Received: 31 January 2023 | Revised: 6 March 2023 | Accepted: 21 March 2023 | Published online: 24 March 2023

RESEARCH ARTICLE

Journal of Data Science and Intelligent Systems
2023, Vol. 1(1) 55–64
DOI: 10.47852/bonviewJDSIS3202701

BON VIEW PUBLISHING

Topological Data Analysis of COVID-19 Using Artificial Intelligence and Machine Learning Techniques in Big Datasets of Hausdorff Spaces

Allan Onyango[1], Benard Okelo[1,*] and Richard Omollo[2]

[1]*Department of Pure and Applied Mathematics, Jaramogi Oginga Odinga University of Science and Technology, Kenya*

[2]*Department of Computer Science and Software Engineering, Jaramogi Oginga Odinga University of Science and Technology, Kenya*

Abstract: In this paper, we carry out an in-depth topological data analysis of COVID-19 pandemic using artificial intelligence and machine learning techniques. We show the distribution patterns of the pandemic all over the world, when it was at its peak, with respect to big datasets in Hausdorff spaces. The results show that the world areas, which experience a lot of cold seasons, were affected most.

Keywords: artificial intelligence, machine learning, topological data analysis, COVID-19, python

1. Introduction

Artificial intelligence (AI) refers to an extensive branch of computer science involved in the theory and development of smart computer systems, having the capability to perform tasks that usually require human intelligence, for example, natural language translation, voice recognition, and visual perception. Machine learning (ML) refers to a discipline of AI that utilizes computer algorithms to learn, experience, adapt, and automatically improve without human programming. One such techniques of ML used to analyze and feature engineer the big datasets (BDSs) is the t-SNE-Stochastic Neighbor Embedding (SNE) unsupervised ML algorithm. Topological data analysis (TDA) is a very important type of big data (BD) analysis which is very useful in AI and ML techniques. It is very important when classifying data in regions with unique characteristics, which can be represented in the sets of topological spaces. TDA is founded under the ubiquitous theory of persistent homology. Some of the pioneer contributors to TDA include Frosini (1992), Robins (1999), Morris (2012), and Edelsbrunner et al. (2002), who founded the notion of how features persist as the data are modified. Nevertheless, the genesis of the term TDA expression appears not to have surfaced till contributions by De Silva & Carlson (2004) and Bremer (2004). Thereafter, Carlsson (2014) became instrumental in the popularization of TDA, establishing the ways topological techniques will remedy challenges encountered while implementing topology to analyze BD. Perea (2019) put up other developments by observing that persistent homology is currently one of the more widely known tools from computational topology and TDA.

Topology and geometry are tools used to investigate highly composite data (Carriere & Oudot, 2018) by creating a compendium of the features of data to uncover hidden attributes within the dataset. Normally, the dataset of interest is often centered around structures that appear challenging to be revealed with traditional methods (Butler, 2016). The major TDA approach for removing "topological noise" is to map the original data to a lower dimensional approximation, acquired through a multi-dimensional assortment (Marr, 2015).

Open sets therefore provide an essential approach to understand nearness of points without a distance element defined in a topological space (Caleb et al., 2019). Other inherent mathematical concepts to understand besides topology include continuity, connectedness, and closeness, which embrace nearness. The problem is that there is not a single story happening in these data (Carriere & Oudot, 2018). We can therefore say these data have much "noise"! The explosive growth in data, voice and video traffic, and ubiquity of social-media content, health records, and many more data sources have been a contributing factor of BD (Dayten, 2020). It was anticipated that the generated data volume could be 44 zettabytes in 2020 as found in Butler (2016).

The vast complex nature of data has propelled technological advancements realized, as well as accelerated increase in bandwidth capacity, processing power, storage capability, and transfer velocity (De Silva & Carlsson, 2004). This is partially due to the technological advancement in high power computing (Colleen, 2021).

*Corresponding Author: Benard Okelo, Department of Pure and Applied Mathematics, Jaramogi Oginga Odinga University of Science and Technology, Kenya. Email: bnyaare@yahoo.com

There is therefore urgent need to establish robust and resilient techniques, to process the BD (Elizabeth, 2017). BD consists of 5 Vs: Value, Variety, Volume, Veracity, and Velocity (Frederic & Bertrand, 2017). The data size to be processed and analyzed constitutes the volume (Vclav et al., 2017). The speed of growth and usage of these data is the velocity. The varied data formats cum types are the variety. Veracity involves accuracy plus analysis of the results of the datasets (Singh et al., 2007). The richness obtained after the processing the dataset is the value. The growing volumes of Voice over IP (VoIP) data traffic, social-media content (IDC, 2012), underscores the requirement of ways of countering the ambiguity innate the finite datasets. Presently, roughly 80% of datasets remains indeterminate. TDA has lately recorded advances in innumerable directions and application disciplines (Tierny, 2006). The fundamental aim of TDA is to extract multi-dimensional rich data features, based on geometry and topology pre-existing in distributed data points as shown in lexico.com (2021) and Oudot (2015). Connections within the data and topological methods have a close affiliation to neural networks between data points,, which reveal insight into this united structure. According to Frederic & Bertrand (2017), most commonly, every other form of TDA revolves around the steps below: firstly, the data sample is presumed as data points which are finite quantified as a metric space $\mathbf{R}^d$. Worth to mention is that the metric choice may be vital to guarantee remarkable topological and geometric features of the dataset. Secondly, to tap more from the fundamental concepts of geometry and topology, a mathematical structure is computed on top of the dataset King (2021), Costa (2017), Tierny (2006), Adler et al. (2010), Morris (2012), and the references therein. This is mostly cases in simplicial complex (SC) or a convention of SCs, which depicts the high dimensional data structures at varied degrees (Vclav et al., 2017).

Thirdly, from these high dimensional data structures built on atop the dataset, topological or geometric information is derived (Data Educations in Schools, 2021). The shape of the data from which we extract the topological/geometrical high dimensional features can either be crude structural summaries or relevant approximations, which need further approaches like persistent homology and visualization (Carlsson, 2009). The resulting topological and geometric information give rise to insightful features and descriptors, into the data, which when injected into further analysis and ML procedures, reveal very rich results and significant meaning, that can be used in other disciplines such as medicine, biology and astrophysics, just to mention a few (Dayten, 2020).

