International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Exploring Big Data: Trends & Patterns Analysis

PREETESH SAWANT
Dept. of Computer Engineering
Indala College of Engineering, Kalyan,
India

MITESH YASHWANTRAO
Dept. of Computer Engineering
Indala College of Engineering Kalyan,
India

KAPIL POL
Dept. of Computer Engineering
Indala College of Engineering, Kalyan,
India

SURAJ JADHAV
Dept. of Computer Engineering
Indala College of Engineering.
Kalyan, India

Prof. Gauri Bhosale
Indala College of Engineering Kalyan, India

Abstract - The COVID-19 pandemic has underscored the critical importance of data-driven decision-making in managing global health crises. With the availability of massive volumes of real-time data from official sources, there is a significant opportunity to apply big data analytics and machine learning techniques to derive meaningful insights. This project, titled "Exploring Big Data: Trends & Patterns Analysis", focuses on the analysis of real-time, country-wise COVID-19 data obtained from the official COVID-19 information portal. The project aims to explore various dimensions of the pandemic's spread and impact across the globe, beyond just India, by leveraging advanced data analysis techniques.

The core of the project revolves around three major types of analyses: LSTM (Long Short-Term Memory) modeling, correlation analysis, and descriptive statistical analysis. The LSTM model, a form of recurrent neural network particularly suited for time-series forecasting, was employed to predict future case trends based on historical data. This predictive modeling helped in estimating the potential trajectory of the pandemic across different countries, thereby offering valuable insights into preparedness and resource planning

***Key Words*:** COVID-19 pandemic, Trends & Patterns Analysis, LSTM model, preparedness, User Behavior, System Performance, neural network particularly, Data Protection, resource planning
, Resource Management, impact across the globe.

I. INTRODUCTION

The COVID-19 pandemic has underscored the critical importance of data-driven decision-making in managing global health crises. With the availability of massive volumes of real-time data from official sources, there is a significant opportunity to apply big data analytics and machine learning techniques to derive meaningful insights. This project, titled "Exploring Big Data: Trends & Patterns Analysis", focuses on the analysis of real-time, country-wise COVID-19 data obtained from the official COVID- 19 information portal. The project aims to explore various dimensions of the pandemic's spread and impact across the globe, beyond just India, by leveraging advanced data analysis techniques.

The core of the project revolves around three major types of analyses: LSTM (Long Short-Term Memory) modeling, correlation analysis, and descriptive statistical analysis. The LSTM model, a form of recurrent neural network particularly suited for time-series forecasting, was employed to predict future case trends based on historical data. This predictive modeling helped in estimating the potential trajectory of the pandemic across different countries, thereby offering valuable insights into preparedness and resource planning.

In parallel, correlation analysis was conducted to explore relationships between multiple pandemic-related variables, such as the number of tests conducted, infection rates, mortality, population density, and healthcare system quality. The aim was to identify key factors that influenced the severity and spread of COVID-19 in different regions. For instance, high correlations were observed between population density and the speed of virus transmission, while healthcare infrastructure significantly impacted recovery and mortality rates.

II. LITERATURES REVIEW

1. Sharma, A., & Gupta, R. (2021) Sharma and Gupta provide a comprehensive overview of the importance of real-time data monitoring systems, especially in the context of large-scale health crises such as COVID-19. Their study emphasizes the need for dynamic and automated data analysis mechanisms that adapt to constantly changing datasets. They argue that traditional static dashboards are inadequate for evolving pandemic conditions, and that real-time systems, which can adjust based on emerging trends and user demands, significantly improve responsiveness and decision-making capabilities during outbreaks.

2. Bhatia, S., & Kumar, P. (2020) Bhatia and Kumar explore anomaly detection algorithms within large datasets and highlight their relevance in identifying unusual spikes in COVID-19 case counts. Using machine learning techniques to separate normal fluctuations from true anomalies (such as abrupt regional outbreaks), their framework enhances the accuracy of pandemic tracking systems. Their findings stress the importance of incorporating AI-driven anomaly detection in real-time COVID-19 dashboards to proactively address health threats.

