A Comparative Analysis of the Major COVID-19 Variant of Concern (VOC) in the World and India

Veena Jokhakar
Veer Narmad South Gujarat University
Surat, Gujarat, India

Tejas Shah
Veer Narmad South Gujarat University
Surat, Gujarat, India

ABSTRACT
The world is aware of how Corona, a virus belonging to the Nidovirus family, impacted it. In 2019, this had an impact on a large number of people across several nations. Later down the line it has been seen that various types of versions of the coronavirus kept evolving, which were different in terms of symptoms and intensity. The WHO-based assessment designated the multiple COVID-19 Variants of Concern (VOCs) and Variants of Interest (VOIs) for causing new waves with increased spread and the need for required public health actions. This paper shows the comparative analysis of major Variants of Concern (VOCs) and their num_sequences and pert_sequences with respect to location and time. The analysis done through Power BI shows the most and least affected countries with a specific variant of concern.

Keywords
COVID-19, Variants of Concern (VOC), Power BI

1. INTRODUCTION
SARS-COV2 evolved from Wuhan, China, all over the world within a few months; it was confirmed to be a severe acute infection that required isolation, sequencing, and phylogenetic analysis. [1]

Since the advent of the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) in late 2019, the virus has seen ongoing genetic change via mutation and recombination. These changes have given rise to multiple viral lineages that differ in transmissibility, virulence, and immune escape potential. To systematically monitor and classify these emerging variants, the World Health Organization (WHO) and other health agencies, such as the Centers for Disease Control and Prevention (CDC), introduced a nomenclature system based on Variants of Concern (VOCs), Variants of Interest (VOIs), and Variants Under Monitoring (VUMs). These variants have been designated based on their assessed potential for expansion and replacement of prior variants, for causing new waves with increased circulation, and for the need for adjustments to public health actions. [2]

The ongoing evolution of SARS-CoV-2 is driven by natural selection, population immunity, and global transmission dynamics. Variants continue to emerge as the virus adapts to selective pressures such as widespread vaccination and antiviral treatments. Consequently, genomic surveillance and timely identification of new variants remain essential components of the global COVID-19 response strategy.

The outbreak of corona-virus led all academicians, scientists, doctors, and the government to not spreading it. The SARS-COV-2 variant is seen as an expansion replacement of prior encountered variants and spreads at an increased pace. In order to comprehend the impact of the COVID-19 virus, we suggest in this paper doing a thorough data analytics investigation. Additionally, we compare and contrast its severity in terms of Variant of Concern (VOC) and Variant of Interest (VOI) with different variants like Alpha, Delta, Kappa, Omicron. We have gathered information from reliable sources that is generally acknowledged by the scientific community in order to accomplish this goal.

A (VOC) is defined as a SARS-CoV-2 strain that shows evidence of increased transmissibility, more severe disease (e.g., higher hospitalization or mortality rates), or significant reduction in neutralization by antibodies generated through vaccination or prior infection. VOCs have posed major public health challenges, driving successive global waves of the COVID-19 pandemic.

List of VOCs:

- Alpha (B.1.1.7) – First detected in the UK (September 2020)
- Beta (B.1.351) – First detected in South Africa (May 2020)
- Gamma (P.1) – First detected in Brazil (November 2020)
- Delta (B.1.617.2) – First detected in India (October 2020)
- Omicron (B.1.1.529 and its sub-lineages) – first detected in South Africa/Botswana (November 2021)

In contrast, a VOI refers to a viral lineage with genetic changes that are suspected or known to affect viral characteristics—such as transmissibility, immune escape, or disease severity—but where the epidemiological impact is not yet fully established. Examples include Eta (B.1.525), Iota (B.1.526), Lambda (C.37), and Mu (B.1.621).

Table 1 summarizes the major Variants of Concern (VOC) for SARS-CoV-2, including their name, year-month, lineages, first detection locations, current status, and notable impacts.[3][4][5]

Table 1. Major VOC of SARS-Covid-19

Year-Month	Variant	Lineage	First Detected	Current Status	Notable Impact Notes
2019-Dec	–	Wuhan-Hu-1 (Reference)	China (Wuhan)	Baseline	First identified human SARS-CoV-2 cases
2020-Sept	Alpha (VOC)	B.1.1.7	United Kingdom	De-escalated Mar 2022	Increased transmissibility, Possible higher severity, Spread globally early 2021
2020-May	Beta (VOC)	B.1.351	South Africa	De-escalated Mar 2022	Immune escape, Increased transmissibility, Moderate global spread
2020-Nov	Gamma (VOC)	P.1	Brazil / Japan	De-escalated Mar 2022	Immune escape, Increased transmissibility, Major Brazilian wave
2020-Oct	Delta (VOC)	B.1.617.2	India	De-escalated Mar 2023	Severe global wave, High transmissibility, Partial immune escape
2021-June	Lambda (VOI)	C.37	Peru	De-escalated Mar 2023	Dominant in Peru, Moderate immune escape
2021-Aug	Mu (VOI)	B.1.621	Colombia	De-escalated Mar 2023	Immune escape, Limited spread
2021-Nov	Omicron (VOC)	B.1.1.529	Botswana/South Africa	Active (2025)	Significant immune escape, Very high transmissibility, Reduced severity
2023-Feb	EG.5 (Eris, VOI)	EG.5 / EG.5.1	China / Global	Active 2024–2025	Moderate growth advantage, Monitored globally
2023-Dec	JN.1 (VOI)	BA.2.86 descendant	Denmark / Israel / USA	Active 2024–2025	Dominant Omicron desc. Strong immune escape.

The earlier variants of Variant of Concern (VOC) had occurred in different places like the UK, South Africa, Brazil, and India and multiple countries respectively named alpha, beta, gamma, and delta earlier in the year 2020 and omicron in the year 2021.

The rest of the paper is organized as follows. Section II covers the related work, which covers the research carried out in the field of major VOCs and prediction and visualization methods. Section III included the methodology, Section IV covers the results and discussion, and the last section concludes the paper.

2. RELATED WORK

Many researchers have explored their idea, invention, analysis of SARS-CoV-2 variants through scientific methodology, medical process, comparison of variants, impact on health, solutions through vaccination, and prediction.

Several VOC mutations were covered in [6] et al., along with the necessity of combination therapy approaches that target the viral cycle and immune host responses.

This review article outlines the different SARS-CoV-2 variations that have surfaced, with a focus on the VOCs that are spreading globally. It also discusses the effects of multiple viral mutations and how these changes affect the virus's characteristics [7].

Five SARS-CoV-2 VOCs with mutations in the S gene (B.1.1.7, B.1.351, P.1, B.1.617.2, and B.1.1.529) have been described by Hirabara et al. in [8]. Additionally, they emphasized the potential for vaccination or evasion of neutralizing antibodies produced by the prior infection, which could lessen the impact of such novel variations during the pandemic.