2. Literature Review

Describing topological points is very intricate due to the nature of BD. This makes it difficult to locate BDS particularly in a general topological space setting. Because of the $T_2 - axiom$ in T_2-space, it is even more difficult to locate these BDS in Hausdorff spaces. In spite of the remarkable efforts put up by traditional techniques in data analysis, they have not always kept up with the exploding data quantity and complexity, since they often depend on exceedingly simplistic assumptions and approximations during computations. Besides, they do not pay attention to the arbitrariness within these TDSs as well as the underlying instability within the topological datasets. Consequently, most of these techniques are exploratory, lacking the efficiency to distinguish what is sometimes called the "topological noise," from information of interest. The Vietoris–Rips complex for a parameter t has been so ubiquitously used to build a useful SC, to mirror the data structure, and utilizes the original data as the vertex set. The bone of contention, however, has always been how to choose the t parameter, such that the Rips complex reveals the structure of the underlying dataset. It is precisely this question that appeals to our conscience of thought toward persistence diagram as a topological signature of the dataset. Two metrics have commonly been used to measure the similarity of those objects: the bottleneck and Wasserstein distances. Each works by matching points of 1 diagram with points of another diagram, while allowing the match to be done with the diagonal if necessary. The Mapper algorithm does have some limitations however. The topology of shape graphs is to a large extent dependent on whether the filter function chosen is linear or nonlinear.

3. Research Methodology

The t-SNE algorithm refers to a feature engineering algorithm to unsupervised ML algorithms (Applied AI) like K-Means. t-SNE can be described as a nonlinear DR algorithm instrumental in highly dimensional data exploration. Dimensionality reduction refers to a linear or nonlinear technique involved in mapping higher dimensional to a low-dimensional space while preserving local features within the primary space. Reducing the dimensions to a lower dimension achieves the preservation of two things: if data points are close by in the high dimensional space, it tries to retain that closeness in small dimension space. If points are far apart, it also tries to keep them a part in a smaller dimensional space. So, it preserves closeness and farness in the space. And it does that by applying attractive forces between points that are close and repulsive forces to points that are far apart. These forces are applied repeatedly to all the points in the space for a number of iterations between points closer and those far apart. t-SNE is therefore important in showing clusters of data that are similar and close. We choose t-SNE because linear dimensionality reduction algorithms like PCA emphasize on positioning contrasting TDPs distantly separated in a lower dimensional space. However, achieving representation of highly dimensional TDSs on a low dimension, indistinguishable data points must be mapped closer together and this can only be achieved by nonlinear algorithms unlike their linear counterparts. t-SNE is an enhancement over the (SNE) algorithm (Vclav et al., 2017). To implement the t-SNE ML algorithm, the python programing language has proved very instrumental in ML, given its vast libraries and frameworks for advanced computing and visualization. The t-SNE ML algorithm will be very applicable in establishing the distribution patterns of the topological data points within a Hausdorff space. The next section describes the python libraries, distributions, and the computing requirements applicable in ML for our study. Python is a popular programming language that is widely used in AI communities. Python possesses simplicity and readability features for its syntax, and this reduces the time taken to test complicated algorithms through minimal code in comparison to the existing languages (Vclav et al., 2017). Python prides itself with a great numbers of rich library modules for ML. Furthermore, python commands overwhelming community of developers globally, who generously share troubleshooting and debugging tips through online platforms. The ubiquitous nature of python has granted its usage in nearly all research institutions as well as commercial applications of deep learning and ML. Jupyter Notebook is a computing notebook environment operated on an interactive web browser. It is a component of the Anaconda Navigator. Some of the python libraries include pandas, sklearn, matplotlib, seaborn, plotly, cu□links, TSNE,

and K-Means. Some of these libraries are resource intensive especially the ones that implement ML algorithm to run through over thousands of rows of data for hundreds of iterations, in just a few minutes. Therefore, the laptop specifications to be used to run these python algorithms must possess a powerful processor of Core i7, 16GB of RAM memory, over 2.3 GHz speed, and a faster 500GB solid-state drive storage. The next section presents the BDSs that will be used in our study. The datasets used in this study are downloaded from Kaggle.com, an open source community of data scientists, ML, and a huge published repository of BD sets. We specifically focus on a collection of time series COVID-19 datasets of cases reported from all countries of all the six world continents (Africa, Europe, Asia, North America, South America, and Oceania) starting February 24th, 2020 until June 28th, 2021. The dataset contains 98,904 rows and 60 columns. The huge volume of this dataset qualifies it as a candidate of a BDS. We denote the COVID-19 dataset as H throughout the study. Besides, python also has the ability to randomly generate larger datasets in a simulated environment for analysis purposes. The next section takes us through the initial cleaning of the raw data and exploratory data analysis (EDA), in order to get maximum output from our dataset. The raw data have to go through several processes before the final analysis. The data are first imported into the python using the Pandas library. We then expose the data through the process of shaping to determine the initial high dimension, and we check the data distribution, wrangling, and initial visualization processes. Subsequently, EDA follows, where we perform structural modification, further analysis, and visualization on the data using line graphs and surface plots using various python libraries. After EDA, we perform a very crucial stage on the data known as feature engineering. This involves feature selection, dimensionality reduction using techniques like backfilling, forward-filling, data encoding, best column formation, variance, and means computations. These processes achieve data cleaning which assists us to manage handling of outliers and excessive null values, which TDA refers to as "topological noise." At this point, the data can finally be exposed to the ML algorithms, either as unsupervised, supervised, semi-supervised, or reinforcement ML algorithms. During our study, we will however restrict ourselves to t-SNE which is a feature engineering algorithm to unsupervised ML algorithm as a branch of applied artificial intelligence.

