Open Peer Review on Qeios

SARS-CoV-2 Virion: A Humane Debacle - An Analytical Approach

Raja Sarath Kumar Boddu[1]

1 Lenora College of Engineering

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.

Abstract

> The World Health Organization (WHO) has declared COVID-19 a pandemic, as the SARS-CoV-2 virus and its variants have spread worldwide, causing coronavirus diseases (COVID-19). COVID-19 is primarily described as an infectious disease that leads to severe acute respiratory syndrome (SARS), later transforming into the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) virion variant. These virion variants have emerged globally with deceptively higher transmissibility and immune evasion capabilities. In this research paper, we propose to compare several ML algorithms to predict COVID-19 mortality using data from various countries and select the best performing algorithm as a predictive tool for decision-making. The study aims to develop a mortality risk prediction for COVID-19 based on ML algorithms that utilize data.

Raja Sarath Kumar Boddu

Senior Member, IEEE

Professor and Principal,

Computer Science Department,

Lenora College of Engineering, India.

Email: iamsarathphd@ieee.org

Keywords: COVID-19, Machine Learning, Mortality, SARS-CoV-2, Virion, Vaccination.

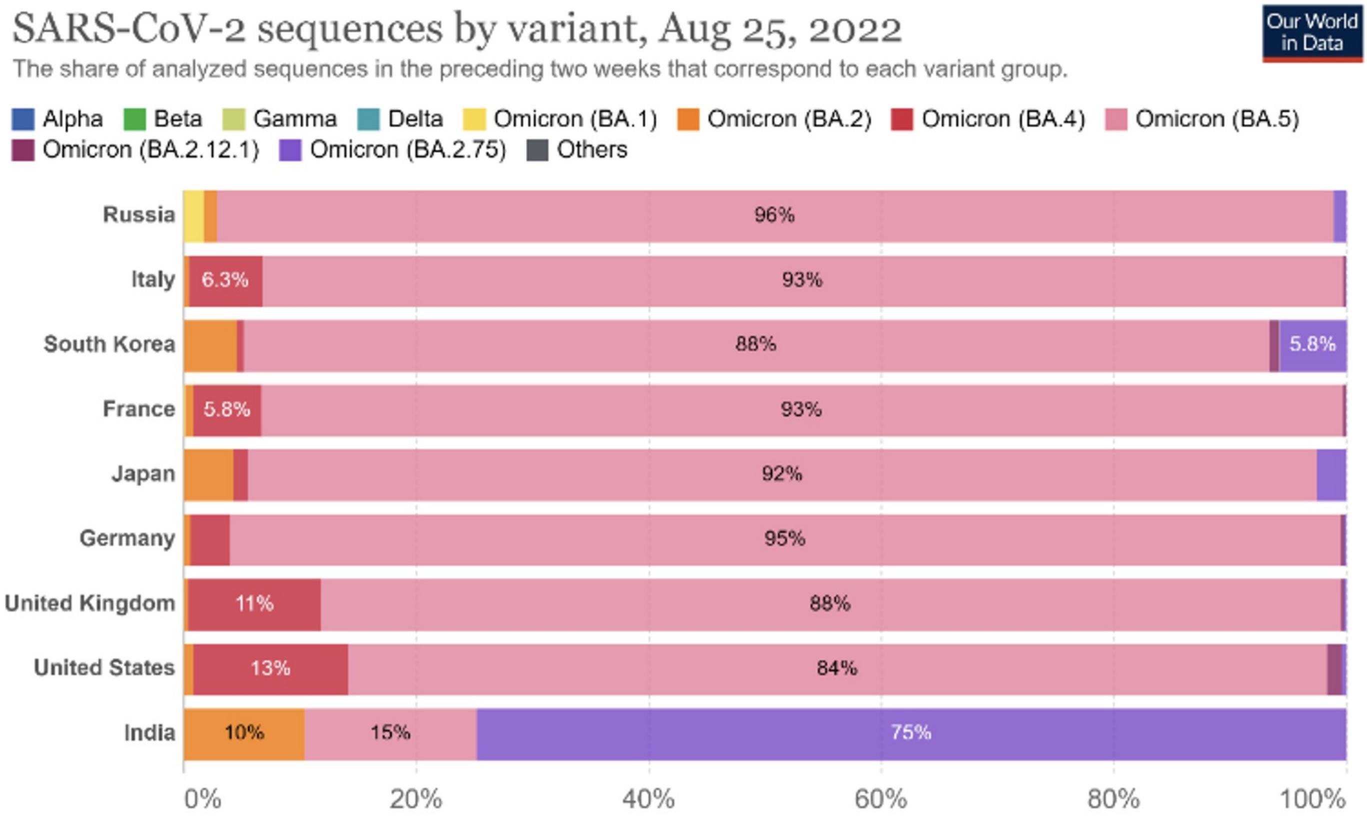
I. Introduction

The COVID-19 pandemic is a global epidemic. Unfortunately, it is a harsh reality that as of October 28, 2022, the World Health Organization (WHO) has reported a total of 626,337,158 confirmed cases of COVID-19, including 6,566,610 deaths worldwide [1]. Furthermore, as of August 19, 2022, a total of 12,814,704,622 vaccine doses have been administered globally [1], providing hope and satisfaction regarding vaccination statistics. However, the situation continues to worsen, with millions of people affected globally. The top ten countries with the highest death tolls are listed in Table 1.

A. SARS-CoV-2 Virion and its Variants

In later stages, the infectious disease causing severe acute respiratory syndrome (SARS) transforms itself into SARS-CoV-2 and further into various variants like Alpha, Beta, Gamma, Delta, Omicron (BA.1), Omicron (BA.2), Omicron (BA.4), Omicron (BA.5), Omicron (BA2.12.1), Omicron (BA2.75), and others. Fig.1 represents the SARS-CoV-2 variants in the selected 10 countries listed in Table 1. The immunity of an individual infected with SARS-CoV-2 could be estimated when the disease symptoms are known to some extent [2].

