Mini Project Report on

Exploring the Factors behind COVID-19 Surge: Predictive Modelling and Analysis

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE & ENGINEERING

Submitted by:

Student Name: Abhay Rautela University Roll No.: 2018059

Under the Mentorship of
Mr.Aniruddha Prabhu
Assistant Professor

Department of Computer Science and Engineering Graphic Era (Hill University)

Dehradun, Uttarakhand January 2023

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the project report entitled "Exploring the factors behind covid 19 Surge: Predictive modelling and Analysis" in partial fulfilment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineering of the Graphic Era (hill University), Dehradun shall be carried out by myself under the mentorship of Mr. Aniruddha Prabhu ,Assistant Professor Department of Computer Science and Engineering, Graphic Era (Hill University), Dehradun.

Name: Abhay Rautela University Roll no.: 2018059

Table of Contents

Chapter 1

Introduction

The epidemic of the new coronavirus SARS-CoV-2 has resulted in a global health catastrophe, impacting millions of people worldwide and stretching healthcare systems to their limits. COVID-19 is mostly spread by inhaling droplets, which are made when a sick person sneezes, breathes, coughs, or speakes. Close contact with sick persons, particularly in crowded indoor settings, significantly increases the chance of transmission. The high transmissibility of COVID-19 has led to community spread, where the virus circulates within the population without a clear source of infection. In the wake of the initial outbreak of COVID-19 in late 2019, the virus has spread swiftly over the world, causing severe illness, a high mortality rate, and unprecedented disruptions to societal and economic activities worldwide.

India is one among the countries most afflicted by the epidemic, with Delhi NCR being one of the most affected regions. The region's densely populated areas, coupled with the high transmissibility of the virus, have contributed to community transmission, and increased the risk of infection. Ensuring sufficient healthcare system capacity is essential for managing and providing adequate care to COVID-19 patients. Governments, healthcare organizations, and scientists have been striving to gain a better insight into the virus' transmission dynamics and identify key factors driving the surge in cases. Accurate forecasting and analysis of these factors can provide valuable insights for public health interventions, resource allocation, and policymaking.

In this research, we leverage predictive modelling techniques to understand the relationships between various parameters, the rise in the number of cases, and the spreading patterns of COVID-19. We collect a dataset from various government sources, including information on temperature, humidity, air quality index, demographic characteristics, medical history, surgical history, and lifestyle habits. In this research, we employ a dataset covering the period from January 2020 to December 2022, encompassing three years of data. Exploratory data analysis techniques are applied to gain insights into the distribution and patterns within the dataset, facilitating the identification of potential relationships between variables. We employ various modelling techniques to analyze and predict COVID-19 cases, including regression models, decision trees, and time series analysis. These models are applied to identify significant relationships between the variables selected and the frequency of COVID-19

cases. Among these models, the LightGBM model consistently demonstrates superior performance, providing accurate predictions and valuable insights.

The LightGBM model is a powerful machine-learning algorithm known for its efficiency and effectiveness in handling large datasets. Feature selection techniques are applied to identify the most influential variables driving the surge in COVID-19 cases, aiding in understanding the underlying factors and their impact. We conduct multiple sensitivity analysis tests to gauge the impact of the individual parameters on the number of COVID-19 cases in the DelhiNCR region. The findings of our investigation show that past medical history, followed by climate variables, has the most significant influence on the number of cases. Demographic information, including age groups, gender, and occupation, also affects the number of cases. The conclusions of this research may assist public health authorities, lawmakers, and medical specialists build realistic methods to restrict the virus's spread and lessen its adverse influence on the population.

.

Chapter 2

Literature Survey

COVID-19's proliferation has resulted in an unparalleled worldwide health disaster, necessitating the discovery of critical variables driving its transmission. This literature review focuses on previously done research that analyses how various environmental, demographic, lifestyle, and medical variables might influence the incidence rate of COVID19 patients across diverse geographical areas.

