Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/69/20241513

Analysis and trend prediction of COVID-19 pandemic data based on big data visualization

Xinyuan Lu

School of Artificial Intelligence, Yanggo University, Fujian, China

2081659950@qq.com

Abstract. Since the outbreak of COVID-19 at the end of 2019, this global public health crisis has profoundly impacted the socio-economic conditions and daily life of countries worldwide. To effectively combat the pandemic, scientists and public health experts rely on vast amounts of data to track the progression of the disease, evaluate the effectiveness of control measures, and predict future trends. Big data technology plays a crucial role in the analysis of pandemic data and trend forecasting. This paper will explore the methods of analyzing COVID-19 pandemic data and the application of trend forecasting.

Keywords: Big Data Analysis, Pandemic, Data Visualization, Control Measures, Future Trends.

1. Introduction

With the rapid advancement of technology, the application of big data visualization techniques is increasingly pervasive across various domains, particularly in the healthcare sector. Against the backdrop of the global COVID-19 pandemic, the predictive analysis of virus spread has become significantly pertinent. This paper visualizes the data pertaining to the evolution of the COVID-19 pandemic, analyzes the characteristics and patterns of virus transmission, and utilizes techniques in data cleansing, preprocessing, and analysis to examine the disease's progression. Concurrently, it forecasts the trajectory of the epidemic, aiming to offer insights for epidemic prevention strategies and decisions concerning economic recovery.

2. Sources and Characteristics of COVID-19 Pandemic Data

2.1. Data Sources

The data on the COVID-19 pandemic mainly comes from the following sources:

The World Health Organization (WHO) provides global statistics on daily confirmed cases, deaths, and recoveries. Johns Hopkins University updates global pandemic data in real-time through its online dashboard and GitHub data repository. National health departments release detailed pandemic data for their countries, including case numbers, vaccination rates, and mortality rates. Additionally, social media and news reports provide real-time captures of public reactions, policy changes, and other dynamic information.

DOI: 10.54254/2755-2721/69/20241513

2.2. Specific Data Analysis

According to the World Health Organization (WHO), as of December 31, 2021, there were over 287 million confirmed cases globally and 5.42 million deaths [1]. As shown in Figure 1, we can identify peaks and troughs of the pandemic, as well as the speed of its spread through time-series analysis of daily confirmed and death cases. We can also compare case numbers across different regions to assess the severity of the pandemic in various countries. For example, Europe and the Americas have higher numbers of confirmed and death cases, while the Asia-Pacific region is relatively lower.

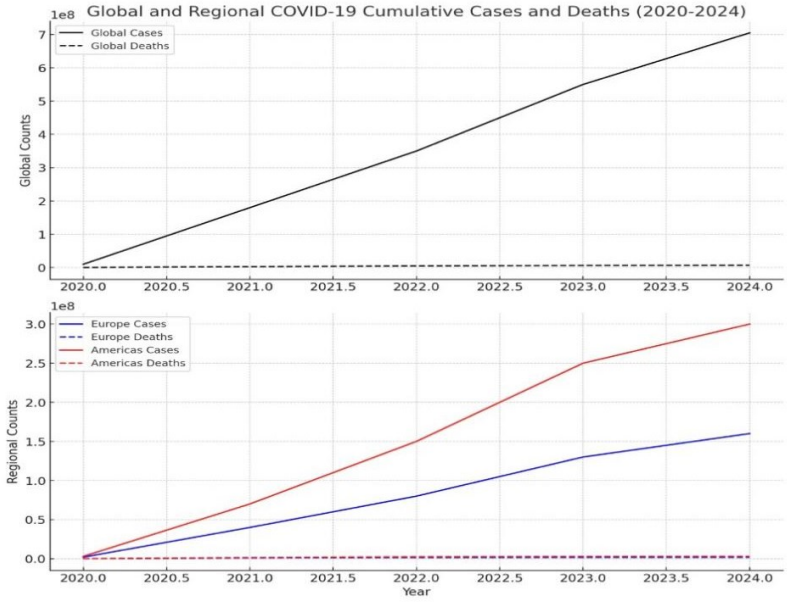

Figure 1. Global and Region COVID-19 Cumulative Cases and Deaths(2020-2024)

