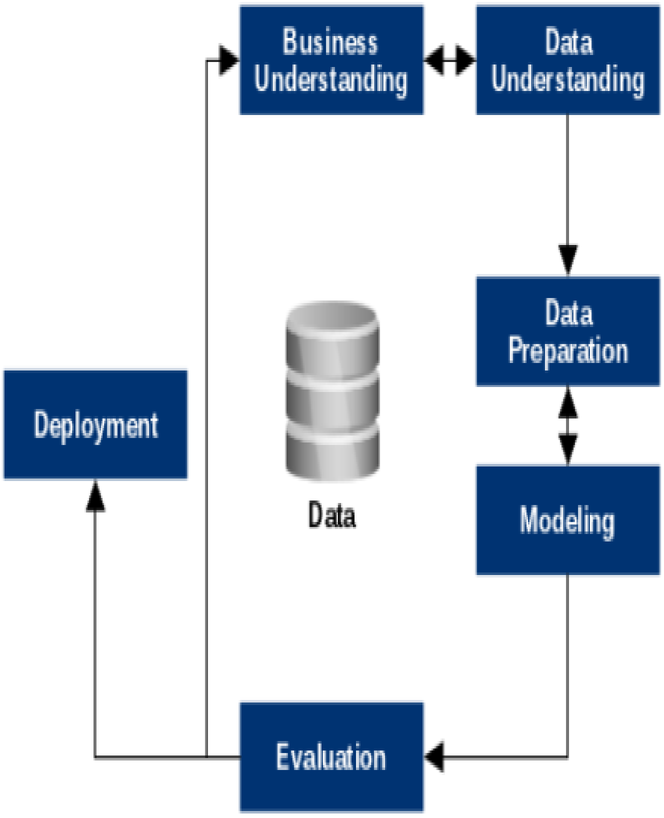


Figure 4.1: CRISP-DM.(Source:https://towardsdatascience.com/data-science-career-reflection based-on-crisp-dm-process-model-aedd8542b019)

Business Understanding

The evaluation of existing research and literature is essential in order to ascertain the resources that are currently accessible and required. The identification of the data mining purpose is a very important part of this process. The initial step in data mining involves specifying the particular type of analysis, such as classification, alongside identifying the key performance indicators for evaluating the data mining process, such as accuracy, precision, or recall. It is imperative to develop a project plan that is mandatory.

Data Understanding

The procedural elements encompassed in this process entail the collection of data from diverse sources, the subsequent extraction of pertinent information, the verification of accuracy, and the assurance of its overall quality. This stage of the research endeavour encompasses a comprehensive depiction of the data, as well as an investigative examination of the statistical patterns and interconnections present within the data employed in the study.

Data Preparation

Prior to selecting data, it is essential to establish clear inclusion and exclusion criteria. The issue of inadequate data quality can be addressed through the process of data cleansing. The construction of derived features is contingent upon the model chosen, as described in the initial step. There exist various different approaches for each of these procedures, which are depending on the available resources and the specific model being utilised.

Modelling

The process entails the selection of a suitable modelling strategy, the formulation of a test case, and the subsequent development of the model. All techniques utilised in the field of data mining hold relevance. Typically, the business problem and the factual circumstances have influence on the decision-making process. The primary focus lies in the process of justifying the selection. In order to create the model, some parameters need to be established. The act of assessing the model by comparing it to predetermined evaluation criteria and thereafter choosing the most suitable ones is deemed to be a suitable approach.

Evaluation

During the evaluation phase, the outcomes are assessed in relation to the predetermined business objectives. Therefore, it is vital to assess the obtained outcomes and formulate further procedures accordingly. An additional contention posits that a comprehensive examination of the entire method is warranted. Currently, in the course of this project, one or more models exhibiting exemplary attributes of data analysis have been developed. Prior to proceeding with the ultimate implementation of the model, a comprehensive evaluation is conducted on the model(s), and the methodologies employed in its development are scrutinised to ascertain its alignment with the business objectives. The basic objective is to evaluate whether there exist any substantial business concerns that have not been sufficiently resolved. Upon reaching the culmination of this phase, it is imperative to arrive at a definitive determination on the utilisation of the outcomes derived from the process of data mining.

Deployment

Typically, the construction of the model does not serve as the culmination of the project. In general, it is necessary to organise and deliver acquired knowledge in a manner that allows the recipient to effectively employ it. The deployment phase can range from a simple task of providing a report to a complex endeavour involving the implementation of a repeatable data mining process, contingent upon the specific requirements. In numerous cases, the responsibility of executing deployment operations lies with the user rather than the data analyst. Regardless, it is imperative to premeditate the necessary measures for effectively implementing the generated models. The process of deployment is comprehensively outlined in the user guide. The deployment phase encompasses activities like as deployment planning, monitoring, and maintenance. Consequently, the subsequent sections of this chapter will adhere to the phases suggested in the CRISP-DM architecture.

4.1 Study Area

Despite the global devastation caused by the COVID-19 pandemic, there has been a limited amount of research conducted on determining the most effective strategies for implementing lockdown measures and travel restrictions. The accuracy of these research studies is likely compromised due to the lack of uniformity in data collection methods across different countries. It is important to note that certain countries may have recorded false data, which introduces a potential bias into the analysis.

The selection of the UK as the research setting is motivated by its limited exploration within the domain of utilising deep reinforcement learning to determine optimal timing for lockdown and travel restriction in order to reduce the spread of COVID-19. Existing studies that have incorporated UK datasets have not adequately utilised relevant data, focusing primarily on predicting cases (such as death, recovery, and confirmed cases) for resource allocation purposes.

4.2 Research Design

Figure 4.2, depicted below, presents the sequential procedures entailed in achieving our research aims and objectives.

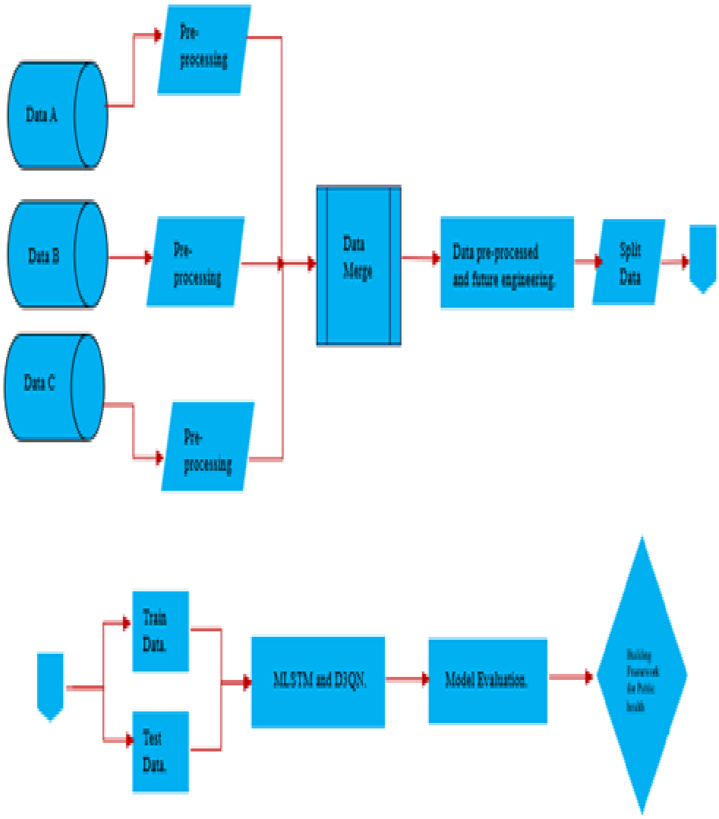


Figure 4.2: Research Process Framework

The subsequent portion of this chapter will provide a more detailed explanation of the flow nodes of the research process framework.

4.3 Data Understanding

This section gives a comprehensive explanation of the collected datasets and a thorough exploratory study of the epidemic COVID-19 dataset.

4.4 Data collection and Description

Data used in this investigation were collected from various sources. These dataset are as follow:

a. The date of the first confirmed case.

- Johns Hopkins coronavirus data repository
- WHO's case reports

b. The time series for the number confirmed, recovery, and death.

- Johns Hopkins coronavirus data repository
- WHO's case reports
- OurworldInData Coronavirus Pandemic (COVID-19)

c. The timing of lockdown and international travel restrictions.

- OurworldInData Coronavirus Pandemic (COVID-19):
- Wikipedia :

d. UK COVID-19 dataset.

- https://coronavirus.data.gov.uk/

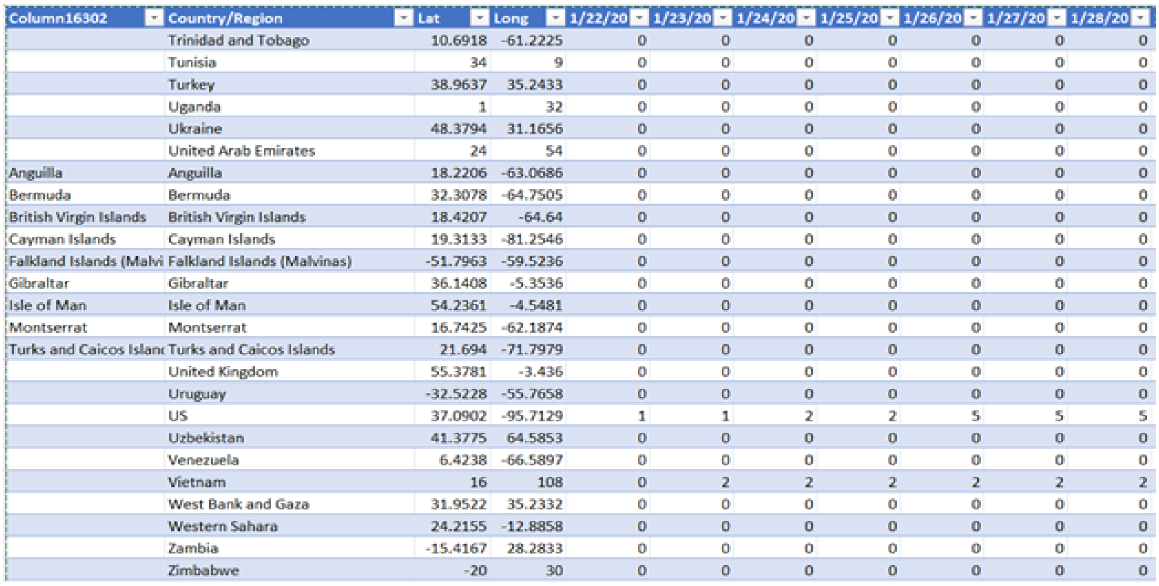
4.4.1 Global COVID-19 epidemic Data Description

This study utilised COVID-19 epidemic data obtained from reputable sources such as the Johns Hopkins coronavirus data repository, the World Health Organisation (WHO), OurworldInData, Wikipedia, and the UK coronavirus repository. The data spanned from 2020 to 2021 and served as the primary source for COVID-19 case information in this research. The data used in this study were selected due to their relevance in providing pertinent information. This would also enhance the credibility of the research, as the data were collected from reliable sources without any instances of information duplication. The Table 5.3, Table 5.4 and Table 5.5 present data on the number of deaths, confirmed cases, and recoveries.

In our approach, we incorporated many features including the number of confirmed infections, recoveries, and fatalities, as well as the rate of change from ech case Additionally, I considered factors such as population size and population demography During each time step, a feature vector of size 5 × 1 was utilised as the state s_t. Moreover, supplementary features are also obtained, encompassing the comparative proportion and cumulative data of those features.

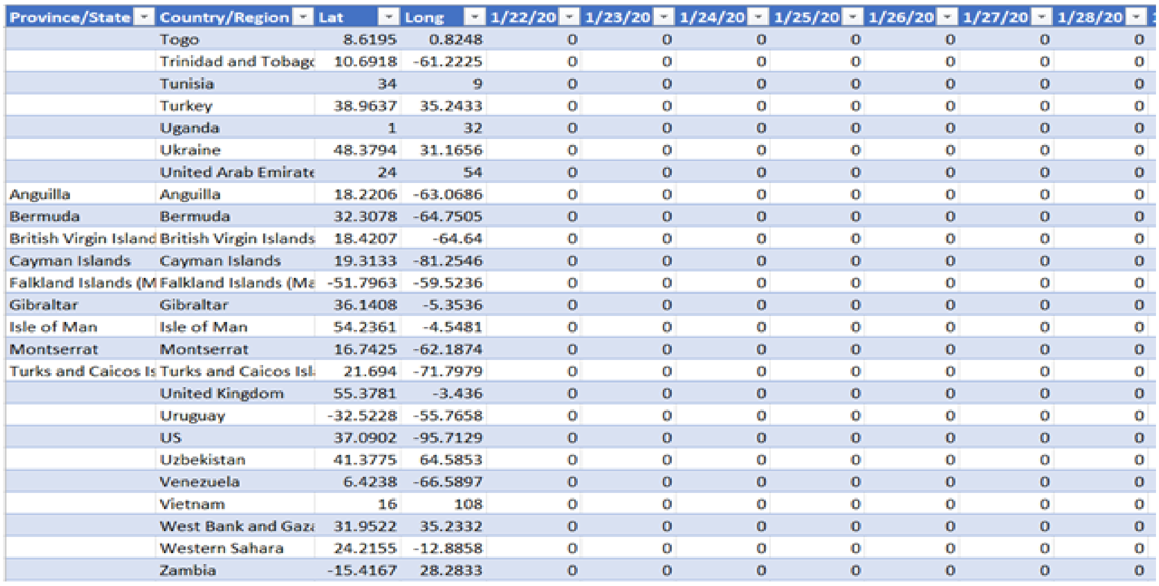
4.4.2 Data Exploratory Analysis

This section presents a concise statistical analysis of the datasets, aiming to facilitate the selection of suitable data processing techniques, feature engineering approaches, and strategies to address class imbalance concerns. Additionally, it aims to address crucial inquiries that could contribute to the development of an optimal policy regarding appropriate lockdown measures and travel restrictions. This section additionally examines the subsequent inquiries: What has been the trajectory of COVID-19 transmission over the past few weeks, and what are the projected trends for confirmed cases, fatalities, and re-



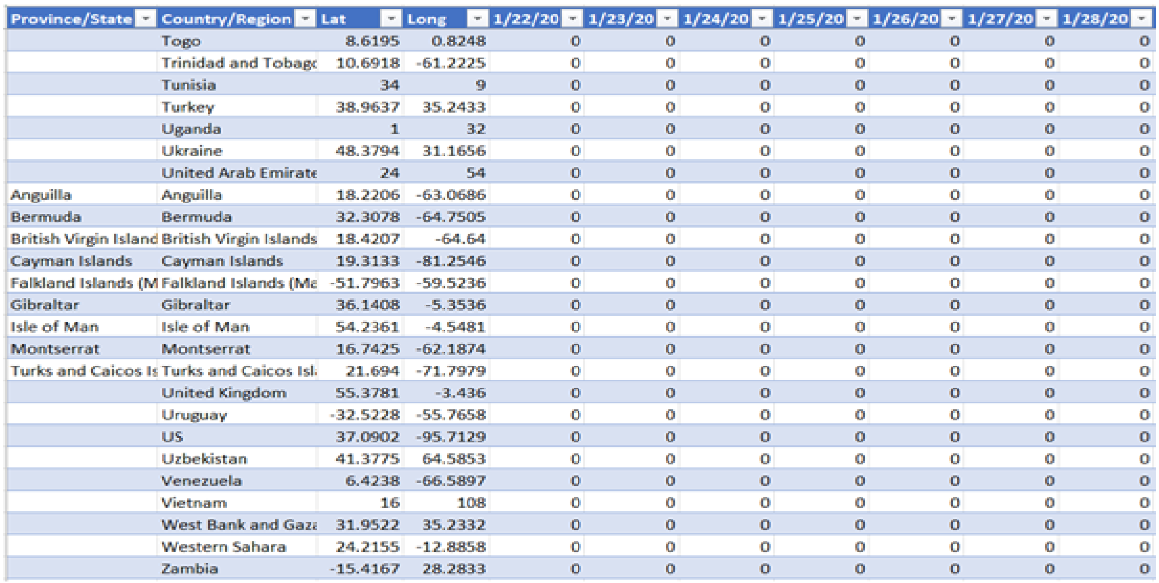
Column16302	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20
	Trinidad and Tobago	10.6918	-61.2225	0	0	0	0	0	0	0
	Tunisia	34	9	0	0	0	0	0	0	0
	Turkey	38.9637	35.2433	0	0	0	0	0	0	0
	Uganda	1	32	0	0	0	0	0	0	0
	Ukraine	48.3794	31.1656	0	0	0	0	0	0	0
	United Arab Emirates	24	54	0	0	0	0	0	0	0
Anguilla	Anguilla	18.2206	-63.0686	0	0	0	0	0	0	0
Bermuda	Bermuda	32.3078	-64.7505	0	0	0	0	0	0	0
British Virgin Islands	British Virgin Islands	18.4207	-64.64	0	0	0	0	0	0	0
Cayman Islands	Cayman Islands	19.3133	-81.2546	0	0	0	0	0	0	0
Falkland Islands (Malvi	Falkland Islands (Malvinas)	-51.7963	-59.5236	0	0	0	0	0	0	0
Gibraltar	Gibraltar	36.1408	-5.3536	0	0	0	0	0	0	0
Isle of Man	Isle of Man	54.2361	-4.5481	0	0	0	0	0	0	0
Montserrat	Montserrat	16.7425	-62.1874	0	0	0	0	0	0	0
Turks and Caicos Islanc	Turks and Caicos Islands	21.694	-71.7979	0	0	0	0	0	0	0
	United Kingdom	55.3781	-3.436	0	0	0	0	0	0	0
	Uruguay	-32.5228	-55.7658	0	0	0	0	0	0	0
	US	37.0902	-95.7129	1	1	2	2	5	5	5
	Uzbekistan	41.3775	64.5853	0	0	0	0	0	0	0
	Venezuela	6.4238	-66.5897	0	0	0	0	0	0	0
	Vietnam	16	108	0	2	2	2	2	2	2
	West Bank and Gaza	31.9522	35.2332	0	0	0	0	0	0	0
	Western Sahara	24.2155	-12.8858	0	0	0	0	0	0	0
	Zambia	-15.4167	28.2833	0	0	0	0	0	0	0
	Zimbabwe	-20	30	0	0	0	0	0	0	0