Through quantitative analysis of viral load, Ron Sender et al. estimated that each infected person carries a viral load of SARS-CoV-2 in bodily fluids of up to 100 billion virions when infected, equivalent to a mass of approximately 0.10 mg. When estimated globally for all infected individuals, this mass could be as much as 10 kg [3].



Source: GISAID, via CoVariants.org CC BY

Note: This share may not reflect the complete breakdown of cases, since only a fraction of all cases are sequenced. Recently-discovered or actively-monitored variants may be overrepresented, as suspected cases of these variants are likely to be sequenced preferentially or faster than other cases.

Figure 1. SARS-CoV-2 variants in the selected 10 countries as of August 25th, 2022

Actually, viruses are nucleoproteins and non-cellular structures with infectious genetic material. Virions are capsid-encapsulated viruses with DNA or RNA molecules. The coronavirus is a cluster of related RNA (RiboNucleic Acid) viruses commonly found in birds and mammals [3]. It has both nucleic acid and protein layers.

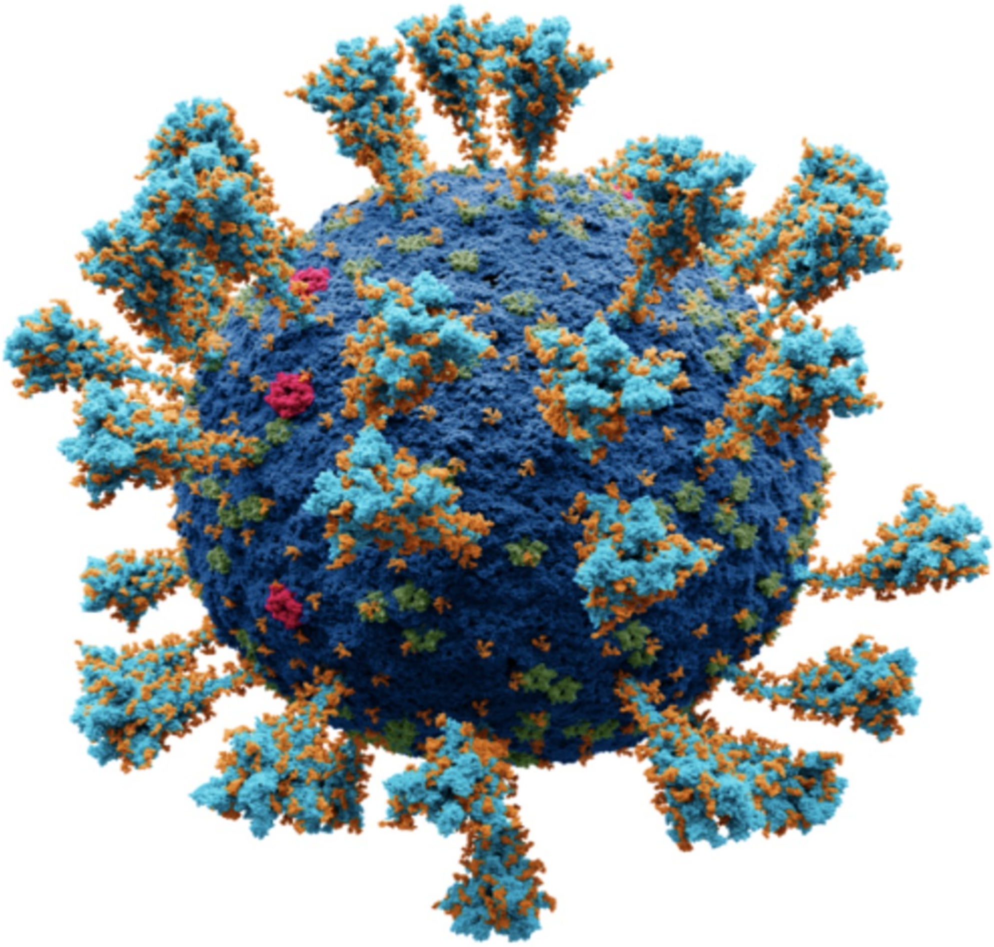


Figure 2. Scientifically accurate atomic model of coronavirus (SARS-CoV-2). Each "ball" represents an atom.