According to data from Johns Hopkins University, on December 31, 2021, there were 1.3 million new confirmed cases globally, including 425,000 in the United States and 206,000 in France. Over 9.3 billion vaccine doses have been administered globally [2]. As shown in Figure 2, based on the changes in daily new confirmed cases, we can predict the trend of the pandemic over the next few days and identify the outbreak points in specific countries or regions. Additionally, by analyzing the relationship between vaccination rates and confirmed case numbers, we can assess the effectiveness of vaccinations. Countries with high vaccination rates show a declining trend in new cases.

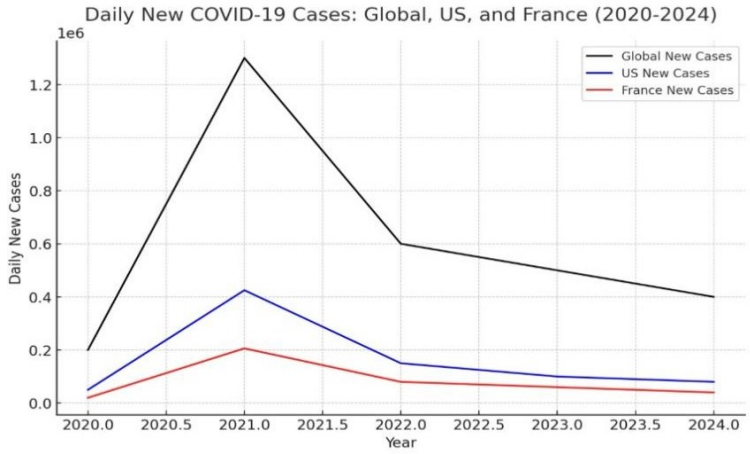

Figure 2. Daily New COVID-19 Cases: Global, US, and France(2020-2024)

DOI: 10.54254/2755-2721/69/20241513

According to health department data, as of December 31, 2021, China had a total of 102,083 confirmed cases, 4,636 deaths, and 2.88 billion vaccine doses administered. The United States had 54.38 million confirmed cases, 833,000 deaths, and 492 million vaccine doses administered [3]. As shown in Figure 3, by analyzing detailed pandemic data from different countries, we can understand the severity of the pandemic and the effectiveness of control measures in each country. For example, the relatively low numbers of confirmed cases and deaths in China indicate effective control measures. Additionally, by assessing the progress of vaccinations in various countries, we can further understand their impact on pandemic control. Countries with high vaccination rates have significantly better control of the pandemic.

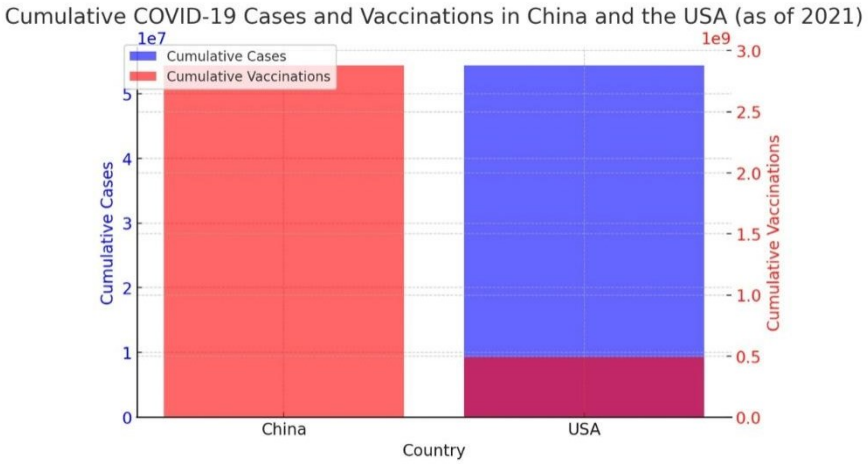

Figure 3. Cumulative COVID-19 Cases and Vaccinations in China and the USA (as of 2021)