4. Results and Discussion

The datasets used in this study are downloaded from Kaggle.com, an open source community of data scientists, ML, and a huge published repository of BD sets. In our coding, the world is represented by a Hausdorff space $\mathbf{X}$ and subspaces of $\mathbf{H}$ represent regions or countries, unless otherwise stated. These subspaces represent BDSs. We specifically focus on a collection of time series COVID-19 datasets of cases reported from all countries of all the six world continents (Africa, Europe, Asia, North America, South America, and Oceania) starting February 24th, 2020 until June 28th, 2021. The dataset contained 98,904 rows and 60 columns. The huge volume of this dataset qualifies it as a candidate of a BDS. The next section describes the connectedness of our dataset. Given $\mathbf{H}_1$, $\mathbf{H}_2 \subseteq \mathbf{H}$, such that $\mathbf{H}_1 \cap \mathbf{H}_2 = \varnothing$, then collections generated by $\mathbf{H}_1$ and $\mathbf{H}_2$ might be equated to nonempty intersections, and these Hausdorff properties come in handy in building a SC. As a result, this result leads to a "multiresolution" map of the TDS. Connectedness of points in a topological space means nonexistence of separation between the points on the topological space. Let $\mathbf{X}$ be our COVID-19 dataset throughout this study. We applying a nonlinear fit on $\mathbf{X}$ by reducing the dimensionality while the geometric structure, shape, and connectivity of our dataset remain preserved.

4.1. 2D line graph visualizations

After EDA and dimensionality reduction on our COVID-19 dataset, denoted by $\mathbf{H}$, we produce a few line graph visualizations to initially reveal the relationship between the total confirmed, recovered, and deaths cases across the six world continents. A visualization of the total confirmed COVID-19 cases in the world is revealed at a glance through the line graph in Figure 1.

Figure 1
Line plot visualization of dataset H revealing the total confirmed COVID-19 infection cases globally

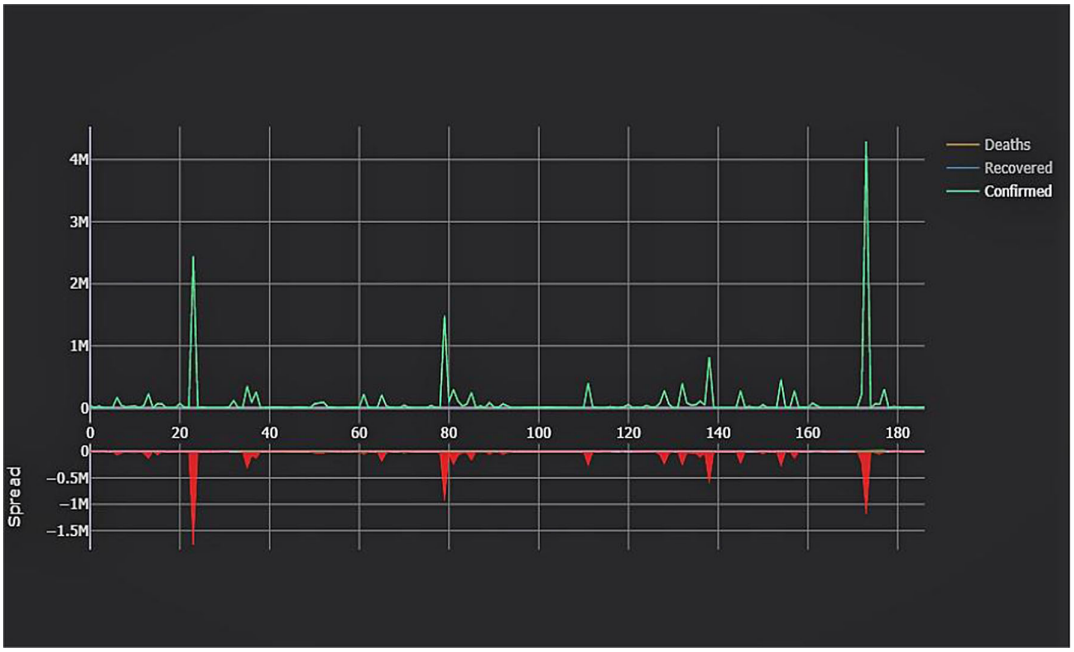


Figure 2 gives a line graph visualization of the total recovered cases in the world. In addition, after performing EDA and dimensionality reduction on our COVID-19 dataset denoted by **H**, we compute surface plot visualizations to reveal topological structure and graphical properties from our dataset **H**. We use python libraries to produce surface plot visualizations which revealed the structured shape between the total confirmed, total recovered and total deaths cases across the world.

The total number of deaths is visualized by the line graph in Figure 3.

Finally, Figure 4 reveals a combined line graph visualization of both the total confirmed, recovered, and death cases of COVID-19 in the world.

Figure 5 displays elevated numbers of confirmed cases globally, slightly lower deaths per hundred cases, and relatively fewer new cases during the period when the data were collected. These