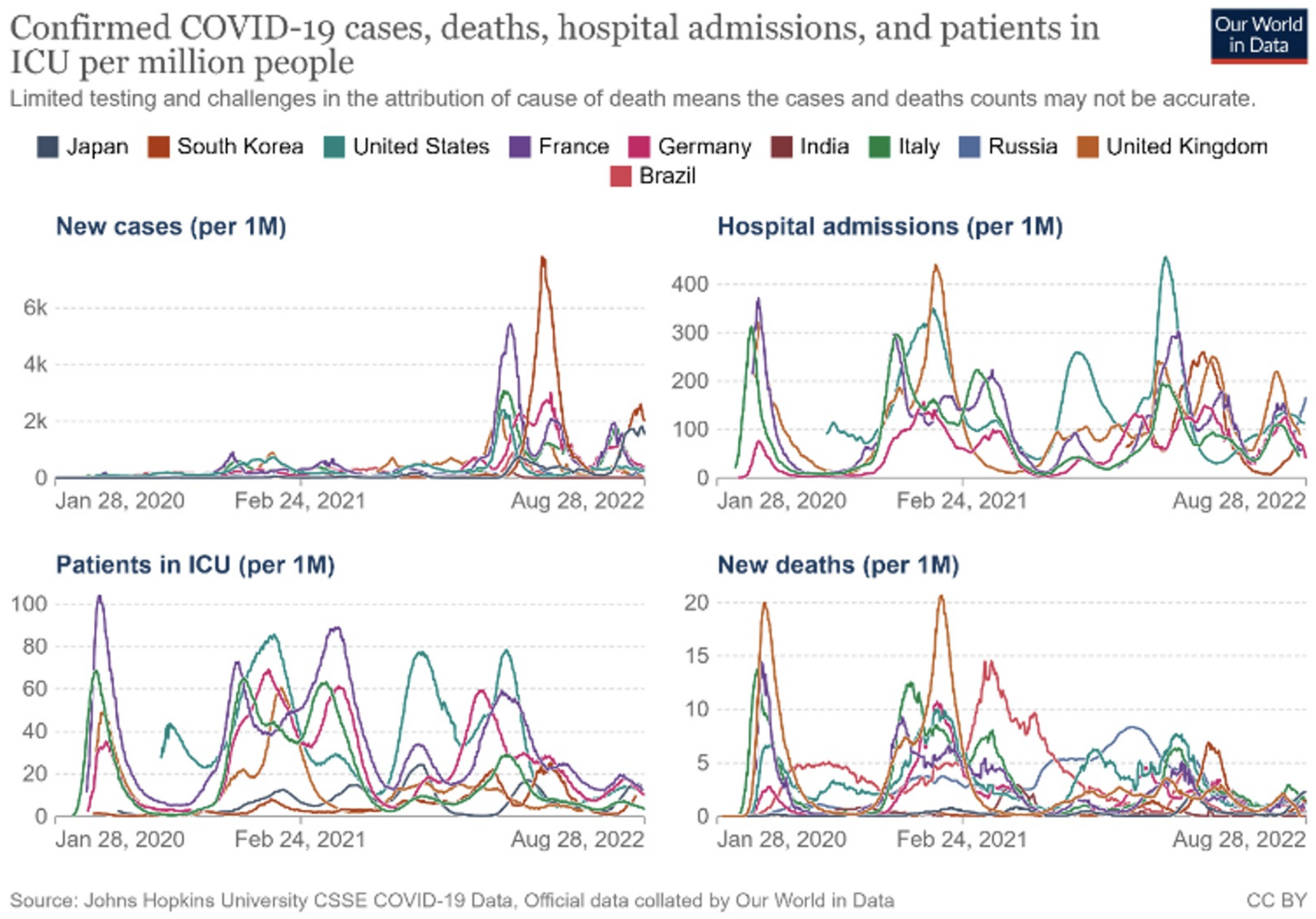
Machine Learning (ML) algorithms are commonly used as alternative methods for classification and prediction and could be utilized as a potential solution for predicting mortality during the COVID-19 pandemic globally [4].

As the COVID-19 pandemic became a nightmare, all countries initiated and implemented countermeasures such as lockdowns, restrictions on movement for public and private conveyance, closure of offices, airports, railways, roadways, and all travel in and out of each country placed under quarantine. Social distancing and physical isolation measures were repeatedly implemented to prevent the spread of COVID-19 on a war footing, as healthcare systems pathetically collapsed in many countries. Additionally, to curb the spread of COVID-19, people were advised to wash their hands thoroughly or use sanitizer, preferably stay at home, and cover their mouth and nose with masks, especially when coughing or sneezing. On the other hand, this pandemic also witnessed a drastic downfall of the global economy due to eccentric activities initiated by several countries. The mortality rate is increasing day by day, demanding an early response to diagnose and prevent the spread of this incurable disease. One critical aspect of the spread of COVID-19 is the lack of specific clinical detection, medications, and treatments, making the situation distressing and terrifying worldwide.

Table 1. Top 10 countries - total confirmed cases, deaths, and recovered cases, and mortality rate by country

COUNTRY	CONFIRMED	DEATHS	CASE-FATALITY	DEATHS/100K POP.
USA	94,190,979	1,043,840	1.1%	316.83
India	44,415,723	527,799	1.2%	38.25
France	34,662,834	154,897	0.4%	237.39
Brazil	34,368,909	683,397	2.0%	321.51
Germany	32,041,350	147,104	0.5%	176.90
United Kingdom	23,708,629	205,414	0.9%	302.59
South Korea	23,026,960	26,618	0.1%	51.92
Italy	21,806,509	175,347	0.8%	290.01
Russia	19,123,501	376,301	2.0%	257.86
Japan	18,531,986	39,047	0.2%	30.87

Despite struggling to handle the origin of the COVID-19 pandemic, the scientific community has managed to change the state of COVID-19 in all aspects from hopelessness to hope. Rigorous global research has been initiated, leading to evidence-backed breakthroughs to contain the unrestricted spread of the COVID-19 pandemic, including the development of precautionary trials and vaccines [4].



Confirmed COVID-19 cases, deaths, hospital admissions, and patients in ICU per million people

Limited testing and challenges in the attribution of cause of death means the cases and deaths counts may not be accurate.

New cases (per 1M)

Hospital admissions (per 1M)

Patients in ICU (per 1M)

New deaths (per 1M)

Source: Johns Hopkins University CSSE COVID-19 Data, Official data collated by Our World in Data

CC BY

Figure 3. Confirmed COVID-19 cases, deaths, hospital admissions, and patients in ICU per million people

The cumulative confirmed deaths during COVID-19 have been officially recorded[5].

This research paper is structured as follows: Section 1, introduction section stated with relevant basic information, facts and figures and data.; Next is the problem statement which states the problem and objective of the research methodology in Section 2. Section 3 is the proposed methodology which states the data, mortality statistics and technique used, architecture and software process flow. Section 4 describes the data, dataset, the steps of the empirical analysis and the discussion of the results; and Section 5 discusses this study’s limitations, conclusion where we have listed our findings and future developments.

II. Research Methodology

In the research methodology, we applied seven machine learning classification techniques to evaluate the dataset: Naive Bayes Classifier, J48 pruned tree Classifier, Bayes Network Classifier, JRip, SVM, RandomForest, and K-NN.

The *Naive Bayes Classifier* is a subset of Bayesian theory used for text classification, which classifies transferred data into a readable format. It discriminates the dataset based on specific features or attributes.