Additionally, the Omicron variant is notable for its antibody escape, or antibody-mediated neutralization resistance, and partial vaccination escape because of numerous important mutations in S-glycoprotein [9].

Authors of [10] discussed autoregressive integrated moving average time series (ARIMA) for forecasting confirmed cases . They further make use of random forest and extra tree classifiers

to measure the accuracy by achieving an accuracy of 90% using random forest and any extra tree classifier with 93%.

A state-of-the-art analysis of the ongoing COVID-19 pandemic has been done through machine learning (ML) and deep learning (DL) methods in the diagnosis, forecasting, and prediction. Moreover, a comparative analysis on the impact of machine learning and other competitive approaches like mathematical and statistical models on the COVID-19 problem has been conducted [11].

In [12], the authors implemented an ML-based enhanced model to predict the possible threat of COVID-19 all over the world, and the algorithm classifies the COVID patients based on several subsets of features and predicts their likelihood of getting affected by this disease. This model uses 20 metrics, including the patient's geographical location, travel history, health record statistics, etc., to predict the severity of the case and the feasible outcome. The model developed using K-Nearest Neighbors (KNN) is effective with a prediction accuracy of 98.34%, Recall of 97% and an F1-Score of 0.97.

In this paper [13], the researchers reviewed the newly born data science approaches to confronting COVID-19, including the estimation of epidemiological parameters, digital contact tracing, diagnosis, policy-making, resource allocation, risk assessment, mental health surveillance, social media analytics, drug repurposing and drug development.

Table 2 shows the literature review for some of the work, which focuses on analysis of COVID-19 data, vaccination, and other data using different tools like Power BI, Tableau, and Python.

Table 2: Literature Review of Analysis of COVID-19 using Different Tools

Authors/Paper	Techniques / Tools	Summary of Paper / Key Findings
Shirke et al. (2023) [14]	Power BI, Data Visualization	Number of people vaccinated for the first and second dose, the gender-wise and the state-wise distribution of the vaccine
Leung, C. K et al. (2020) [15]	Data Visualization	Big data visualization and analytic tools to analyse COVID-19 data
Kaufmann, T. N. (2021) [16]	Data Mining and Visualization	Data analysis and visual analytics of COVID-19 data
Rajeevan, S. et al. (2024) [17]	Tableau, Data Visualization	Visualization of COVID-19 data in India using Tableau, emphasizing key metrics such as the total number of cases, the distribution of confirmed cases by age group and gender
Bijay Halder et al. (2023) [18]	Statistical Analysis	Statistical analysis to calculate the mortality rate, ratio between active and death cases, active cases and recovered cases, recovered and death cases in India
Clement, F. et al. (2020) [19]	Data Visualization, Python	Data driven visualization of COVID-19 using Python with interactive dashboard
Jokhakar V. et al.(2020) [20]	Power BI, Data Visualization	Analysis of COVID-19 for death and recovery in India
Akhtar N. et al. (2020) [21]	Tableau, Data Visualization	Dashboards, models, interactive visualization creation using Tableau to discover COVID-19 data patterns
Das (2025) [22]	ARIMA, LSTM and Power BI	Analysis through Machine learning model ARIMA, LSTM and visualization of cases through Power BI

It can be observed from the above review that the analysis of major VOC for COVID-19 has not been done with visualization tools. Vaccinatig85on data is analyzed with different tools.

3. METHODOLOGY

In this work, we have used the visualization module of Microsoft Power Platforms, a professional software, to analyze the collected data and develop visualization dashboards about the corona-virus disease. Our methodology consists in creating descriptive models of the COVID-19 outbreak using statistical charts to understand the number of sequences and percent sequence of major VOC and VOI by applying the steps of data extraction, transformation, and loading. We develop our analysis at three levels, namely, at the country level, at the region level, and at the continent level. Each level provides different granularity towards understanding the distribution of the disease around the world. The descriptive model provides different types of statistical charts, including bar charts and geographic maps to represent different features of the COVID-19 outbreak.

We have selected Covid-19 variants from Kaggle [23], which shows virus variant evolution (including Delta and Omicron). Following are the key components of this data set [23]:

Variant: The specific genetic lineage of the organism (e.g., Alpha, Delta, Omicron for COVID-19).

Num_sequences: The raw number of individual samples that have been sequenced and matched to that particular variant.

Location and Date: The geographical region and time frame for

which the data was collected.

Perc_sequences (Percentage of sequences): Often included alongside the raw count, this represents the proportion of a specific variant's sequences relative to the total number of sequences processed in that location and time period, providing context on its prevalence [23].

3.1 ETL Process

Data warehousing, analytics and machine learning pipelines all use the ETL process (Extract, Transform, Load), a basic workflow for data integration and administration. It helps businesses to transfer data in a clean, organized, and usable format from several sources into a single location (such as a data warehouse or data lake).

3.1.1 Extract

The extraction phase is the first step of the ETL process, where raw data is collected from various sources. The following Fig.1 shows the extraction phase of the Covid-19 data set.

3.1.2 Transform

The transformation phase converts the data set into a consistent set using the Power Query Builder. During this phase, the data is cleaned, aggregated, and formatted according to business rules. This is a crucial step because it ensures that the data meets the quality standards we require for accurate analysis of major VOC. Fig. 2 shows the changing to the data type process.After that step aggregation is performed for the attributes location and variant as shown in Fig. 3. The next step includes the splitting of column transformation. We are here selecting the date that is delimited. This formed three new dimensions after splitting as shown in Fig. 4. This transformation is applied on location and variant grouped by year, hence showcasing location-wise variant data grouped by year as shown in Fig. 5.

3.1.3 Loading of Dataset

After we perform the dataset selection and perform the various steps of transformations as required for the analysis and preparation of data, we move forward for the next steps of loading of data as shown in Fig. 6. This displays the formal data set to preview before loading, we load it by clicking on load to load the data set.