Table 4.1: Samples of confirmed cases



Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20
	Togo	8.6195	0.8248	0	0	0	0	0	0	0
	Trinidad and Tobagc	10.6918	-61.2225	0	0	0	0	0	0	0
	Tunisia	34	9	0	0	0	0	0	0	0
	Turkey	38.9637	35.2433	0	0	0	0	0	0	0
	Uganda	1	32	0	0	0	0	0	0	0
	Ukraine	48.3794	31.1656	0	0	0	0	0	0	0
	United Arab Emirate	24	54	0	0	0	0	0	0	0
Anguilla	Anguilla	18.2206	-63.0686	0	0	0	0	0	0	0
Bermuda	Bermuda	32.3078	-64.7505	0	0	0	0	0	0	0
British Virgin Island	British Virgin Islands	18.4207	-64.64	0	0	0	0	0	0	0
Cayman Islands	Cayman Islands	19.3133	-81.2546	0	0	0	0	0	0	0
Falkland Islands (M	Falkland Islands (Ma	-51.7963	-59.5236	0	0	0	0	0	0	0
Gibraltar	Gibraltar	36.1408	-5.3536	0	0	0	0	0	0	0
Isle of Man	Isle of Man	54.2361	-4.5481	0	0	0	0	0	0	0
Montserrat	Montserrat	16.7425	-62.1874	0	0	0	0	0	0	0
Turks and Caicos Is	Turks and Caicos Isl	21.694	-71.7979	0	0	0	0	0	0	0
	United Kingdom	55.3781	-3.436	0	0	0	0	0	0	0
	Uruguay	-32.5228	-55.7658	0	0	0	0	0	0	0
	US	37.0902	-95.7129	0	0	0	0	0	0	0
	Uzbekistan	41.3775	64.5853	0	0	0	0	0	0	0
	Venezuela	6.4238	-66.5897	0	0	0	0	0	0	0
	Vietnam	16	108	0	0	0	0	0	0	0
	West Bank and Gaza	31.9522	35.2332	0	0	0	0	0	0	0
	Western Sahara	24.2155	-12.8858	0	0	0	0	0	0	0
	Zambia	-15.4167	28.2833	0	0	0	0	0	0	0

Table 4.2: Samples of death cases



Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20
	Togo	8.6195	0.8248	0	0	0	0	0	0	0
	Trinidad and Tobagc	10.6918	-61.2225	0	0	0	0	0	0	0
	Tunisia	34	9	0	0	0	0	0	0	0
	Turkey	38.9637	35.2433	0	0	0	0	0	0	0
	Uganda	1	32	0	0	0	0	0	0	0
	Ukraine	48.3794	31.1656	0	0	0	0	0	0	0
	United Arab Emirate	24	54	0	0	0	0	0	0	0
Anguilla	Anguilla	18.2206	-63.0686	0	0	0	0	0	0	0
Bermuda	Bermuda	32.3078	-64.7505	0	0	0	0	0	0	0
British Virgin Island	British Virgin Islands	18.4207	-64.64	0	0	0	0	0	0	0
Cayman Islands	Cayman Islands	19.3133	-81.2546	0	0	0	0	0	0	0
Falkland Islands (M	Falkland Islands (Ma	-51.7963	-59.5236	0	0	0	0	0	0	0
Gibraltar	Gibraltar	36.1408	-5.3536	0	0	0	0	0	0	0
Isle of Man	Isle of Man	54.2361	-4.5481	0	0	0	0	0	0	0
Montserrat	Montserrat	16.7425	-62.1874	0	0	0	0	0	0	0
Turks and Caicos Is	Turks and Caicos Isl	21.694	-71.7979	0	0	0	0	0	0	0
	United Kingdom	55.3781	-3.436	0	0	0	0	0	0	0
	Uruguay	-32.5228	-55.7658	0	0	0	0	0	0	0
	US	37.0902	-95.7129	0	0	0	0	0	0	0
	Uzbekistan	41.3775	64.5853	0	0	0	0	0	0	0
	Venezuela	6.4238	-66.5897	0	0	0	0	0	0	0
	Vietnam	16	108	0	0	0	0	0	0	0
	West Bank and Gaz	31.9522	35.2332	0	0	0	0	0	0	0
	Western Sahara	24.2155	-12.8858	0	0	0	0	0	0	0
	Zambia	-15.4167	28.2833	0	0	0	0	0	0	0

Table 4.3: Samples of recovery cases

coveries in the future? The responses to these questions were included in the results section.

On March 11, 2020, the declaration of the COVID-19 pandemic was made by the World Health Organisation (WHO). The rapid spread of the disease has had significant implications for healthcare systems in several countries, including Italy, Spain, France, the United States, and the United Kingdom (Velásquez and Lara, 2020). The precise modelling and prediction of the quantity of confirmed and recovered cases of COVID-19 is crucial in comprehending the situation and aiding policymakers in implementing measures to mitigate or halt its spread.

In light of the global spread of the COVID-19 pandemic, there is a pressing need for real-time assessments of epidemiological data in order to provide the population with a robust strategy to combat the illness. Since the emergence of the COVID-19 pandemic, there has been a global and zealous pursuit of its objectives (Punn et al., 2020). As of July 27, 2020, the global number of confirmed COVID-19 cases stood at 828,508,485, with the United

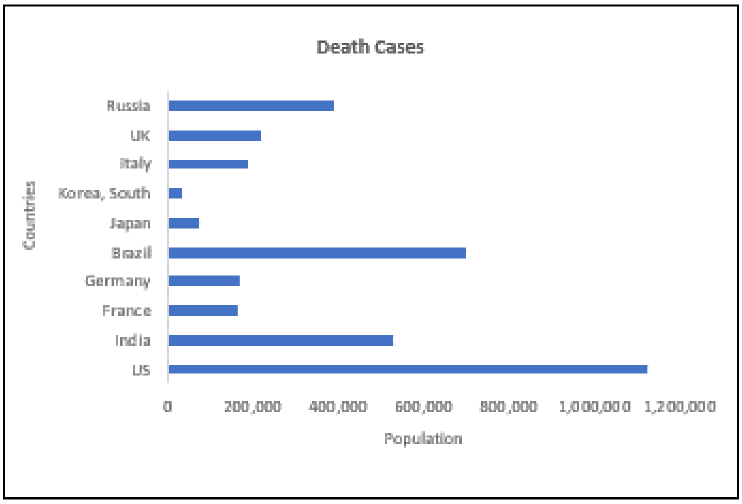


Figure 4.3: Death cases

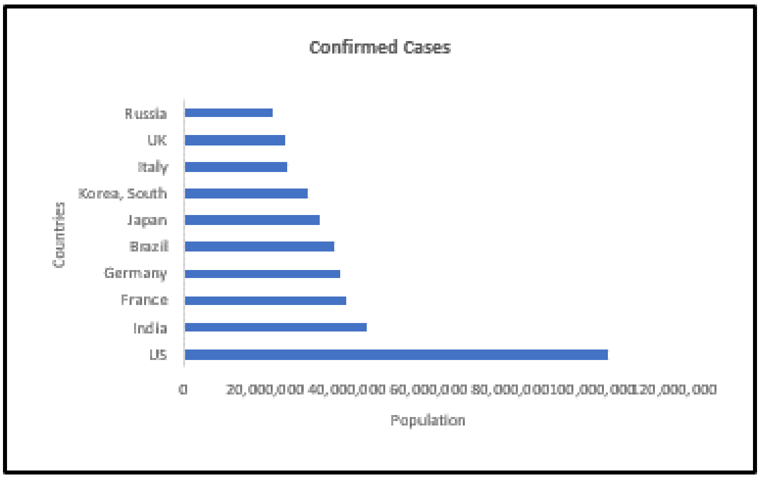


Figure 4.4: Confirmed cases

Kingdom reporting 313,798 cases. Among these instances, there were 43,384,903 deaths globally, with the UK reporting 46,706 deaths. Additionally, there were 388,408,229 recovered cases worldwide, with the UK reporting 267,092 recoveries. These figures are based on the data available to us . Figures 5.8,5.9,5.10 and 5.11 illustrate the temporal patterns seen in the number of confirmed cases and fatalities attributed to the Covid-19 pandemic.

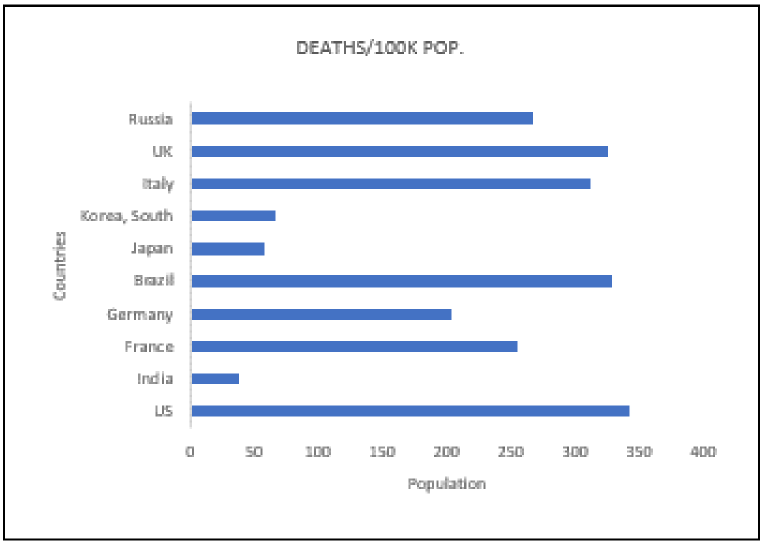


Figure 4.5: Death per 100K population

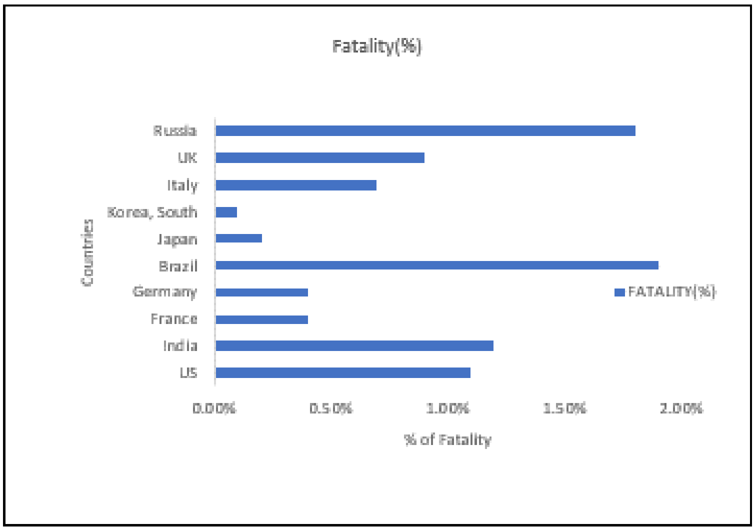


Figure 4.6: Fatility rate

Additionally, there were 388,408,229 recovered cases worldwide, with the UK reporting 267,092 recoveries. These figures are based on the data available to us . Figures 4.8,4.9,4.10 and 4.11 illustrate the temporal patterns seen in the number of confirmed cases and fatalities attributed to the Covid-19 pandemic.

4.5 Data Pre-processing

This section aims to provide an analysis of both globally COVID-19 epidemic data and the UK COVID-19 dataset. Before the merger process, the two datasets underwent cleaning procedures that focused solely on tasks such as standardising variable names, imputing missing data, and performing feature engineering. Consequently, the aforementioned datasets were utilised for the purpose of training both the MLSTM and the duelling DQN.

A Deep Reinforcement Learning (DRL) technique was employed to enhance the forecasting of future prevalence trends of COVID-19. This was achieved by including the DRL algorithm into the Long Short-Term Memory (LSTM) model, specifically targeting improvements in the prediction accuracy of recovery, confirmation, and mortality rates. This allows for the tracking of illness severity as it progresses. Based on the established facts regarding the severity and prevalence of the disease, it is now possible to determine an optimal policy for implementing lockdown measures and travel restrictions. This can be achieved through the utilisation of a trained dataset. The data preprocessing stage involved the utilisation of Scikit-learn, whereas the model training phase employed the TensorFlow framework.

4.6 Merge Datasets

The process involved utilising linear interpolation techniques to estimate the number of COVID-19 cases in the United Kingdom based on the index case date in each country. Subsequently, the extracted UK cases were integrated with the appropriate dataset specific to the United Kingdom from the global COVID-19 data set. Subsequently, the data was aggregated by calculating the average value over a three-day period. The purpose of this action was to mitigate the influence of potential bias resulting from delayed reporting and fluctuating viral testing capabilities. The decision was made to utilise numerical data over the previous three days as opposed to daily figures, as each time stamp necessitated time-sensitive data. Simultaneously, this measure was implemented in order to mitigate potential bias resulting from delayed reporting and fluctuations in viral testing capacity on weekends. After applying data normalisation for feature scaling, the dataset was separated into training, validation, and test sets in the ratio of 6:2:2 of trained, validate and test dataset.

4.7 Feature Engineering

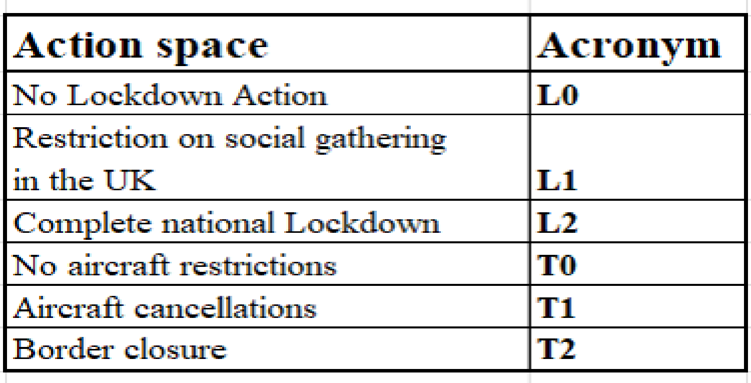
Feature engineering is an essential procedure in the realm of data analysis and machine learning, wherein one generates novel features (variables) based on pre-existing ones or modifies the current features to enhance the efficacy of the models. The implementation of well-designed feature engineering techniques has the potential to enhance the accuracy, interpretability, and efficiency of models. The following are few prevalent methodologies and strategies employed in the process of feature engineering:

The dataset exhibits a class imbalance in the target variable, as seen by a preliminary study. This research utilises the up-sampling method to address class imbalance concerns, as it has demonstrated high performance across different numbers of variables and is suitable for our dataset's size (Willette et al., 2022). Various strategies, such as undersampling and oversampling, exist for addressing class imbalance concerns. However, the up-sampling method is specifically chosen for this research.

Addressing outliers during the process of feature engineering is a crucial undertaking in the preparation of data for machine learning models. The presence of outliers can have a substantial impact on the performance and accuracy of models, thereby necessitating the appropriate handling of such data points. Regarding this research, the outliers were addressed through the utilisation of normalisation and scaling techniques, specifically the Min-Max strategy.

4.8 Action Space

In response to the implementation of lockdown measures and travel restrictions, I established a 3x3 action space. A tripartite system of lockdown measures was created, consisting of three distinct stages. There are three levels of measures implemented in response to the current situation: Level 0, which involves no action being taken; Level 1, which entails restrictions on social gatherings; and Level 2, which involves a complete national lockdown. In terms of the travel policy, there are three possible scenarios: no action (T0), aircraft cancellations (T1), and border closures (T2).