Furthermore, according to social media and news reports, by analyzing posts on Twitter, Facebook, and other social media platforms, we can understand public emotional reactions to the pandemic, such as panic, anxiety, and hope. On December 31, 2021, discussions about the Omicron variant became a hot topic. News reports provide real-time updates on changes in control policies in various countries, such as lockdowns, travel restrictions, and vaccination policies. For example, in December 2021, France implemented stricter epidemic prevention measures in response to a surge in cases. As vaccination rates increase, public anxiety gradually decreases. By analyzing the relationship between policy changes and pandemic data, we learn that strict lockdown measures can significantly reduce confirmed cases [4].

In summary, by analyzing the data from the aforementioned sources, we can gain a comprehensive understanding of the development trends and influencing factors of the COVID-19 pandemic. Using big data technology and data mining methods, we can extract valuable information from massive datasets to provide a scientific basis for the formulation of public health policies.

2.3. Characteristics of the Data

COVID-19 pandemic data is characterized by its real-time nature, diversity, large scale, and complexity. The data is updated frequently, often daily, to enable real-time tracking of changes in the pandemic. The types of data are diverse, including the number of confirmed cases, deaths, recoveries, and vaccination rates. The volume of data is substantial, covering detailed information from multiple countries and regions worldwide. The sources of the data are varied, and the formats differ, necessitating preprocessing and cleaning to ensure data accuracy and consistency.

DOI: 10.54254/2755-2721/69/20241513

3. Data Processing and Analysis Methods

3.1. Data Cleaning and Preprocessing

Before analyzing data, it is essential to clean and preprocess the raw data. This includes data integration, where data from different sources are merged to ensure consistency and completeness. Next is handling missing values, which involves methods such as interpolation, deletion, or imputation to ensure the continuity and integrity of the data. This is followed by anomaly detection, where outliers within the data are identified and addressed to prevent them from impacting the analysis results. Finally, data normalization is carried out to facilitate comparisons and analyses across different datasets.

3.2. Data Analysis Techniques

For analyzing COVID-19 pandemic data, commonly used techniques include descriptive statistical analysis, which provides basic statistical descriptions such as mean, median, and standard deviation to understand the fundamental characteristics of the data. Time series analysis is used to examine the data's temporal sequence, revealing the pandemic's patterns and trends over time. Spatial analysis is employed to investigate the distribution of the pandemic across different geographical areas, identifying hotspots and paths of transmission. Machine learning algorithms are utilized for predicting pandemic trends and risk assessments, including classification, regression, and clustering methods.

4. Common Data Visualization Types

As shown in Figure 4, common visualization types used in COVID-19 data analysis include line charts to display trends over time, which are suitable for analyzing daily changes in new confirmed, death, and recovery cases. Heatmaps illustrate the distribution of the pandemic across geographical spaces and are suitable for analyzing the severity of the pandemic in different regions. Bar charts compare pandemic data across different countries or regions, such as total confirmed cases and total death counts. Pie charts show the proportion of different categories of data, such as the age distribution of confirmed cases [5].

Figure 4. Commonly Used Visualization Charts

DOI: 10.54254/2755-2721/69/20241513

5. COVID-19 Pandemic Trend Prediction

As of December 31, 2021, there were over 287 million confirmed cases globally, with 5.42 million deaths. During this period, the United States and France reported 425,000 and 206,000 new confirmed cases, respectively, indicating that the pandemic was still spreading at a high rate. To predict future pandemic trends, researchers have utilized various data sources and prediction models.