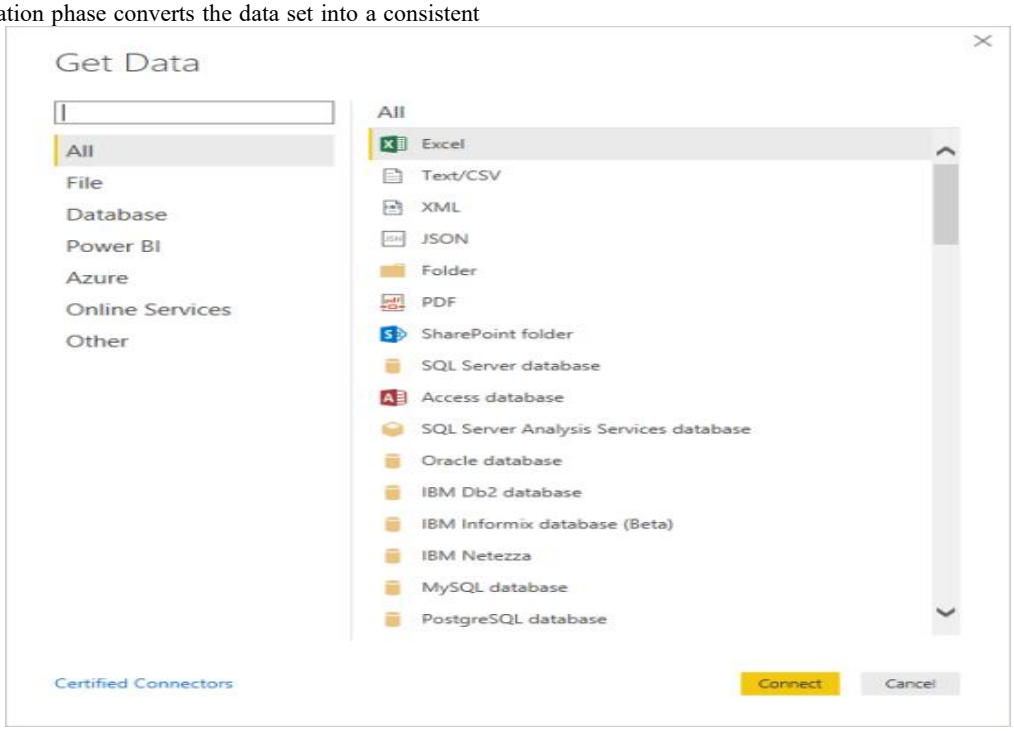


Fig 1: Extraction of Data Set

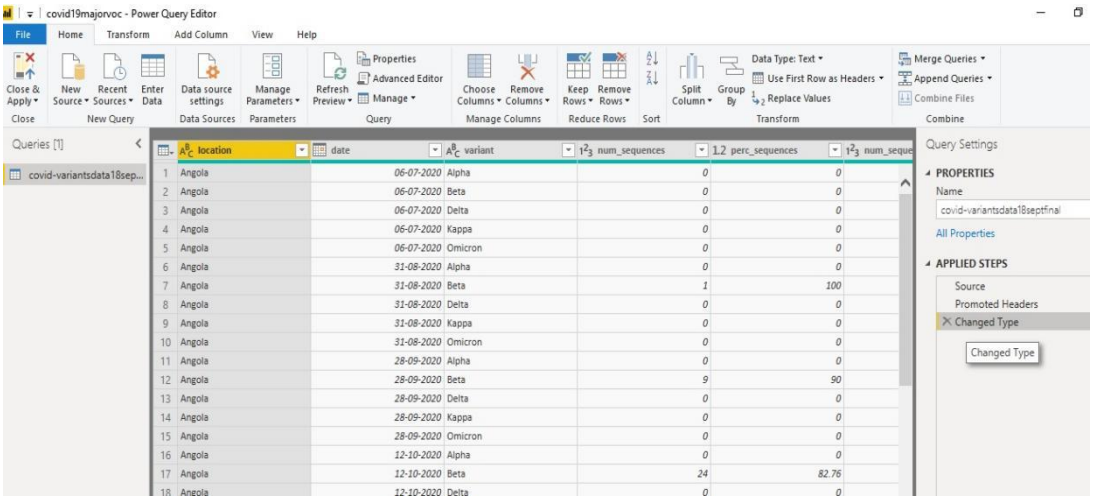


Fig 2: Change Data Type

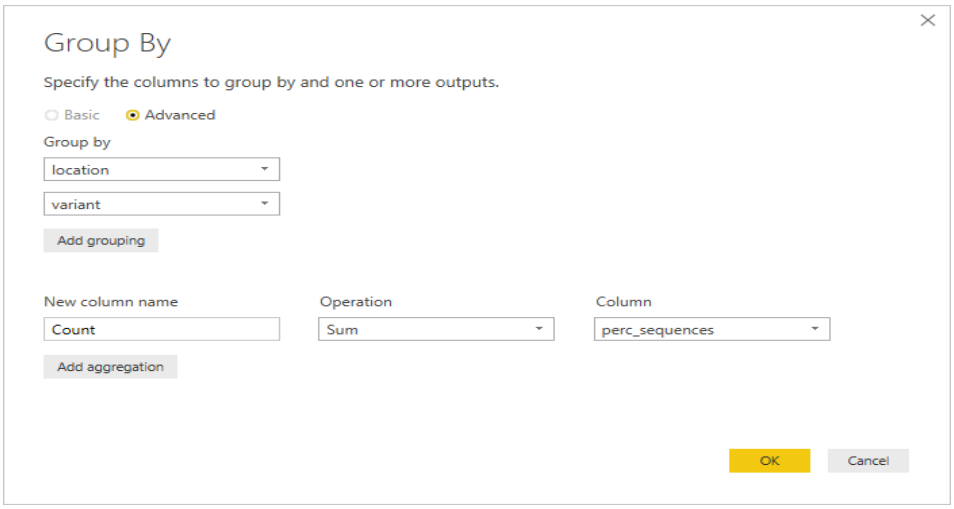


Fig 3: Group by

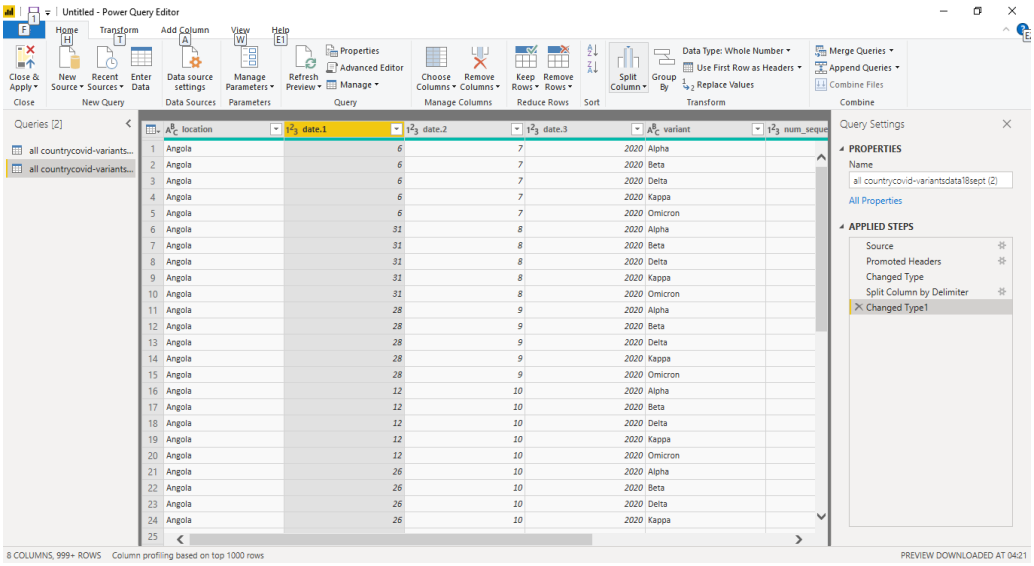


Fig 4: Splitting of Date Columns

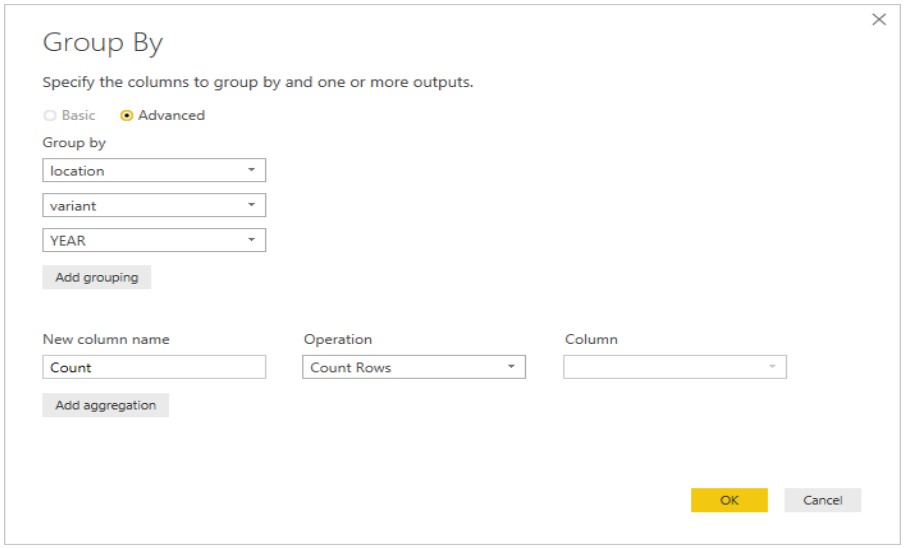
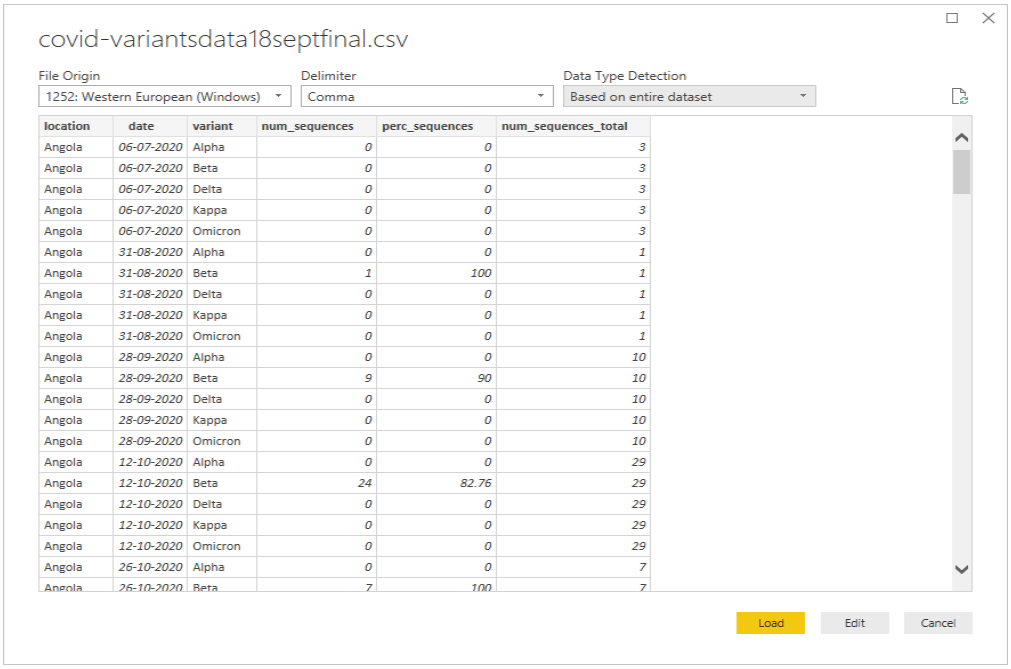


Fig 5: Aggregation of location, variant, year



location	date	variant	num_sequences	perc_sequences	num_sequences_total
Angola	06-07-2020	Alpha	0	0	3
Angola	06-07-2020	Beta	0	0	3
Angola	06-07-2020	Delta	0	0	3
Angola	06-07-2020	Kappa	0	0	3
Angola	06-07-2020	Omicron	0	0	3
Angola	31-08-2020	Alpha	0	0	1
Angola	31-08-2020	Beta	1	100	1
Angola	31-08-2020	Delta	0	0	1
Angola	31-08-2020	Kappa	0	0	1
Angola	31-08-2020	Omicron	0	0	1
Angola	28-09-2020	Alpha	0	0	10
Angola	28-09-2020	Beta	9	90	10
Angola	28-09-2020	Delta	0	0	10
Angola	28-09-2020	Kappa	0	0	10
Angola	28-09-2020	Omicron	0	0	10
Angola	12-10-2020	Alpha	0	0	29
Angola	12-10-2020	Beta	24	82.76	29
Angola	12-10-2020	Delta	0	0	29
Angola	12-10-2020	Kappa	0	0	29
Angola	12-10-2020	Omicron	0	0	29
Angola	26-10-2020	Alpha	0	0	7
Angola	26-10-2020	Beta	7	100	7

Fig 6: Loading of Data set

4. RESULTS and DISCUSSION

Using Microsoft Power BI, we examined the covid-variants.csv data set and found the crucial insights pertaining to the widespread of the virus pandemic. We depict different analysis by portraying the different visualizations.

Based on the data collected from Kaggle, WHO and other legitimate sources, 4th quarter of 2020 was the initial months of the start of onset of the corona pandemic, which rose exponentially in the year 2021, mainly in the months of March, April, May, June, and July, incurring huge loss of life throughout the country in India and other countries and continents like USA,Africa etc.

The following section consists of various charts that show the comparison and analysis of major VOCs.

As shown in Fig. 7, the study shows that the virus was not present up to quarter 3 in the year 2020. The delta variant was initiated in Qtr 4 of 2020, and the kappa variant, which is a VOI, was active in Qtr 1 and Qtr 2 of 2021. With the inception of the year 2021, the Delta variant was having major prominence, while the Alpha variant was spread widely as compared to Omicron and Kappa. The Omicron variant is considered a VOC and was found in Qtr 4 of 2021.

It can be observed from Fig. 8 and Fig. 9 that all the major VOCs are increased in the year 2021 in terms of number of sequences and percent of sequences. The delta has reached the highest level in perc_sequences in the year 2021.