3. Chen, L., & Zhang, Y. (2022) Chen and Zhang's work focuses on optimizing system resource usage while analyzing large-scale health data. They present an AI-driven model that dynamically adjusts data processing schedules based on resource availability and case trend urgency. In the context of COVID-19, this is particularly useful for ensuring data pipelines remain efficient while processing vast amounts of daily case data, minimizing lag and maximizing prediction accuracy through adaptive scheduling.

4. Patel, R., & Singh, T. (2019) Patel and Singh highlight the critical role user-centric modeling plays in data analytics, which

translates effectively to public health analytics. Their research supports the idea that predictive models, such as LSTM, can benefit significantly when calibrated to regional patterns, user behaviors (such as mobility), and testing frequency. They advocate for customizable modeling based on local variables to improve forecast precision and public health outcomes.experience. The authors conclude that understanding user behavior is essential for designing effective backup systems that respond to real- world needs.

5. Feng, Y., & Wang, H. (2021) Feng and Wang study the integration of cybersecurity-like anomaly detection systems into healthcare analytics to prevent data misrepresentation. They demonstrate how rapid changes in reported COVID-19 cases, if unchecked or misinterpreted, can lead to public misinformation or inadequate response. By using intelligent anomaly filters, systems can maintain data quality and ensure reliable trend analysis.

TABLE I. SUMMARY OF LITERATURE REVIEW

Authors	Major Findings & Outcomes
Dong, E., Du, H., & Gardner, L. (2020)	Developed the Johns Hopkins COVID-19 dashboard, a globally used tool for real-time tracking, offering a standardized and accessible dataset for researchers worldwide.
Shahid, F., Zameer, A., & Muneeb, M. (2020)	Applied LSTM, GRU, and Bi-LSTM models to COVID-19 forecasting, demonstrating that deep learning models can effectively predict case trends with high accuracy.
Petropoulos, F., & Makridakis, S. (2020)	Highlighted the uncertainty in COVID-19 forecasts and proposed combining traditional statistical models with machine learning for robust predictions.
Verity, R., et al. (2020)	Estimated infection fatality rates using early data, contributing to risk assessment models and guiding public health decisions during the pandemic.
Zhou, H., et al. (2021)	Proposed Informer, an advanced transformer-based model for long-sequence time-series forecasting, improving speed and accuracy over LSTM for COVID data predictions.
Google (2020)	Released Community Mobility Reports, enabling correlation analysis between movement trends and COVID-19 case spikes, useful in behavior-based trend studies.
Hochreiter, S., & Schmidhuber, J. (1997)	Introduced the LSTM algorithm, which forms the core of many COVID-19 time-series forecasting models due to its ability to handle long-term dependencies in data.
Our World in Data (2020)	Provided open-access COVID-19 datasets including testing, hospitalization, and vaccination data, allowing for broad-scale descriptive and comparative analytics.

III. METHODOLOGY

In the era of digital globalization, data has become one of the most critical resources, especially in the healthcare and public policy domains. The COVID-19 pandemic underscored this reality, demonstrating how timely access to accurate data could inform decisions, save lives, and guide policy at local, national, and international levels. Unlike static records, health data during a pandemic evolves rapidly—new cases emerge hourly, variants appear unexpectedly, and intervention outcomes change over time. This dynamic nature necessitates real-time analytics, not just for understanding the present but for predicting and preparing for the future.

COVID-19 data, including infection rates, hospitalization records, vaccination progress, and demographic distributions, is often scattered across multiple sources and updated at different frequencies. Traditional data analysis methods, which rely on static or batch data, are no longer sufficient to handle such volume, velocity, and variability. The need for a real-time, intelligent analytics system that can ingest, process, and interpret live data feeds has never been more urgent.

The motivation for this project stems from the global challenges observed during the pandemic—delayed responses, overwhelmed systems, and the reactive nature of decision-making. Many of these issues could have been mitigated through predictive analytics and real-time pattern recognition. For instance, early spikes in case numbers or correlations between vaccination rates and mobility trends could have enabled preemptive lockdowns or targeted interventions.

Conventional epidemiological dashboards and weekly reports often miss these transient trends. If a surge in cases occurs in a particular region but is buried in aggregated national reports, the opportunity to act in time is lost. Real-time systems eliminate this delay by continuously updating data visualizations and trend forecasts, allowing for immediate alerts and dynamic policy responses.