The *J48 Decision Tree* algorithm produces a graphical tree frequently used for classification purposes. It follows a divide and conquer approach, where instances are divided into sub-ranges based on the values of the attributes.

The *Bayes Network Classifier* provides a compact, flexible, and interpretable representation of a joint probability distribution [6]. It is useful for predicting the likelihood of known causes contributing to an event that occurred.

JRip implements a propositional rule learner called "Repeated Incremental Pruning to Produce Error Reduction" (RIPPER) and uses sequential covering algorithms to create ordered rule lists. The algorithm involves four stages: Growing a rule, Pruning, Optimization, and Selection. In Weka, it is referred to as JRip, and it is a basic incremental reduced-error pruning algorithm.

Support Vector Machines (SVM) is a supervised learning method used for classifying and predicting data. In the classification task, training and testing are involved with the data, where each instance contains the target values in training. SVM provides a set of supervised learning models enabling accurate prediction for classification and regression tasks [7]. Support vectors are information points that lie closest to the decision surface, also known as hyperplanes. *RandomForest (RF)* is another algorithm that produces a tree, but it generates several trees from random samples in the dataset. The final result is based on the majority of the outcomes from the developed trees, resulting in improved accuracy due to this ensemble of trees selecting the most suitable class. *k-Nearest Neighbors (KNNs)* are considered one of the best "simplest" supervised machine learning algorithms, extensively studied in pattern recognition and used as a case-based learning method for classification and regression problems. KNN relies on labeled data to provide the correct output for unlabeled data [8].

In the classification technique, the training of the dataset is implemented first on the seven different algorithms mentioned above. Once the training process is completed, the performance evaluation phase becomes essential for evaluating the model's performance. The performance of each developed model is evaluated using sensitivity, specificity, accuracy, precision, and ROC performance metrics. The automation machine learning tool, *Waikato Environment for Knowledge Analysis (WEKA)* [9], has been used for data cleansing, pre-processing, feature extraction, and classification of the dataset, and to simulate the models.

A. Model development

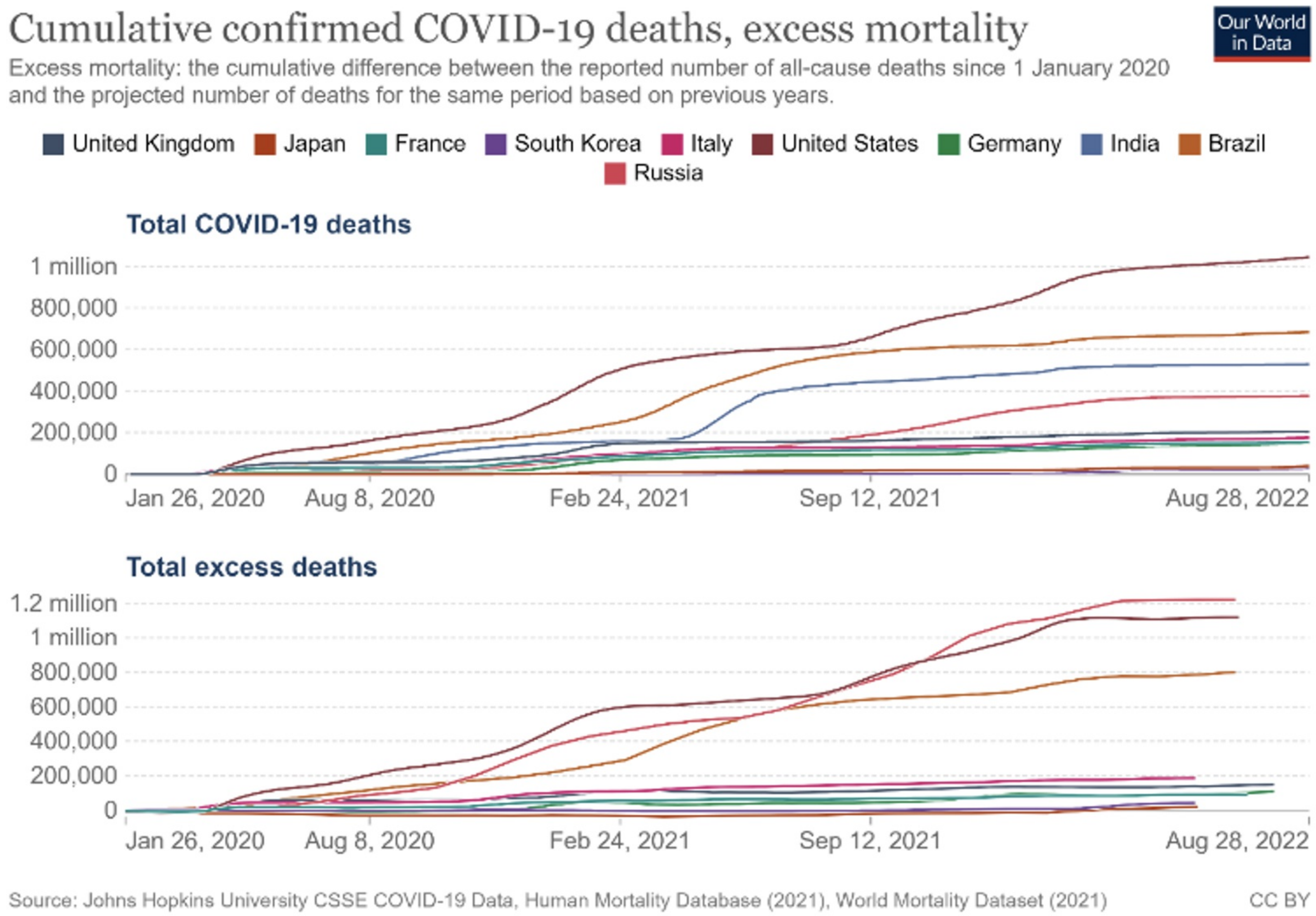
Before constructing the mortality prediction model, selection criteria are based on related studies in the field with thorough reviews. The model is then constructed using the seven ML algorithms: Naive Bayes, J48, Bayes Net, JRip, SVM, RandomForest, and K-NN.

