Journal of Artificial Intelligence and Metaheuristic

♦ ISSN: 2833-5597 Vol: 10, No. 02, 2025 ♦ PP. 09–14

Article DOI: https://doi.org/10.54216/JAIM.100202



FULL LENGTH ARTICLE

Hierarchical Clustering of Global COVID-19 Statistics: Comparative Insights from Pandemic Indicators

Noor Razzaq Abbas[1,*] **Ghassan AL-Thabhawee**[2] **Isam Bahaa Aldallal**[3]
Mostafa Abotaleb[4] **Klodian Dhoska**[5]

[1] Al-Furat Al-Awsat Technical University, Technical Institute of Najaf, Najaf, Iraq
[2] Sciences of Mathematics, Computer Sciences, College of Health and Medical Techniques-Kufa, Al-Furat Al-Awsat Technical University, Kufa, Iraq
[3] Department of Electrical and Computer Engineering, Altinbas University, Istanbul, Turkey
[4] Engineering School of Digital Technologies, Yugra State University, Khanty-Mansiysk, 628012, Russia
[5] Department of Production and Management, Polytechnic University of Tirana, 1001, Tirana, Albania

Emails: noor.hachame@atu.edu.iq · gmohammed@atu.edu.iq · isam.aldallal@gmail.com · abotalebmostafa@bk.ru · kdhoska@fim.edu.al

Received: February 22, 2025 Revised: April 18, 2025 Accepted: June 10, 2025 ⋆ Corresponding author

ABSTRACT

Hierarchical clustering is applied to global COVID-19 data from the Worldometer tracker, whose final update was April 13, 2024, to partition countries into clusters based on epidemiological indicators. Total cases, deaths, recoveries, active cases, tests, population, and per-million measures are standardized and analyzed using agglomerative hierarchical clustering with Ward linkage. The evaluation yields an average Silhouette score of 38.5%, a Davies–Bouldin value of 0.87, and a Calinski–Harabasz value of 77.6. Dendrograms and principal component analysis (PCA) projections identify four clusters, reflecting differences in pandemic impact and response. The analysis finds that high-burden clusters account for almost 45% of global deaths, while low-burden clusters predominate in more than 40% of nations with fewer than 100,000 accumulated cases. The results illustrate how hierarchical clustering can support comparative public-health analysis by providing quantitative estimates of epidemiological patterns across countries.

Keywords: Hierarchical Clustering ▪ COVID-19 ▪ Agglomerative Clustering ▪ Silhouette Score ▪ PCA ▪ Epidemiology ▪ Machine Learning ▪ Unsupervised Learning ▪ Cluster Analysis ▪ Global Health

1. INTRODUCTION

The COVID-19 pandemic has prompted extensive research into clustering algorithms for modeling transmission patterns, country-level heterogeneity, and epidemiological dynamics. Studies have demonstrated the value of unsupervised learning, particularly clustering, for distinguishing country-level patterns and characterizing outbreak severity.

James and Menzies examined COVID-19 trajectories in the United States during the second surge using chaos theory, highlighting the role of time-series modeling in identifying epidemic waves [14]. Zarikas et al. applied clustering to country-level COVID-19 cases and epidemiological variables to examine similarities in outbreak patterns. Their findings show how clustering can reveal latent patterns in large datasets and support comparisons of national responses [15].

Other methods have also been explored. Mahmoudi et al. used fuzzy clustering to compare COVID-19 spread rates among high-risk countries, accounting for uncertainty and

© 2025 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

overlapping cluster membership [16]. Pasin examined countries in terms of cases and deaths, while Ramadan et al. used multivariate analysis of daily statistics to track changing pandemic indicators [17, 18]. Together, these studies support using multiple variables to represent the scale and temporal development of the pandemic.

Time-series clustering has been used to characterize how epidemic curves evolve across countries and how outbreak peaks synchronize [19]. Research on pandemic responses has also emphasized the influence of democracy, capacity, coercion, and political institutions on intervention strategies and epidemiological outcomes [20]. Global preparedness measures, including the Global Health Security Index and the epidemic preparedness index, can complement outcome-based clustering by describing health-system readiness [21, 22].