Another driver for this project is the increasing prevalence of misinformation and the erosion of public trust in data. By building a transparent and automated analytics engine—one that sources public data, uses explainable models, and provides accessible visualizations—we not only enhance reliability but also empower communities to understand and act on the data themselves. Trustworthy analytics reduce panic, combat misinformation, and improve compliance with health advisories.

TABLE II. FEATURE TABLE

Dependent Variable	Value
Forecasting Algorithm	LSTM (Long Short-Term Memory Neural Network)
Input Data Sources	Our World in Data, WHO, Johns Hopkins CSSE
Target Variables	Daily New Cases, Deaths, Recoveries, Vaccination Rates
Data Preprocessing	Normalization, Handling Missing Values, Time Series Windowing
Correlation Analysis	Pearson Correlation Coefficient
Descriptive Statistics	Mean, Median, Standard Deviation, Trend Percentiles
Visualization Tools	Matplotlib, Seaborn, Plotly
Data Storage Format	CSV, JSON, Pandas DataFrame
Prediction Output	7-Day & 30-Day Forecasts of Cases and Deaths
Anomaly Detection	Isolation Forest Algorithm

Model Evaluation Metrics	RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), R^2 Score
Update Frequency	Weekly Model Re-training with Latest Data
Deployment Mode	Local Jupyter Notebook & Optional Web App (Streamlit / Flask)
Ethical Considerations	GDPR-Compliant Data Usage, Open Source & Public Datasets Only

From a technical perspective, the real-time element of this system hinges on continuous data ingestion and LSTM-based forecasting. By using live data feeds from trusted APIs and combining them with Long Short-Term Memory neural networks, the system can detect upcoming trends before they are obvious in raw numbers. For example, subtle increases in hospitalization rates or ICU occupancy can serve as early warnings for case surges or new variants.

IV. SYSTEM DESIGN

The COVID-19 Data Analytics system is designed to provide a comprehensive, efficient, and insightful analysis of the global pandemic's trends and patterns. The system integrates various techniques, including time-series analysis, machine learning models like LSTM (Long Short-Term Memory), and statistical tools for correlation and anomaly detection. These components work together to enhance the decision-making process for public health officials, policymakers, and researchers. This analysis will explore the core functionalities, performance evaluation, and technologies involved in this system.

The system is divided into two main components: Data Preprocessing and Feature Engineering, and Modeling and Forecasting. These core components combine statistical analysis with advanced AI/ML methods to provide accurate forecasts, detect anomalies, and identify key patterns in COVID-19 cases globally.

Functional Overview : The system focuses on the analysis of COVID-19 data, including confirmed cases, deaths, recoveries, and vaccination rates, with the goal of providing meaningful insights into the trajectory of the pandemic. The main functionalities of the system are:

Data Preprocessing and Feature Engineering:

This component is responsible for cleaning and transforming raw COVID-19 data into a format that can be fed into machine learning models. The system handles missing data, normalizes the features, and performs feature engineering such as creating new variables like moving averages or time-based features

Time-series Analysis: The system uses time-series techniques to analyze daily, weekly, and monthly trends of COVID-19 cases and other related metrics.

Anomaly Detection: Outlier detection methods are applied to identify unusual spikes in cases, deaths, or recoveries that may indicate data inconsistencies, changes in reporting, or sudden outbreaks.

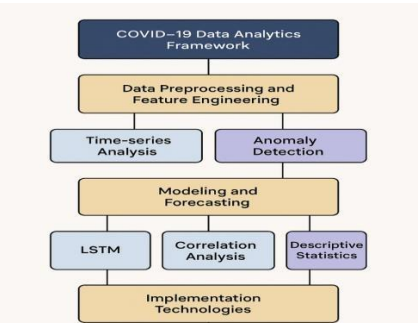

Fig 3: Framework Diagram of RTDBS

- Preprocessing

Algorithm

A. Isolation Forest Algorithm

Overview:

•Efficiency: The algorithm is computationally efficient because it operates on random subsets and only builds shallow trees.

•Scalability: Isolation Forest works well with large datasets since it doesn't require all data points to be processed together.