B. Cross-validation

A 10-fold cross-validation process is used to evaluate the performance of the classification models. In this process, the original samples are randomly divided into 10 sub-samples of approximately equal size. One of the sub-samples is used as the validation dataset for testing the models, while the remaining nine sub-samples are used as training datasets. This cross-validation method is repeated 10 times, with each sub-sample being used sequentially for validation. The validation results from the ten experimental models are then combined to derive performance metrics such as sensitivity, specificity, accuracy, precision, and ROC based on the testing results.

III. Mortality Statistics

Mortality varies by time and location, and its measurement is affected by well-known biases that have been intensified during the COVID-19 pandemic. The chart below shows the number of deaths either per 100 confirmed cases or per 100,000 population, indicating the countries most affected by COVID-19 worldwide. Countries at the top of this figure have the highest number of deaths proportionally to their COVID-19 cases or population [10].


Cumulative confirmed COVID-19 deaths, excess mortality

Excess mortality: the cumulative difference between the reported number of all-cause deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years.

Total COVID-19 deaths

Total excess deaths

Source: Johns Hopkins University CSSE COVID-19 Data, Human Mortality Database (2021), World Mortality Dataset (2021)

CC BY

Figure 4. Cumulative confirmed COVID-19 deaths, excess mortality

The diagonal lines on the chart (Fig. 5) correspond to different case fatality ratios. Countries falling on the uppermost lines have the highest observed case fatality ratios. Points with a black border correspond to the 20 most affected countries by COVID-19 worldwide, based on the number of deaths.

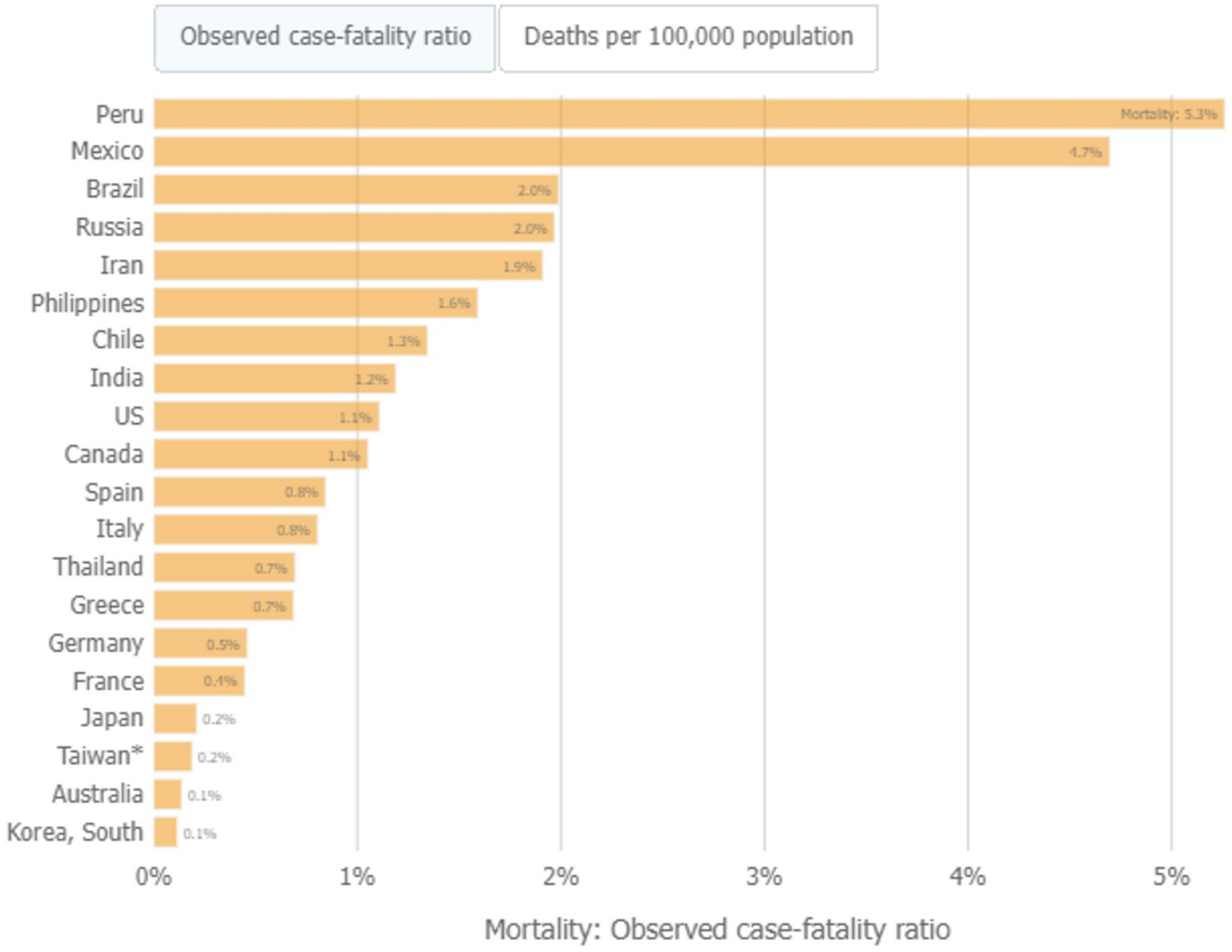


Figure 5. Mortality - Observed case-fatality ratio

IV. Dataset

The dataset, worldometer_data, has been taken from Kaggle[11] and is used for the analysis and prediction of mortality during COVID-19. The dataset contains 16 features with 209 country records. These features make it a widely utilized database for researchers globally, providing information on mortality in various countries, region-wise details, total cases, total deaths, and the mortality rate. These attributes are used in the present work to study infection and mortality progression.