Action space	Acronym
No Lockdown Action	L0
Restriction on social gathering in the UK	L1
Complete national Lockdown	L2
No aircraft restrictions	T0
Aircraft cancellations	T1
Border closure	T2

Table 4.4: Action space

This work posits the utilisation of a discrete action space for the determination of intensity levels pertaining to lockdowns and travel limitations, as opposed to conventional approaches. The study conducted by Julius and Deshendran (2019) found that the Deep Q-Network (DQN) shown superior performance compared to the Deep Deterministic Policy Gradient (DDPG) algorithm when the number of episodes was restricted. This indicates that the DQN-based recommendation engine exhibits greater adaptability and personalization compared to a system that operates in a continuous action space (Kailiang, 2022). However, the Dueling DQN method, which was used in our model, demonstrated superior performance in properly handling the discrete output space.

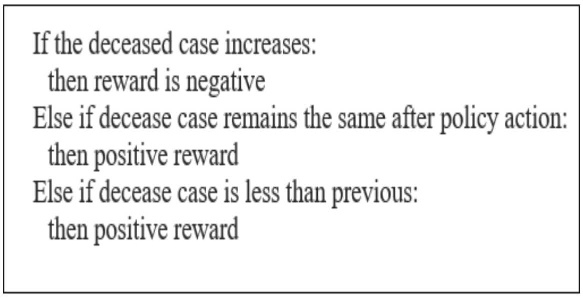
4.9 State Space

The disclosure of trends in diseases is a crucial aspect of state observations, as these characteristics inherently include the analysis of time series data. If we solely provide a number of infections for the present day, the agent will lack any indication of whether this metric is improving or worse. The agent is endowed with the capacity to deduce the sequential relationship between variables in the multivariate time series data by integrating prior knowledge obtained from preceding days. This aids the agent in making a more knowledgeable determination regarding the escalation or relaxation of particular control measures in the long run. In the course of our experimental study, we administered one single state input including the feature values observed during the preceding ten days.

4.10 Reward Defination

In order to assist the agent in its pursuit of identifying the optimal policy, it is imperative to devise a well-designed incentive function. In the case in point, the objective would entail managing the ongoing pandemic crisis through the implementation of daily adjustments to the extent of lockdown measures and travel restrictions. From a logical standpoint, it is desirable to implement measures to mitigate the rise in deaths or diseases, as this would yield an unfavourable outcome in terms of rewards within this context. Conversely, it is imperative to promote certain actions and then provide positive incentives when they effectively mitigate the number of infections and deaths, keeping them below a predetermined level. The reward system employed inside our surroundings consists of two components, namely death and recovery. Each of these components possesses its own distinct system of rewards and punishments, thereby constituting the complete framework.

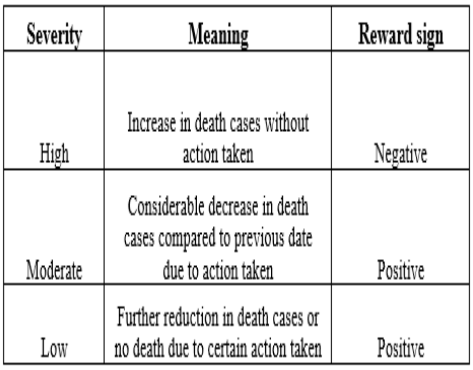
Subsequently, the total daily rewards can be determined by aggregating the two components of the reward. The following table 4.1 presents a visual representation of the overall sign orientation pertaining to the underlying justification for the establishment of rewards. This orientation is determined by the gravity of death cases, and the same rationale can be applied to instances of recovery. However, the sign will alter in this particular scenario. This research proposes the utilisation of discrete action space to decide the severity level



```
If the deceased case increases:
  then reward is negative
Else if decease case remains the same after policy action:
  then positive reward
Else if decease case is less than previous:
  then positive reward
```

Figure 4.7: Reward distribution: This shows the IFELSE distribution of the reward according to change in cases

of lockdowns and travel restrictions. When compared to a continuous action space, this approach is deemed more dependable, given the assumption that the epidemic will ultimately end.



Severity	Meaning	Reward sign
High	Increase in death cases without action taken	Negative
Moderate	Considerable decrease in death cases compared to previous date due to action taken	Positive
Low	Further reduction in death cases or no death due to certain action taken	Positive

Table 4.5: Designing the reward sign orientation for death severity is one example. Death and recovery follow the same logic, with the latter assuming the reverse sign. Because the sickness is contagious, the number of deaths will almost certainly increase if no action is done. On the other hand, if the situation improves as a result of particular acts, the effect may take a few days to become apparent, and the reward will finally be positive and therefore encouraged by the agent.

In accordance with this overriding guideline, the following is how we designed the reward function rt. It is important to keep in mind that the relative weights of these two components can be adjusted to suit our needs. This research factored into account the effect of the COVID-19 on the quality of life and devastation on the economy. The good quality of life and good economy is denoted by 10 why the otherwise is denoted by 0. It is worth noted that the quality of life and economy have so many factory to determine how good they are, however, for this research work , I arbitrarily chosen 0 and 10 and bad and good respectively. Thus, we have the equation below:

$$c_0 = cf + \frac{1}{ld} \tag{4.1}$$

$$c_0 = cf + \frac{1}{tr} \tag{4.2}$$

$$c_1 = cf + \frac{1}{ld} \tag{4.3}$$

$$c_1 = cf + \frac{1}{tr} \tag{4.4}$$

$$r_t = r_t^{rc} + r_t^{dt} + r_t^{cf} \tag{4.5}$$

$$r_t^i = \begin{cases} -c_0 - c_1 \times (s_{t+1}^t - s_t^i) & \text{if} & (s_{t+1}^t > s_t^i) \text{ and } (s_t^i > 0) \\ -0.5xc_0 + c_1 \times (s_{t+1}^t - s_t^i) & \text{if} & (s_{t+1}^t = s_t^i) \text{ and } (s_{t+1}^i \neq 0) \\ \text{and} & ((a_{t+1}^{lockdown} > 0) \text{ or} & (a_{t+1}^{travel\_ban} > 0)) \\ c_0 - c_1 \times (s_{t+1}^t - s_t^i) & \text{if} & (s_{t+1}^t < s_t^t) \\ -0.5xc_1 \times (s_{t+1}^t - s_t^i) & \text{otherwise} & \\ \text{for } i \text{ in } rc \text{ and } dt & & \end{cases}$$

(4.6)

$$r_t^i = \begin{cases} c_0 + c_1 \times (s_{t+1}^t - s_t^i) & \text{if} & (s_{t+1}^t > s_t^i) \text{ and } (s_t^i > 0) \\ 0.5xc_0 + c_1 \times (s_{t+1}^t - s_t^i) & \text{if} & (s_{t+1}^t = s_t^i) \text{ and } (s_{t+1}^i \neq 0) \\ \text{and} & ((a_{t+1}^{lockdown} > 0) \text{ or} & (a_{t+1}^{travel\_ban} > 0)) \\ -c_0 + c_1 \times (s_{t+1}^t - s_t^i) & \text{if} & (s_{t+1}^t < s_t^t) \\ -c_1 \times (s_{t+1}^t - s_t^i) & \text{otherwise} & \\ \text{for } i \text{ in } cf & & \end{cases} \tag{4.7}$$

Cf :confirmed cases, Ld.: lockdown=0.5 tr: travel restriction, Lq: living quality= 0 or 10 and econ: Economy = 0 or 10.
rc: recovery cases,cf: confirmed cases,and dt death cases

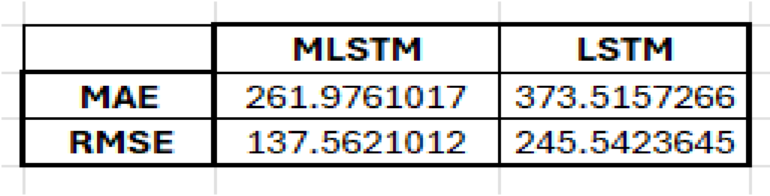
r_t=reward at timestamp t
s_t=state at timestamp t
c_0,c_1 = constant value
$a_t^{lockdown}$,$a_t^{travel_ban}$ =lockdown and travel ban

4.11 Model Bulding

The assessment of the algorithmic performance entailed a comparison of the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics for both the LSTM and MLSTM models. Metrics are utilised to assess the accuracy of expected case trends and evaluate their performance. Furthermore, an evaluation was conducted to determine the convergence of the Moving Average Reward (MAR) for the trained and validated datasets of the D3QN method.

4.11.1 Building Model

Two separate algorithms (LSTM and D3QN) were trained independently, with and without hyper-parameter adjustment, using the previously prepared cleansed COVID-19 dataset.



	MLSTM	LSTM
MAE	261.9761017	373.5157266
RMSE	137.5621012	245.5423645

Table 4.6: Loss function errors result

4.12 Model Evaluation

The evaluation of the performance of each algorithm involved comparing the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for both the LSTM and MLSTM models. The metrics are used to deterterming how well performed the predicted trends of cases. Also, in the case of D3QN , the convergence in the Moving Average Reward (MAR) was assessed for the trained and validated data of the Duelling Deep Q-Network (D3QN) algorithm.

The algorithms presented for trend prediction, namely LSTM and MLSTM, have the capability to assess the effectiveness of learning by utilising the following metrics: The two often used metrics in evaluating the accuracy of a predictive model are the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE). Figure 4.8 illustrates the comparative analysis of evaluation metrics. The table 4.6 shows the difference the measure of the average absolute difference between the predicted and actual values in the case of MAE and in the case of RMSE ,it penalizes large errors. Based on the observed results, the MLSTM system is suggested due to its comparatively lower error rate when compared to the MAE. Deep learning techniques are employed to construct a predictive system for forecasting future COVID-19 cases.

This research does projections on instances that have been confirmed, patients that have recovered, and cases that have resulted in death. The escalating number of fatalities and

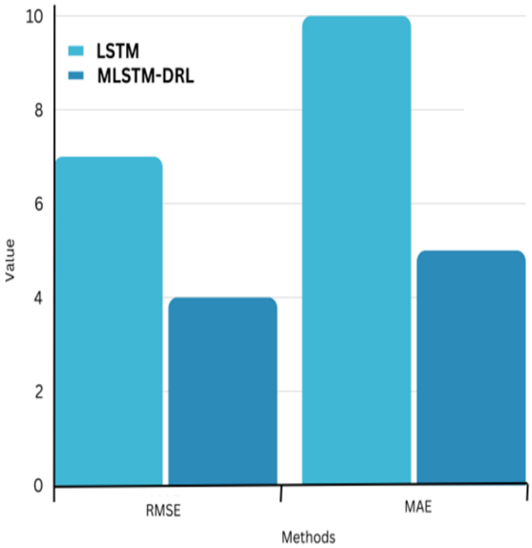


Figure 4.8: Loss function comparison between MAE and RMSE methods

documented instances of infection presents a disconcerting situation for the global community. The precise magnitude of the impact of the COVID-19 pandemic on the population of the United Kingdom remains uncertain. However, we endeavours to approximate the projected impact on the population in terms of new cases of recovery, confirmed cases, and fatalities during the upcoming days. Consequently, this process contributes to the enhancement of public health and facilitates the allocation of resources by the government. Furthermore, it plays a crucial role in determining the severity of COVID-19, thereby assisting in the formulation of appropriate policies pertaining to lockdown measures and travel restrictions, since the correlation coefficient between the original and forecasted data of confirmed and death cases is 0.999.

4.12.1 Baseline Policy Comparison Evaluation

This study aims to assess the efficacy of the D3QN algorithm in applying lockdown and travel limitations by utilising the average moving reward as a metric. The evaluation will be conducted through a baseline comparison, with the objective of identifying the algorithm that achieves the largest cumulative rewards for the ideal policy.

The utilisation of the average moving reward is a commonly utilised metric in the field of reinforcement learning for the purpose of evaluating the effectiveness of an agent over a specific duration. The metric measures the average cumulative reward obtained by the agent within a moving time frame of episodes or steps. The application of this metric is commonly utilised to reduce the influence of unpredictable reward signals and generate a more reliable assessment of an agent's effectiveness.

The V-D D3QN algorithm is an enhanced version of the D3QN algorithm that integrates the concept of value distributions. Instead of making an estimation of the Q-value as a singular scalar, the V-D D3QN algorithm does an estimation of the complete probability distribution encompassing all potential values of the Q-function. This has the potential to enhance the algorithm's efficacy in managing uncertainty. Based on the experimental findings obtained through the utilisation of the D3QN model. In the context of the experiment, the discounting factor was assigned a value of 0.99. The maximum number of episodes was set to 100, while the maximum mean reward per 100 steps was set to 10000.

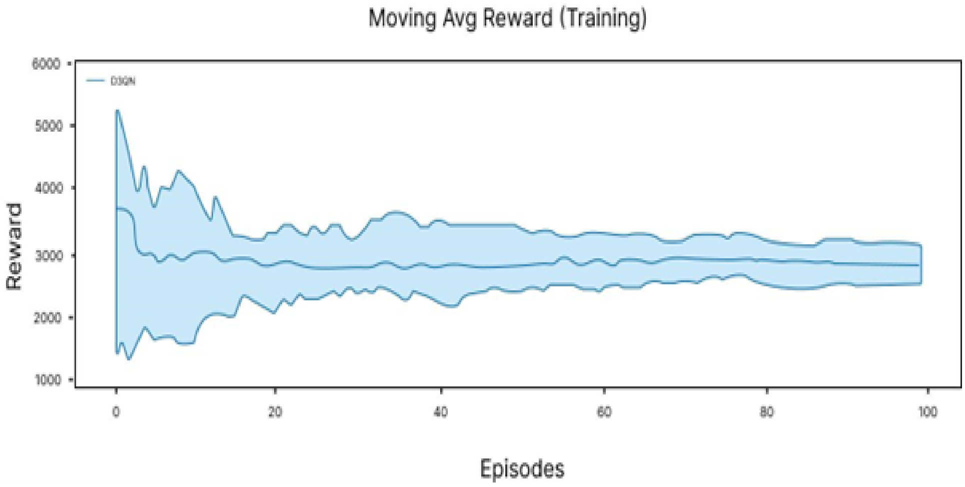


Figure 4.9: Moving average rewards for D3QN algorithm

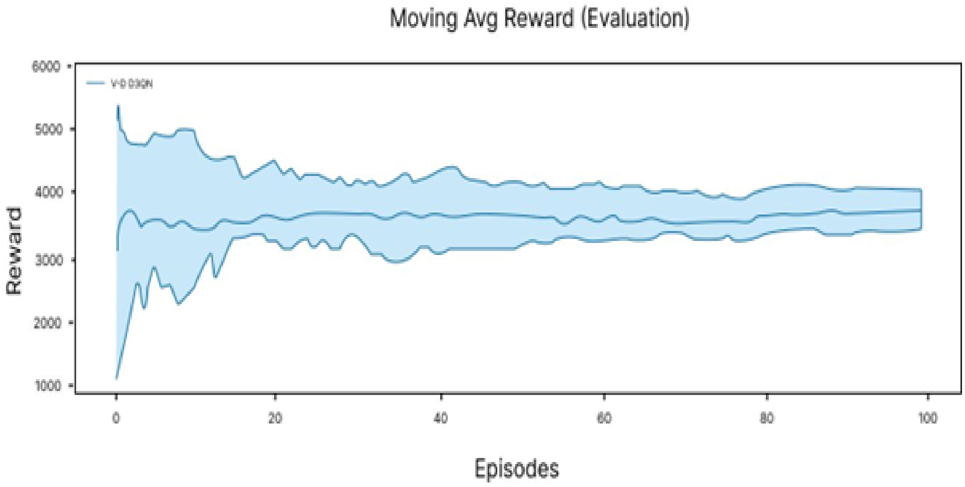


Figure 4.10: Moving average rewards for V-D D3QN algorithm

Additionally, the learning rate for both the policy and value networks was set to 0.0001. The replay buffer size was configured to be 10000, whereas the batch size was set to 32. The value of the soft update parameter was modified to 0.005. Figures 4.9 and 4.10 display the moving average rewards for both the training and evaluation phases. It is observed that the evaluation performance rapidly stabilises at approximately 3000 and 4000 during the initial episodes, and gradually improves thereafter. This indicates that the D3QN algorithm is proficient and successful in capturing the enduring systematic patterns present in the data.