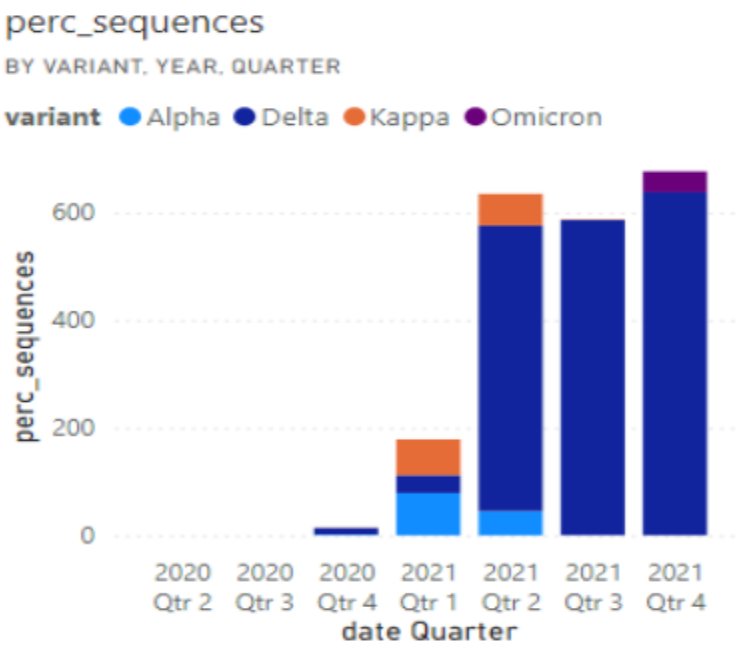


Fig 7: Perc_sequences of variant by quarter

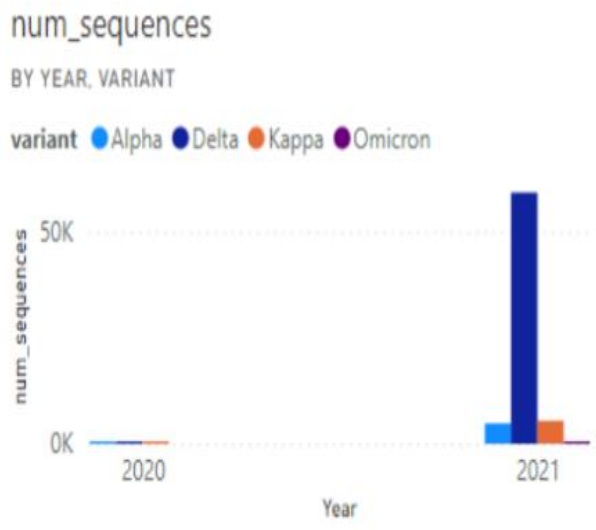


Fig 8: Num_sequences Year

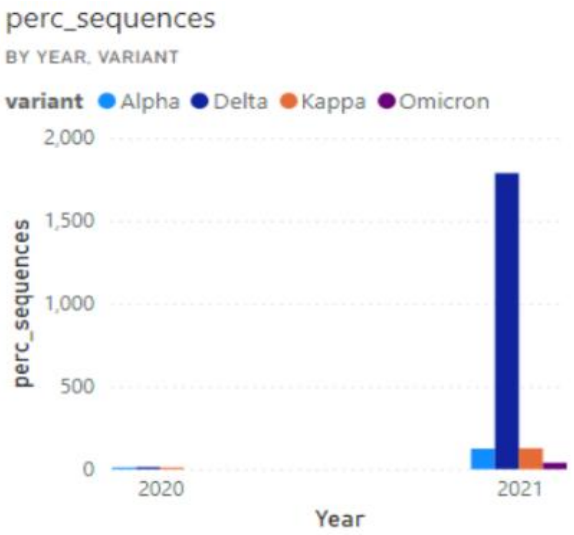


Fig 9: Perc_sequences Year

The report shown in Fig. 10 and Fig. 11 displays the propagation of different variants in terms of Perc_Sequences, showing the Delta variant had reached the highest count during the months of May and December, whereas Alpha and Kappa were at the lowest, followed by Omicron. This figures out the various variants and percentage sequences for each month. The above report in Fig. 12 shows the continent-wise count of variants, depicting the highest in South Africa with the Beta variant and United Kingdom with the Delta variant.Finally, we conclude by looking at the visuals that the outreach of the virus burst was highest in the year 2021 in various continents. The visual shown in Fig. 14 depicts that the lowest count of the variants was observed in German, Denmark and Japan.