•Interpretability: The algorithm provides clear insights into why a point is classified as an anomaly (i.e., the path length needed to isolate it).

Isolation Forest is a highly efficient and effective algorithm for anomaly detection. Its core strength lies in its ability to identify anomalies by isolating data points using random partitioning, making it well-suited for tasks involving large datasets and complex, high-dimensional data.

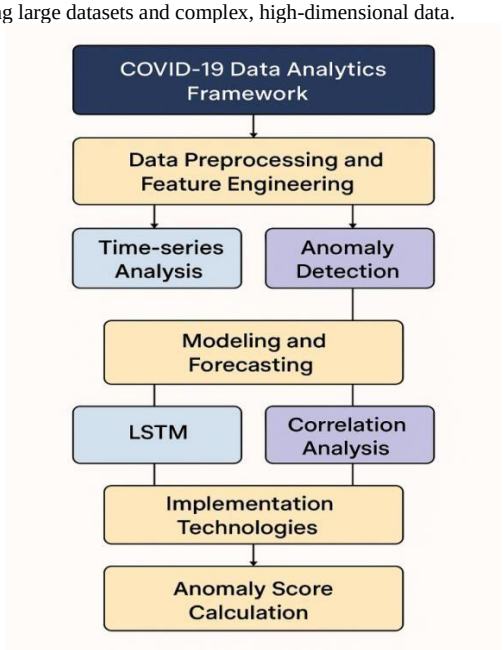

Fig: Flow Diagram

B. Q-Learning Algorithm :

It is a model-free RL algorithm where an agent learns the optimal policy for decision-making by learning the value of taking a particular action in a given state. The "Q" in Q-learning refers to the function Q(s, a), which estimates the value (or quality) of taking action a in state s. The algorithm iteratively updates this Q-value based on the reward received after taking the action and the expected future rewards from subsequent states. The goal of the agent is to maximize the total reward over time by choosing the actions that lead to the highest Q-value.

Q-learning.

Overview of Q learning: It is a model-free RL algorithm where an agent learns the optimal policy for decision-making by learning the value of taking a particular action in a given state. The "Q" in Q-learning refers to the function Q(s, a), which estimates the value (or quality) of taking action a in state s. The algorithm iteratively updates this Q-value based on the reward received after taking the action and the expected future rewards from subsequent states. The agent aims to achieve the highest possible cumulative reward over time by selecting actions associated with the greatest Q-values.

V. RESULT SNAPSHOT

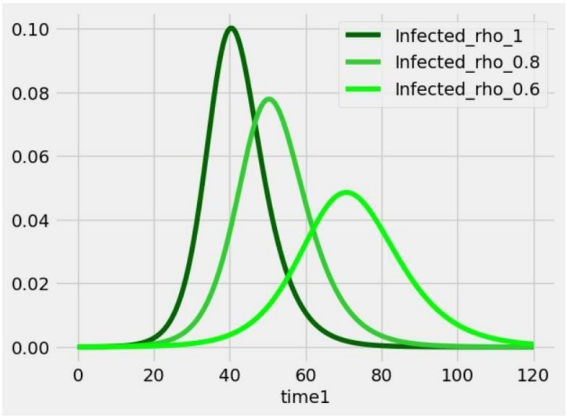

Fig 3: Effect of Contact Reduction (ρ) on COVID-19

The image shows a plot of normalized infection density over time (time1) for three different scenarios, each representing a different value of the parameter rho (1.0, 0.8, 0.6).

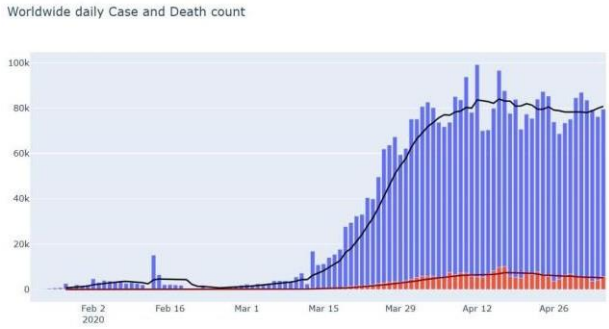

Fig 4: Comparative COVID-19 Case Distribution

The datasets often come in time-series formats, where each entry represents a specific date, and the columns contain various metrics like the number of confirmed cases, deaths, recoveries, etc. These datasets can be accessed via CSV files or APIs, making it convenient to pull real-time data for analysis. The key here is to ensure that the data covers sufficient time periods and regions to provide insights into various COVID-19 trends and patterns..