1.Liu et al. examined the atmospheric properties of SARSCoV-2 and reported low concentration of viral RNA in hospital wards but larger amounts in patient toilet regions. They suggested the potential for aerosol transmission and emphasized the importance of ventilation and sanitization

2.Morawska and Cao emphasised the airborne dispersion of COVID-19 in confined situations. They emphasized the need to acknowledge this route of infection and recommended ventilation as a control measure

3 .Kumari et al. proposed a mathematical model, SEIAQRDT, which accurately predicted COVID-19 cases by considering asymptomatic individuals. Their model improved control strategies by classifying infected individuals based on symptoms

4. A compartmental model was created by Khajanchi et al. to study Covid-19 transmission in India. They estimated model parameters, performed short- term predictions, and investigated optimal control strategies to reduce transmission

5. Juneja et al. studied the effect of external factors on COVID-19 spread in Delhi, Kerala, and Tamil Nadu. Their analysis considered temperature and humidity as influencing factors for policymakers .

6.Several studies highlight the relationship between weather patterns and the transmission of the virus. The study by Gupta et al. (2020) examines the influence of climate on COVID-19 cases in India and finds that transmission is subject to a specific climate pattern.

7. Similar results are revealed by Bashir et al. (2020), who analyse the association between meteorological conditions and COVID-19 in the US capital and conclude that air quality, minimum temperatures, and average temperatures are all highly connected to the pandemic

8. The COVID-19 pandemic has also had significant environmental impacts. Hammad et al.

(2023) discuss the positive effects of the pandemic, such as reduced air, water, and noise pollution, but also highlight negative impacts, such as increased release of microcontaminants and biomedical waste generation

9. Additionally, the pandemic has affected various aspects of society, including the social dimension. Goswami and Neog (2023) explore the social consequences of COVID-19, including psychological effects, social crises, and emerging trends in the social economy. Air quality has been a significant concern during the pandemic.

10.Ali and Islam (2020) study the effects of air pollution on COVID-19 illnesses and death, focusing on the role of PM2.5 and nitrogen dioxide in higher infection and mortality rates

11. Another study by Kaloni et al. (2022) analyzes the implications of the COVID-19 lockdown on the air's cleanliness in New Delhi, India, and finds a large drop in pollution amounts during the restricted period

12. Finally, the article by Kumar, Khandelwal, and Gadhwal (2022) discusses the nexus between global environmental problems, climate change, human health, and COVID-19. It emphasizes the critical need for mitigation solutions to address the difficulties presented by climate change to human health, food, and water security, and contamination of the environment.

13. The COVID-19 pandemic's nonpharmaceutical interventions (NPIs) are thoroughly reviewed by Perra (2021). The study covers various aspects, such as travel restrictions, social isolation, and lockdowns, showing the effect of these measures on disease spread. In addition, the study analyses a wide range of literature to understand the effectiveness of NPIs and provides insights into future challenges and opportunities.

14.Abirami and Kumar (2022) compare machine learning models for disease detection and prediction. They discovered that supervised learning approaches, particularly classification models, are more accurate than unsupervised learning methods in predicting COVID-19

15. Shinde et al. (2020) provides a review of forecasts models for COVID-19. It discusses the role of various parameters in pandemic forecasting and addresses technical and generic challenges associated with these approaches.

16. Kumar and Susan (2020) focus on time series projection models, especially ARIMA and Prophet, to predict the spread of COVID-19. They assess the effectiveness of the models using different error metrics and indicate that the ARIMA model is more efficient for predicting the rate of COVID-19 frequency.

17. Mandayam et al. (2020) use regression models, specifically Linear Regression and

Support Vector Regression, to forecast future COVID-19 cases.

18.Machine learning techniques are used by Malki et al. (2021) to forecast the propagation of COVID-19. Their proposed model accurately predicts proven cases and suggests that COVID-19 infections will greatly drop in the first week of September 2021.
19.Rajan Gupta, Pandey, and Pal (2021) perform a comparison study of different demographic and machine/deep learning models for COVID-19 forecasts. They evaluate these models using different evaluation criteria across ten regions and provide insights into the suitability of different models based on the flattening of cases and growth curve patterns
20. Univariate and multivariate time series models were discussed for COVID-19 forecasting by Sreehari et al. in 2021.