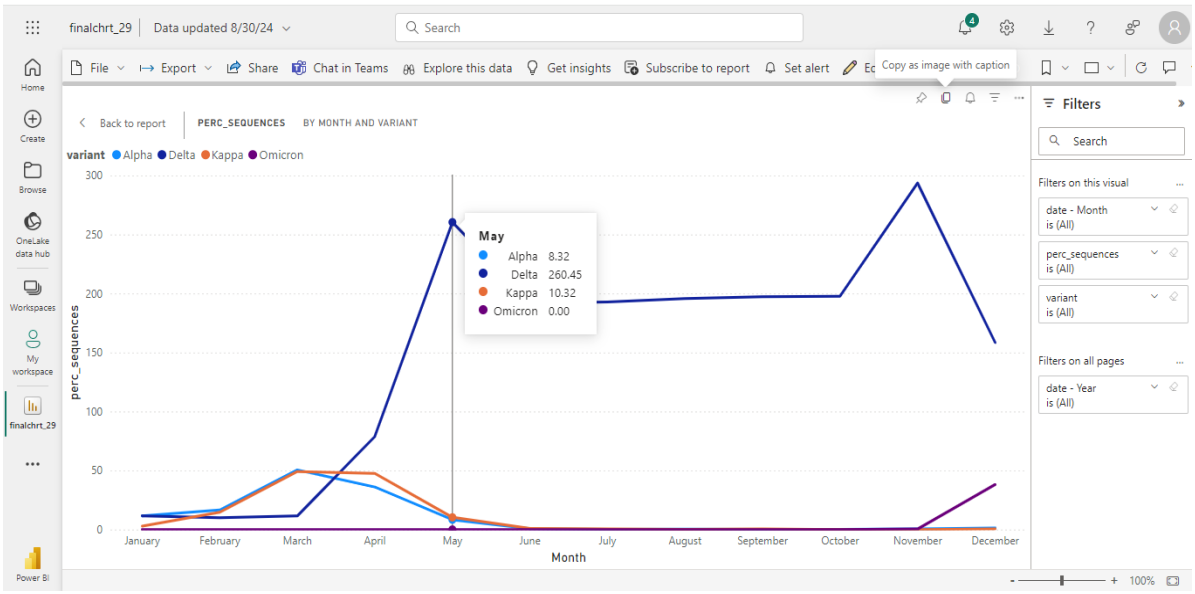


Fig 10: Perc_sequences by month and variant-May

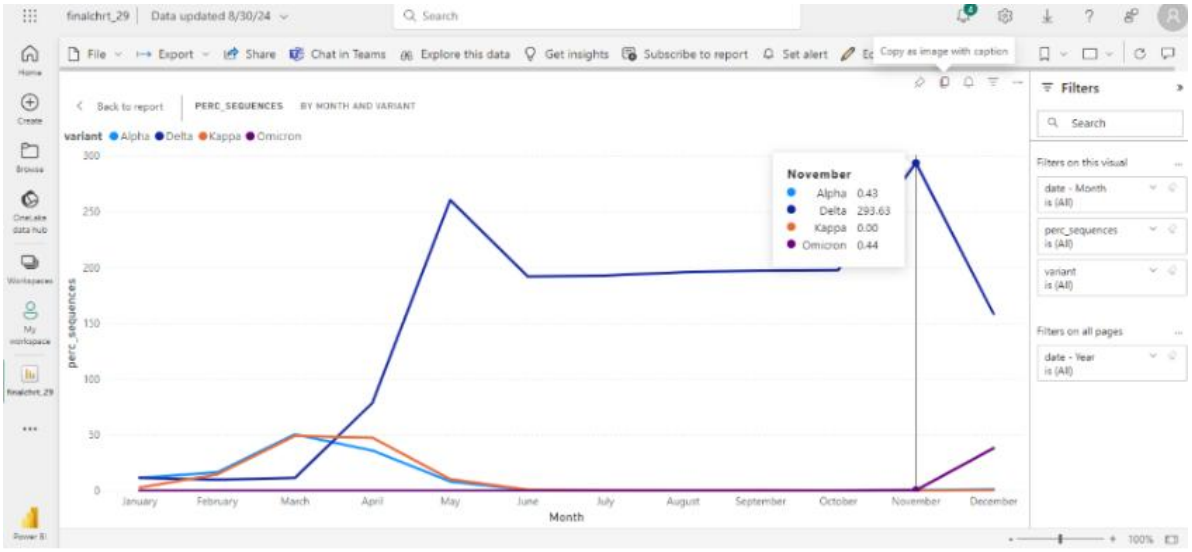


Fig 11: Perc_sequences by month and variant-November

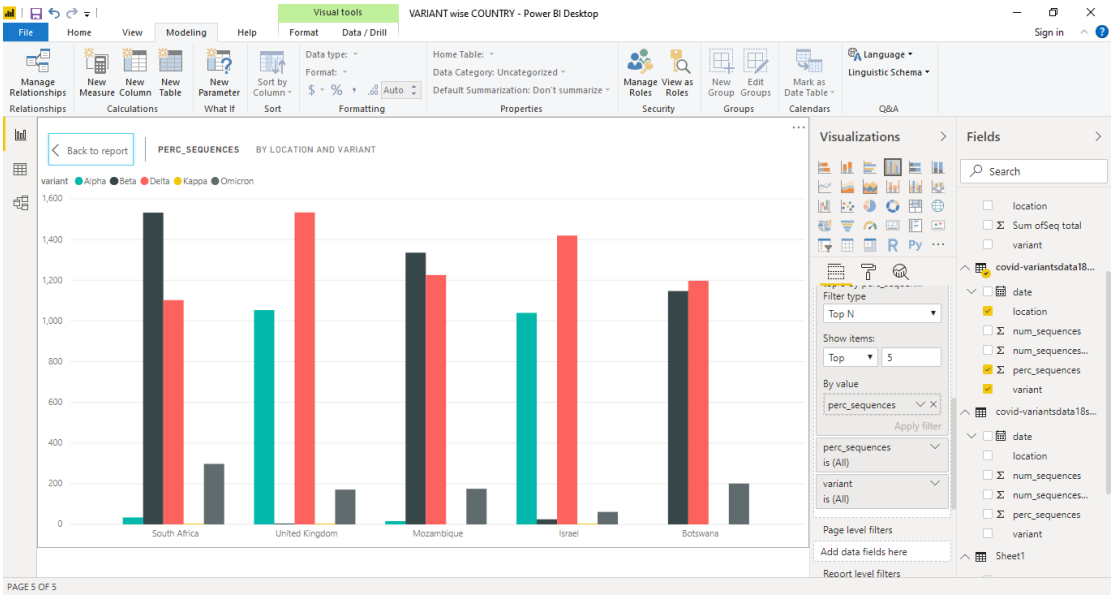


Fig 12: Top 5 locations with Variant

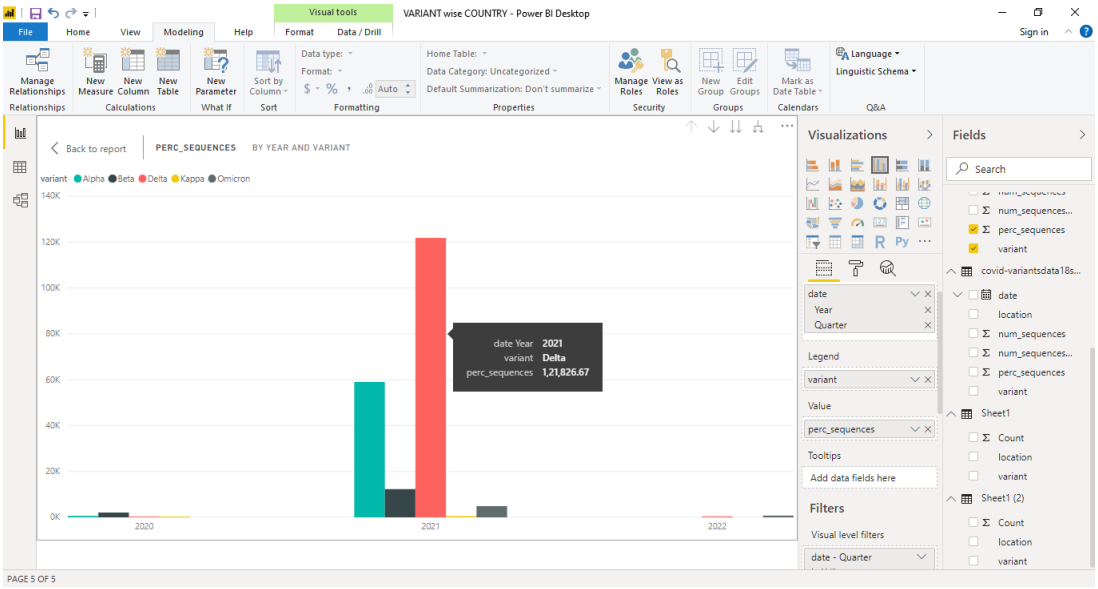


Fig 13: perc_sequences Delta Variant

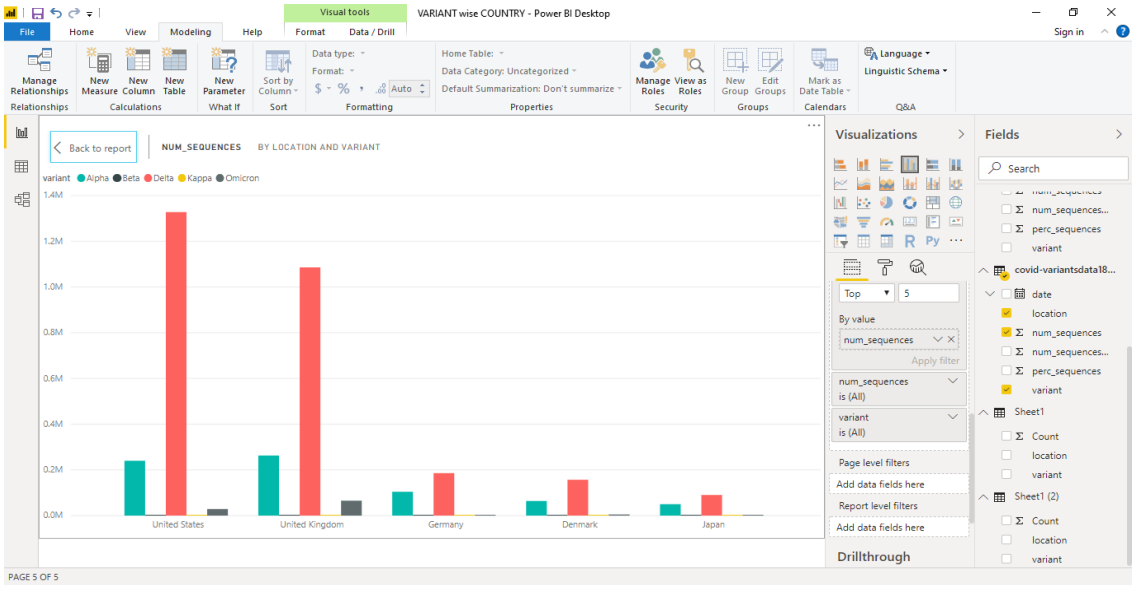


Fig 14: num_sequences lowest in location