A. Simulation

The dataset contains 209 instances representing countries, and 16 attributes representing country/region, continent, population, total cases, new cases, total deaths, new deaths, total recovered, new recovered, active cases, serious, critical, total Cases/1m pop, deaths/1m pop, total tests, tests/1m pop, and WHO region. Once the dataset is loaded into the model, its resultant plot matrix is shown in Fig. 6. The plot matrix displays the correlation between the attributes of the dataset. The points will fall into a line or curve if the variables are connected. The closer the points are to the axis, the stronger the connection.

Table 2. Major accuracy measures from the 10-fold cross-validation performance of the models built using the supervised machine learning algorithms.

Algorithm/ performance	Accuracy	Precision	Recall	F-Measure	ROC
Naive Bayes	49.45%	0.581	0.495	0.489	0.745
Bayes Net	76.08%	0.733	0.761	0.745	0.935
JRip	73.91%	0.704	0.739	0.716	0.884
SVM	77.17%	**0.803**	0.772	0.770	0.894
RandomForest	77.17%	0.722	0.772	0.737	**0.942**
K-NN	76.08%	0.762	0.761	0.756	0.858
J48	**78.80%**	0.772	**0.788**	**0.778**	0.887

V. Result Analysis

As mentioned before, the proposed methodology consists of seven different classification algorithms: Naive Bayes, J48, Bayes Net, JRip, SVM, RandomForest, and K-NN. When the K-NN algorithm is applied, the total number of instances with correctly classified instances is 140, while the incorrectly classified instances are 44, resulting in an accuracy of 76.087%. The precision and recall are 0.762 and 0.761, respectively. For the Bayes Net algorithm, the total number of instances with correctly classified instances is 140, and there are 44 incorrectly classified instances. The accuracy is 49.4565%, with precision and recall of 0.733 and 0.761, respectively. Regarding the Naive Bayes algorithm, the total number of instances with correctly classified instances is 91, and there are 93 incorrectly classified instances, resulting in an accuracy of 68.53%. The precision and recall are 0.581 and 0.495, respectively. For the RandomForest algorithm, the total number of instances with correctly classified instances is 142, and there are 42 incorrectly classified instances, with an accuracy of 77.1739%. The precision and recall are 0.722 and 0.772, respectively. Similarly, for the SVM algorithm, the total number of instances with correctly classified instances is 142, and there are 42 incorrectly classified instances, with an accuracy of 77.1739%. The precision and recall are 0.803 and 0.772, respectively. When the J48 algorithm is applied, the total number of instances with correctly classified instances is 145, and there are 39 incorrectly classified instances, with an accuracy of 78.8043%. The precision and recall are 0.772 and 0.788, respectively. For the JRip algorithm, the total number of instances with correctly classified instances is 136, and there are 48 incorrectly classified instances, with an accuracy of 73.913%. The precision and recall are 0.704 and 0.739, respectively [12], as per the results listed in Table 2.