5.1. Prediction Models

Time series models are used to analyze the trends in daily new confirmed cases. Based on data from December 2021, the ARIMA model predicts that the global daily new confirmed cases may remain high in the coming weeks, but the growth rate may gradually slow down. Machine learning models, which incorporate vaccination rates, social distancing policies, and seasonal factors, are used. The Random Forest model predicts fluctuations in new cases in some countries as vaccination rates increase and new variants spread [6]. For instance, the model predicts a new peak in the pandemic in the United States and Europe in early 2022, but the overall trend is expected to stabilize gradually.

According to Figure 5, the infectious disease model simulates the dynamics of susceptible, exposed, infected, and recovering populations. By inputting current vaccination data and prevention and control measures, the model predicts a gradual decline in the number of new confirmed cases and deaths globally in the coming months, especially in countries with high vaccination rates.

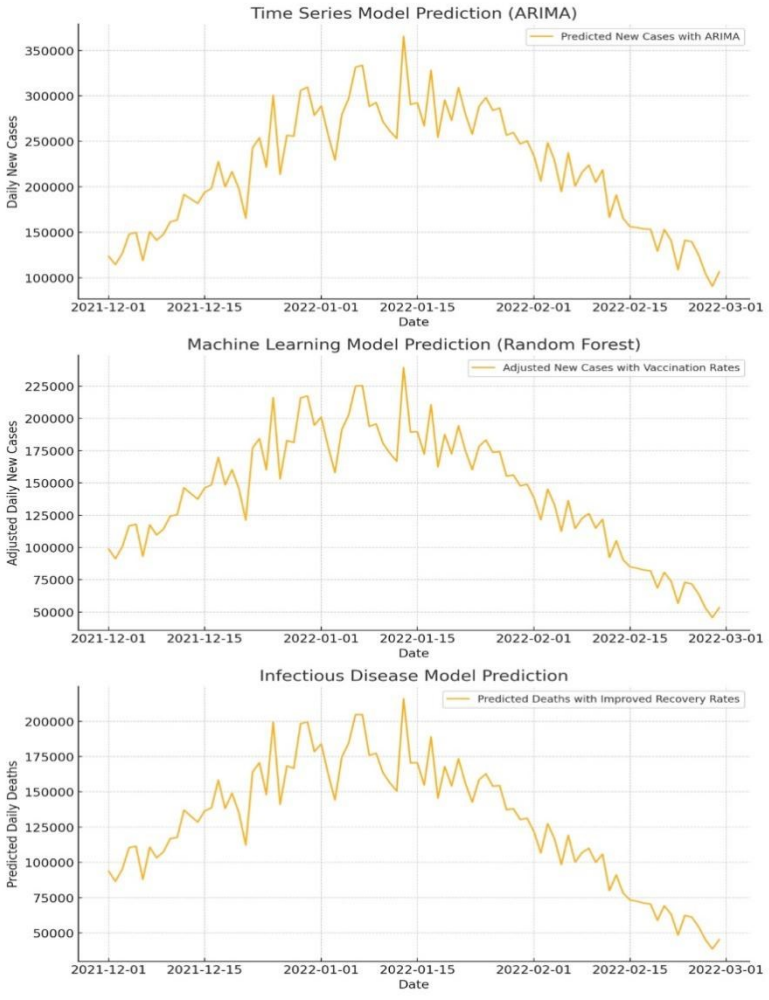

Figure 5. COVID-19 Pandemic Prediction Models

DOI: 10.54254/2755-2721/69/20241513

5.2. Actual Prediction

Combining the above models and data, we predict that in the short term (January to February 2022), the daily new confirmed cases globally will continue to fluctuate, with some countries potentially experiencing peaks due to the spread of the new Omicron variant. Case numbers in the United States and Europe may remain high, but the rate of increase is expected to slow with the broader distribution of booster vaccines [7].