Figure 2
Line plot visualization of dataset H revealing COVID-19 data globally

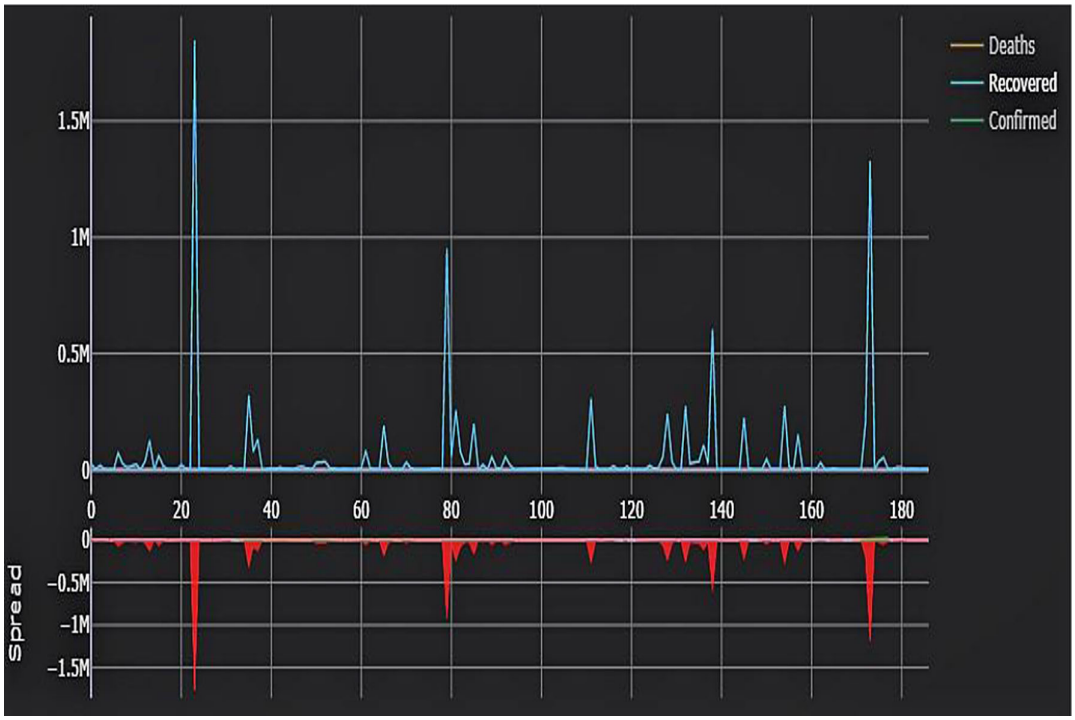


Figure 3
Line plot visualization of dataset H revealing the total confirmed death cases globally

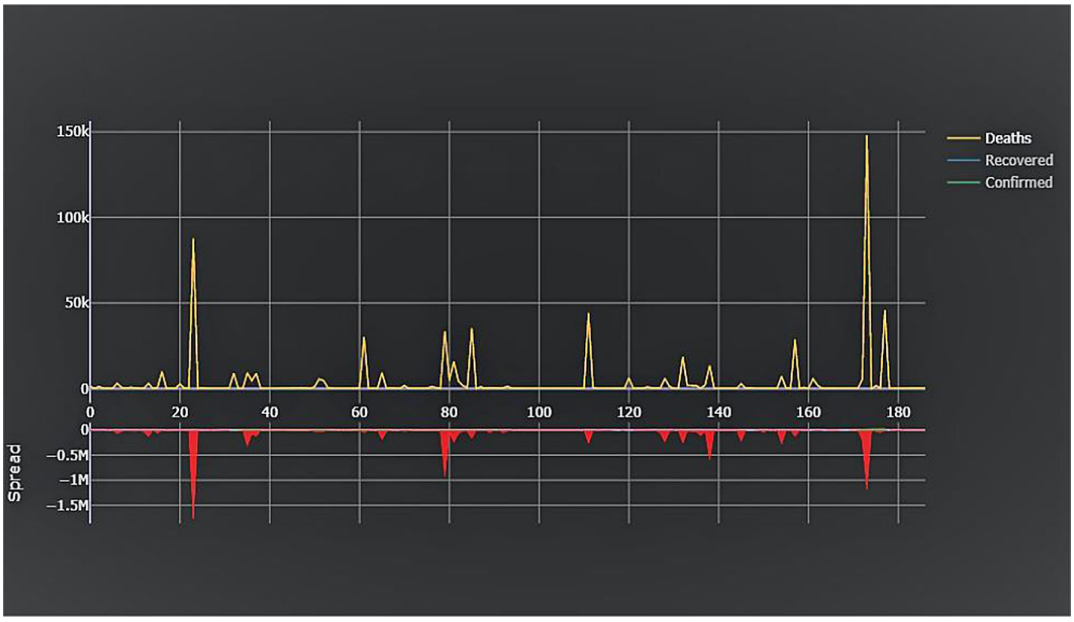
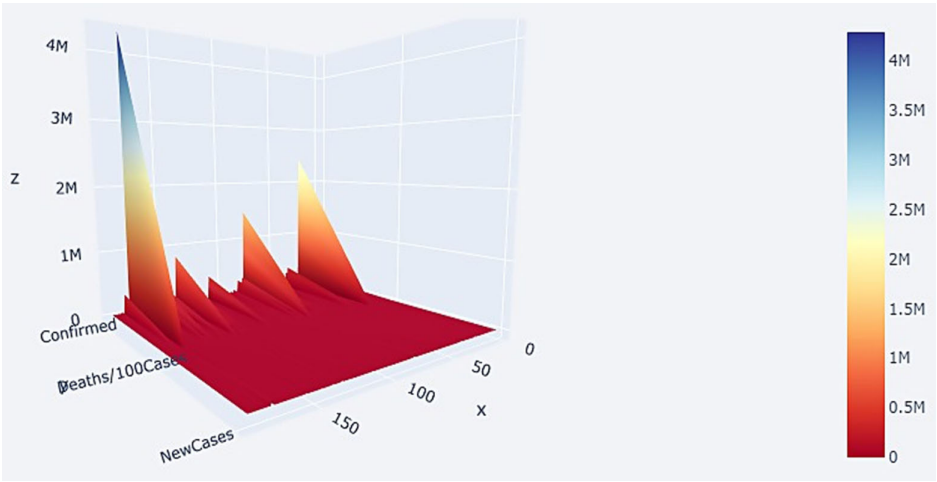


Figure 4
A combined line plot visualization of dataset H revealing the total confirmed, recovered, and death cases globally

Figure 5
A 3D surface plot visualization of the COVID-19 dataset displaying higher numbers of confirmed cases globally, slightly lower deaths per hundred cases, and relatively fewer new cases globally



clusters reveal valuable information about the underlying features such that when some initial conditions of the experiment are fulfilled, the data points are more likely to assemble in a certain manner. Thereafter, notable clusters may begin to emerge, varying from those achieved with smaller radii, basically encompassing them. As the radius of the balls increases, a single component of **H** is finally observed for the very first time. The component can among other things be a loop, a chain, have holes or loops, have multiple flares, or another structure. So as the radius r increases toward r^*, the balls around the points a_1 and a_2 intersect and the final edge emerges to complete the loop.

Figure 6 visualizes a surface plot of the COVID-19 dataset displaying higher numbers of new cases globally, slightly moderate new death cases, and relatively fewer death cases globally, during the period of data collection.