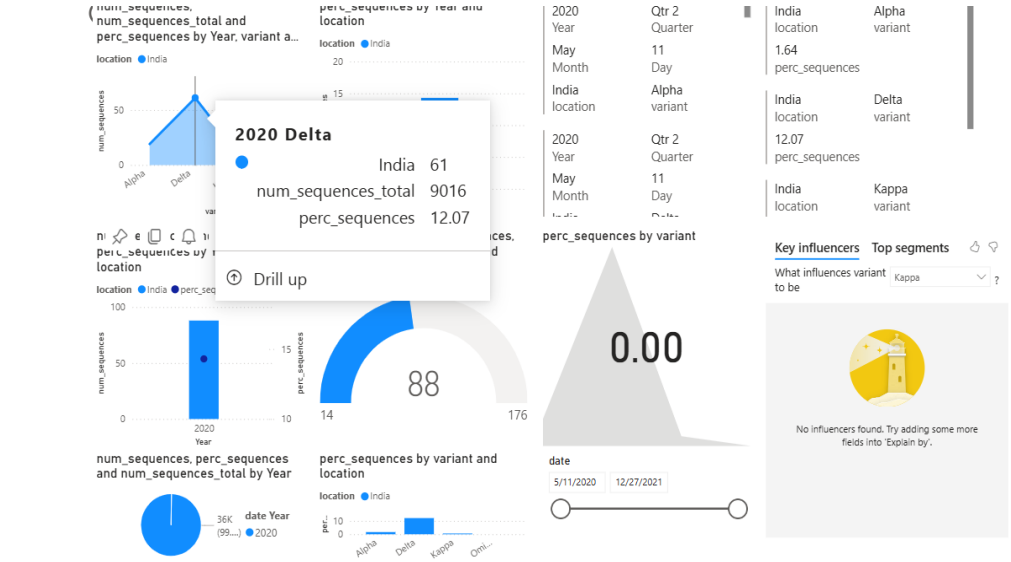


Fig 15: Dashboard of All Variants

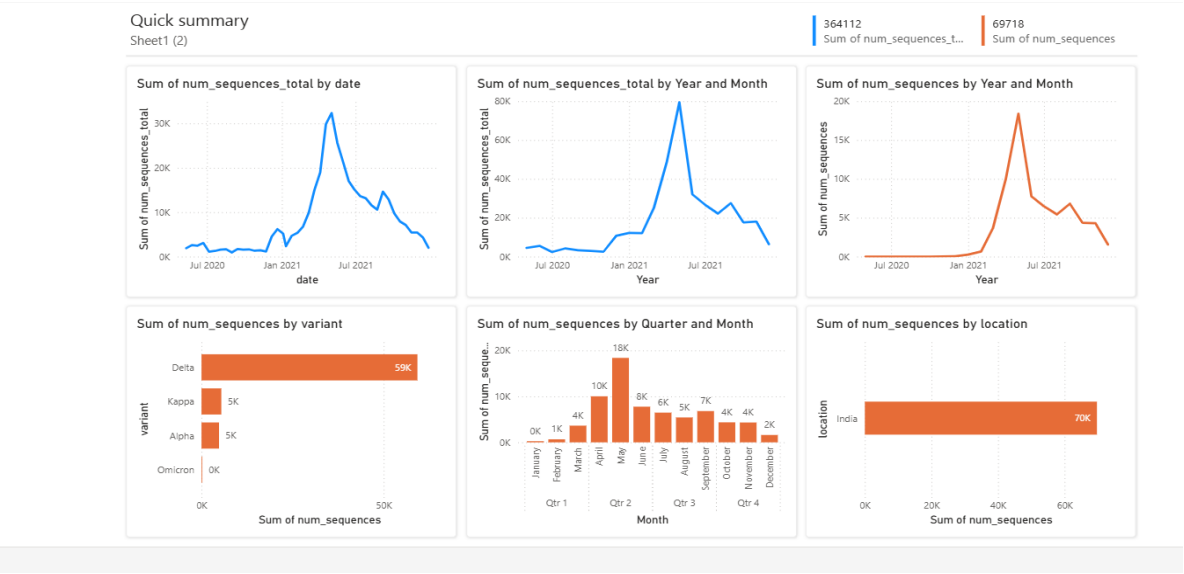


Fig 16: Dashboard - Quick Summary

We also generated a combined dashboard shown in Fig. 15 and Fig. 16 depicting the summary of various displays indicating the start of the pandemic. This also highlights the Delta variant being the most serious variant of concern as per our analysis.