Medium-term (March to June 2022): The global epidemic situation is expected to gradually level off as vaccinations and natural infection rates increase. Especially in countries with high vaccination rates, the number of new cases and severe disease rates are expected to decline significantly. Outbreaks in parts of Asia-Pacific and Africa still require close monitoring, as vaccination rates in these regions are relatively low.

In the long term (second half of 2022 and beyond): The global epidemic is expected to be brought under control without the emergence of new highly contagious variants. The effectiveness of national policies and public health measures will continue to influence the development of the epidemic. Further optimization and popularization of vaccines and therapeutics will provide strong support for the complete control of the epidemic.

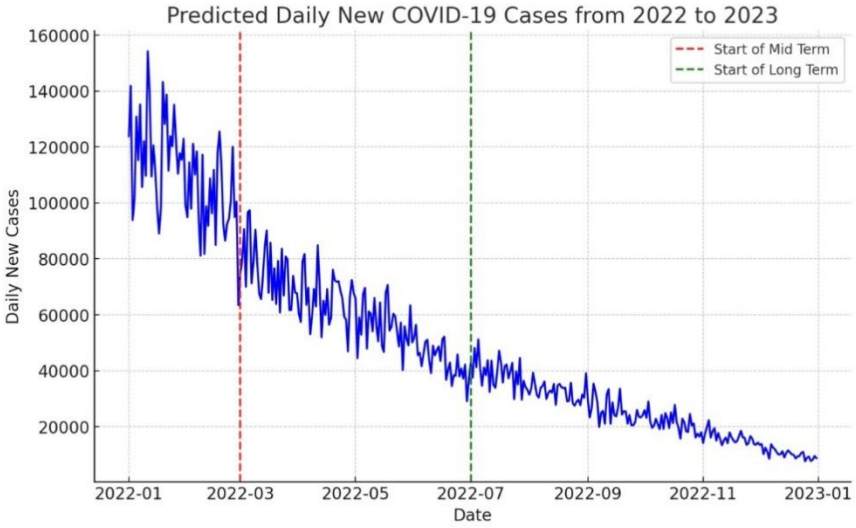

Figure 6. Predicted Daily New COVID-19 Cases from 2022-2023

Through the application of multiple data mining and prediction models in Figure 6, we analyze and predict the future trend of the new crown pneumonia epidemic. While the outbreak may still fluctuate in the short term, it is expected to be effectively contained in the medium to long term as global vaccination rates increase and public health measures continue to be implemented. These predictions provide an important reference for public health policymakers to formulate and adjust prevention and control strategies to defeat the epidemic.

6. Case Studies

6.1. Global Epidemic Trends

Most countries experienced multiple peaks in 2020 and 2021. Line chart analysis provides a clear view of the timing and peaks of these peaks. In the United States, for example, according to Johns Hopkins University, the first peak of the epidemic was reached in December 2020 and January 2021, with the number of new confirmed cases per day exceeding 300,000 at one point. In August 2021, with the spread

DOI: 10.54254/2755-2721/69/20241513

of the Delta variant, the United States once again ushered in a peak of the epidemic, with the number of new confirmed cases per day again exceeding 200,000 [8].

As vaccinations rolled, the number of new confirmed cases and deaths dropped significantly in most countries. This trend can be seen in the time series plot, as shown in Figure 7, in the case of Israel, which has one of the highest vaccination rates in the world. According to Our World in Data, as of June 2021, more than 60% of Israel's population had received two doses of the vaccine, and with that was a significant drop in the number of new confirmed cases and deaths.