Figure 7 displays a surface plot visualization of the COVID-19 dataset displaying extremely higher numbers of new deaths globally, moderately higher new recovery cases, and relatively lower deaths per hundred cases glob ally. Each TDP is enclosed around a ball at the center at a point with a radius ϵ. While ϵ increases in size, the cloud gradually ceases to appear as secluded points, but slowly gains shape. As the output increases in size, a single shapeless blob emerges. These events eventually generate a SC, which starts with vertices as TDPs. An edge is inserted between the two balls as they intersect. Meanwhile, a face bounded by the three edges is added as 3 balls intersect. With the onset of these

Figure 6
A 3D surface plot visualization of the COVID-19 dataset displaying higher numbers of new cases globally, slightly moderate new death cases, and relatively fewer death cases globally

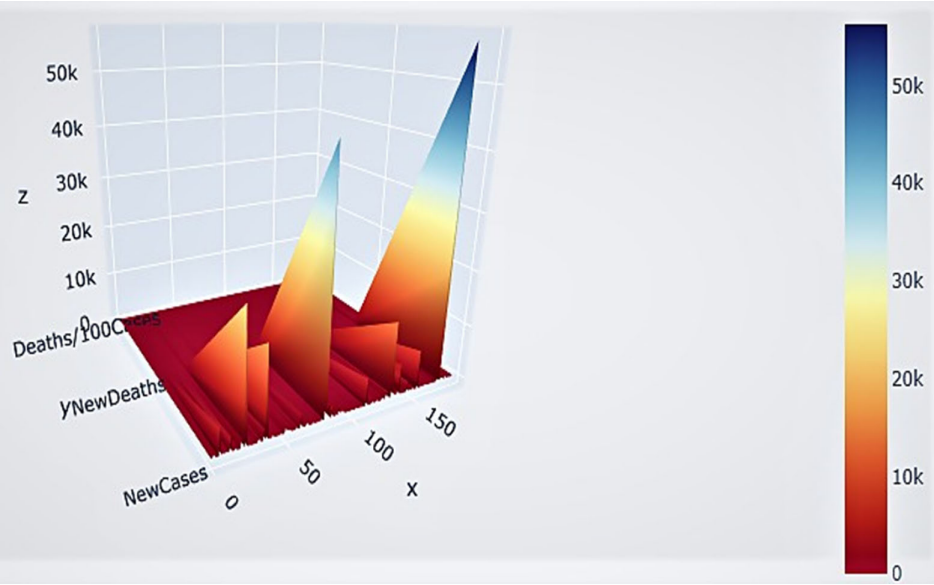
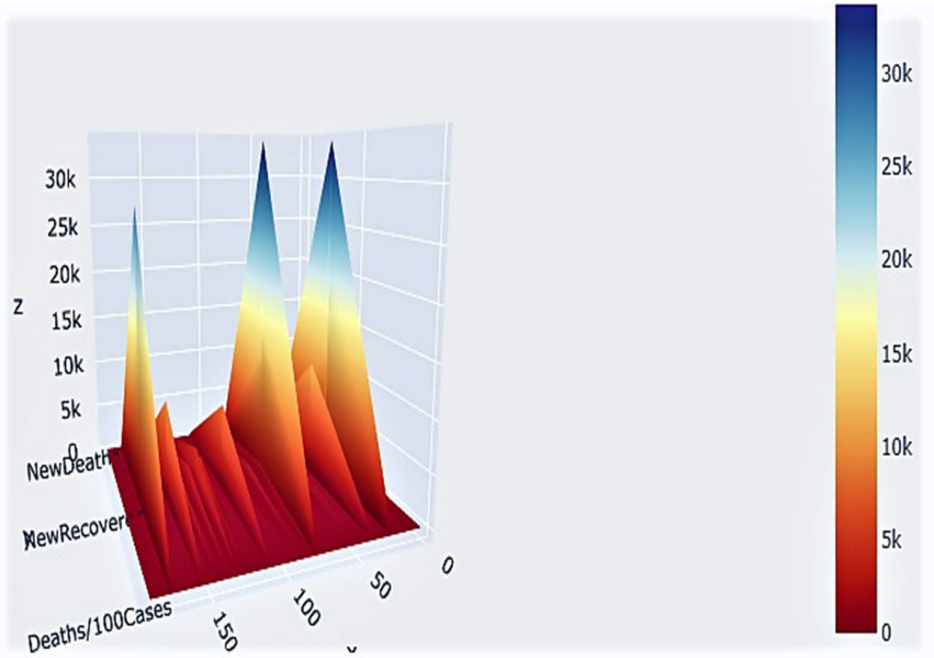


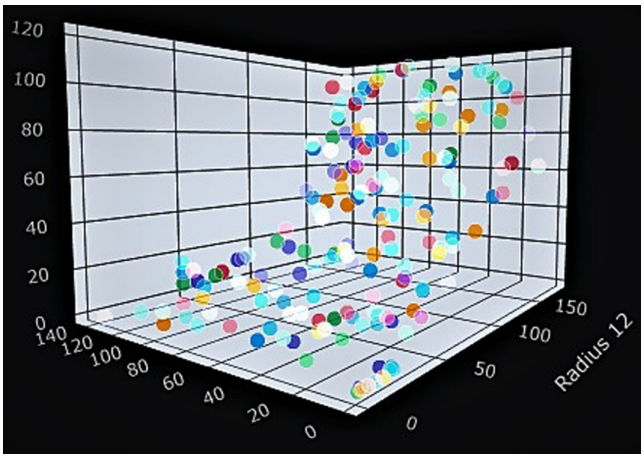
Figure 7
A 3D surface plot visualization of the COVID-19 dataset displaying extremely higher numbers of new deaths globally, moderately higher new recovery cases, and relatively lower deaths per hundred cases globally



events, highly dimensional n-faces with n + 1 intersecting emerges. Taking **H** and draw a ball $\mathbf{H}(a_1, r)$ around each point a_1 for some small initial radius r (Figure 8). As r increases at intervals, we can establish how **H** connects at different radii by observing different angle of snapshots of the 3-D outputs.

We obtain a three-dimensional scatter plot nonlinear random shape of the data at an initial default radius of 12 as shown in Figure 8. At this point, we will keenly observe the initial intersection of the closures of $\mathbf{H}(a_1, r)$ and $\mathbf{H}(a_2, r)$ on a point within the plane. Without loss of generality, we can confidently state that a_1 and a_2 are $2r$ distanced apart. Connecting them with an edge, we get a visual graph. When eventually this radius is attained, further increment of the radius reveals no further structure for a significant time period. We therefore conclude that the ranges of the radius are $r > r^*$ as shown.