Chapter 3

Methodology

3.1 Data Collection and Preprocessing.

The data was collected from recognized international organizations like the World Health Organisation (WHO) and official sources within the Indian government, including statistics and health departments. The collected data included parameters such as temperature, humidity, air quality index, senior citizen population, gender diversity, labor/hazardous work percentage, prior medical and surgical history, cardiovascular and gastrointestinal diseases, smoking and alcohol habits, and the athletic individual's percentage. This dataset covers a significant period of the COVID-19 pandemic, ranging from January 2020 to December 2022.

Before modeling, any dataset abnormalities must be identified and corrected to achieve the best performance. In this research, the dataset underwent preprocessing to handle outliers, missing data, and other irregularities. The dataset's null values were first replaced with the median values for the associated feature as part of the preparation procedure. Additionally, data augmentation techniques were employed to ensure a robust testing dataset. Data augmentation also helped to ensure accurate and reliable predictions without overfitting the dataset. After the data preprocessing and augmentation steps, the dataset was ready for our analysis. [Fig 2] illustrates a segment of the dataset, offering a glimpse into its overall structure and characteristics.

3.2 Exploratory Data Analysis

Understanding the traits and patterns in the dataset is crucially dependent on the exploratory data analysis (EDA). In this study, several key steps were implemented to extract insightful information from the data gathered. The dataset was transformed into a time series format to analyze the temporal aspects of the data. Graphs and visualizations were plotted for each variable [Fig 3], allowing for a better understanding of the trends, patterns, and seasonality present in the data. Standardization techniques, such as scaling the data using methods like StandardScaler(), were applied to bring the variables to a consistent scale. This process ensured that each variable's contribution to the analysis was unbiased and comparable. To

uncover underlying components and trends, seasonal decomposition [Fig 4] was performed, which involved decomposing the time series data into its seasonal, trend, and residual components. This decomposition aided in identifying recurring patterns and understanding the long-term and short-term variations within the data.

3.3 Splitting The Dataset For Training and Testing

To assess the prediction models' performance and generalization, the dataset was separated into training and testing sets based on key considerations. The testing data was carefully selected to cover a duration of 120 days, ensuring an adequate representation of different time periods and capturing various trends and patterns within the dataset. This duration allows for a comprehensive evaluation of the model's aptitude to forecast COVID-19 cases over a meaningful time span. The training set, comprising the remaining data, was utilized for training the models on historical patterns and relationships, enabling them to learn and capture the underlying dynamics. By conducting rigorous testing on the independent testing set, the model's performance can be accurately assessed, including its ability to generalize to unseen data and handle future scenarios. The splitting of the dataset for training and testing strikes a balance between capturing temporal dependencies and providing a robust evaluation framework, ultimately enhancing the reliability and applicability of the research findings.

3.4 Model Selection

The selection of an appropriate model is a critical step to accurately predict COVID-19 cases and analyze the factors driving the surge. Different models were assessed for their performance and applicability in light of the features of the dataset and the study's goals. Time series analysis, decision trees, regression models, and other machine learning techniques are among the models investigated. However, after thorough consideration, the Light Gradient Boosted Machine (LightGBM) model was determined to be the best option because of its higher performance in handling big datasets, effective computing, and excellent predictive skills. The LightGBM model handles the dataset's temporal nature and accommodates the variables' multidimensionality, making it ideal for capturing complex relationships and interactions. By utilizing the LightGBM model, the research aims to leverage its advanced boosting techniques, such as gradient boosting and decision tree-based learning, to achieve accurate predictions and meaningful insights into the factors behind the COVID-19 surge.

E. Predictive Analysis

Utilizing the selected Light Gradient Boosted Machine
(LightGBM) model, the goal is to correctly predict COVID19 cases and gain greater insights into the reasons causing the surge. The trained model, fine-tuned on the training dataset, will be applied to the independent testing dataset to assess its performance in predicting future COVID-19 cases. The model's predictions will shed light on the multifaceted dynamics of COVID-19 transmission, aiding in the development of targeted treatments, public health interventions, and resource allocation strategies. The findings of the predictive analysis will contribute to evidence-based decision- making, enabling effective control and prevention of COVID-19.