Chapter 5

Result and Discussion

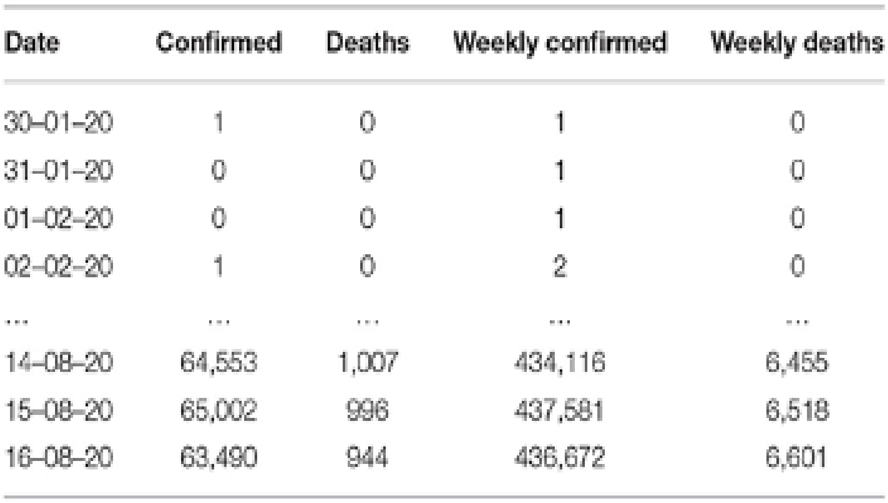
5.1 Introduction

This section is the crucial part of our research , since it discusses about the result and interpretation to those findings.

5.2 Result of the Analysis Performed with the Modified (LSTM)

The study incorporates summary tables pertaining to regular time intervals, encompassing data on the count of cases as well as deaths recorded within a specified number of days following the onset of the pandemic. The data analysis encompassed the period from January 22, 2020, which marks the date of the initial global report of the first COVID-19 case, to March 30, 2021. The dataset was divided into two subsets, with 70% of the data used for training purposes and the remaining 30% utilised for prediction and validation.

The provided data consists of granulated samples representing daily time series for con-



Date	Confirmed	Deaths	Weekly confirmed	Weekly deaths
30-01-20	1	0	1	0
31-01-20	0	0	1	0
01-02-20	0	0	1	0
02-02-20	1	0	2	0
...	...	...	...	...
14-08-20	64,553	1,007	434,116	6,455
15-08-20	65,002	996	437,581	6,518
16-08-20	63,490	944	436,672	6,601

Table 5.1: Daily and weekly death and confirmed cases for COVID-19 in the UK

firmed cases, death cases, and recovered cases. Additionally, Table 5.1 displays the daily and weekly data for confirmed cases and death cases. Figs 5.1,5.2,5.3,5.4,5.5 and 5.6 present a comparison of the confirmed, death, and recovered cases between the original and predicted figures from the onset of the pandemic into the future. In this particular instance, a significant correlation was seen between the actual cases and the projected cases. The presented data illustrates a comparison between the confirmed and recovered cases of death, both in their original and forecasted values. This comparison highlights the observable trend of a little increase in the confirmed, recovered, and death rates on certain days.

The COVID-19 pandemic is still alive and poses a major threat to people's health worldwide in a short period of time. This study proposes a deep learning-based prediction technique for predicting COVID-19 risk. The system examines the real daily dataset and use deep learning algorithms to provide forecasts for the coming days. The research presented here determines the optimal activation function for MLSTM by applying a deep reinforcement learning algorithm to maximise the prediction outcomes. When compared

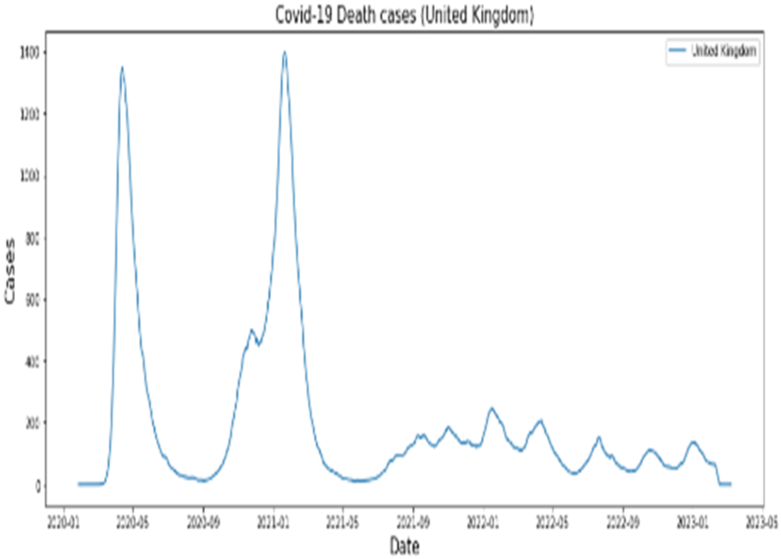


Figure 5.1: Death cases

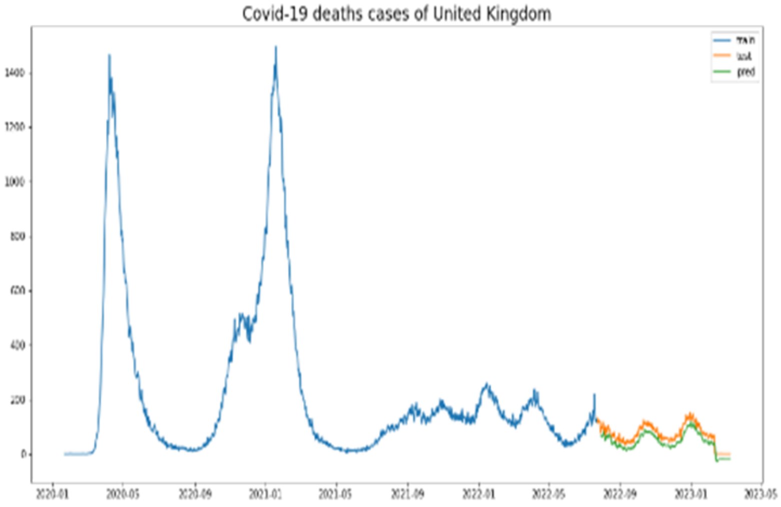


Figure 5.2: Death Prediction cases

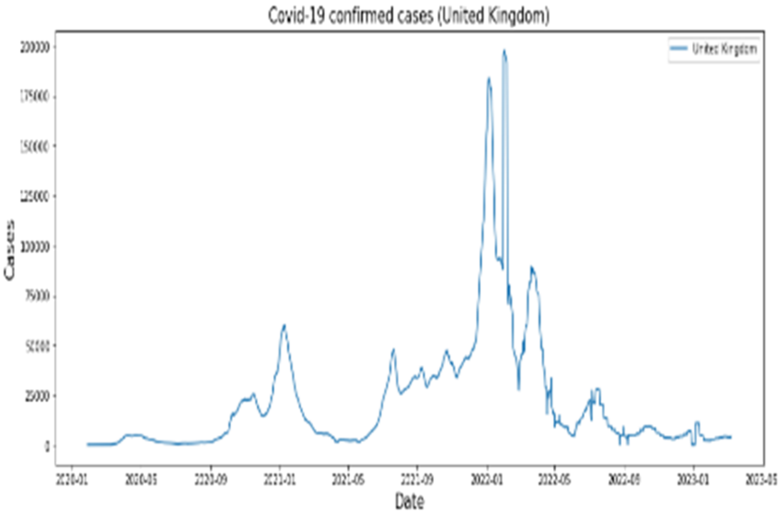


Figure 5.3: Confirmed cases

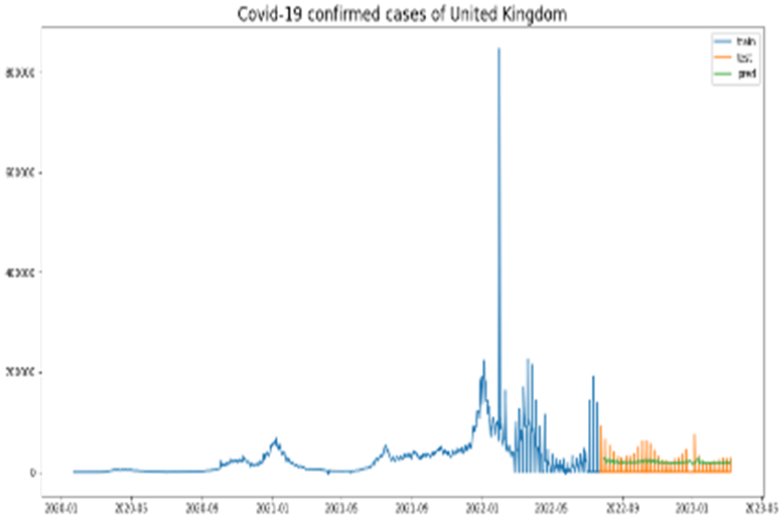


Figure 5.4: Confirmed Prediction cases

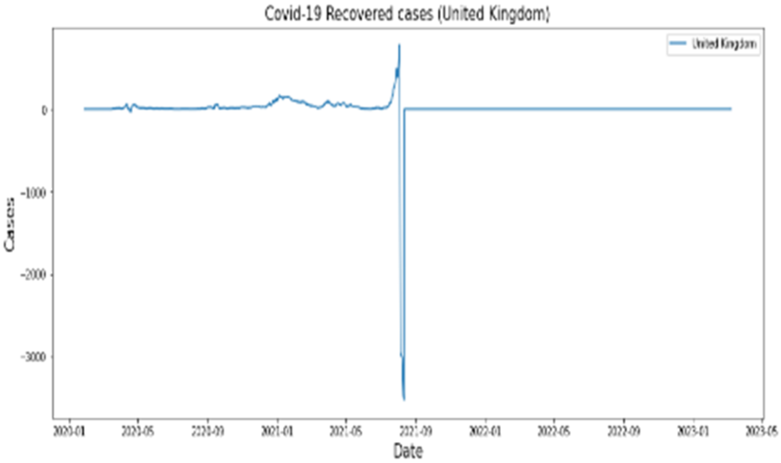


Figure 5.5: Recovery cases

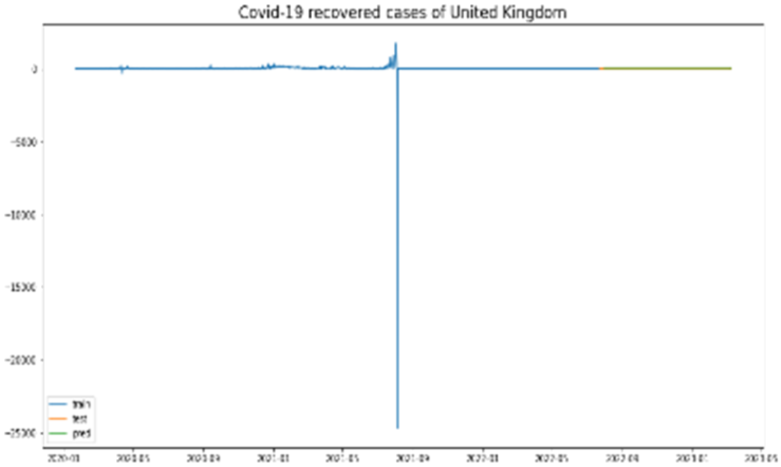


Figure 5.6: Recovery Prediction cases

to well-known contemporary algorithms such as LSTM. The findings of the study show that the DL technique is helpful in predicting COVID-19 cases in the future. Overall, the model's predictions are consistent with the trends of the virus cases; this may help us understand the virus and stop its spread. The forecast offered by this study may be of significant use in addressing the COVID-19 scenario by quick measures and educated assessments. To prevent additional COVID-19 and social media platform proliferation, future COVID-19 and social media platforms should be recognised using a semi-supervised hybrid architecture.

Two distinct measures were employed to assess the modification of the Long Short-Term Memory (LSTM) model, as elucidated in the preceding Chapter 4, Section 4.12.

5.3 Lockdown and Travel restrictions Optimal Policy for COVID-19 Pandemic control

Applying empirical disease data pertaining to COVID-19 in the United Kingdom spanning from January 2, 2020 to January 2, 2021, with the initial case being documented on January 31, 2020, we conducted experimental investigations to assess the effectiveness of our proposed model.

In order to facilitate the agent's search for the optimal policy, it is imperative to develop a well-designed incentive function. In the above scenario, the objective would entail managing the ongoing pandemic by implementing daily modifications to the extent of lockdown measures and travel limitations. From a cognitive perspective, individuals would naturally be inclined to mitigate the occurrence of elevated mortality rates or illnesses, as such circumstances would yield unfavourable outcomes. Conversely, it is imperative to promote specific actions and thereafter provide rewards for successfully mitigating the number of infections and death cases below a predetermined level. The reward system utilised inside our surroundings consists of two components, namely death and recovery. Each of these components possesses its own distinct system of rewards and punishments, thereby constituting the complete framework.

Subsequently, the aggregate reward for each day can be computed by summing the two components of the reward. The table that presents a visual representation of the overall sign orientation pertaining to the underlying justification for the establishment of awards can be seen in Chapter 5, section 5.10 . This orientation is determined by the level of severity in cases of death, and the same line of reasoning may be applied to instances of recovery. However, the sign will alter in this particular scenario. This study proposes the utilisation of discrete action space to decide the severity level of lockdowns and travel restrictions. When compared to a continuous action space, that was proposed by (Liu K, 2022) , a DDPG implementation, this approach demonstrates less reliability, particularly when taking into account the eventual ending of the epidemic.

5.3.1 Results

The suggested dueling DQN was trained and the findings were examined using data from three unique time periods related to the epidemic. The time periods under consideration can be categorised as follows: the initial three months, the entire duration, and the most recent three months.

In Figs 5.7,5.8 and 5.9, the agent was trained exclusively on data from the initial three months. As a result, the agent initially advised maintaining stringent regulation of both local and international policy. Nevertheless, commencing in mid-March, there was a discernible decline in the magnitude of the agent's suggested measures. In late March, ideas were put out regarding policy at a foundational level. During the interim period, the policies put forward in the past three months have attained a level of agreement with the policies suggested by the public health, as depicted in Figs 5.10, 5.11 and 5.12. To provide more clarification, the proposed lockdown policy exhibited a marginally lesser degree compared to the current public health policy. However, in terms of travel restrictions, the present public health policy shown a little higher degree.