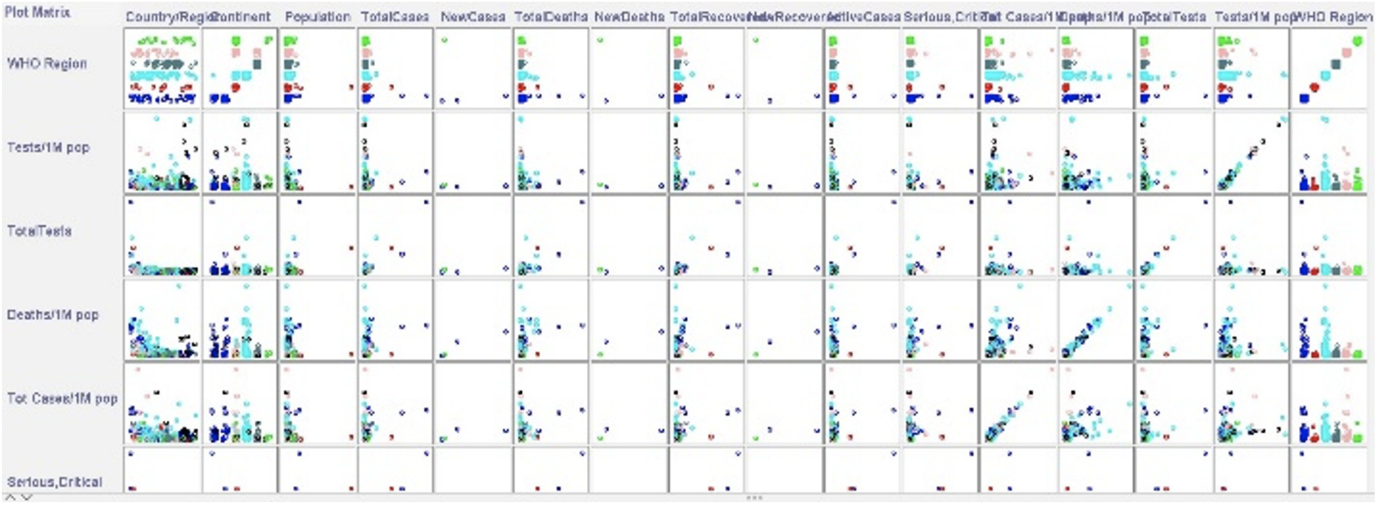
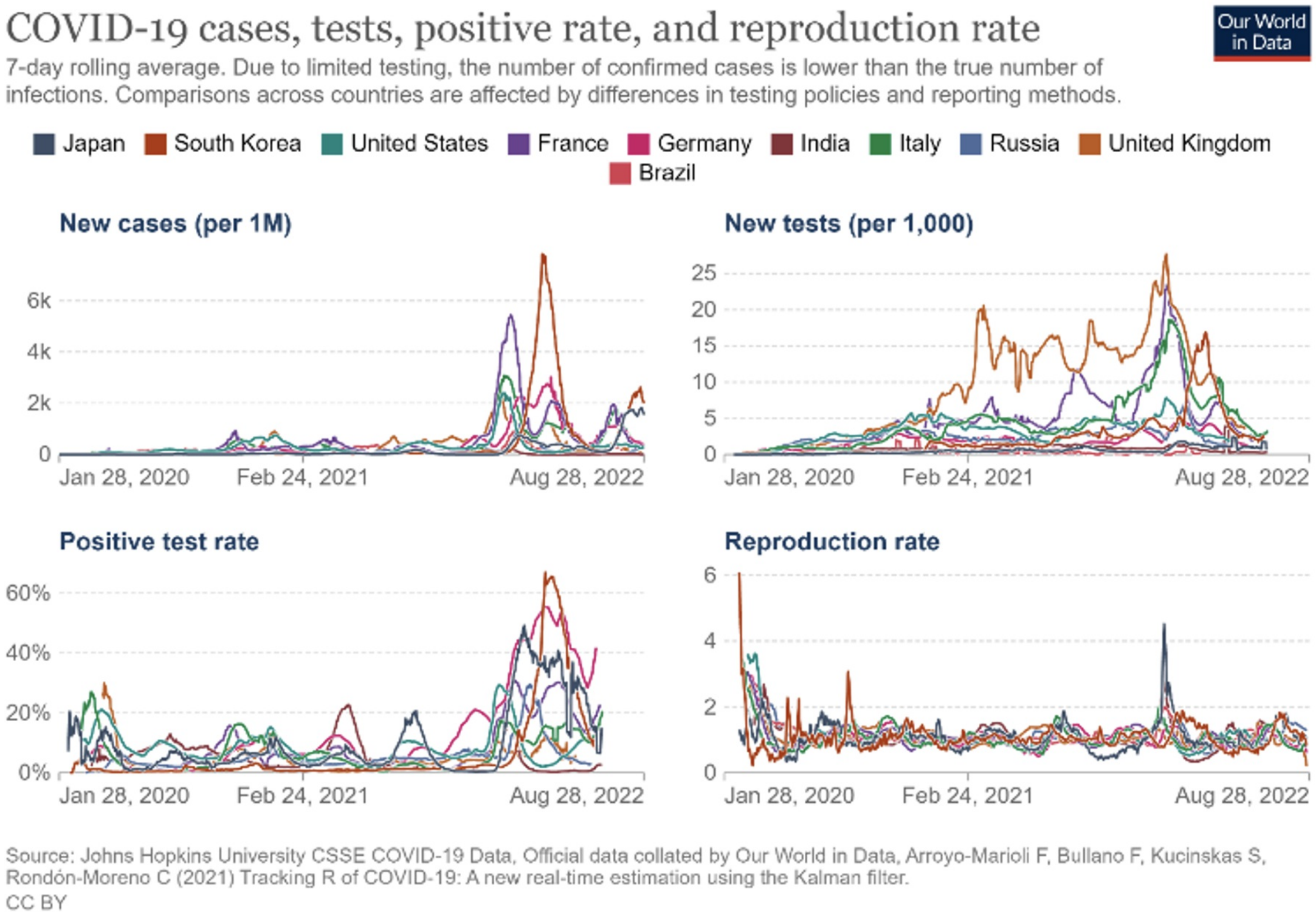


Figure 6. Plot Matrix of Proposed dataset

A. Developing and evaluating models

This research paper aimed to develop and validate ML models based on the most relevant features in determining the risk of COVID-19 mortality. To achieve this, Naive Bayes Classifier, J48 pruned tree Classifier, Bayes Network Classifier, JRip, SVM, RandomForest, and K-NN models were implemented using a dataset. The experimental results, shown in Table 2, indicated that J48 had the best performance among the other seven ML techniques, with an accuracy of 78.80%, precision of 77.20%, and an ROC of around 88.70%. Different studies have also evaluated the application of ML techniques in predicting mortality in COVID-19 patients. The selected features were used as inputs for developing ML-based models for severity, deterioration, and mortality risk analysis in COVID-19 patients. The proposed algorithms demonstrated the ability to predict mortality rates with ROC, accuracy, precision, sensitivity, and specificity rates [13]. J48 ML algorithm performance has been reported as the best among all other ML algorithms tested.