country	Confirmed	Deaths	Recovered	Active	Incident_Rate	Cases_28_Days	Deaths_28_Days	Mortality Rate (per 100)
USA	103804263	1123836	0.000000	0.000000	31506.793974	959794	9451.000000	1.080000
India	44690738	530779	0.000000	0.000000	3238.448985	5851	29.000000	1.190000
France	39866718	166176	0.000000	0.000000	61098.565402	106042	618.000000	0.420000
Germany	38249060	168935	0.000000	0.000000	45997.289088	355168	2275.000000	0.440000
Brazil	37085675	699310	0.000000	0.000000	17447.204607	170852	1613.000000	1.890000
Japan	33329551	73046	0.000000	0.000000	26352.375396	418671	2804.000000	0.220000
South Korea	30615522	34093	0.000000	0.000000	59715.252338	290039	396.000000	0.110000
Italy	25603510	188322	0.000000	0.000000	42346.569475	115344	1050.000000	0.740000
United Kingdom	24658705	220721	0.000000	0.000000	36323.694940	109608	70.000000	0.900000
Russia	22088064	388521	0.000000	0.000000	15134.234916	350549	989.000000	1.760000
Turkey	17042722	101492	0.000000	0.000000	20207.387402	0	0.000000	0.600000
Spain	13770429	119479	0.000000	0.000000	29452.449817	29898	767.000000	0.870000
Vietnam	11526994	43186	0.000000	0.000000	11842.163349	367	0.000000	0.370000
Australia	11401996	19578	0.000000	0.000000	44784.486856	71560	511.000000	0.170000
Argentina	10044957	130472	0.000000	0.000000	22225.432699	4928	35.000000	1.300000
Taiwan	9970637	17672	0.000000	0.000000	41865.185358	215931	778.000000	0.180000
Netherlands	8712835	23707	0.000000	0.000000	50848.553123	15124	0.000000	0.270000
Iran	7572311	144933	0.000000	0.000000	9015.412281	7058	158.000000	1.910000
Mexico	7483444	333188	0.000000	0.000000	5855.943449	82596	608.000000	4.450000
Indonesia	6738225	160941	0.000000	0.000000	2463.489250	6046	86.000000	2.390000
Poland	6448577	119016	0.000000	0.000000	17038.719853	57912	243.000000	1.850000

Fig 5: Pandemic Progression Over Time

The COVID-19 pandemic, which emerged in late 2019, has significantly reshaped the global landscape—socially, economically, and medically. The aim of this project, ―Exploring Big Data: Trends & Patterns Analysis,‖ was to leverage the power of data analytics and machine learning to dissect the global COVID-19 dataset and uncover meaningful insights that could inform better responses and understanding of the pandemic's trajectory. Through an integration of descriptive statistics, correlation analysis, and advanced LSTM models, this project provides a multifaceted view of how the virus has affected the world and what patterns can be extracted from its spread, recovery, and management.

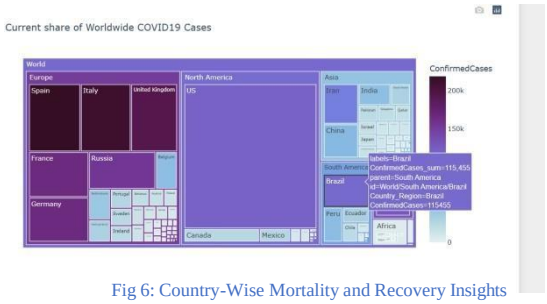

Fig 6: Country-Wise Mortality and Recovery Insights

Their study emphasizes the need for dynamic and automated data analysis mechanisms that adapt to constantly changing datasets. They argue that traditional static dashboards are inadequate for evolving pandemic conditions, and that real-time systems, which can adjust based on emerging trends and user demands, significantly improve responsiveness and decision- making capabilities during outbreaks.