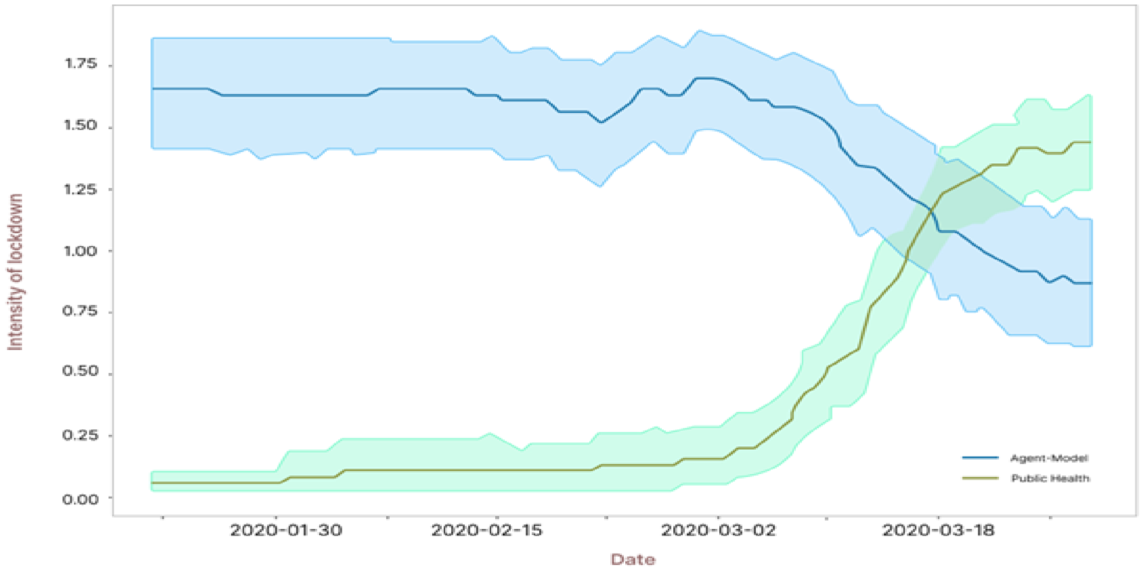


Figure 5.7: Lockdown 1st-3 months

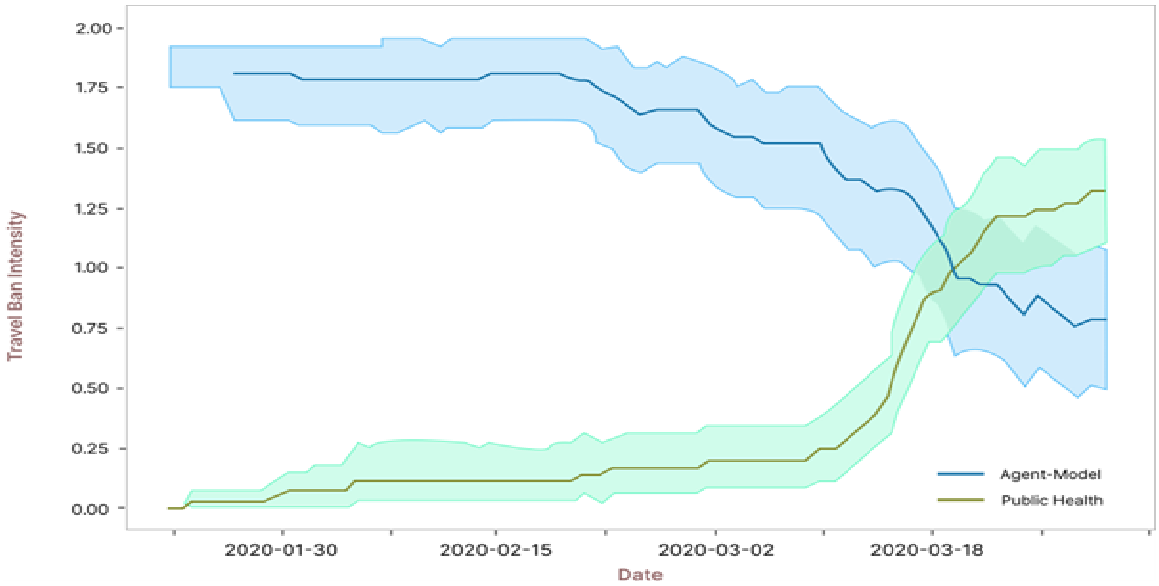


Figure 5.8: Travel ban for 1st-3 months

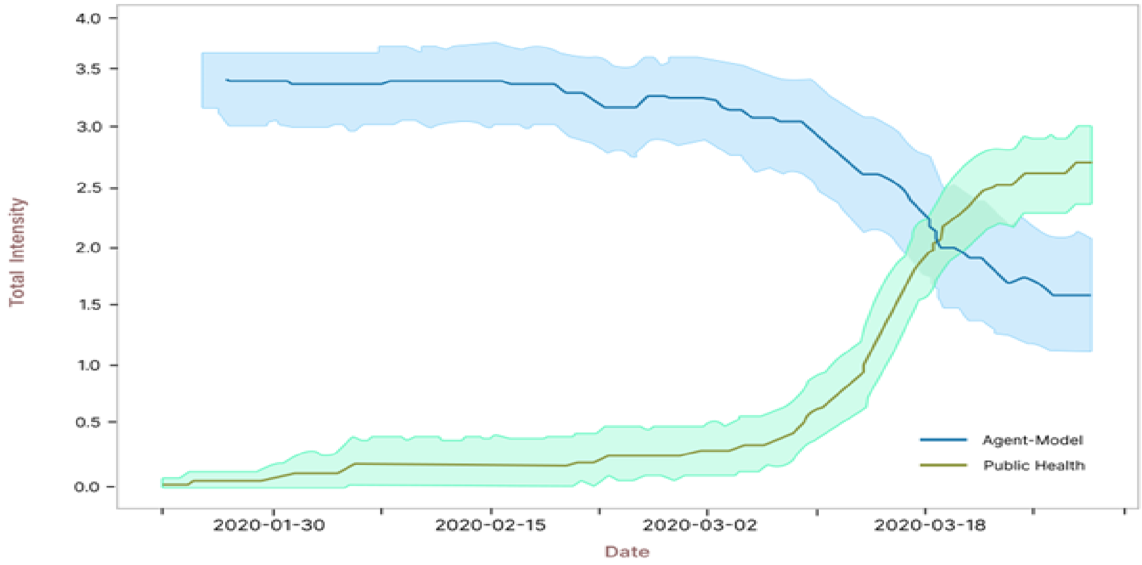


Figure 5.9: Total Intensity for 1st-3 months

Figs 5.13,5.14 and 5.15 illustrate the total intensity level of lockdown and travel restriction measures implemented by the public health authorities and endorsed by our agent using deep reinforcement learning over the duration of the pandemic. The implementation of these policies was prompted by the findings made by our agent regarding the proliferation of the pandemic. In the majority of instances, the agent recommended implementing a policy of lockdown or travel restrictions at level one significantly earlier than the point at which it was officially adopted by public health authorities, as indicated in figure 5.15. Our agent has suggested the implementation of a minimum degree of lockdown or travel restrictions in the United Kingdom. Despite the fact that the index case date occurred in late March, the agent suggested that the initial policy, regardless of its level, should have been implemented in late January. Conversely, the agent advised a postponement of policy implementation in the United Kingdom, whereas other public health experts advocated for prompt action, despite the absence of exponential growth in the number of cases.

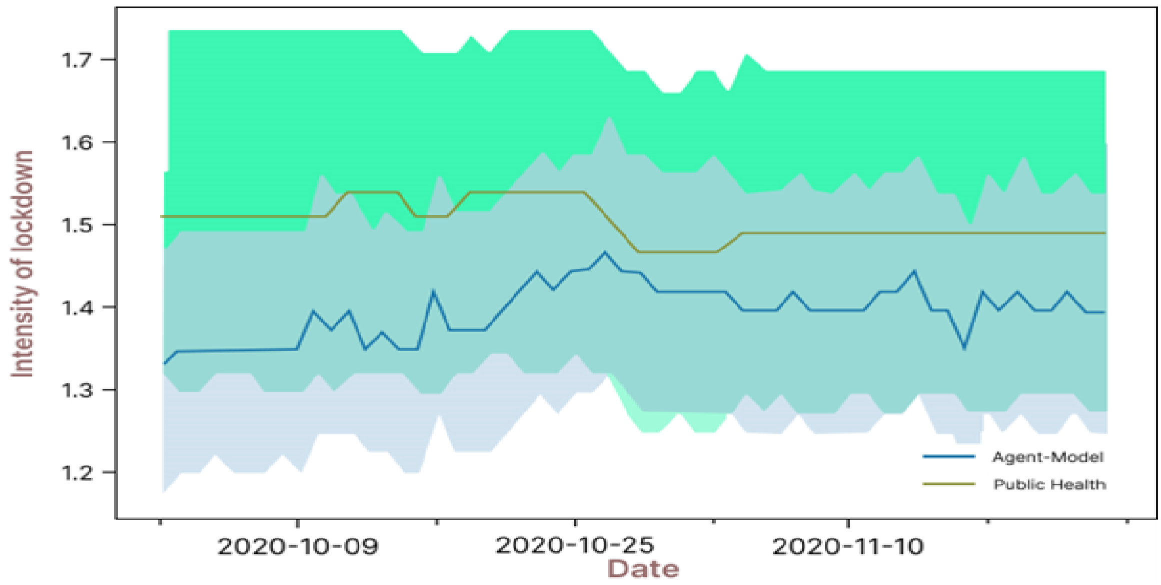


Figure 5.10: Lockdown for most last 3months

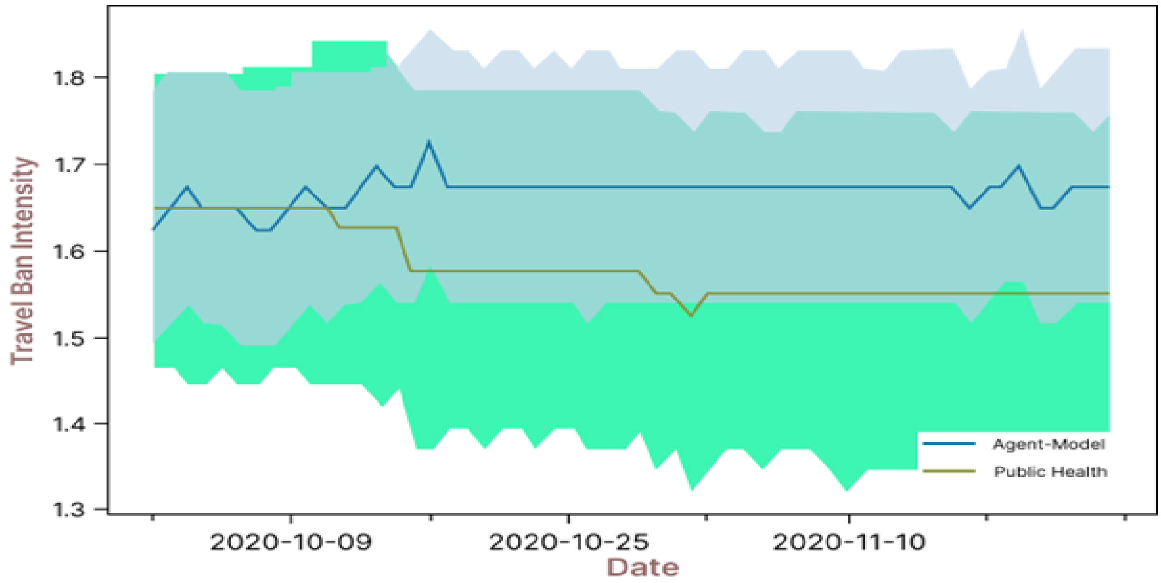


Figure 5.11: Travel ban for most last 3months

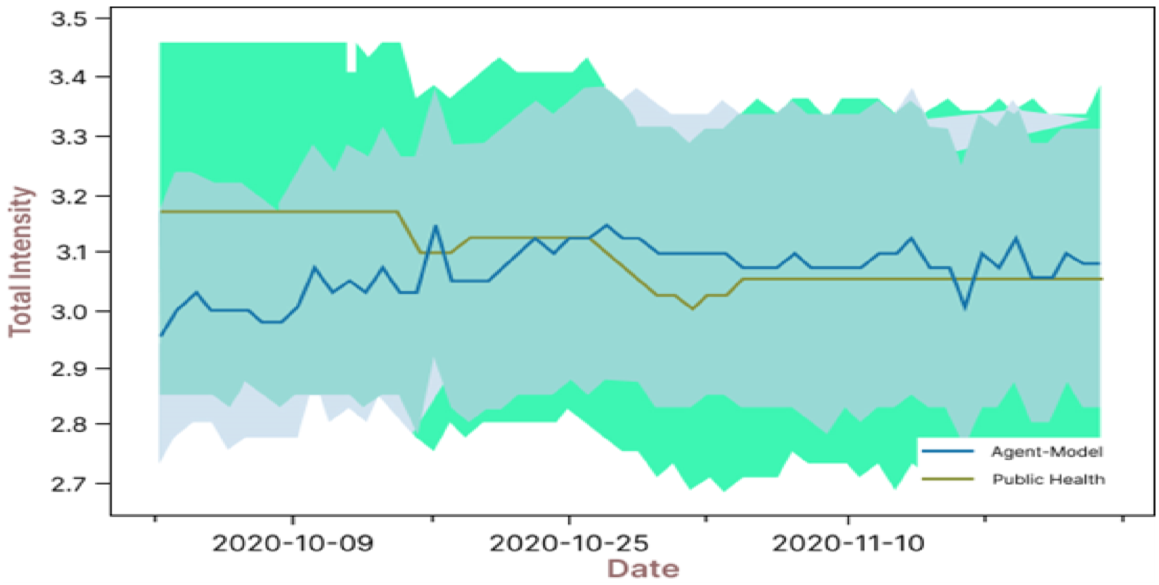


Figure 5.12: Total Intensity for most last 3months

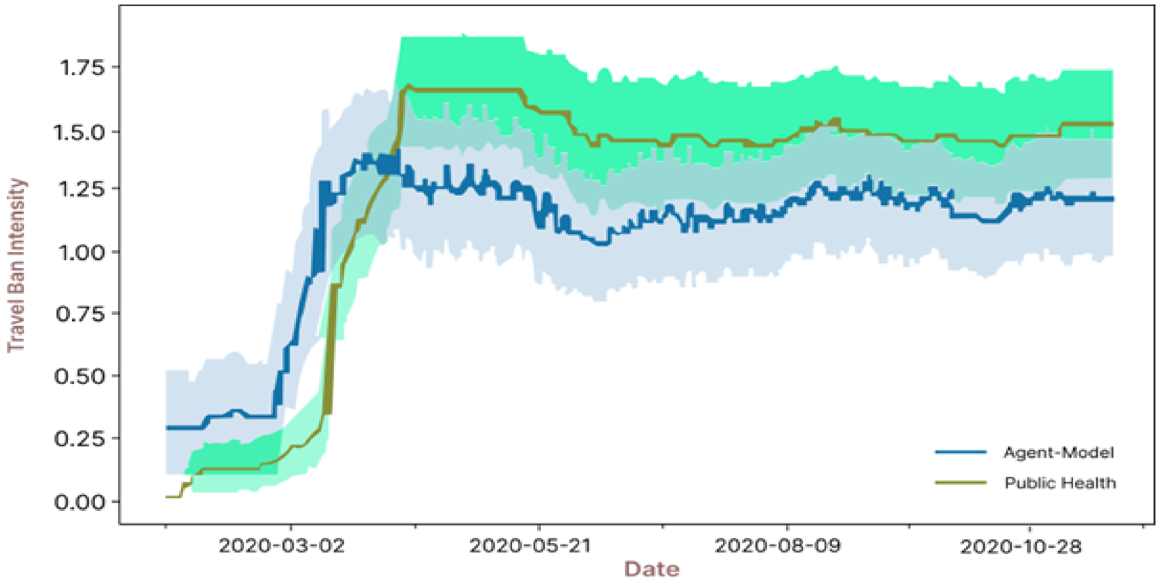


Figure 5.13: Lockdown for overtime period

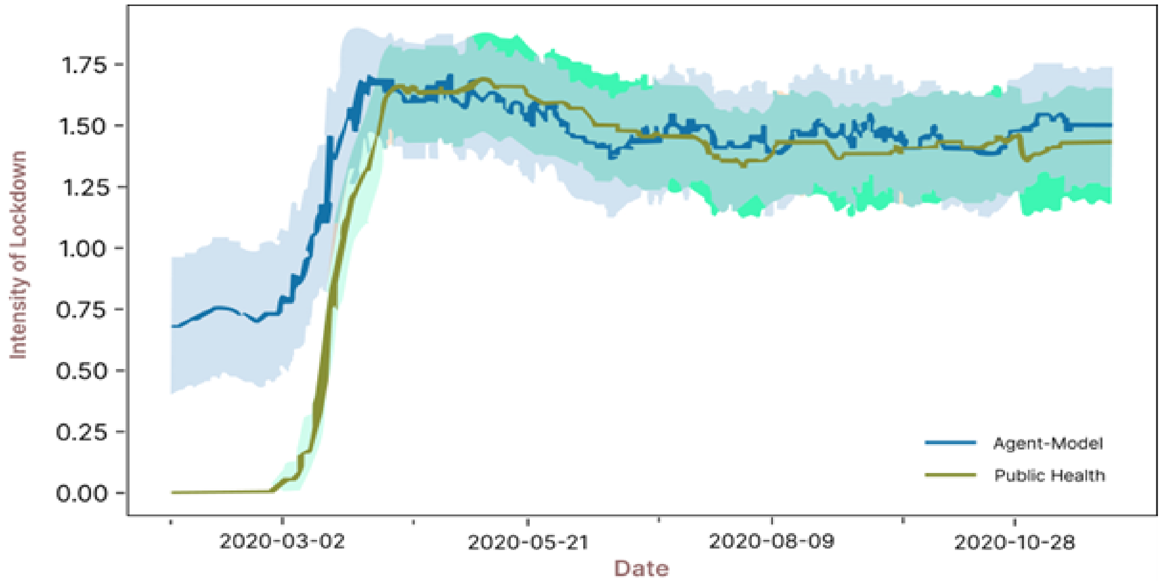


Figure 5.14: Travel ban for overtime period

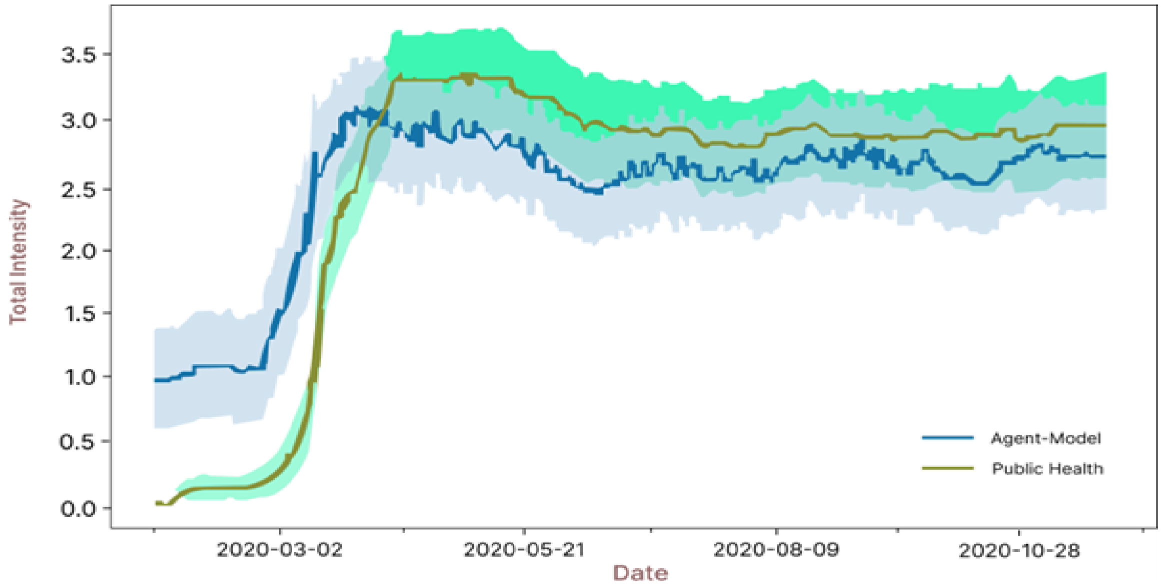


Figure 5.15: Total Intensity for overtime period

5.4 Discussion

This study presents a novel approach for training an agent to determine optimal timing and intensity of lockdown measures and travel restrictions for United Kingdom. The proposed method utilises Deep Reinforcement Learning (DRL) and leverages a dataset extracted from the United Kingdom's COVID-19 epidemiological data and global COVID-19 dataset from Johns Hopkins coronavirus data repository, specifically focusing on United Kingdom information. In order to ascertain the optimal course of action for individual states over a given period, we conducted a temporal analysis of policy implementation in relation to varying levels of crisis severity. This analysis was carried out using deep reinforcement learning techniques, specifically focusing on continuous state spaces and incentive structures. The algorithm we developed primarily recommended a comparatively less stringent approach to lockdown measures and travel restrictions in response to the COVID-19 pandemic, in contrast to the measures actually implemented by public health authorities. The recommendation put forth by our agent was to implement a lockdown at an earlier stage of the epidemic with a greater degree of severity.