The project has undergone the following process:

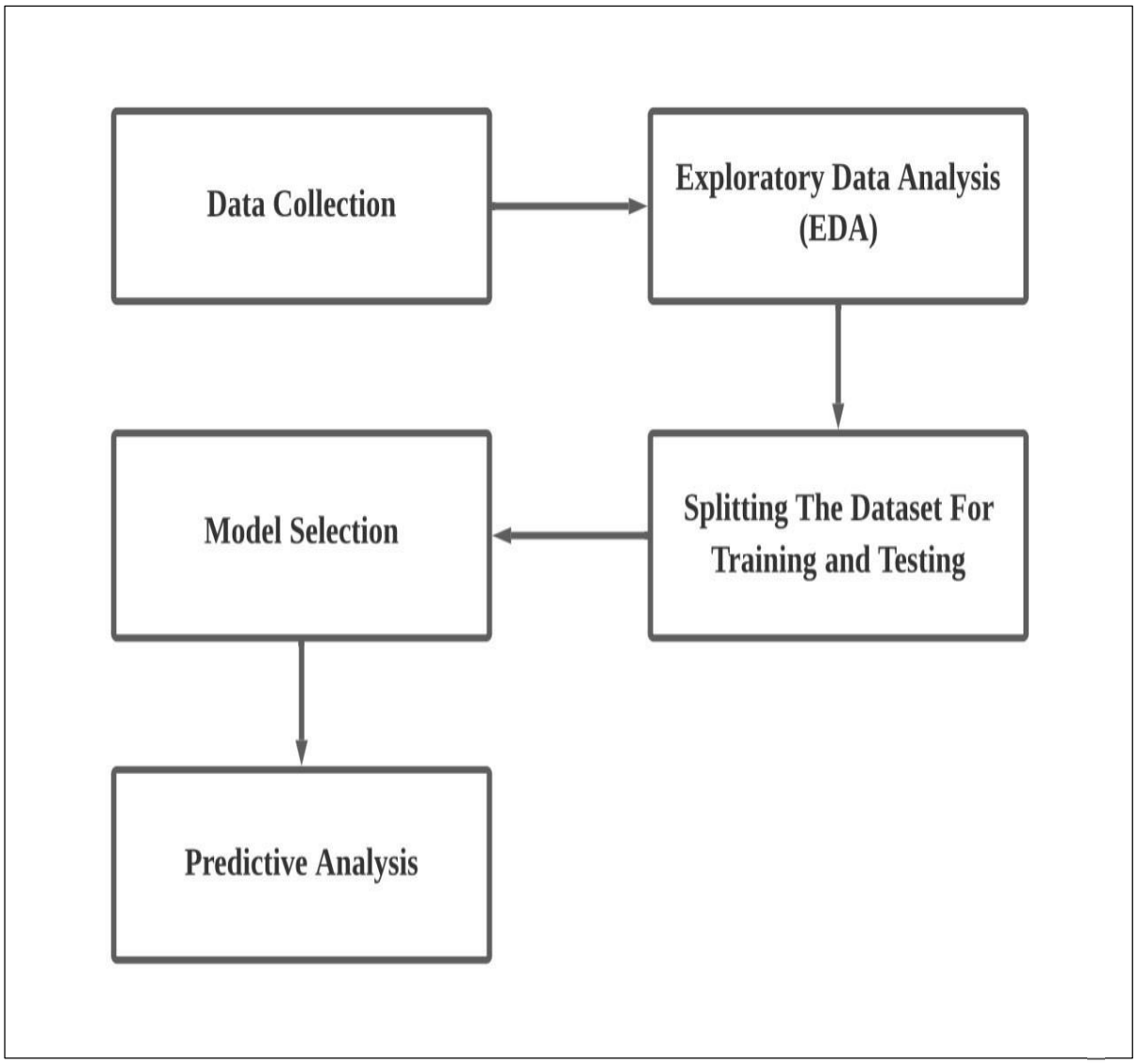

Fig 3.1 System Architecture

4 Result And Discussion

The prediction modeling and analysis performed in this study provide useful insights into the factors adding to the surge in COVID-19 cases in the Delhi-NCR region. The results show important connections and trends that help in understanding the dynamics of COVID-19 spread and guide tailored medicines, public health measures, and resource allocation techniques. It is found that temperature has a substantial effect on COVID- 19 cases [Fig 5]. The research also identified previous medical history, climate variables, and labor/hazardous work percentage as the most determining factors. The analysis of demographic data of COVID-19 patients reveals that elderly individuals are more susceptible to the virus. Additionally, smoking and alcohol habits, as well as a lack of physical activity, have a direct impact on the virus' transmission. The LightGBM model is employed to predict the number of cases and evaluate the predictive power of the dataset. In this paper, the model's performance is measured in the terms of mean absolute error(MAE) [Fig 6], mean squared error(MSE), and R-Squared(R2). The model is applied to the testing dataset after being trained on the training dataset, and it performs excellently with an accuracy of 95.4%.

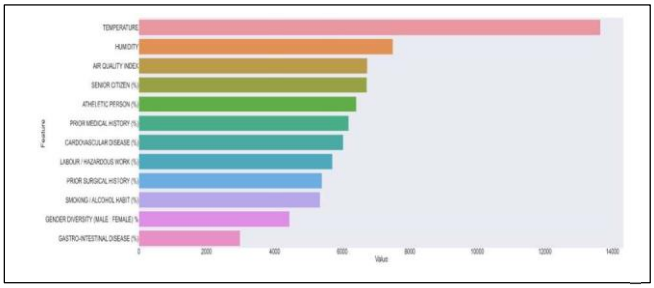

Fig. 5. LightGBM Model Features Importance

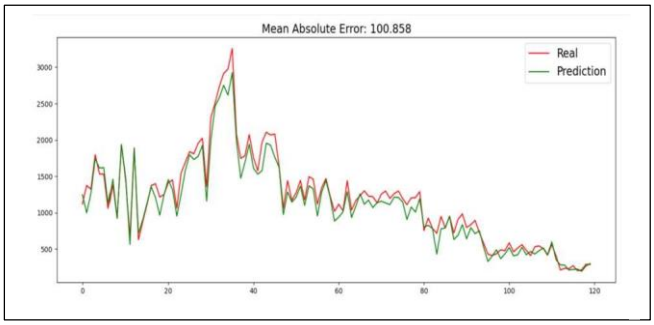

Fig. 6. Real v/s Prediction

Chapter 5

Conclusion and Future Work

5.1 FUTURE WORKS

The analysis of the dataset revealed several important findings regarding the factors influencing the surge in COVID-19 cases. These findings can serve as a foundation for further research and analysis.

- Integration of additional variable: While this study considered a wide range of variables, there may be other elements that add to the spread of COVID-19. Future studies could explore the inclusion of variables such as vaccination rates, mobility data, and socioeconomic indicators to gain a more comprehensive understanding of the pandemic's dynamics.