In this work, we have used Power BI, a power platform software, to analyze the collected data and develop visualization dashboards about the corona virus disease.

Our methodology consists of creating descriptive models of the corona virus outbreak using statistical charts to understand the nature of the spread and its impact. We develop our analysis at three levels, namely, at the country level, at the region level, and continent level. Each level provides different granularity towards understanding the distribution of the disease around the world.

The descriptive model provides different types of statistical charts, including bar charts, geographic maps, line charts, and cluster graphs to represent different features of the COVID-19 outbreak.

5. CONCLUSION

This paper focuses on the chronological emergence and analysis of the number of sequences of major VOCs from 2019 to 2024, highlighting their implications. This comparative study shows the spike in the number of sequences in VOC and not in VOI. Through Power BI, the dashboard has shown unique charts that reveal the increase of major VOC. Our analysis through Power BI showed that the Delta variant, Beta variant and Omicron variant had the highest number of sequences in UK, South Africa respectively. The last variant, Omicron, does not have a major impact on the severity part compared to Alpha. Future work that could be performed in extension to this work can be vaccine emergence related with respective country, effectiveness and results of the vaccines, also focusing on the side effects of each of them with the relative time to recover or cure. Vaccination data can be linked to the major VoCs and VoIs. Further analysis could be performed on the relationship or association between the vaccine's content and the health parameters by applying machine learning techniques. This can be very useful for the counter measures for combating in case of further development of such viruses.

6. REFERENCES

[1] Biswas A., Bhattacharjee U., Chakrabarti A.K., Tewari D.N., Banu H., Dutta S., 2020. Emergence of Novel Coronavirus and COVID-19: Whether to Stay or Die Out? Crit. Rev. Microbiol. 2020;46:182–193.

[2] World Health Organization, "WHO Coronavirus (COVID-19) Dashboard," WHO, 2024. [Online]. Available: https://data.who.int/dashboards/covid19/cases.

[Accessed: 22-Jul-2024].

[3] National Library of Medicine, "Characteristics of SARS-CoV-2 Variants of Concern", 2025 [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9866527/ [Accessed: 22-Aug-2025]

[4] Wikipedia, "SARS-CoV-2", 2025 [Online]. Available: https://en.wikipedia.org/wiki/SARS-CoV-2 [Accessed: 15-10-2025]

[5] National Library of Medicine, "The Emerging concern of SARS-CoV-2 Variants of Concern", 2025 [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8224338/ [Accessed: 22-Aug-2025]

[6] 1. Khateeb, J., Li, Y., & Zhang, H. 2021. Emerging SARS-CoV-2 variants of concern and potential intervention approaches. Critical Care, 25, Article 244

[7] Chavda VP, Patel AB, Vaghasiya DD. 2022. SARS-CoV-2 variants and vulnerability at the global level. J Med Virol, 94(7):2986–3005

[8] Hirabara SM, Serdan TDA, Gorjao R, et al. 2022. SARSCOV-2 variants: differences and potential of immune evasion. Front Cell Infect Microbiol. 11:781429

[9] R.K. Mohapatra, R. Tiwari, A.K. Sarangi, M.R. Islam, C. Chakraborty, K. Dhama. 2022. Omicron (B. 1.1. 529) variant of SARS-CoV-2: concerns, challenges, and recent updates, J. Med. Virology

[10] Painuli, D., Mishra, D., Bhardwaj, S., & Aggarwal, M. 2021. Forecast and prediction of COVID-19 using machine learning. In *Data Science for COVID-19* (pp. 381-397). Academic Press.

[11] Swapnarekha, H., Behera, H. S., Nayak, J., & Naik, B. 2020.

Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review. *Chaos, Solitons & Fractals*, *138*, 109947.

[12] Rohini, M., Naveena, K. R., Jothipriya, G., Kameshwaran, S., & Jagadeeswari, M. 2021. A comparative approach to predict corona virus using machine learning. In *2021 international conference on artificial intelligence and smart systems (ICAIS)* (pp. 331-337). IEEE.

[13] Zhang, Q., Gao, J., Wu, J. T., Cao, Z., & Dajun Zeng, D. 2022. Data science approaches to confronting the COVID-19 pandemic: a narrative review. *Philosophical Transactions of the Royal Society A*, *380*(2214), 20210127.

[14] Shirke, A., Sanas, A., Patil, A., Wani, M., & Shelke, M. 2023. Analysis of COVID-19 Vaccination Data in India using Power BI. International Journal of Technology Engineering Arts Mathematics Science Vol. 3, No. 1

[15] Leung, C. K., Chen, Y., Hoi, C. S., Shang, S., Wen, Y., & Cuzzocrea, A. 2020. Big data visualization and visual analytics of COVID-19 data. In 2020 24th international conference information visualisation (iv) (pp. 415-420). IEEE.

[16] Leung, C. K., Kaufmann, T. N., Wen, Y., Zhao, C., & Zheng, H. 2021. Revealing COVID-19 data by data mining and visualization. In International Conference on Intelligent Networking and Collaborative Systems (pp. 70-83). Cham: Springer International Publishing.

[17] Rajeevan, S., Ramachandran, S., & Poulose, A. 2024. Tableau-driven data analysis and visualization of covid-19 cases in india. In 2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT) (pp. 1-6). IEEE.

[18] Halder, B., Bandyopadhyay, J., & Banik, P. 2023. COVID-19 pandemic: a health challenge for commoners during the first unlock phase in India. Journal of Public Health, 31(3), 427-433.

[19] Clement, F., Kaur, A., Sedghi, M., Krishnaswamy, D., & Punithakumar, K. 2020. Interactive data driven visualization for COVID-19 with trends, analytics and forecasting. In 2020 24th International Conference Information Visualisation (IV) (pp. 593-598). IEEE.

[20] Veena Jokhakar, Kamlendu Pandey, Ronak Panchal; "An Investigation on Coronavirus (COVID-19) Pandemic for India using Power BI", International Journal of Creative Research Though (IJCRT), ISSN: 2320-2882, Vol-8, Issue 4, April 2020.

[21] Akhtar, N., Tabassum, N., Perwej, A., & Perwej, Y. (2020). Data analytics and visualization using Tableau utilitarian for COVID-19 (Coronavirus). *Global Journal of Engineering and Technology Advances*.

[22] Das, A. 2025. Predicting COVID-19 Cases in India Using ARIMA, Prophet, LSTM and Data Analysis Using Power BI. In: Namasudra, S., Kar, N., Patra, S.K., Taniar, D. (eds) Data Science and Network Engineering. ICDSNE 2024. Lecture Notes in Networks and Systems, vol 1165. Springer.

[23] Dataset - "Covid-19 Variants Worldwide Evolution", 2024. [Online]. Available: https://data.who.int/dashboards/covid19/cases. [Accessed: 22-Jul-2024].