The agent ultimately acquiesced to the policies of public health, though. Moreover, it aligned with the recommendations of public health authorities to mitigate the stringency of lockdown measures during the final phases of the pandemic. The significance of this matter lies in the need to carefully consider the early implementation of policies aimed at mitigating the impact of COVID-19, taking into account the potential consequences on economic, social, and health-related aspects. Sustaining prolonged lockdowns and border closures poses considerable challenges, despite the implementation of punitive measures such as fines and jail to deter non-compliance with public health regulations.

Furthermore, the agent provided counsel against consistently advocating for a comprehensive and complete lockdown. Instead, the agent advocated the implementation of an initial low-intensity lockdown, such as limitations on public meetings or the promotion of online e-learning, as a means of alerting and safeguarding the population.

When comparing the initial implementation to the subsequent official implementation, the agent proposed the imposition of travel restrictions as a minimum measure. In contrast to the projections made by travel agents, it is noteworthy that public health authorities have implemented and upheld more rigorous travel restrictions. The implications of the algorithm and the findings of the study suggest that the agent does not endorse protracted high intensity lockdowns and travel restrictions for the sake of UK public health. The outputs of the suggested technique were compared and assessed in the result of UK government result time line as shown in figure 5.16, the full duration, the most recent three months, and the beginning three months. This analysis was supported by Figures 5.7-5.12.

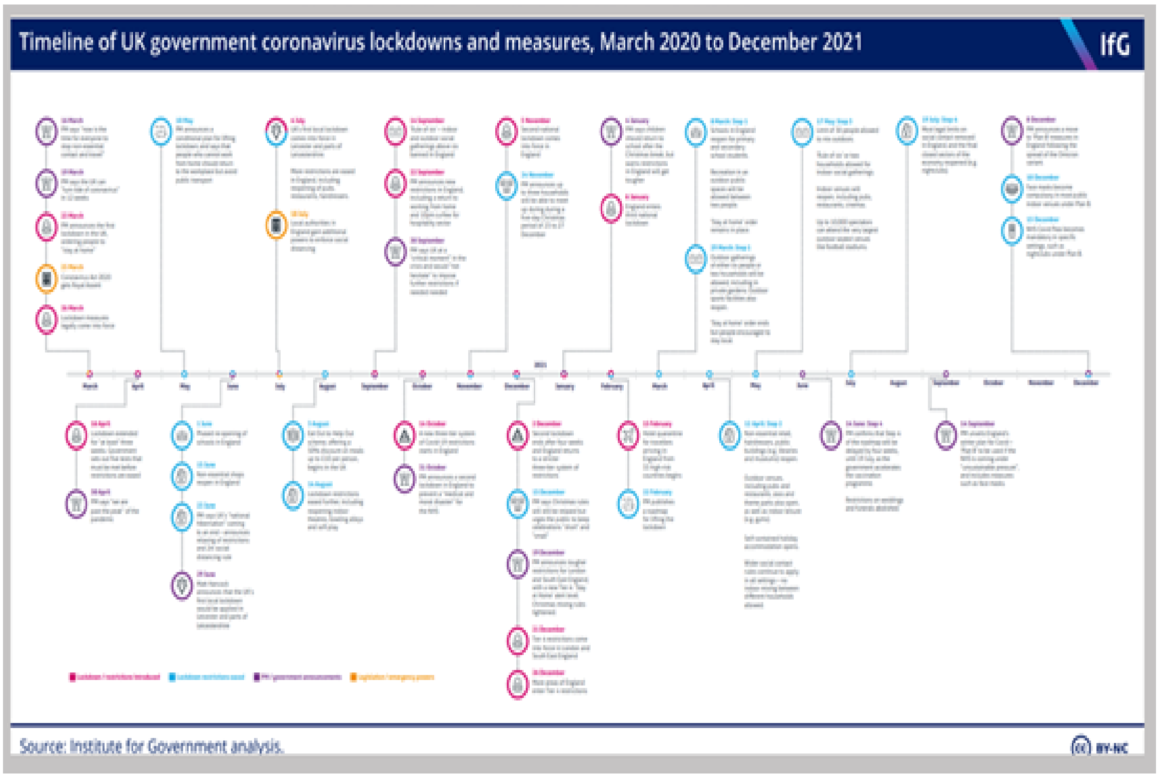
According to Figure 5.10-5.12, the agents first recommended the implementation of stringent regulatory policies, which were afterwards advised to be gradually relaxed starting in mid-March. This shift in policy intensity coincided with the utilisation of algorithm training based on data gathered during the initial three months. Based on an analysis of pandemic data spanning the last three months, Figures 5.13-5.15 illustrates that agents have reached a certain level of consensus about public health decisions, which is equivalent to the policies proposed throughout the entire period (Figure 5.7-5.9). However, there were inconsistencies seen in relation to travel restrictions. During the latter half of the research period, the policy on travel restrictions exhibited a little more lenient approach compared to that of public health authorities.

Nevertheless, the travel restriction policy implemented by the agent exhibited a slightly more stringent approach compared to the government's policy. This was mostly due to the fact that the agent only took into account data from the preceding three months. Consequently, no immediate action was taken, as depicted in Figures. This phenomenon can be attributed to the observation that in the case of Figures 5.7-5.9, the implementation of lockdown measures and travel restrictions, which were acquired over the duration of the epidemic, were promptly and intensively enforced during its initial stages, with subsequent adherence to these policies being largely sustained thereafter. In summary, the aforementioned findings, specifically Figures 5.7-5.9 and Figures 5.13-5.15, along with the data shown in Figures 5.10-5.13 for the initial three-month period, collectively suggest that initiating a programme at an early stage might effectively mitigate various forms of losses.

The research training of a reinforcement learning agent is contingent upon the accurate definition of a reward function. The reward functions serve as a weak indicator. In contrast to supervised learning, the incentive signal in reinforcement learning solely communicates the immediate input from the environment, without explicitly offering the proper answer to the current learning problem. The pursuit of immediate gains without considering the transition dynamics in the environment might lead to premature and suboptimal convergence. Hence, it is imperative that the reward definition takes into account the specific attributes of the environment in which the agent operates, enabling it to effectively learn and optimise its actions over an extended planning horizon.

In the process of seeking improvement, the learning algorithm may encounter challenges arising from inadequately specified reward functions, which might potentially lead to confusion or misguidance of the agent in certain circumstances. The article outlines the design of a customised reward function that incentivizes timely disease management and discourages adverse outcomes resulting from inadequate control measures. We suggest that this definition adequately fulfils the primary objective of disease prevention. Opting for a cautious and risk-averse approach to control at the earliest opportunity offers greater benefits compared to implementing the optimal policy suggested by our agent. This decision may prevent the potential consequence of incurring excessive costs during the implementation phase.

Chapter 4 subsection 4.12.1 shows the evaluation of the algorithm performance.


Timeline of UK government coronavirus lockdowns and measures, March 2020 to December 2021
IfG
BY-NC

Source: Institute for Government analysis.

Figure 5.16: COVID-19 UK Government Lockdown timeline

Chapter 6

Conclusion and Recommendations

6.1 Introduction

This chapter will examine the implications of the study's findings regarding the potential of DRL as a tool for determining the most effective policy for implementing lockdown measures and travel restrictions in the United Kingdom. The findings drawn in this study were derived from the stated aims, research problems, and obtained outcomes. This research will elucidate the consequences of the aforementioned findings and then provide an analysis of the resulting recommendations. The recommendations were derived from the study's conclusions and aligned with its intended objectives.

6.2 An Overview of the Research

The present study employed the widely accepted data mining process known as CRISP-DM in order to accomplish the predetermined objectives and effectively tackle the research difficulties at hand. This study aims to investigate the application of deep reinforcement learning as an agent in determining the optimal timing for implementing lockdown measures and travel bans. The study adopts an exploratory, descriptive, and contextual qualitative approach to gain a comprehensive understanding of this topic. The selection of research methodology and approach for this study was based on identified gaps in prior research and the findings of a thorough systematic review. The subsequent discussion pertains to the research issues, objectives, and outcomes derived from the data analysis. The research inquiries were centred on the subsequent research problems:

- Inadequate planning while selecting the best method to finding the optimal policy for lockdown and travel restriction in the case of COVID-19 in the United Kingdom, which might be because of insufficient data exploration on the best agent , would lead to the feature engineering required to develop the most effective machine learning models.
- The majority of studies utilised simulated data rather than real-world COVID-19 data.

The aforementioned objectives were formulated in order to effectively tackle the identified concerns.

- A comprehensive exploratory analysis was conducted to investigate the potential relationship between various features utilised for the development of an optimal policy. The analysis also recommended the most suitable pre-processing, transformation, and engineering techniques that should be implemented prior to constructing the optimal policy for an effective lockdown and travel restriction model.
- Having to identify the future trends for all cases firstly, let us to knew the severity of the disease and also helps us to know weather the agent is acting properly when administered to find the appropriate period for both lockdown and travel ban. In lieu

of a continuous action space that does not anticipate the conclusion of the pandemic, this study proposes the use of a novel discrete state and action representation based reinforcement learning paradigm.

6.3 Future Research

- The formulation of rewards in deep reinforcement learning (DRL) is a crucial component in the development and training of a proficient reinforcement learning agent. The reward function is a fundamental component that delineates the optimisation objectives for an agent and holds significant influence over the agent's behavioural patterns. The formulation of the reward function has a significant impact on both the learning dynamics and the overall performance of the agent. Thus, building an improved reward system to improve the optimal policy for lockdown and travel restriction.

- Explore the area of using deep reinforcement learning with regards to optimal vaccine distribution to different population based on various factors, such as infefection rates,population density, and healthcare infrastructure. The application of reinforcement learning (RL) in vaccine distribution systems enables the optimisation of vaccine allocation, resource management, and delivery processes. The COVID-19 pandemic has brought to the forefront the significance of effectively disseminating vaccines on a substantial magnitude. This my future research aims to discuss the potential application of Reinforcement Learning (RL) techniques in the context of vaccination distribution.

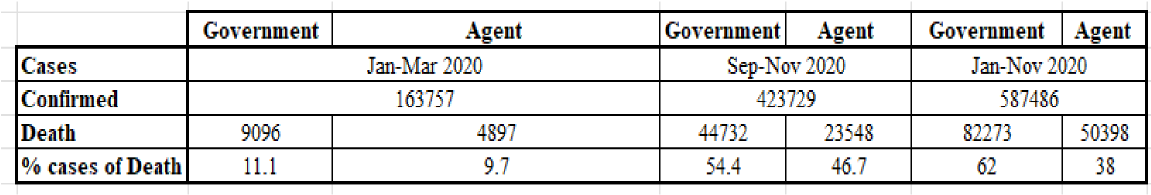
6.4 Research Limitation

This study examined COVID-19 epidemiology data as a primary source of information to determine the most effective policy for implementing lockdowns and travel restrictions. Nevertheless, there is a restricted availability of UK COVID-19 data, which can only be accessed through the UK government website and the National Office of Statistics. The public has access to exclusively aggregated data pertaining to COVID-19. Therefore, the research is dependent on global epidemiological data. During the initial stages of the COVID-19 pandemic, obtaining a sufficiently enough sample size posed challenges in order to achieve more precise outcomes.

The scope of this study is confined to the examination of COVID-19 data. However, it is worth noting that additional data pertaining to the specific aspects of quality of life and socio-economic factors during the pandemic might be organised and incorporated to facilitate a more comprehensive analysis.

6.5 Conclusion

This research employed deep reinforcement learning to assess the effectiveness of implementing a lockdown and travel restrictions in limiting the COVID-19 epidemic. Our study demonstrates that it is possible to train an algorithm to enable an agent to make informed decisions regarding the optimal strategy for maximising the expected value of total rewards over time in the United Kingdom and its territories. This is achieved by utilising data obtained from the local population, as well as epidemiological data specifically related to COVID-19. By engaging in comprehensive discussions regarding the conceptualization of the action space, state space, and reward function in reinforcement learning, we undertake a study of the underlying framework of the optimal control problem. This enables us to



	Government	Agent	Government	Agent	Government	Agent
Cases	Jan-Mar 2020		Sep-Nov 2020		Jan-Nov 2020	
Confirmed	163757		423729		587486	
Death	9096	4897	44732	23548	82273	50398
% cases of Death	11.1	9.7	54.4	46.7	62	38

Table 6.1: Showing cases and fatalities in three different months mode

ascertain the appropriate approach for resolving the situation. In this work, we provide a theoretical rationale for the value and policy networks of the Dueling DQN model, as well as evidence supporting its superior performance in comparison to alternative models. Our aim is to showcase the significant impact of this model. The demonstration of this was achieved through the utilisation of the available data. In contrast to the strategies implemented by public health and government authorities, the key proposition put out by the agent was the implementation of early lockdowns and travel restrictions as a means to mitigate the impact of COVID-19.

If the public health and government had followed the recommendations provided by our agent, as indicated in figures 5.1-5.12, which proposed implementing earlier lockdown measures and travel restrictions, it is estimated that an approximately total of 50,398 that is 47% of lives may have been spared from January to November 2020, based on the available dataset as per the period under consideration.

6.5.1 Novelty

This is the first research of its sort to be conducted in the United Kingdom that involved the use of deep reinforcement learning to the management of the spread of COVID-19.

Reference List

1. Aldridge, R. W., Lewer, D., Katikireddi, S. V., Mathur, R., Pathak, N., Burns, R., ... Hayward, A. (2020). Black, Asian and Minority Ethnic groups in England are at increased risk of death from COVID-19: indirect standardisation of NHS mortality data. Wellcome Open Research, 5(88), 1-17.

2. Ali I, Iryna P, Haneya NQ et al. AI4COVID-19: AI enabled preliminary diagnosis for COVID-19 from cough samples via an app. Inform Med Unlock 2020; 20: 100378.

3. Allington, D., Duffy, B., Wessely, S., Dhavan, N., Rubin, J. (2020). Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. Psychological Medicine, 51(10), 1763-1769.

4. Baldi P (2012) Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of the international conference on machine learning (ICML), pp 37–50

5. Barr D. Same storm, different boats. 2020. Available: latest/https/we-are not-all-in-the-same-boat

6. Barto A, G and S. Mahadevan, S. Recent advances in hierarchical reinforcement learning. Discrete Event Dynamic Systems, 13(4):341–379, 2003.

7. Bastiaanssen, W., Molden, D., Makin, I. (2000). Remote sensing for irrigated agriculture: examples from research and possible applications. Agricultural water management, 46(2), 137-155.

8. Becky M. COVID-19 and artificial intelligence: protecting health-care workers and curbing the spread. Lancet Digital Health 2020; 2: 4.e166–4.e167.

9. Bellazzi R, Abu-Hanna A. Data mining technologies for blood glucose and diabetes management. J

10. Bellemare MG, Dabney W, Munos R (2017) A distributional perspective on reinforcement learning. In: Proceedings of the 34th international conference on machine learning, vol 70, pp 449–458. .

11. Biavardi N. G., "Being an Italian medical student during the COVID-19 outbreak," International Journal of Medical Students, vol. 8, no. 1, pp. 49–50, 2020. BMJ Open 2022;12:e048279. doi: 10.1136/bmjopen-2020-048279

12. Bohg, J., Hausman, K., Sankaran, B., Brock, O., Kragic, D., Schaal, S., and Sukhatme, G. S. (2017). Interactive perception: Leveraging action in perception and perception in action. IEEE Transactions on Robotics.