Clinical and epidemiological evidence provides further context for interpreting clusters. Reviews of COVID-19 pathophysiology, transmission, diagnosis, and treatment, together with studies of hospitalized patients and excess mortality, help explain differences in mortality burden among country groups [23, 24, 25]. Overall, clustering offers an unsupervised means of identifying similarities and contrasts in pandemic impact, while preparedness and clinical indicators can improve interpretation of the resulting groups.

2. RELATED WORK

The COVID-19 pandemic has driven sustained investigation of clustering methods for identifying patterns of spread, national-level heterogeneity, and epidemiological dynamics. Prior work has shown that unsupervised learning can distinguish country-level trends and help characterize outbreak severity.

James and Menzies analyzed the second COVID-19 surge in the United States through chaos theory, emphasizing the usefulness of time-series modeling for identifying epidemic waves. Zarikas et al. grouped countries using COVID-19 case data and epidemiological variables to study similarities in outbreak trends and national responses [14, 15].

Fuzzy clustering has been used to compare transmission rates in high-risk countries while representing uncertainty and overlapping cluster membership [16]. Country comparisons based on deaths and cases, as well as multivariate analysis of daily statistics, have also been used to examine international mortality, morbidity, and changing pandemic indicators [17, 18].

For temporal analysis, a cluster-based dual-evolution model has been applied to multivariate COVID-19 time series to describe epidemic-curve development and the synchronization of outbreak peaks [19]. Other work has examined how governance, national capacity, and coercion shape pandemic responses [20]. The Global Health Security Index and epidemic preparedness measures provide system-readiness indicators that can be considered alongside epidemiological outcomes [21, 22].

Clinical studies further inform interpretation of country clusters. Research on COVID-19 pathophysiology and transmission, patient characteristics and outcomes, and pandemic-associated excess mortality provides context for differences in disease burden across groups [23, 24, 25]. Taken together, prior methods—from fuzzy clustering to time-series analysis—complement traditional epidemiological models by revealing patterns of similarity and contrast. Incorporating preparedness and clinical indicators can make cluster results more interpretable and support further study of global health vulnerabilities and strengths.

3. DATA AND METHODOLOGY

3.1 Dataset

The dataset used in this analysis is a Kaggle-distributed copy of Worldometer's country-level cumulative COVID-19 tracker, whose final update was April 13, 2024 [26]. It contains 231 reporting entries: 229 countries and territories, plus the Diamond Princess and MS Zaandam cruise ships. Its 21 attributes represent absolute and relative pandemic measures, including total confirmed cases, deaths, recoveries, active cases, serious or critical cases, and testing numbers. To support comparisons among countries with different population sizes, the dataset also includes cases, deaths, active cases, and tests per million people, as well as population and continent.

Preprocessing was required to address missing values, normalize numeric attributes, and correct formatting inconsistencies. Figure 1 summarizes the procedure. In the raw Worldometer table, string-formatted numbers, commas, and special characters were cleaned, and variables were converted to numeric form. Missing values were imputed using the median. Feature selection then retained nine epidemiological indicators: total cases, deaths, recoveries, active cases, cases per million, deaths per million, active cases per million, tests per million, and population.

3.2 Hierarchical Clustering

Hierarchical clustering is an unsupervised machine-learning technique that organizes observations into nested clusters according to similarity. Let the dataset be defined as

$$\mathscr{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}, \qquad \mathbf{x}_i \in \mathbb{R}^m, \tag{1}$$

where n is the number of countries or samples and m is the number of epidemiological variables. The algorithm starts with each observation in its own cluster and sequentially merges clusters until all observations belong to a single cluster. Its aim is to minimize within-cluster distance and maximize separation between clusters.