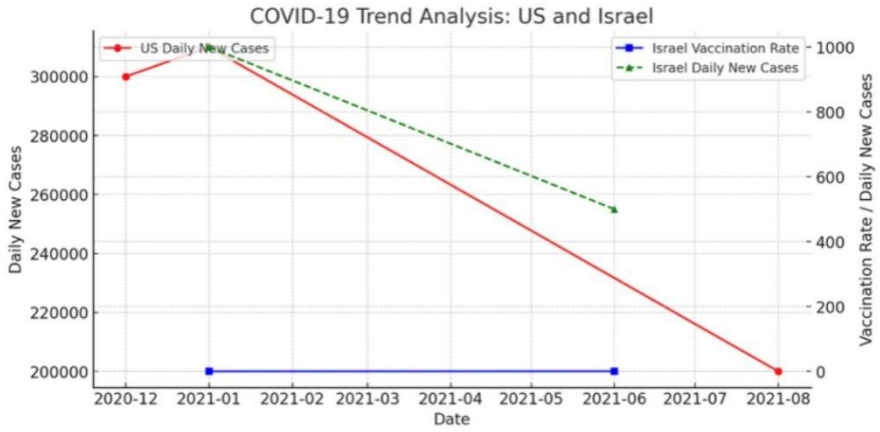

Figure 7. COVID-19 Trend Analysis: US and Israel

6.2. Geospatial analysis

In terms of hot spots, such as the United States, India, Brazil, and other countries, the epidemic is more serious, and the number of confirmed cases and deaths remains high. Through the heat map, you can clearly see the hot spots of the epidemic in these countries and help identify the key prevention and control areas. India, for example, had more than 400,000 new daily confirmed cases in May 2021, making it a global hotspot.

In terms of transmission paths, spatiotemporal analysis can be used to find out the transmission paths of the epidemic. For example, in early 2020, the outbreak first broke out in Wuhan, China, then spread to Europe, with Italy and Spain as the hardest hit areas, and then to the Americas, especially the United States and Brazil. In 2021, the Delta variant was first detected in India and then spread to several countries around the world. Geographic information systems (GIS) can be used to track the path and speed of the spread of the virus, helping public health authorities to take targeted prevention and control measures [9].

6.3. Social Media Data Analytics

By analyzing data related to the epidemic on social media, it is possible to capture the public's reaction and mood changes to the epidemic:

Using natural language processing technology to analyze text data on social media such as Twitter and Facebook, it is possible to detect changes in public attitudes towards the epidemic at different stages. For example, when the pandemic first broke out in early 2020, the public's mood was dominated by panic and anxiety. As the pandemic progressed and vaccines rolled out, the mood on social media gradually shifted to hope and optimism. Through sentiment analysis, changes in public sentiment can be quantified and compared with pandemic data to reveal the impact of the pandemic on public mental health [10].

DOI: 10.54254/2755-2721/69/20241513

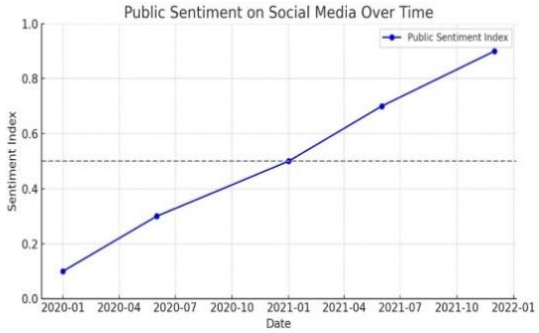

Figure 8. Public Sentiment on Social Media Over Time

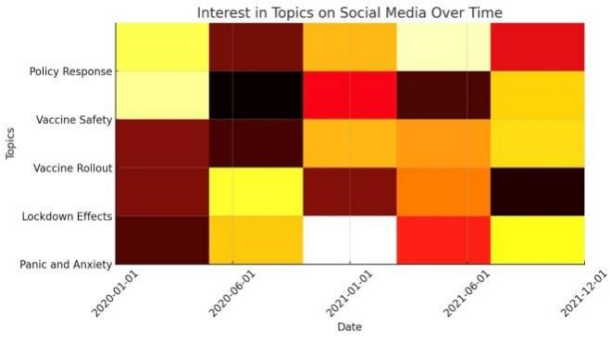

Figure 9. Interest in Topics on Social Media Over Time

According to Figure 8, the line chart shows the trend of public sentiment on social media from the beginning of 2020 to the end of 2021.

