GR-391

Insight into current COVID-19 variants and data science applications

Department of Computer Science & Department of Information Technology

ABSTRACT

COVID-19 has undergone several mutations resulting in the emergence of new variants such as Alpha, Beta, Gamma, Delta, and Omicron. Data science plays a vital role in understanding the spread of COVID-19 and its variants. The emergence of new COVID-19 variants has raised concerns about the effectiveness of existing vaccines and treatments. This poster provides insights into the current COVID-19 variants and the data science applications used to monitor and understand their spread. We will showcase the use of SQL to manage and analyze genomic data of SARS-CoV-2 variants, Power BI for visualization and tracking of COVID-19 cases and deaths, and machine learning algorithms for variant classification and prediction of transmission. The poster highlights the importance of integrating data science tools in COVID-19 research for effective variant surveillance, risk assessment, and mitigation strategies. The data used in this project was obtained from the GISAID website.

INTRODUCTION

COVID-19 is a respiratory illness caused by the SARS-CoV-2 virus that was first identified in Wuhan, China, in December 2019, and has since spread globally, leading to a pandemic. SARS-CoV-2 belongs to the CoV's latest evolutionary branch and primarily spreads through respiratory droplets and microscopic particles expelled during sneezing, coughing, or talking by an infected person. The symptoms of COVID-19 can range from mild to severe and include fever, cough, fatigue, loss of taste or smell, and difficulty breathing. The virus spread globally, causing damage to various organ systems in both humans and other vertebrate animals. The COVID-19 pandemic has had adverse effects on the social, mental, and physical health, economies, and global disparities. The virus can cause mild to severe illness, and people above the age of 60 or those with underlying medical issues are more vulnerable. The virus undergoes genetic mutations during genome replication, leading to the emergence of various genetically distinct virus variants. A lineage or set of lineages may be categorized as a variant of concern (VOC), variant of interest (VOI), or variant being monitored (VBM) based on shared traits, features that may call for public health intervention. The alpha variant (B.1.1.7) of the SARS-CoV-2 virus was first identified in the UK in September 2020,Beta variant (B.1.351) in South Africa in May 2020,Gamma variant (P.1)in Brazil in November 2020, Delta variant (B.1.617.2) in India in December 2020 and the omicron variant (B.1.1.529) on South Africa in November 2021 and this variants are also classified as a variant of concern.

	VOC Omicron GRA (B.1.1.529+BA.*)	2-11	2022-12-18	2022-12-25	2023-01-01	2023-01-08	2023-01-15	2023-01-22	2023-01-29	2023-02-05	2023-02-12	2023-02-19	2023-02-26	2023-03-05	2023-03-12	2023-03-19
Aruba	count	11	20	31	28	37	24	25	9	8						
	total	11	20	31	28	37	24	25	9	8						
Australia	count	1,034	1,736	1,117	1,037	1,305	1,244	1,016	743	866	793	838	793	673	396	7
	total	1,038	1,737	1,118	1,039	1,307	1,246	1,018	743	870	794	841	812	702	418	8
Austria	count	1,712	2,934	2,689	2,491	2,328	2,306	2,067	2,262	2,551	2,650	2,955	2,866	2,357	1,555	13
	total	1,712	2,934	2,689	2,491	2,328	2,306	2,068	2,264	2,551	2,650	2,955	2,867	2,357	1,555	13
Azerbaijan	count															
	total															
Bahrain	count	6	76	17	44	48	83	86	101	111	48					
	total	6	76	17	44	48	86	89	101	111	50					
Bangladesh	count				5	2	1									
	total				5	2	1									
Barbados	count	2	1		1	4	12	7	3	3						
	total	2	1		1	4	12	7	3	3						
Belarus	count															
	total															
Belgium	count	452	465	377	249	336	240	206	169	250	360	290	295	366	184	27
	total	470	466	378	251	336	240	206	170	250	360	291	298	368	187	27
Belize	count															