The Euclidean distance between two observations is

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{f=1}^{m} (x_{if} - x_{jf})^2}. \tag{2}$$

This is a commonly used measure for epidemiological data. Alternatives include the Manhattan distance,

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{f=1}^{m} |x_{if} - x_{jf}|, \tag{3}$$

and cosine similarity,

$$\mathrm{sim}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \, \|\mathbf{x}_j\|}. \tag{4}$$

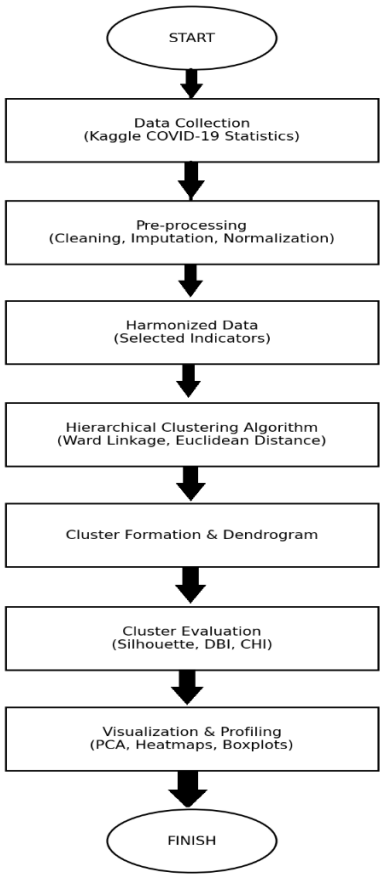


Figure 1. Data preprocessing workflow.

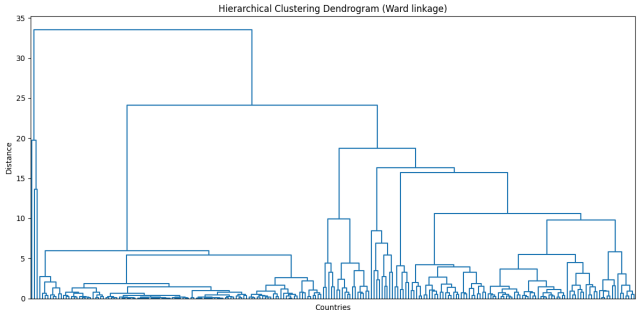


Figure 2. Hierarchical clustering dendrogram using Ward linkage.

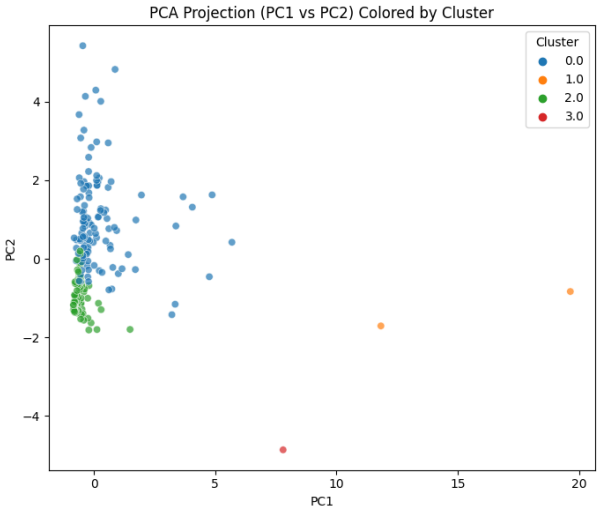


Figure 3. PCA projection (PC1 versus PC2), colored by cluster.

Linkage criteria define how distances between clusters are measured. Single linkage uses the minimum pairwise distance,

$$d(A,B) = \min_{\mathbf{x}\in A,\,\mathbf{y}\in B} d(\mathbf{x},\mathbf{y}), \tag{5}$$

whereas complete linkage uses the maximum:

$$d(A,B) = \max_{\mathbf{x}\in A,\,\mathbf{y}\in B} d(\mathbf{x},\mathbf{y}). \tag{6}$$

Average linkage takes the mean of the pairwise distances,

$$d(A,B) = \frac{1}{|A||B|} \sum_{\mathbf{x}\in A} \sum_{\mathbf{y}\in B} d(\mathbf{x},\mathbf{y}), \tag{7}$$

balancing the extremes of single and complete linkage. Ward's criterion instead minimizes the increase in within-cluster variance:

$$\Delta(A,B) = \frac{|A||B|}{|A|+|B|} \|\bar{\mathbf{x}}_A - \bar{\mathbf{x}}_B\|_2^2, \tag{8}$$

where $\bar{\mathbf{x}}_A$ and $\bar{\mathbf{x}}_B$ are the cluster centroids.