Figure 8
3D scatter plot of the sample H of data points at initial default radius of 12



Before we apply the t-SNE algorithm, we first compute a quick visualization of three categories within the global COVID-19 dataset. These groups include the cardiovascular death rate, diabetes prevalence, and the pop ulation aged 70 years and older.

The 3D surface plot on Figure 9 visualizes a common relationship between the three groups.

From the dataset, we also compute a 3D visualization to reveal the shape of three more categories, that is, total confirmed, total recovered, and the total deaths among the six continents as shown in Figure 10.

After the COVID-19 dataset is subjected to the t-SNE algorithm, a feature engineering algorithm to unsupervised ML algorithms, we obtain the following three-cluster distributions: at the global, continental (Africa), and country (Kenya) levels. Within the clustered distributions, a shape is revealed outlining the distribution pattern of the data points within the COVID-19 datasets. Also included, are outliers (topological "noise" (TN)). The outliers include extremely high or extremely low figures within the datasets. Also noted within the distribution spaces are

Figure 9
3D surface plot visualization revealing how cardiovascular death rate, diabetes prevalence, and the population aged 70 years and older are related

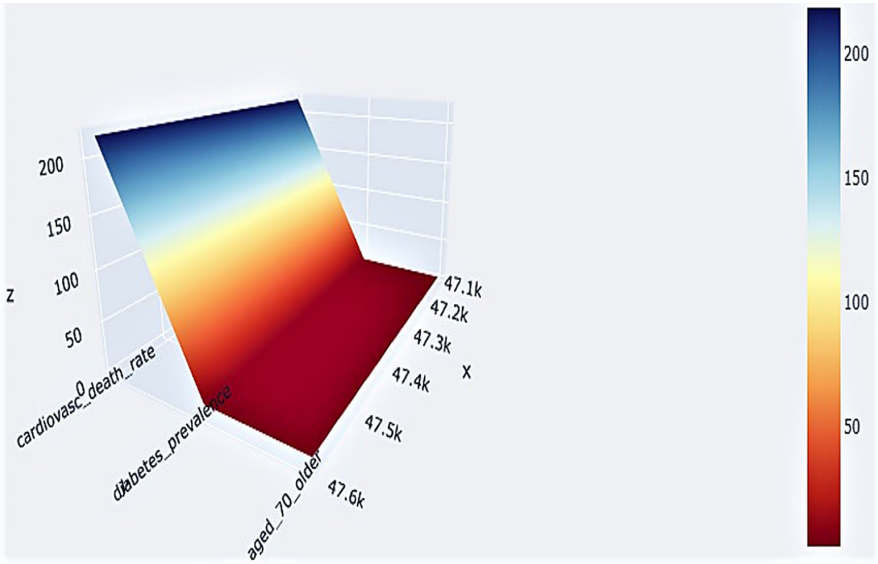


Figure 10
3D surface plot visualization revealing how the means of total confirmed, total recovered, and the total deaths among the six continents are related

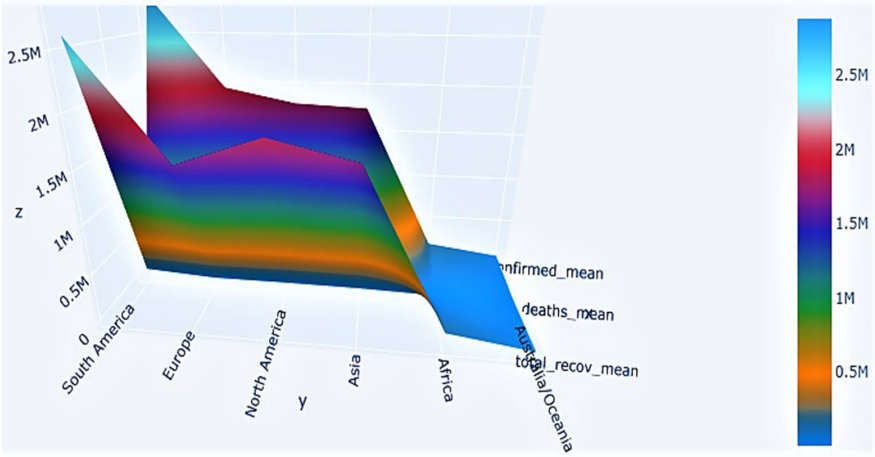
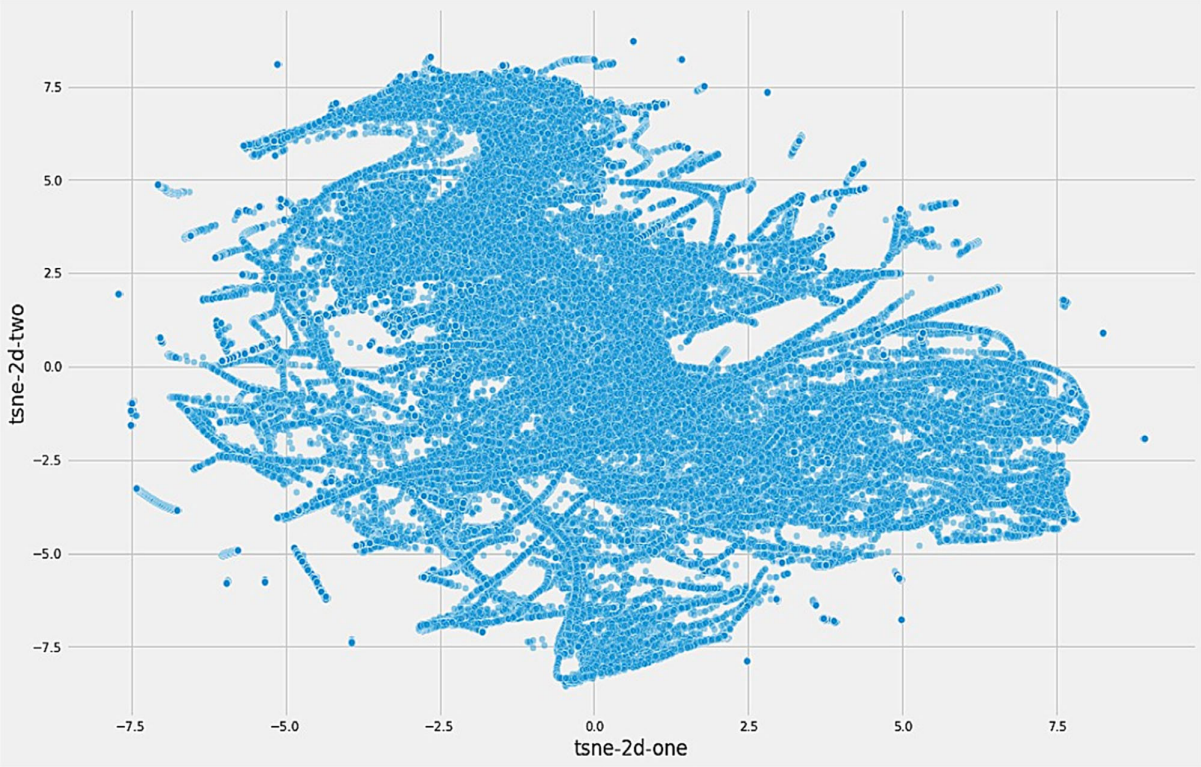