13. BRACHMAN, R. J. ANAND, T. The Process of Knowledge Discovery in Databases: A First Sketch. KDD workshop, 1994. 1-12.

14. Bubar, K. M., Reinholt, K., Kissler, S. M., Lipsitch, M., Cobey, S., Grad, Y. H. (2021). Model-informed COVID-19 vaccine prioritization strategies by age and serostatus. Science, 371(6532), 916-921.

15. Caicedo, J. C. and Lazebnik, S. (2015). Active object localization with deep reinforcement learning. In the IEEE International Conference on Computer Vision (ICCV)

16. Chinazzi M, Davis JT, Ajelli M, Gioannini C, Litvinova M, Merler S, Piontti AP, Mu K, Rossi L, Sun K, Viboud C, Xiong X, Yu H, Halloran ME, Ira M.Longini Jr., Vespignani A. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science. 2020;368(6489):395–400 pmid:32144116

17. Chakraborty B, and Moodie, E, Statistical methods for dynamic treatment regimes. Springer, 2013.

18. Chinazzi, M. et al. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19)

19. Chinazzi, M., Davis, J. T., Ajelli, M., Gioannini, C., Litvinova, M., Merler, S., ... Viboud, C. (2020). The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. Science, 368(6489), 395-400.

20. Cho, K., Van Merri¨enboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259, 2014.

21. Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In NIPS 2014 Deep Learning and Representation Learning Workshop. COVID-19 epidemic in China. Science (80-.). 368, 638–642 (2020). 7. Khadilkar, H., Ganu, T. Seetharam, D. P. Optimising Lock.

22. Dagan, N., Barda, N., Kepten, E., Miron, O., Perchik, S., Katz, M. A., ... Balicer, R. D. (2021). BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. New England Journal of Medicine, 384(15), 1412-1423.

23. Davies, N. G., Kucharski, A. J., Eggo, R. M., Gimma, A., Edmunds, W. J., Jombart, T. (2020). Effects of non-pharmaceutical interventions on COVID-19 cases, deaths, and demand for hospital services in the UK: a modelling study. The Lancet Public Health, 5(7), e375-e385.

24. Davillas A, Jones AM. The COVID-19 pandemic and its impact on inequality of opportunity in psychological distress in the UK. SSRN. 2020.

25. De Pillis L, G and Radunskaya A, “The dynamics of an optimally controlled tumor

model: A case study," Mathematical and Computer Modelling, vol. 37, no. 11, pp. 1221–1244, 2003.

26. countries with Arima models. medrXiv. (2020).

27. Deng, L., Yu, D. (2014). Deep learning: methods and applications. Foundations and Trends in Signal Processing, 7(3-4), 197-387.

28. Dilsizian S. E and Siegel E. L. "Artificial intelligence in medicine and cardiac imaging: harnessing big distribution," Journal of Biomedical Engineering, vol. 26, no. 5, pp. 953–958, 2009.

29. Docherty, A. B., Harrison, E. M., Green, C. A., Hardwick, H. E., Pius, R., Norman, L., ... Semple, M. G. (2020). Features of 20133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. bmj, 369.

30. Dong T, Benedetto U, Sinha S, et al Deep recurrent reinforced learning model to compare the efficacy of targeted local versus national measures on the spread of COVID-19 in the UK

31. EClinica lMedicine (2020), https://doi.org/10.1016/j.eclinm.2020.100457 Effects. ArXiv Published Online First: 5 April. ⟨http://arxiv.org/abs/1504.01132⟩ (Accessed 10 May 2016). energy-aimed timetable rescheduling problem. Energies 2019; 12: 183461.

32. Engineering in Medicine and Biology Society. Annual Conference, pages 2978–2981, 2016. epidemic in China. Science (80-.). 368, 638–642 (2020)

33. Ernst E, Geurts P, and L. Wehenkel, "Tree-based batch mode reinforcement learning," Journal of Machine Learning Research, vol. 6, no. Apr, pp. 503–556, 2005.

34. EU Open Data Portal: FAO. (2009). How to Feed the World in 2050. Rome: Food and Agriculture Organization of the United Nations.

35. Ferretti, L., Wymant, C., Kendall, M., Zhao, L., Nurtay, A., Abeler-Dörner, L., ... Hinch, R. (2020). Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing. Science, 368(6491), eabb6936.

36. A. C. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. Nature, 584(7820), 257-261.

37. Fortunato M, Azar MG, Piot B, Menick J, Osband I, Graves A, Mnih V, Munos R, Hassabis D, Pietquin O, et al (2017) Noisy networks for exploration. arXiv:170610295.

38. Fujimoto S, van Hoof H, Meger D (2018) Addressing function approximation error in actor-critic methods. arXiv:180209477.

39. Gao J., Zheng P., Jia Y., Chen H., Mao Y., Chen S., et al. "Mental health problems and social media exposure during COVID-19 outbreak," PLOS One, vol. 15, no. 4, p. e0231924, 2020. Gebbers, R., Adamchuk, V. I. (2010). Precision agriculture and food security. Science, 327(5967), 828-831.

40. Glucose Control Diabetic Patients. In: Proceedings of the International Mechanical Engineering Congress and Exposition. 2015 Presented

41. Greenhalgh, T., Wherton, J., Shaw, S., Morrison, C. (2020). Video consultations for COVID-19. BMJ, 368, m998.

42. Ha D, Schmidhuber J (2018) Recurrent world models facilitate policy evolution. In: Advances in neural information processing systems, pp 2450–2462

43. Haarnoja T, Zhou A, Abbeel P, Levine S (2018) Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. arXiv:180101290.

44. Haarnoja, T., Hartikainen, K., Abbeel, P., and Levine, S. Latent space policies for hierarchical reinforcement learning. International Conference on Machine Learning (ICML), 2018a.

45. Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. International Conference on Machine Learning (ICML), 2018b.

46. Hale S, Webster S, Petherick A, Phillips T, Kira B. Variation in government responses to COVID-19 version 6.0, England: Blavatnik School of Government; May 25, 2020. [Working Paper]Available from www.bsg.ox.ac.uk/covidtracker

47. He, K., Zhang, X., Ren, S., and Sun, J. (2016d). Deep residual learning for image recognition. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

48. Health-Related Quality of Life in Tehran Lipid and Glucose Study (TLGS). Int J Endocrinol Metab. 2017;15(4): e14548. Published 2017 Oct 23. doi:10.5812/ijem.14548

49. Heess N, Sriram S, Lemmon J, Merel J, Wayne G, Tassa Y, Erez T, Wang Z, Eslami S, Riedmiller M, et al (2017) Emergence of locomotion behaviours in rich environments. arXiv:170702286.

50. Hemdan, E., Shouman M., Karar M. (2020). COVIDX-Net: A Framework of Deep Learning Classifiers

51. Hinch, R., Probert, W., Nurtay, A., Kendall, M., Wymant, C., Hall, M., ... Kucharski, A. (2020). Effective configurations of a digital contact tracing app: A report to NHSX. London: Imperial College London.

52. Hochreiter, S. and Schmidhuber, J. Long short-term memory. Neural Computation, 1997.

53. HOSSAIN, S., ABTAHEE, A., KASHEM, I., HOQUE, M. M. SARKER, I. H. Crime prediction using spatio-temporal data. International Conference on Computing Science, Communication and Security, 2020a. Springer, 277-289.

54. Hu F, Jiang J, Yin P. Prediction of potential commercially inhibitors against SARS-CoV-2 by multi-task deep model. arXiv [Preprint] arXiv:2003.00728. (2020).

55. Hu Y. Intersecting ethnic and native–migrant inequalities in the economic impact of the COVID-19 pandemic in the UK. Research in Social Stratification and Mobility. 2020; 68: 100528. https://doi.org/10. 1016/j.rssm.2020.100528 PMID: 32834346.

56. Huang H. McCullagh P. Clack N. Harper R., Feature selection and classification model construction on Type 2 Diabetic Patient's Data, 2004.

57. Huang Y., Wu Y., and Zhang W., "Comprehensive identification and isolation policies have effectively suppressed the spread of COVID-19," Chaos, Solitons Fractals, vol. 139, p. 110041, 2020.

58. Humphrey K, "Using reinforcement learning to personalize dosing strategies in a simulated cancer trial with high dimensional data," 2017.

59. Imai, N., Cori, A., Dorigatti, I., Baguelin, M., Donnelly, C. A., Riley, S., ... Ferguson, N. M. (2020). Report 3: Transmissibility of 2019-nCoV. Imperial College London, 25. Intel Neurosci 2019; 2019: 1–19. International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE .

60. Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In the International Conference on

Machine Learning (ICML).

61. Ishimwe, R., Abutaleb, K., Ahmed, F. (2014). Applications of thermal imaging in agriculture—A review. Advances in Remote Sensing, 3(3), 128.

62. J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). IEEE Transactions on Autonomous Mental Development, 2(3):230–247, 2010.

63. Jaderberg, M., Mnih, V., Czarnecki, W., Schaul, T., Leibo, J. Z., Silver, D., and Kavukcuoglu, K. (2017). Reinforcement learning with unsupervised auxiliary tasks. In the International Conference on Learning Representations (ICLR).

64. Jaiswal AK, Tiwari P, Kumar S, Gupta D, Khanna A, Rodrigues JJ. Identifying pneumonia in chest X-rays: a deep learning approach. Measurement. (2019).

65. Jalali-Farahani S, Amiri P, Bakht S, Shayeghian Z, Cheraghi L, Azizi F. Socio-Demographic Determinants of Health-Related Quality of Life in Tehran Lipid and Glucose Study (TLGS). Int J Endocrinol Metab. 2017 Oct 23;15(4):e14548. doi: 10.5812/ijem.14548. PMID: 29344034; PMCID: PMC5750782.

66. Jalalimanesh A, Haghighi H,S, Ahmadi A, and Soltani M, “Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning,” Mathematics and Computers in Simulation, vol. 133, pp. 235–248, 2017.

67. Javad M, Agboola SO, Jethwani K, Zeid A, Kamarthi S. A Reinforcement Learning-Based Method for Management of Type 1 Diabetes: Exploratory Study. JMIR Diabetes. 2019 Aug 28;4(3):e12905. doi: 10.2196/12905. PMID: 31464196; PMCID: PMC6737889.

68. Jia L, Li K, Jiang Y, Guo X. Prediction and analysis of coronavirus disease 2019. arXiv [Preprint] arXiv:2003.05447. (2020).

69. Jie, Z., Liang, X., Feng, J., Jin, X., Lu, W. F., and Yan, S. (2016). Tree-structured reinforcement learning for sequential object localization. In the Annual Conference on Neural Information Processing Systems (NIPS).

70. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine Learning and Decision Support in Critical Care. Proc IEEE Inst Electr Electron Eng. 2016 Feb;104(2):444-466. doi: 10.1109/JPROC.2015.2501978. Epub 2016 Jan 25. PMID: 27765959; PMCID: PMC5066876.

71. Joseph B, Alexandra L, Katherine HP et al. Mapping the landscape of artificial intelligence applications against covid-19, 2020,

72. Kamilaris, A., Gao, F., Prenafeta-Boldú, F. X., Ali, M. I. (2016). Agri-IoT: A semantic framework for Internet of Things-enabled smart farming applications. 3rd World Forum on Internet of Things (WF-IoT) (págs. 442-447). Reston, VA, USA: IEEE.

73. Kamilaris, A., Kartakoullis, A., Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, 143(1), 23-37.

74. Kass-Hout, T. A., Alhinnawi, H., Wehbe, R. (2020). Public health surveillance systems: recent advances in their use and evaluation. Annual Review of Public Health, 41, 137-155.

75. Kavadi DP, Patan R, Ramachandran M, Gandomi AH. Partial derivative nonlinear global pandemic machine learning prediction of covid 19. Chaos Solitons Fractals. (2020).

76. Kerimray A., Baimatova N., Ibragimova O. P., Bukenov B., Kenessov B., Plotitsyn P., et al. "Assessing air quality changes in large cities during COVID-19 lockdowns:

The impacts of traffic-free urban conditions in almaty, kazakhstan," Science of The Total Environment, vol. 730, p. 139179, 2020.

77. Khadilkar H, Ganu T, Seetharam DP. Optimising Lockdown Policies for Epidemic Control using Reinforcement Learning. arXiv Preprint arXiv200314093. 2020.

78. Kingma D and J. Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015

79. Kitzes, J., Wackernagel, M., Loh, J., Peller, A., Goldfinger, S., Cheng, D., Tea, K. (2008). Shrink and share: humanity's present and future Ecological Footprint. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 363(1491), 467-475.

80. Kober J and Peters J R, "Policy search for motor primitives in robotics," in Advances in Neural Information Processing Systems, 2009, pp. 849–856.

81. Komorowski M, Gordon A, Celi L. A, and Faisal A. A. "Markov Decision Process to suggest optimal treatment of severe infections in intensive care". In Neural Information Processing Systems Workshop on Machine Learning for Health, December 2016.

82. Kong, X., Xin, B., Wang, Y., and Hua, G. (2017). Collaborative deep reinforcement learning for joint object search. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

83. Korea Centres for Disease Control and Prevention. The updates on novel Coronavirus infection in Korea. https://www.cdc.go.kr (2020).

84. Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In the Annual Conference on Neural Information

Processing Systems (NIPS)

85. Kwak GH, Ling L, Hui P (2021) Deep reinforcement learning approaches for global public health strategies for COVID-19 pandemic. PLOS ONE 16(5): e0251550.

86. Larremore, D. B., Wilder, B., Lester, E., Shehata, S., Burke, J. M., Hay, J. A., ... Parker, R. (2020). Test sensitivity is secondary to frequency and turnaround time for COVID-19 surveillance. medRxiv.

87. Learning and decision support in critical care," Proceedings of the IEEE, vol. 104, no. 2, pp. 444–466, 2016.

88. Learning for Sepsis Treatment", e-prints, 2017.

89. LeCun, Y., Bengio, Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10).

90. LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature, 521:436–444.

91. Lee, Seunghoon Lee, Young. (2020). Improving Emergency Department Efficiency by Patient

92. Leite H., Hodgkinson I. R., and Gruber T., "New development:'Healing at a distance'—telemedicine and COVID-19," Public Money Management, pp. 1–3, 2020.

93. Levine, S., Finn, C., Darrell, T., and Abbeel, P. (2016a). End-to-end training of deep visuomotor policies. The Journal of Machine Learning Research, 17:1–40.

94. Lewer, D., Braithwaite, I., Bullock, M., Aldridge, R. W., Hayward, A. C. (2020). Occupation and COVID-19 mortality: A record-linkage study. International Journal of Epidemiology, 49(6), 1895-1902.

95. Li Y. Reinforcement learning applications. arXiv Prepr arXiv190806973. 2019.

96. Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In the International world wide web conference(www).

97. Liaghat, S., Balasundram, S. K. (2010). A review: The role of remote sensing in precision agriculture. American journal of agricultural and biological sciences, 5(1), 50-55.

98. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, Silver D, Wierstra D (2015) Continuous control with deep reinforcement learning. arXiv:150902971

99. Liu Y, Logan B, Liu N, Xu Z, Tang J, Wang Y. Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data. Healthc Inform. 2017 Aug;2017:380-385. doi: 10.1109/ICHI.2017.45. PMID: 29556119; PMCID: PMC5856473.

100. Li L, Qin L, Xu Z, Yin Y, Wang X, Kong B, Bai J, Lu Y, Fang Z, Song Q, Cao K, Liu D, Wang G, Xu Q, Fang X, Zhang S, Xia J, Xia J. Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy. Radiology. 2020 Aug;296(2):E65-E71. doi: 10.1148/radiol.2020200905. Epub 2020 Mar 19. PMID: 32191588; PMCID: PMC7233473.

101. Lucas, T. C., Lse, R., University, C. (2021). The Impact of Urbanization on COVID-19: Evidence from the United Kingdom. Environmental and Resource Economics, 79(4), 621-651.

102. Mahsa OMJ, Stephen OA, Kamal J et al. A reinforcement learning–based method for management of type 1 diabetes: exploratory study. JMIR Diabet 2019; 4: 3e12905.

103. Malavika B, Marimuthu S, Joy M, Nadaraj A, Asirvatham ES, Jeyaseelan L. Forecasting COVID-19 epidemic in India and high incidence states using SIR and logistic

growth models. Clin Epidemiol Glob Health. (2021).

104. Malik MI. Machine learning the phenomenology of COVID-19 from early infection dynamics. Medrxiv, 2020,

105. Malik MI. Machine learning the phenomenology of COVID-19 from early infection dynamics. Medrxiv, 2020,

106. Martens J and Grosse R. Optimizing neural networks with kronecker-factored approximate curvature. In International conference on machine learning, pages 2408–2417, 2015.