Figure 1. Raw Data

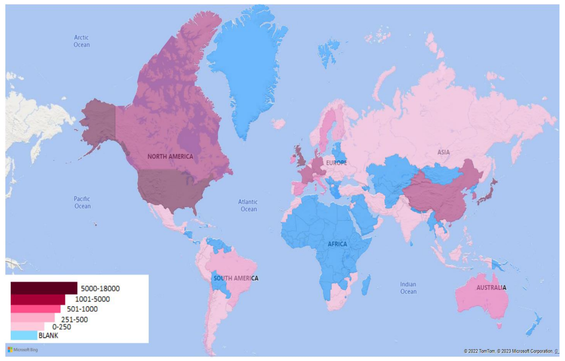

Figure 3. Map distribution of Omicron Variant in 2023

Results | Messages

	Id	Country	Month	Total count
1	1	Algeria	January	2
2	2	Algeria	January	3
3	3	Algeria	January	5
4	4	Algeria	January	7
5	5	Algeria	February	1
6	6	Algeria	February	2
7	7	American Samoa	January	4
8	8	Andorra	January	3
9	9	Anguilla	January	1
10	10	Antigua and Barbuda	January	1

Figure 2. Prepared SQL Data

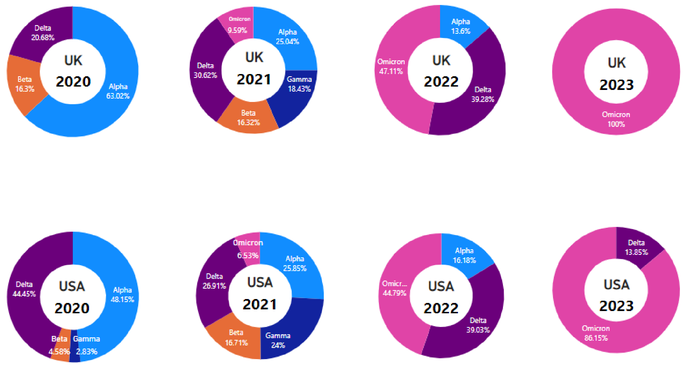

Figure 4. Illustration of Dominant Variant of UK and USA with respect to year 2020 -2023

The WHO's Director-General, in the 73rd World Health Assembly held on 19 May 2020, expressed concern about the pandemic's morbidity, mortality, and advised governments to adopt and execute COVID-19 pandemic responses unique to their country situation, and mobilize necessary resources. The pandemic has had 3,732,046 confirmed cases and 261,517 fatalities as of May 6, 2020. The SARS-CoV-2 virus omicron BA.5 subvariant dominated from July to November 2022, significantly outperforming earlier variations in terms of neutralization escape.

METHODOLOGY

New variants of the COVID-19 virus have emerged, some of which are more contagious and potentially more deadly than the original strain. Analyzing the genetic changes of these variants is crucial in informing public health responses and vaccine development. SQL can be used to query databases of sequenced viral genomes to identify specific mutations associated with each variant and track the spread of the virus over time. Power BI can be used to visualize and analyze data on the spread of the virus and the effectiveness of public health interventions, integrating data from multiple sources to gain insights into how the virus is spreading and how well public health interventions are working. Machine learning techniques can be applied to COVID-19 data to predict future trends and identify factors that contribute to the spread of the virus. For example, by analyzing data on the number of COVID-19 cases and deaths over time, machine learning algorithms can predict future trends and identify patterns in the spread of the virus, helping public health officials prepare for future surges. These data science applications are crucial in informing public health responses and developing effective interventions to control the pandemic.

CONCLUSION

As the COVID-19 pandemic continues, predicting future cases is difficult due to various factors affecting the virus spread. To address this, research on the mechanism of SARS-CoV-2 Spike protein binding with ACE2 using Machine Learning Techniques like Biotite, PyTorch, and Torch Drug can aid in understanding the binding behavior. GCN with mutual attention can be used to predict protein-protein interactions, leading to new treatments and drugs. Therefore, investing in further research on protein-protein interactions using Machine Learning Techniques is crucial for managing COVID-19 and other viral diseases. Data science also plays a critical role in tracking COVID-19 variants, guiding public health interventions and vaccine distribution.

KENNESAW STATE UNIVERSITY

Author(s): Ravi Potlapalli, Thanusha Sai Ande

Advisor: Dr. Chloe Yixin Xie