Figure 11
t-SNE generated distribution pattern of COVID-19 data points within statistics of the world continents



condensed points and sparsely distributed data points. The global distribution pattern is revealed in Figure 11.

You can find our code at the links: https://colab.research.google.com/drive/1uLNJm6NDtfPu3td8gTB1-nANTRM5JdLZ?usp=sharing However, to run the code to view the visualizations, you might need to install Anaconda Navigator, a python distribution library, for ML, install the relevant libraries, and then download the CSV datasets from the kaggle.com repository.

Figure 12 reveals the distribution pattern of the data points within Africa as a continent.

Figure 12
t-SNE generated distribution pattern of COVID-19 data points within statistics of the African continent

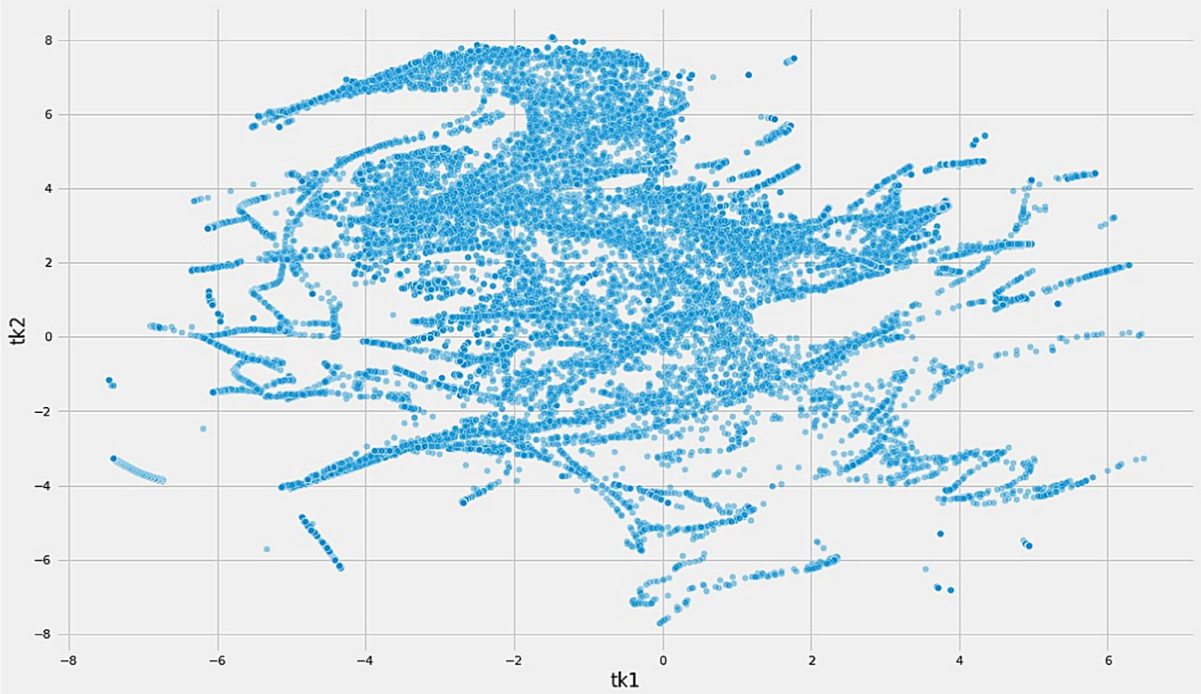
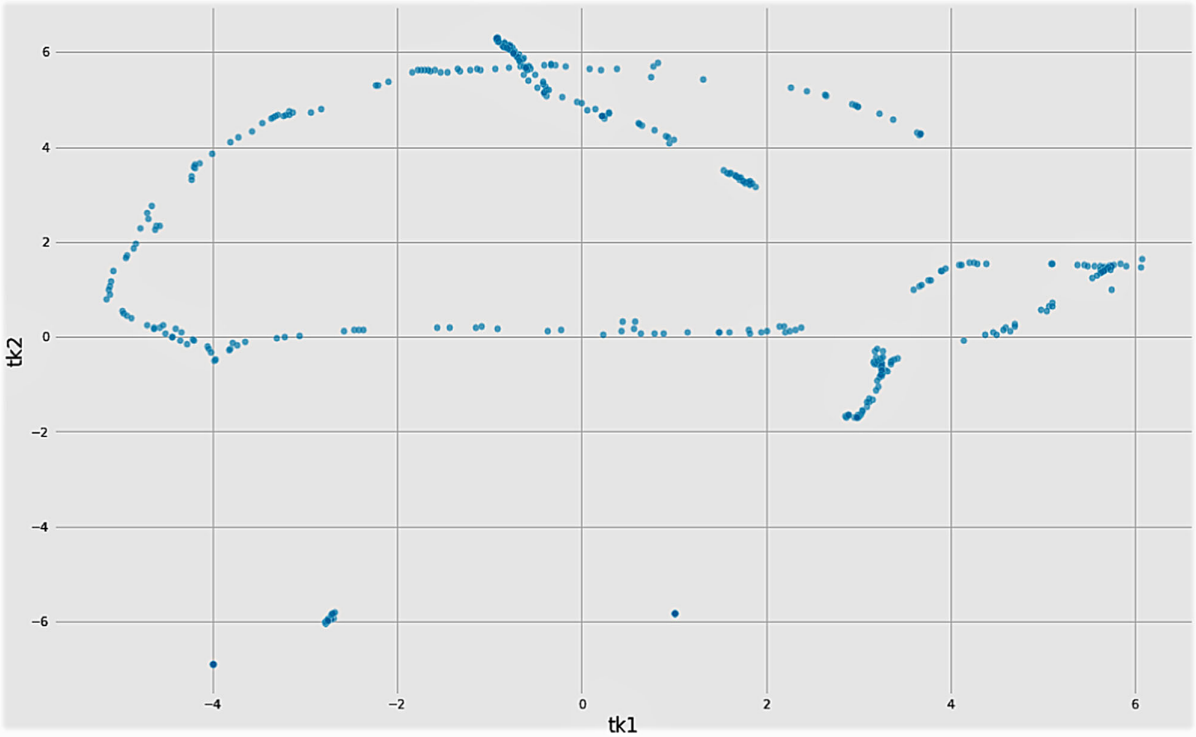