The sequence of merges is recorded in a dendrogram, whose branch heights correspond to linkage distances. A fixed set of k clusters can be obtained by cutting the dendrogram at threshold τ:

$$\mathscr{C} = \{C_1, C_2, \ldots, C_k\}, \qquad d(C_i, C_j) > \tau \quad (i \neq j), \tag{9}$$

for distinct clusters [27, 28, 29, 30, 31, 32, 33, 34].

Figure 2 shows the Ward-linkage dendrogram for the global COVID-19 data. Its leaves represent individual countries, and the vertical links indicate the distances at which clusters merge. Short links connect countries with similar epidemiological characteristics, while taller links connect more dissimilar groups. Natural break points in the tree provide a basis for selecting the number of clusters.

4. RESULTS

The analysis was conducted in Python. Pandas, NumPy, and scikit-learn were used for preprocessing, normalization, and clustering; matplotlib and seaborn were used to produce dendrograms, correlation heatmaps, and cluster profiles.

The PCA projection of the first two components is shown in Figure 3. Four clusters are distinguished by color. Cluster 0 (approximately 50% of countries) and Cluster 2 (approximately 45%) account for most observations, while Clusters 1 and 3 (approximately 1% each) appear as extreme outliers.

Figure 4 shows the first and third principal components. Clusters 0 and 2 again contain most countries, whereas Clusters 1 and 3, together comprising less than 2% of the dataset, appear as outliers. This pattern indicates that the clustering is stable across these dimensions.

Figure 5 compares principal components 2 and 3. Clusters 0 and 2 together include about 95% of countries and form the dense center of the projection. The remaining observations belong to Clusters 1 and 3, representing uncommon pandemic profiles.

Figure 6 presents boxplots of seven COVID-19 indicators

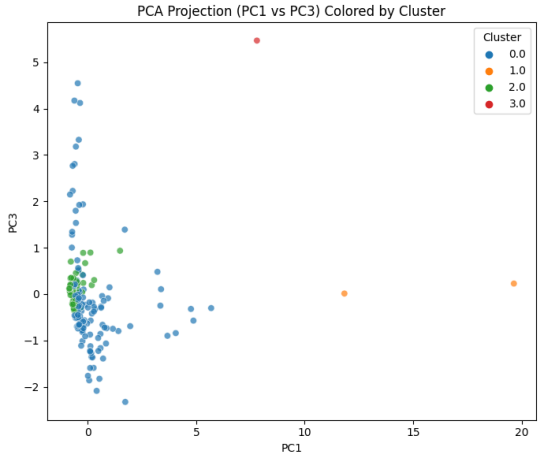


Figure 4. PCA projection (PC1 versus PC3), colored by cluster.

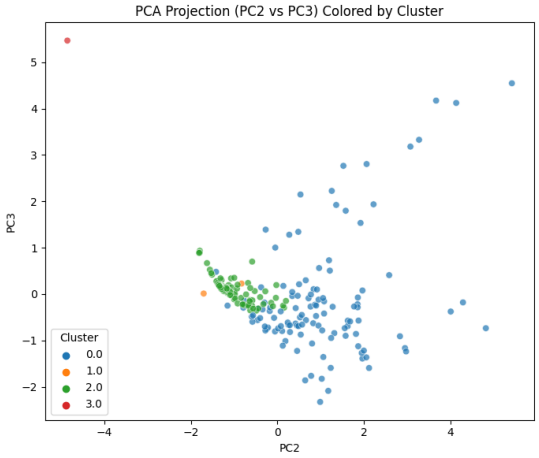


Figure 5. PCA projection (PC2 versus PC3), colored by cluster.

across the four clusters. Cluster 1 (approximately 1% of countries) has exceptionally high total cases, reaching tens of millions. Clusters 0 and 2 contain most countries and generally have lower case counts, while Cluster 3 represents an intermediate but consistent profile. Total deaths follow a similar pattern: despite its small size, Cluster 1 includes countries with more than one million fatalities, whereas Clusters 0 and 2 have lower median counts and varied pandemic effects.