- Evaluation of intervention strategies: The prediction models can be expanded to assess the effectiveness of different intervention strategies, such as lockdown measures, mask requirements, and vaccine programmes. By incorporating data on these interventions, policymakers can make informed decisions regarding the implementation and timing of various measures.
- Regional analysis: This research focused on the DelhiNCR region; however, different regions may exhibit distinct patterns and factors contributing to COVID-19 transmission. Future work could involve expanding the analysis to other regions within India or even globally, allowing for a more nuanced understanding of the pandemic's impact.
- Long-tern predictions: The predictive analysis in this research covered a 120-day testing period. Extending the prediction horizon to longer timeframes can provide insights into the potential trajectory of the pandemic and facilitate proactive planning and resource allocation.

5.2 CONCLUSION

This research paper successfully explored the factors behind the surge in COVID-19 cases through predictive modeling and analysis. By utilizing a diverse set of variables, the study shed light on the multifaceted dynamics of COVID19 transmission. The Light Gradient Boosted Machine (LightGBM) model proved to be an effective tool for accurately predicting COVID-19 cases and capturing complex relationships within the dataset. The findings contribute to evidence-based decision-making by informing targeted treatments, public health interventions, and resource allocation strategies. The future works outlined above can further enhance our understanding of the pandemic and facilitate the development of proactive and effective measures to control and prevent COVID-19. By continuing to analyze and explore the factors influencing transmission, we can better protect public health and mitigate the impact of the pandemic.