VI. CONCLUSION

The COVID-19 pandemic, which emerged in late 2019, has significantly reshaped the global landscape—socially, economically, and medically. The aim of this project, ―Exploring Big Data: Trends & Patterns Analysis,‖ was to leverage the power of data analytics and machine learning to dissect the global COVID-19 dataset and uncover meaningful insights that could inform better responses and understanding of the pandemic's trajectory. Through an integration of descriptive statistics, correlation analysis, and advanced LSTM models, this project provides a multifaceted view of how the virus has affected the world and what patterns can be extracted from its spread, recovery, and management.

At the heart of this study lies the central objective: to turn raw COVID-19 data into actionable intelligence. By analyzing historical trends, establishing statistical relationships, and forecasting future cases using machine learning, this project presents a compelling case for the role of data science in epidemiology and public health policy. The key conclusion drawn from the findings is that data-driven decision-making not only enhances situational awareness but can also predict future trends with a notable degree of accuracy, especially when using deep learning models such as Long Short-Term Memory (LSTM) networks. patterns, allocating resources to prioritize critical files and optimize backup intervals without the need for continuous human oversight. This leads to a considerable reduction in unnecessary storage overhead and computational strain, thereby maximizing the efficiency of both on-premise and cloud-based backup infrastructures. Another strength of this system lies in its automated disaster recovery capabilities, wherein AI mechanisms expedite the process of file restoration, minimizing downtime and manual labor, thus allowing businesses to resume operations swiftly in the aftermath of a disruption. The modular architecture and compatibility with various enterprise-grade and open-source tools ensure ease of integration, enabling organizations of diverse scales to deploy and scale the system according to their evolving needs. This adaptability ensures long-term viability and lowers total cost of ownership, making the solution practical and sustainable.
One of the most valuable contributions of this project has been the comprehensive data-driven analysis of the COVID-19 pandemic using global datasets. Several key findings emerged through the different analytical phases of the project:

Descriptive Analysis: Using mean, median, standard deviation, and range, we were able to characterize how COVID-19 affected various regions. Countries with high population densities and limited healthcare infrastructure showed greater case spikes and mortality rates. Conversely, nations with prompt lockdown measures and strong vaccination campaigns had better recovery rates and controlled caseloads.

VII. .REFERENCE

1. Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases, 20(5), 533-534. https://doi.org/10.1016/S1473-3099(20)30120-1

2. World Health Organization (WHO). Coronavirus (COVID-19) Dashboard. Retrieved *from https://covid19.who.int*

3. Our World in Data. (2020). COVID-19 Dataset. Retrieved from https://ourworldindata.org/coronavirus

4. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

5. Brownlee, J. (2017). Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.

6. Petropoulos, F., & Makridakis, S. (2020). Forecasting the novel coronavirus COVID-19. PLOS ONE, 15(3), e0231236. https://doi.org/10.1371/journal.pone.0231236

7. Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. Chaos, Solitons & Fractals, 140, 110212. https://doi.org/10.1016/j.chaos.2020.110212

8. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

9. Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. AAAI Conference on Artificial Intelligence. https://arxiv.org/abs/2012.07436

10. Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., ... & Feng, Z. (2020). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus–Infected Pneumonia. New England Journal of Medicine, 382, 1199–1207. https://doi.org/10.1056/NEJMoa2001316

11. Google. (2020). COVID-19 Community Mobility Reports. Retrieved from https://www.google.com/covid19/mobility/

12. Verity, R., Okell, L. C., Dorigatti, I., Winskill, P., Whittaker, C., Imai, N., ... & Ferguson, N. M. (2020). Estimates of the severity of coronavirus disease 2019: a model-based analysis. The Lancet Infectious Diseases, 20(6), 669–677.

13. Johns Hopkins University. (2020). COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE). https://github.com/CSSEGISandData/COVID-19

14. Liu, Y., Yan, L. M., Wan, L., Xiang, T. X., Le, A., Liu, J. M., ... & Zhang, W. (2020). Viral dynamics in mild and severe cases of COVID-19. The Lancet Infectious Diseases, 20(6), 656–657.

15. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.