Cluster 1 also has the greatest active-case burden, with many countries exceeding one million active cases. The cases-per-million distributions for Clusters 0 and 2 have broad ranges. Cluster 1 has the highest deaths per million, while Cluster 0 has a broader, more scattered distribution. Testing per million is most dispersed in Cluster 0, including countries with more than 20 million tests per million people; Cluster 2 has more moderate and consistent testing. The population plot highlights the exceptionally large population represented in Cluster 3, while most countries in Clusters 0 and 2 have more moderate population sizes.

Figure 7 quantifies cluster membership. Cluster 0 contains approximately 50% of countries and Cluster 2 approximately 45%, while Clusters 1 and 3 each contain fewer than 2%. Thus, the global dataset is dominated by two epidemiological patterns, with a small number of countries differing substantially.

Figure 8 shows normalized mean indicator values across the

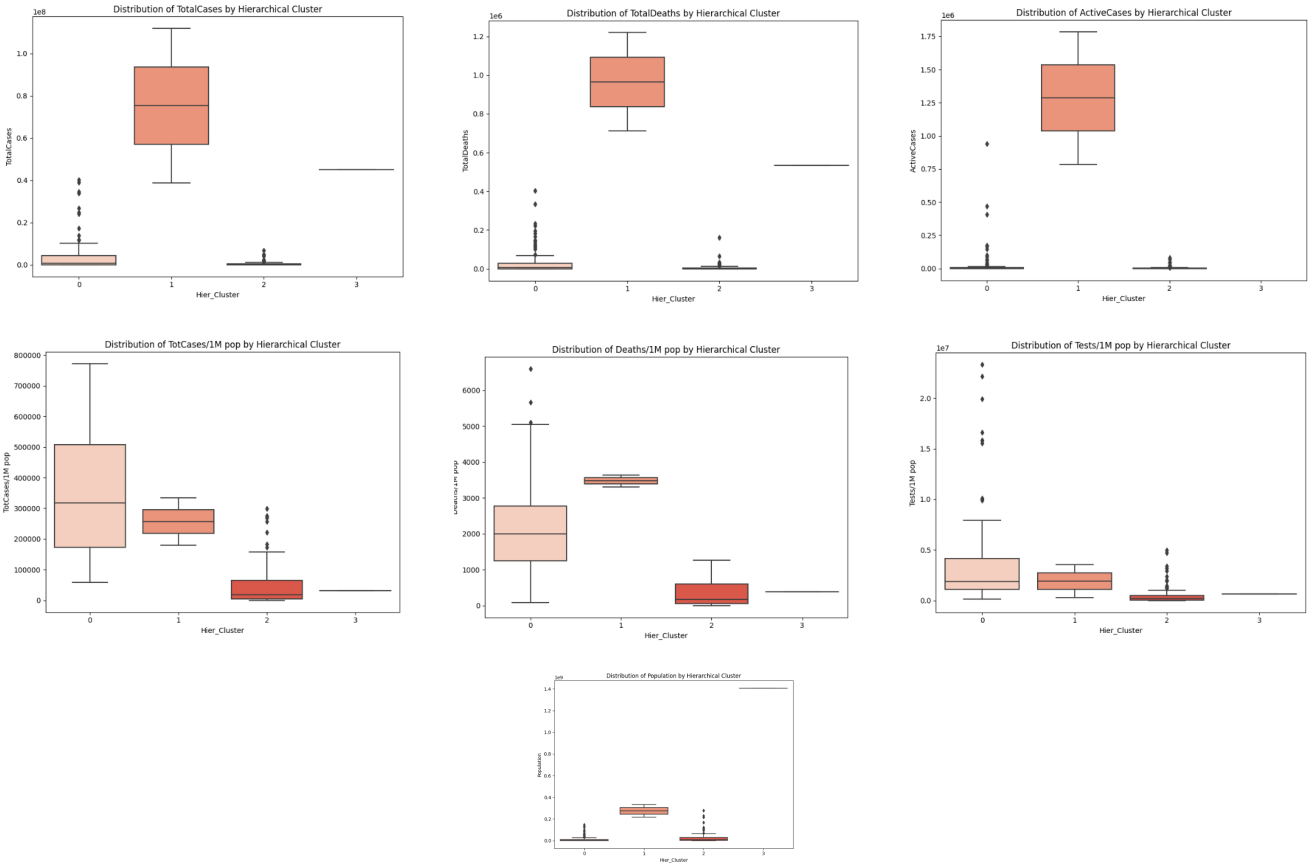


Figure 6. Distributions of COVID-19 indicators by hierarchical cluster.