References

1.Liu, Y., Ning, Z., Chen, Y. et al. Aerodynamic analysis of SARS-CoV2 in two Wuhan hospitals. Nature 582, 557–560 (2020). https://doi.org/10.1038/s41586-020-2271-3

2.Lidia Morawska, Junji Cao, Airborne transmission of SARS-CoV-2: The world should face the reality, Environment International, Volume 139, 2020, 105730, ISSN 0160-4120, https://doi.org/10.1016/j.envint.2020.105730

3.Kumari, P., Singh, H.P. & Singh, S. SEIAQRDT model for the spread of novel coronavirus (COVID-19): A case study in India. Appl Intell 51, 2818–2837 (2021). https://doi.org/10.1007/s10489-020-01929-4

4.Khajanchi, S., Sarkar, K. & Banerjee, S. Modeling the dynamics of COVID-19 pandemic with implementation of intervention strategies. Eur. Phys. J. Plus 137, 129 (2022). https://doi.org/10.1140/epjp/s13360-022-02347-w

5.Juneja, N., Sunidhi, Kaur, G., Kaur, S. (2022). Impact of Environmental Factors on COVID19 Transmission Dynamics in Capital New Delhi Along with Tamil Nadu and Kerala States of India. In: Dua, M., Jain, A.K., Yadav, A., Kumar, N., Siarry, P. (eds) Proceedings of the International Conference on Paradigms of Communication, Computing and Data Sciences. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-16-5747-4_36

6.Gupta, A., Pradhan, B. & Maulud, K.N.A. Estimating the Impact of Daily Weather on the Temporal Pattern of COVID-19 Outbreak in India. Earth Syst Environment 523–(2020). https://doi.org/10.1007/s41748-020-00179-1

7.Bashir MF, Ma B, Bilal, Komal B, Bashir MA, Tan D, Bashir M. Correlation between climate indicators and COVID-19 pandemic in New York, USA, Science of The Total Environment, Volume 728, 2020, 138835, ISSN 0048-9697, https://doi.org/10.1016/j.scitotenv.2020.138835

8.Hammad, H.M., Nauman, H.M.F., Abbas, F. et al. Impacts of COVID19 pandemic on environment, society, and food security. Environ Sci Pollut Res (2023). https://doi.org/10.1007/s11356-023-25714-1

9.Goswami, R., Neog, N. (2023). COVID-19: An Insight into Social Dimension. In: The Handbook of Environmental Chemistry. Springer, Berlin, Heidelberg. https://doi.org/10.1007/698_2023_996

10.Parvin R. A Statistical Investigation into the COVID-19 Outbreak Spread. Environmental Health Insights. 2023;17. doi:10.1177/11786302221147455

11.Venkatesh S Amin, Dr. Abhishek N, Dr. Abhinandan Kulal et al. Covid-19 and Dynamic Changes in Learning Environment: A Perceptional Study, 09 January 2023, PREPRINT (Version 1) available at Research Square. https://doi.org/10.21203/rs.3.rs-2454785/v

12.Ali N, Islam F. The Effects of Air Pollution on COVID-19 Infection and Mortality-A Review on Recent Evidence. Front Public Health. 2020 Nov 26;8:580057. https://doi.org/10.3389/fpubh.2020.580057

13.Dewansh Kaloni, Yee Hui Lee, Soumyabrata Dev, Air quality in the New Delhi metropolis under COVID-19 lockdown, Systems and Soft Computing, Volume 4, 2022, 200035, ISSN 2772-9419, https://doi.org/10.1016/j.sasc.2022.200035

14.Kumar, A., Khandelwal, S.G., Gadhwal, N. (2022). Global Environmental Problems: A Nexus Between Climate, Human Health and COVID 19 and Evolving Mitigation Strategies. In: Arora, S., Kumar, A., Ogita, S., Yau, Y.Y. (eds) Innovations in Environmental Biotechnology. Springer, Singapore. https://doi.org/10.1007/978-981-16-4445-0_3

15.Perra N. Non-pharmaceutical interventions during the COVID-19 pandemic: A review. Phys Rep. 2021 May 23;913:1-52. https://doi.org/10.1016/j.physrep.2021.02.001