The sentiment index ranges from 0 (very negative) to 1 (very positive). It can be seen that over time, public sentiment has gradually shifted from panic and anxiety to hope and optimism.

Figure 9 shows the level of public interest in various topics on social media at different points in time. From left to right, they represent different time points, and from top to bottom, they represent different topics, such as pandemic panic, lockdown impact, vaccine rollout, vaccine safety, and policy response. The hotter the color of the patch, the more attention the topic has at that point in time.

These charts help policymakers and researchers understand how the public's perception of the pandemic is changing and what topics are of concern, so they can better adjust public health strategies and communications.

7. Conclusions and outlook

7.1. Conclusion

This paper discusses the methods and applications of big data mining technology for the analysis and trend prediction of COVID-19 pandemic data. Through the processing, analysis, and prediction of global epidemic data, we can better understand the development trend and influencing factors of the epidemic, and provide a scientific basis for the formulation of public health policies.

7.2. Outlook

In the future, with the continuous development of big data technology and data mining tools, we can process and analyze massive epidemic data more efficiently. At the same time, combined with more data sources, such as genomic data and environmental data, the transmission mechanism and mutation of the virus can be studied in depth, providing more comprehensive and accurate guidance for epidemic prevention and control.

In the context of the pandemic, the application of big data and data mining technology has not only improved our data analysis capabilities but also promoted scientific decision-making and policy formulation in the field of public health. It is hoped that the research and discussion in this paper can provide a valuable reference for researchers and policymakers in related fields.

References

[1] World Health Organization. COVID-19 Dashboard. [Internet]. 2020 [cited 2024 May 30]. Available from: https://covid19.who.int/

[2] Johns Hopkins University. COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE). [Internet]. 2020 [cited 2024 May 30]. Available from: https://github.com/CSSEGISandData/COVID-19

DOI: 10.54254/2755-2721/69/20241513

[3] Tableau Software. Tableau COVID-19 Data Hub. [Internet]. 2020 [cited 2024 May 30]. Available from: https://www.tableau.com/covid-19-coronavirus-data-resources
[4] Microsoft Power BI. COVID-19 Tracking Report. [Internet]. 2020 [cited 2024 May 30]. Available from: https://powerbi.microsoft.com/en-us/blog/covid-19-tracking-report/
[5] McKinsey & Company. How data visualization supports rapid crisis decision-making. [Internet]. 2020 [cited 2024 May 30]. Available from: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/how-data-visualization-supports-rapid-crisis-decision-making
[6] Ahmed I, Ahmad M, Jeon G, Piccialli F. A framework for pandemic prediction using big data analytics. Big Data Research. 2021 Jul 15;25:100190.
[7] Alsunaidi SJ, Almuhaideb AM, Ibrahim NM, Shaikh FS, Alqudaihi KS, Alhaidari FA, Khan IU, Aslam N, Alshahrani MS. Applications of big data analytics to control COVID-19 pandemic. Sensors. 2021 Mar 24;21(7):2282.
[8] Sengupta S, Mugde S, Sharma G. Covid-19 pandemic data analysis and forecasting using machine learning algorithms. MedRxiv. 2020 Jun 26:2020-06.
[9] Clement F, Kaur A, Sedghi M, Krishnaswamy D, Punithakumar K. Interactive data-driven visualization for COVID-19 with trends, analytics and forecasting. In2020 24th International Conference Information Visualisation (IV) 2020 Sep 7 (pp. 593-598). IEEE.
[10] Pan Z, Nguyen HL, Abu-Gellban H, Zhang Y. Google trends analysis of covid-19 pandemic. In2020 IEEE International Conference on Big Data (Big Data) 2020 Dec 10 (pp. 3438-3446). IEEE.