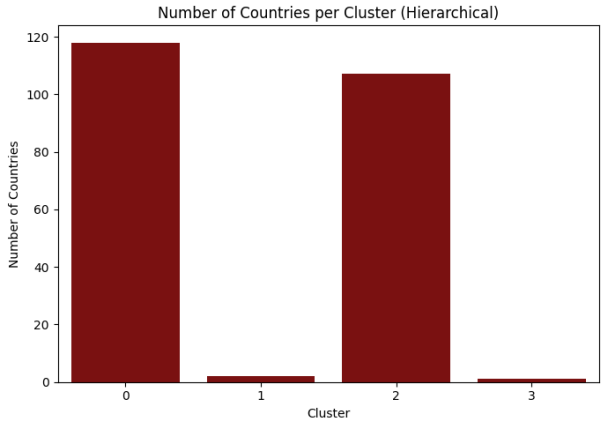


Figure 7. Number of countries in each hierarchical cluster.

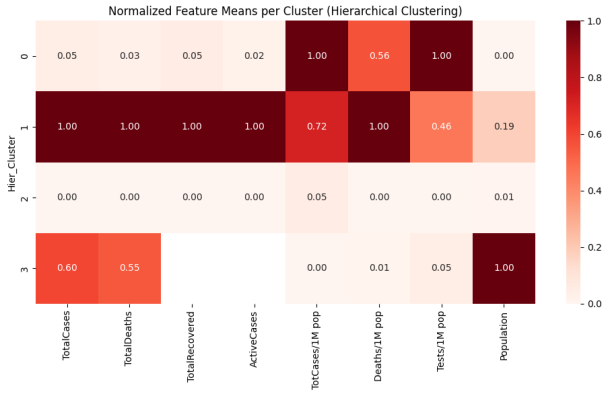


Figure 8. Normalized feature means per cluster for hierarchical clustering.

four clusters, with values ranging from 0 (minimum) to 1 (maximum). Cluster 0 has normalized values of 1.0 for cases per million, deaths per million, and tests per million, while its absolute case, death, and recovery counts are low (approximately 0.05), indicating smaller countries with high population-adjusted impact. Cluster 1 has the highest normalized values for most absolute measures, including cases, deaths, recoveries, and active cases; cases per million are 0.72 and deaths per million are 1.0, while testing intensity is moderate at 0.46. Cluster 2 has values close to zero across indicators, reflecting low reported impact. Cluster 3 has relatively high normalized case (0.60) and death (0.55) counts and the highest normalized population (1.0), indicating a highly populated country with substantial absolute counts but more contained per-capita impact.