Figure 13
t-SNE generated sparse distribution pattern of COVID-19 data points of the country, Kenya, in African continent



Finally, Figure 13 reveals the distribution pattern of the data points of the Kenyan portion of the global portion of the COVID-19 dataset.

5. Conclusion

In this paper, we have carried out an in-depth TDA of COVID-19 pandemic using AI and ML techniques. We have shown the distribution patterns of pandemic all over the world when it was at its peak. The results show that the world areas which experience a lot of cold seasons were affected most. We have managed to demonstrate the distribution patterns of topological data points within a Hausdorff space, using 3D visualizations and application of t-SNE ML algorithm clusters of the dataset from all the six World continents, African continent, and the country Kenya. From the distribution of the real life COVID-19 dataset, the coronavirus situation was densely distributed in Winter-prone regions like Europe, United States of America, and Canada.

Recommendations

From this study, we recommend that characterization of topological data points can be considered in normal spaces. We further recommend that location of BDSs can be carried out in normal spaces. Finally, we recommend that establishment of distribution patterns of topological data points can be done in normal spaces using data from other fields like health, business, and social media.

Acknowledgement

The authors are grateful to the reviewers for their useful comments which helped to improve this work.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

Adler, R. J., Bobrowski, O., Borman, M. S., Subag, E., & Weinberger, S. (2010). Persistent homology for random fields and complexes. In *Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown, 6,* 124–144. Institute of Mathematical Statistics.

Bremer, P. (2004). A topological hierarchy for functions on triangulated surfaces. *IEEE Transactions on Visualization and Computer Graphics*, *10*, 385–396.

Butler, D. (2016). A world where everyone has a robot: Why 2040 could blow your mind. *Nature*, *530*, 75–91.

Caleb, G., Olaf, S., Giovanni, P., & Manish, S., (2019). Generating dynamical neuroimaging spatiotemporal (DyNeuSR) using topological data analysis. *Network Neuroscience, 3*(3), 763–778.

Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, *46*, 255–308.

Carlsson, G. (2014). Topological pattern recognition for point cloud data. *Acta Numerica*, *23*, 289–368.

Carriere, M., & Oudot, S. (2018). Structure and stability of the one-dimensional mapper. *Foundations of Computational Mathematics, 18*, 1333–1396.

Costa, J. P. (2017). Topological data analysis and applications. In *40th International Convention on Information and Communication Technology, Electronics and Micro-electronics*, 558–563.

Data Education in Schools. (2021). What is data? Retrieved from: https://dataschools.education/about-data-literacy/what-is-data/

Dayten, S. (2020). Introductory topological data analysis. *Department of Mathematics and Statistics*, *University of Victoria*.

De Silva, V., & Carlsson, G. E. (2004). Topological estimation using witness complexes. In Eurographics Symposium on Point-Based Graphics, 157–166

Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2002). Topological persistence and simplification. *Discrete & Computational Geometry*, *28*, 511–533.

Elizabeth, M. (2017). A users guide to topological data analysis. *Journal of Learning Analytics*, *4*, 47–61. https://doi.org/10.18608/jla.2017.42.6

Frederic, C., & Bertrand, M. (2017). An introduction to topological data analysis: fundamental and practical aspects for data scientists. *National Institute for Research in Computer Science and Control*.

Frosini, P. (1992). Measuring shapes by size functions. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, *1607*, 122–133.

IDC (2012). Worldwide big data technology and services. *20122015 Forecast*, *1*, 233–485.

King, A. (2021). 7 benefits to using big data for small businesses. Retrieved from: http://www.industriuscfo.com/7-benefits-using-big-data/

Marr, B. (2015). Big data: Using smart big data, analytics and metrics to make better decisions and improve performance. USA: John Wiley & Sons.

Morris, S. A. (2012). *Topology without tears*. USA: Springer Verlag.

Oudot, S. (2015). Persistence theory: From quiver representations to data analysis. *American Mathematical Society*, *1*, 209.

Perea, J. (2019). A brief history of persistence. *Morfismos*, *23*, 1–16.

Robins, V. (1999). Towards computing homology from finite approximations. In *Topology Proceedings*, *24*(1), 503–532.

Singh, G., Mémoli, F., & Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *Eurographics Symposium on Point-Based Graphics*, *2*, 091–100.

Tierny, J. (2006). Introduction to topological data analysis. Sorbonne Universits. Retrieved from: https://www-apr.lip6.fr/~tierny/stuff/teaching/tierny_topologicalDataAnalysis.pdf

Vclav, S., Jana, N., Fatos, X., & Leonard, B., (2017). Geometrical and topological approaches to big data. *Future Generation Computer Systems*, *67*, 286–296.

How to Cite: Onyango, A., Okelo, B., & Omollo, R. (2023). Topological Data Analysis of COVID-19 Using Artificial Intelligence and Machine Learning Techniques in Big Datasets of Hausdorff Spaces. *Journal of Data Science and Intelligent Systems* 1(1), 55–64, https://doi.org/10.47852/bonviewJDSIS3202701