107. Mathe, S., Pirinen, A., and Sminchisescu, C. (2016). Reinforcement learning for visual object detection. In the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

108. Millett G. A., Jones A. T., Benkeser D., Baral S., Mercer L., Beyrer C., et al. "Assessing differential impacts of COVID-19 on Black communities," Annals of Epidemiology, 2020.

109. Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A., Banino, A., Denil, M., Goroshin, R., Sifre, L., Kavukcuoglu, K., Kumaran, D., and Hadsell, R. (2017). Learning to navigate in complex environments. In the International Conference on Learning Representations (ICLR).

110. Mnih V, Badia A, P, Mirza M, Graves V, Lillicrap T, Harley T, Silver D, and Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In International Conference on Machine Learning, pages 1928–1937, 2016.

111. Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (2016) Asynchronous methods for deep reinforcement learning. In: International

Conference on Machine Learning (ICML), pp 1928–1937.

112. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D. Human-level control through deep reinforcement learning. Nature. 2015 Feb 26;518(7540):529-33. doi: 10.1038/nature14236. PMID: 25719670.

113. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D (2015) Human-level control through deep reinforcement learning.

114. Mnih, V., Badia, A. P., Mirza, M., Graves, A., Harley, T., Lillicrap, T. P., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In the International Conference on Machine Learning (ICML).

115. Mnih, V., Heess, N., Graves, A., and Kavukcuoglu, K. (2014). Recurrent models of visual attention. In the Annual Conference on Neural Information Processing Systems (NIPS).

116. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602, 2013

117. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540):529–533.

118. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. Nature, 2015. monitor people following guidelines to avoid COVID-19. J Sensor 2020; 2020: 1–15

119. Moy N., Antonini M., Kyhlstedt M., Paolucci F. Categorising policy technology interventions for a pandemic: a comparative and conceptual framework [Working paper] (June 2020). Available at SSRN: https://ssrn.com/abstract= 3622966.

120. Muhammad S, Tanveer Z, Sajid AK. Decision tree classification: ranking journals using IGIDI. J Inform Sci 2019; 46: 3325–3339.

121. Nagabandi A, Kahn G, Fearing RS, Levine S (2018) Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, Piscataway, pp 7559–7566

122. Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. Annu Int Conf IEEE Eng Med Biol Soc. 2016 Aug;2016:2978-2981. doi: 10.1109/EMBC.2016.7591355. PMID: 28268938. Ngabo D,

123. Dong W, Ibeke E, Iwendi C, Masabo E. Tackling pandemics in smart cities using machine learning architecture. Math Biosci Eng. (2021).

124. Ngo, Phuong Wei, Susan Holubová, Anna Muzik, Jan Godtliebsen, Fred. (2018). Reinforcement-Learning Optimal Control for Type-1 Diabetes. .

125. Osband I, Van Roy B, and Wen Z. Generalization and exploration via randomized

value functions. In International Conference on Machine Learning, pages 2377–2386, 2016.

126. Our World in Data: https://ourworldindata.org/coronavirus-source-data

127. Padmanabhan R, Meskin N, and Haddad W,M, “Reinforcement learning-based control of drug dosing for cancer chemotherapy treatement,” Mathematical biosciences, vol. 293, pp. 11–20, 2017.

128. Page, M. J. et al. (2021). The PRISMA statement: An updated guideline for reporting systematic reviews. PLoS Medicine, 18(3), e1003583.

129. Pal R, Sekh AA, Kar S, Prasad DK. Neural network-based country wise risk prediction of COVID-19. Appl Sci. (2020).

130. Pan, D., Sze, S., Minhas, J. S., Bangash, M. N., Pareek, N., Divall, P., ... Bell, D. (2020). The impact of ethnicity on clinical outcomes in COVID-19: A systematic review. EClinicalMedicine, 23, 100404.

131. pandemic. (2020) Patel P, Hiam L, Sowemimo A, Devakumar D, McKee M. Ethnicity and COVID-19. BMJ. 2020; 369. .

132. Petropoulos F, Makridakis S. Forecasting the novel coronavirus COVID-19. PLos ONE. (2020) 15:e0231236. doi: 10.1371/journal.pone.0231236.

133. Pierce M, McManus S, Hope H, Hotopf M, Ford T, Hatch SL, et al. Mental health responses to the COVID-19 pandemic: A latent class trajectory analysis using longitudinal UK data. The Lancet Psychiatry. 2021; 8: 610–619. .

134. Punn NS, Sonbhadra SK, Agarwal S. COVID-19 epidemic analysis using machine learning and deep learning algorithms. MedRxiv. (2020)

135. Racanière S, Weber T, Reichert D, Buesing L, Guez A, Rezende DJ, Badia AP, Vinyals O, Heess N, Li Y, et al (2017) Imagination-augmented agents for deep reinforcement learning. In: Advances in neural information processing systems, pp 5690–5701

136. Raghu A, Komorowski M Celi L. A, Szolovits P, and Ghassemi M. "Continuous state-space models for optimal sepsis treatment" - a deep reinforcement learning approach. arXiv preprint Raghu, A. et al. Deep reinforcement learning for sepsis treatment. arXiv Prepr. arXiv1711.09602 (2017).

137. Rajaei, A. Vahidi-Moghaddam, A. Chizfahm, and M. Sharifi, "Control of malaria outbreak using a non-linear robust strategy with adaptive gains," IET Control Theory Applications, vol. 13, no. 14, pp. 2308–2317, 2019.

138. Rajkumar R. P., "COVID-19 and mental health: A review of the existing literature," Asian journal of psychiatry, vol. 52, p. 102066, 2020.

139. Rao, Y., Lu, J., and Zhou, J. (2017). Attention-aware deep reinforcement learning for video face recognition. In the IEEE International Conference on Computer Vision (ICCV). Reinforcement Learning. arXiv Prepr. arXiv2003.14093 (2020)

140. Remuzzi, A., Remuzzi, G. (2020). COVID-19 and Italy: what next? The Lancet, 395(10231), 1225-1228.

141. Rismanchian, F., Lee, Y.H. Process Mining–Based Method of Designing and Optimizing the Layouts of Emergency Departments in Hospitals. HERD 2016, 10, 105–120.

142. Russell, T. W., Hellewell, J., Abbott, S., Gimma, A., Jarvis, C. I., van Zandvoort, K., ... Edmunds, W. J. (2020). Estimating the infection and case fatality ratio for

coronavirus disease (COVID-19) using age-adjusted data from the outbreak on the Diamond Princess cruise ship, February 2020. Eurosurveillance, 25(12), 2000256.

143. Rustam F, Reshi AA, Mehmood A, Ullah S, On BW, Aslam W, et al. COVID19 future forecasting using supervised machine learning models. IEEE Access. (2020)

144. Rustam F, Reshi AA, Mehmood A, Ullah S, On BW, Aslam W, et al. COVID19 future forecasting using supervised machine learning models. IEEE Access. (2020).

145. Salgotra R, Gandomi M, Gandomi AH. Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. Chaos Solitons Fractals. (2020). Salimans T, Ho J, Chen X, Sidor S, and Sutskever I. Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864, 2017.

146. SALTZ, J. S. HOTZ, N. Identifying the most common frameworks data science teams use to structure and coordinate their projects. 2020 IEEE International Conference on Big Data (Big Data), 2020. IEEE, 2038-2042.

147. Saxena, L., Armstrong, L. (2014). A survey of image processing techniques for agriculture. Perth, Australia: Proceedings of Asian Federation for Information Technology in Agriculture, Australian Society of Information and Communication Technologies in Agriculture.

148. Schaul T, Quan J, Antonoglou I, Silver D (2015) Prioritized experience replay. Preprint, arXiv:1511.05952.

149. Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized experience replay. arXiv preprint arXiv:1511.05952, 2015b. Scheduling Using Deep Reinforcement Learning. Healthcare. 8. 77. 10.3390/ healthcare8020077.

150. Schmidhuber J, Hochreiter S. Long short-term memory. Neural Comput. (1997)

151. Schulman J, Moritz P, Levine S, Jordan M, and Abbeel P, High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015.

152. Schulman J, Levine S, Abbeel P, Jordan M, and Moritz P. Trust region policy optimization. In International Conference on Machine Learning, pages 1889–1897, 2015.

153. Schulman J, Levine S, Abbeel P, Jordan M, Moritz P (2015) Trust region policy optimization. In: International Conference on Machine Learning (ICML), pp 1889–1897.

154. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O (2017) Proximal policy optimization algorithms. arXiv:170706347.

155. Schulman, J., Duan, Y., Ho, J., Lee, A., Awwal, I., Bradlow, H., Pan, J., Patil, S., Goldberg, K., and Abbeel, P. Motion planning with sequential convex optimization and convex collision checking. International Journal of Robotics Research (IJRR), 2014.

156. Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. International Conference on Machine Learning (ICML), 2015a.

157. Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015b.

158. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

159. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al (2016) Mastering the game of go with deep neural networks and tree search. Nature 529(7587):484

160. Silver D, Hubert T, Schrittwieser J, Antonoglou I, Lai M, Guez A, Lanctot M, Sifre L, Kumaran D, Graepel T, et al (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 362(6419):1140–1144

161. Souch, C., Cossins, L. (2020). The Impact of Population Density and Transport Networks on COVID-19 Infections in the UK. BMC Public Health, 20(1), 1-10.

162. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15:1929–1958.

163. Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In the Annual Conference on Neural Information Processing Systems (NIPS).

164. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT Press, Cambridge

165. Sutton RS, Barto AG (2018) Reinforcement learning: an introduction. MIT Press, Cambridge.

166. Sutton RS, Barto AG. Reinforcement learning: An introduction. 2011

167. Sutton RS, McAllester DA, Singh SP, Mansour Y (2000) Policy gradient methods for reinforcement learning with function approximation. In: Advances in neural information processing systems, pp 1057–1063.

168. Sutton, R. S. Barto, A. G. Reinforcement learning: An introduction. (2011) 2019-20 coronavirus

169. Takens, F “Detecting strange attractors in turbulence,” in Dynamical Systems and urbulence, Warwick 1980, ser. Lecture Notes in Mathematics, 1981, pp. 366–381.

170. Teke, M., Deveci, H. S., Haliloğlu, O., Gürbüz, S. Z., Sakarya, U. (2013). A short survey of hyperspectral remote sensing applications in agriculture. Istanbul, Turkey: 6th International Conference on Recent Advances in Space Technologies (RAST), IEEE.

171. The Centre for Systems Science and Engineering, Johns Hopkins. Coronavirus COVID-19 Data Repository. https://github.com/CSSEGISandData/COVID-19(2020). United Nations. United Nations Data Retrieval System. ttps://data.un.org(2020).

172. Theocharous, G., Thomas, P. S., and Ghavamzadeh, M. (2015). Personalized ad recommendation systems for life-time value optimization with guarantees. In the International Joint Conference on Artificial Intelligence (IJCAI).

173. Thornton, J. (2020). Covid-19: AE visits in England fall by 25

174. Tian, H. et al. An investigation of transmission control measures during the ¦rst 50 days of the to Diagnose COVID-19 in X-Ray Images. arXiv:2003.11055.

175. Togaçar M, Ergen B, Cömert Z. Application of breast cancer diagnosis based on a combination of convolutional neural networks, ridge regression and linear discriminant analysis using invasive breast cancer images processed with autoencoders. Med Hypotheses. (2020).

176. Tommi J, Jordan MI and Satinder SP. Convergence of stochastic iterative dynamic programming Treatment Regimes on Medical Registry Data,” 2017 IEEE Interna-

tional Conference on Healthcare Informatics (ICHI), Park City, UT, 2017, pp. 380-385, doi: 10.1109/ICHI.2017.45.

177. Tyagi, A. C. (2016). Towards a Second Green Revolution. Irrigation and Drainage, 65(4), 388- 389 type 2 diabetes patients' data, 20017, pp 251-262.Zeki, Tawfik Saeed, et al. "An expert system for diabetes diagnosis." American Academic Scholarly Research Journal 4.5 (2012):

178. Van H, Guez A, and Silver P, Deep reinforcement learning with double q-learning. CoRR, abs/1509.06461, 2015.

179. Van Hasselt H, Guez A and Silver D, "Deep Reinforcement Learning with Double Q-learning", ArXiv150906461 Cs, Dec. 2015, [online] Available:.

180. Van Hasselt, H. Double Q-learning. Advances in Neural Information Processing Systems (NeurIPS), 2010.

181. Van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double Q-learning. AAAI Conference on Artificial Intelligence, 2016.

182. Vapnik V, Golowich V, S, and Smola A,J, "Support vector method for function approximation, regression estimation and signal processing," in Advances in Neural Information Processing Systems, 1997, pp. 281– 28.

183. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 2017

184. Velásquez R,M, Lara J,V. Forecast and evaluation of COVID-19 spreading in USA with reduced-space Gaussian process regression. Chaos Solitons Fractals. (2020) 136:109924.

185. Vinceti M, Filippini T, Rothman KJ, Ferrari F, Goffi A, Maffeis G, Orsini N. Lockdown timing and efficacy in controlling COVID-19 using mobile phone tracking. EClinicalMedicine. 2020 Aug;25:100457. doi: 10.1016/j.eclinm.2020.100457. Epub 2020 Jul 13. PMID: 32838234; PMCID: PMC7355328.

186. W. H. Organization, Preventing chronic diseases: a vital investment. World Health Organization, 2015.

187. Walker, P. G. T., Whittaker, C., Watson, O. J., Baguelin, M., Ainslie, K. E. C., Bhatia, S., ... Ghani, A. C. (2021). The impact of COVID-19 and strategies for mitigation and suppression in low-and middle-income countries. Science, 369(6511), 413-422.

188. Wang, L., Wong, A. (2020). COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest Radiography Images. arXiv:2003.09871.

189. Wang, X., Dong, Y., Thompson, W.D. et al. Short-term local predictions of COVID-19 in the United Kingdom using dynamic supervised machine learning algorithms. Commun Med 2, 119 (2022). https://doi-org.salford.idm.oclc.org/10.1038/s43856-022-00184-7

190. Wang, Z., Schaul, T., Hessel, M., Hasselt, H., Lanctot, M., and Freitas, N. Dueling network architectures for deep reinforcement learning. International Conference on Machine Learning (ICML), 2016.

191. Watkins CJ, Dayan P (1992) Q-learning. Mach. Learn. 8(3–4):279–292.

192. Welleck, S., Mao, J., Cho, K., and Zhang, Z. (2017). Saliency-based sequential image attention with multiset prediction. In the Annual Conference on Neural Information Processing Systems (NIPS).

193. WHO, “Anticipating emerging infectious disease epidemics”, 2016.

194. Willette, A.A., Willette, S.A., Wang, Q. et al. Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study. Sci Rep 12, 7736 (2022). https://doi-org.salford.idm.oclc.org/10.1038/s41598-022-07307-z

195. Williamson, E. J., Walker, A. J., Bhaskaran, K., Bacon, S., Bates, C., Morton, C. E., ... Goldacre, B. (2020). OpenSAFELY: factors associated with COVID-19 death in 17 million patients. Nature, 584(7821), 430-436.

196. WIRTH, R. HIPP, J. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000. Manchester, 29-39

197. World Health Organization. Novel Coronavirus – Situation .

198. Xu Y., Wang W. The prevalence and control of diabetes in Chinese adults, JAMA 310(9) (2013) 948-

199. Yang G, Zhang F, Gong C, Zhang S. Application of a Deep Deterministic Policy Gradient Algorithm for Energy-Aimed Timetable Rescheduling Problem. Energies. 2019; 12(18):3461. .

200. Yan S, He L, Yixuan W et al. Theory and application of audio-based assessment of cough. J Sensor 2018; 2018: 9845321.

201. Yang Z, Zeng Z, Wang K, Wong SS, Liang W, Zanin M, et al. Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. (2020)

202. Yu C, Liu J, Nemati S. Reinforcement learning in healthcare: a survey. arXiv Prepr arXiv190808796. 2019

203. Zambrano-Monserrate M. A., Ruano M. A., and Sanchez-Alcalde L., "Indirect effects of COVID-19 on the environment," Science of the Total Environment, vol. 728, p. 138813, 2020.

204. Zhao Y, Kosorok M,R, and D. Zeng, "Reinforcement learning design for cancer clinical trials," Statistics in Medicine, vol. 28, no. 26, pp. 3294–3315, 2009.

205. Liu Y, Logan B, Liu N, Xu Z, Tang J, Wang Y. Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data. Healthc Inform. 2017 Aug;2017:380-385. doi: 10.1109/ICHI.2017.45. PMID: 29556119; PMCID: PMC5856473.

206. Zhou C., Su F., Pei T., Zhang A., Du Y., Luo B., et al. "COVID-19: Challenges to GIS with big data," Geography and Sustainability, vol. 1, no. 1, pp. 77–87, 2020.

207. Zhou M, Kan M-Y (2021) The varying impacts of COVID-19 and its related measures in the UK: A year in review. PLoS ONE 16(9): e0257286. .