5. CONCLUSION

This study applied Ward-linkage hierarchical clustering to global COVID-19 data from Worldometer's tracker, last updated April 13, 2024, to identify country-level patterns of pandemic impact. Preprocessing, normalization, and clustering yielded four groups representing different levels of case burden, mortality, testing intensity, and demographic characteristics. The dendrogram revealed natural splits in the data, and PCA projections showed separation among clusters. Cluster validity scores were a Silhouette measure of 38.5%, a Davies–Bouldin Index of 0.87, and a Calinski–Harabasz Index of 77.6.

Cluster profiling identified two large groups: Cluster 0, containing approximately 50% of countries, and Cluster 2, containing approximately 45%. Clusters 1 and 3, together representing approximately 2% of countries, comprised outliers with unusually high case counts, testing intensity, or population characteristics. Correlation analysis indicated strong relationships among deaths, total cases, and recoveries, while per-capita measures aligned closely with mortality rates.

REFERENCES

[1] World Health Organization, "Joint external evaluation tool: International health regulations," Online, 2018.

[2] M. Moore, B. Gelfeld, A. Okunogbe, and C. Paul, "Identifying future disease hot spots: Infectious disease vulnerability index," *RAND Health Quarterly*, vol. 6, no. 5, 2017.

[3] Melinda Moore, Bill Gelfeld, Adeyemi Okunogbe, and Christopher Paul, "Identifying future disease hot spots: Infectious disease vulnerability index," RAND Corporation, Santa Monica, CA, RAND Research Report RR-1605-OSD, 2016.

[4] World Health Organization, "WHO director-general's opening remarks at the media briefing on COVID-19 – 11 march 2020," Online, 2020.

[5] Worldometer, "Coronavirus graphs: Worldwide cases and deaths," Online, 2024.

[6] T. Hale, N. Angrist, R. Goldszmidt *et al.*, "A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)," *Nature Human Behaviour*, vol. 5, pp. 529–538, 2021.

[7] B. Oppenheim, M. Gallivan, N. K. Madhav *et al.*, "Joint external evaluation (JEE) scores," Supplementary material for the Epidemic Preparedness Index study, 2019.

[8] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, pp. 236–244, 1963.

[9] D. T. Jamison, L. J. Lau, K. B. Wu *et al.*, "Country performance against COVID-19: Rankings for 35 countries," *BMJ Global Health*, vol. 5, p. e003047, 2020.

[10] S. Kumar, "Use of cluster analysis to monitor novel coronavirus-19 infections in maharashtra, india," *Indian Journal of Medical Sciences*, vol. 72, pp. 44–48, 2020.

[11] P. Sengupta, B. G. Ghosh, and S. SenRoy, "An analysis of COVID-19 clusters in india – two case studies on nizamuddin and dharavi," Research Square, 2020.

[12] M. L. F. Nascimento, "A multivariate analysis on spatiotemporal evolution of COVID-19 in brazil," *Infectious Disease Modelling*, vol. 5, pp. 670–680, 2020.

[13] Anonymous, "K-means clustering of COVID-19 cases in indonesia's provinces," in *Proceedings of the International Conference on Global Optimization and Its Applications, November 21–22*, 2020.

[14] N. James and M. Menzies, "COVID-19 in the united states: Trajectories and second surge behavior," *Chaos*, vol. 30, p. 091102, 2020.

[15] V. Zarikas, S. G. Poulopoulos, Z. Gareiou *et al.*, "Clustering analysis of countries using the COVID-19 cases dataset," *Data in Brief*, vol. 31, p. 105787, 2020.

[16] M. R. Mahmoudi, D. Baleanu, Z. Mansor *et al.*, "Fuzzy clustering method to compare the spread rate of COVID-19 in the high risks countries," *Chaos, Solitons & Fractals*, vol. 140, p. 110230, 2020.

[17] O. Pasin, "Clustering of countries in terms of deaths and cases of COVID-19," *Journal of Health and Social Sciences*, vol. 5, pp. 587–594, 2020.

[18] A. Ramadan, A. Kamel, A. Taha *et al.*, "A multivariate data analysis approach for investigating daily statistics of countries affected with COVID-19 pandemic," *Heliyon*, vol. 6, p. e05575, 2020.

[19] N. James and M. Menzies, "Cluster-based dual evolution for multivariate time series: Analyzing COVID-19," *Chaos*, vol. 30, p. 061108, 2020.

[20] M. M. Kavanagh, R. Singh, and S. R. Democracy, "Democracy, capacity, and coercion in pandemic response: COVID-19 in comparative political perspective," *Journal of Health Politics, Policy and Law*, vol. 45, pp. 997–1012, 2020.

[21] Global Health Security Index, "Global health security (GHS) index 2019," Online, 2019.

[22] B. Oppenheim, M. Gallivan, N. K. Madhav *et al.*, "Assessing global preparedness for the next pandemic: Development and application of an epidemic preparedness index," *BMJ Global Health*, vol. 4, p. e001157, 2019.

[23] W. J. Wiersinga, A. Rhodes, A. C. Cheng *et al.*, "Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): A review," *JAMA*, vol. 324, pp. 782–793, 2020.

[24] S. Richardson, J. S. Hirsch, J. Narasimhan *et al.*, "Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the new york city area," *JAMA*, vol. 323, pp. 2052–2059, 2020.

[25] D. M. Weinberger, J. Chen, T. Cohen *et al.*, "Estimation of excess deaths associated with the COVID-19 pandemic in the united states, march to may 2020," *JAMA Internal Medicine*, vol. 180, pp. 1336–1344, 2020.

[26] Worldometer, "COVID-19 coronavirus pandemic," Worldometer COVID-19 tracker, 2024.

[27] T. A. Alhasani, H. Alkattan, A. Ali, E.-S. M. El-Kenawy, and M. M. Eid, "A comparative analysis of methods for detecting and diagnosing breast cancer based on data mining," *Journal of Artificial Intelligence and Metaheuristics*, vol. 4, no. 2, pp. 8–17, 2023.

[28] D. S. Khafaga, H. Alkattan, and A. A. Subhi, "Evaluating the effect of optimized voting using hybrid particle swarm and grey wolf algorithm on the classification of the zoo dataset," *Journal of Artificial Intelligence and Metaheuristics*, vol. 2, no. 1, pp. 8–15, 2022.

[29] H. K. Al-Mahdawi, M. Abotaleb, H. Alkattan, E.-S. M. El-Kenawy, and E. M. Mohamed, "Solving the inverse initial value problem for the heat conductivity equation by using the picard method," *Journal of Artificial Intelligence and Metaheuristics*, vol. 2, no. 2, pp. 46–55, 2022.

[30] H. Alkattan, S. K. Towfek, and M. Y. Shams, "Tapping into knowledge: Ontological data mining approach for detecting cardiovascular disease risk causes among diabetes patients," *Journal of Artificial Intelligence and Metaheuristics*, vol. 4, no. 1, pp. 8–15, 2023.

[31] N. R. Abbas, H. Alkattan, H. Rabiei-Dastjerdi, M. Saber, and M. M. Eid, "Monthly solar prediction using machine learning: Diyala governorate, iraq as a case study," *Journal of Artificial Intelligence and Metaheuristics*, vol. 5, no. 2, pp. 41–46, 2023.

[32] H. K. Al-Mahdawi, A. A. Subhi, H. Alkattan, M. Saber, M. M. Eid, A. A. S. Al-Mahdawi, and J. A. M. Al-Saddaee, "Solving initial value problem in composite materials for heat equation," *Journal of Artificial Intelligence and Metaheuristics*, vol. 6, no. 1, pp. 8–17, 2023.

[33] T. Q. Al-Seyday, T. Qenawy, H. Alkattan, and A. Khaled, "Securing DNS over HTTPS: A machine learning study on traffic classification using DoHBrw-2020," *Journal of Artificial Intelligence and Metaheuristics*, vol. 7, no. 2, pp. 73–81, 2024.

[34] A. Gupta and R. Kumar, "A hybrid deep learning approach for brain stroke detection using MRI images," *Journal of Medical Systems*, vol. 46, no. 3, pp. 1–10, 2022.