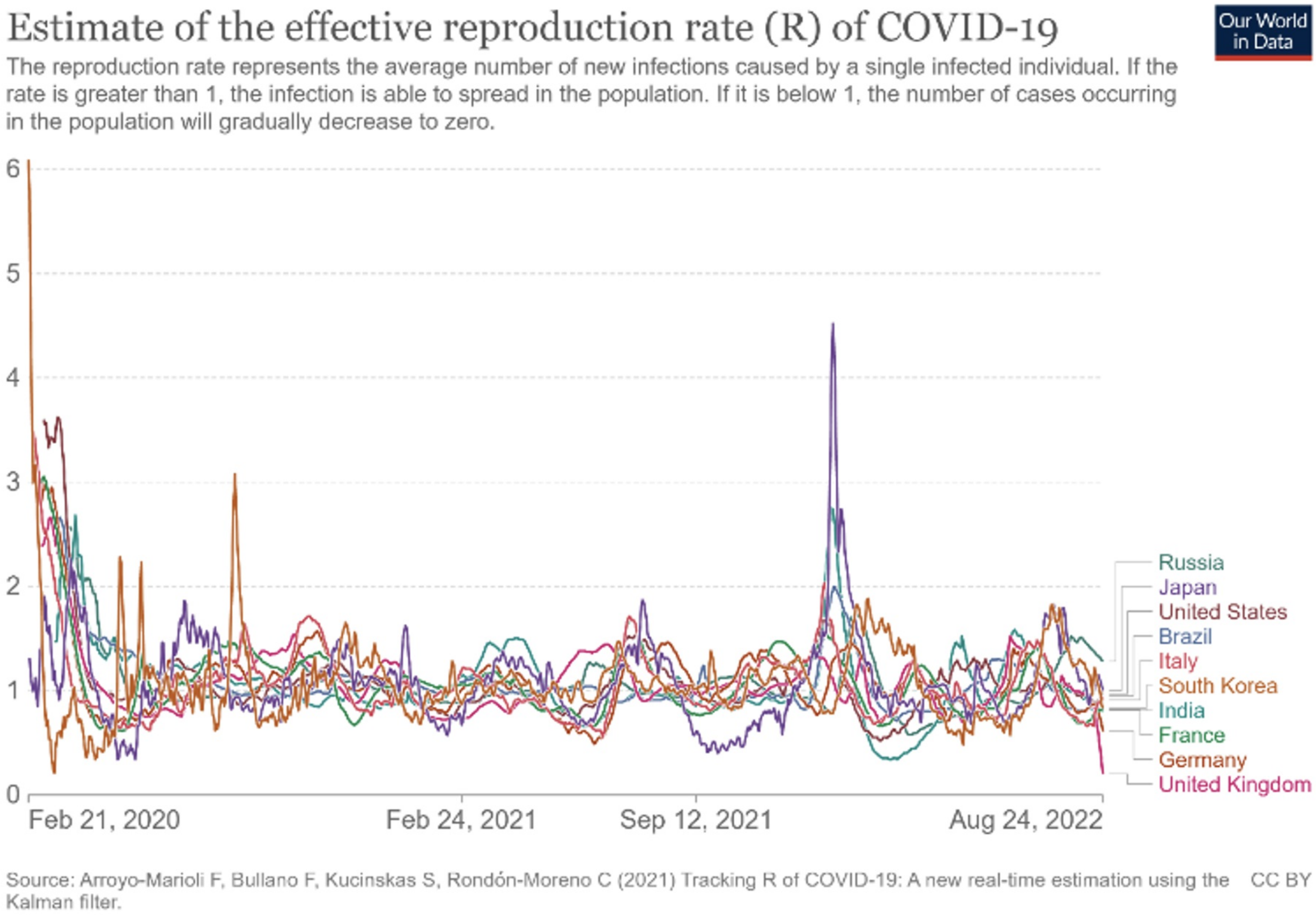


COVID-19 cases, tests, positive rate, and reproduction rate

7-day rolling average. Due to limited testing, the number of confirmed cases is lower than the true number of infections. Comparisons across countries are affected by differences in testing policies and reporting methods.

Source: Johns Hopkins University CSSE COVID-19 Data, Official data collated by Our World in Data, Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C (2021) Tracking R of COVID-19: A new real-time estimation using the Kalman filter.
CC BY

Figure 7. COVID-19 cases, tests, positive rate, and reproduction rate



Estimate of the effective reproduction rate (R) of COVID-19

The reproduction rate represents the average number of new infections caused by a single infected individual. If the rate is greater than 1, the infection is able to spread in the population. If it is below 1, the number of cases occurring in the population will gradually decrease to zero.

Source: Arroyo-Marioli F, Bullano F, Kucinskas S, Rondón-Moreno C (2021) Tracking R of COVID-19: A new real-time estimation using the Kalman filter. CC BY

Figure 8. Estimate of the effective reproduction rate (R) of COVID-19

VI. Conclusion

This research paper serves the purpose of predicting disease spread abnormalities and high recovery rates in humans. The COVID-19 pandemic has had a tremendous impact on people's lives globally, and the number of infected patients has significantly increased. Results have shown that controlling the spread of infection is crucial. However, the vaccination strategy for specific population groups has not been described or evaluated assuming that the number of vaccines is sufficient. Among the six ML algorithms, the J48 model demonstrated the best classification accuracy. The proposed model can be effectively used to predict the mortality risk of COVID-19.

VII. Future Work

The future work aims to provide better significant results using several other machine learning models to find estimates that help online datasets, clinicians, medical, and governmental organizations prepare in real-time for future pandemic diseases. It is strongly considered that with large and public databases, researchers can develop better ML models to accurately detect and predict COVID-19. In the future, the performance of our model will be enhanced if tested on more classification techniques with larger and diverse datasets.

Acknowledgments

I would like to acknowledge the help of the public databases, worldometer_data.csv (https://www.kaggle.com/datasets/imdevskp/corona-virus-report?select=worldometer_data.csv), and for graphical representation (https://www.worldometers.info/coronavirus/) for providing the data used in this research paper.

Declarations

Ethical Approval

I, Raja Sarath Kumar Boddu, affirm that for this manuscript, "SARS-CoV-2 virion, debacle humane: An Analytical Approach," the following criteria have been fulfilled:

- This material represents the authors' original work and has not been previously published elsewhere.
- The paper is not under consideration for publication elsewhere.
- The paper accurately reflects the authors' research and analysis.
- Proper credit is given to co-authors and co-researchers for their meaningful contributions.
- The results are appropriately placed in the context of prior and existing research.
- All sources used are properly disclosed with correct citations. Any direct quotations are indicated using quotation marks and proper references are provided.
- I have been actively involved in substantial work leading to the paper and take public responsibility for its content.

Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Authors' Contributions

After publishing a couple of research papers on the coronavirus during the lockdown, I was curious to know the total weight of the corona virion spread globally, even though it may be very little, possibly in nanograms. I conducted extensive research and found relevant information in one of the research papers listed in the references [3]. The fact is that during peak infection, an infected person carries an estimated 1 billion to 100 billion virions, with a total mass not exceeding 0.1 mg and a diameter size ranging from 20 nm to as large as 500 nm.

Driven by my enthusiasm for research, I decided to write a full research paper on how the corona virion devastated mankind. I started by researching and writing the general introduction, literature search, and more. Collecting data took considerable time. In this process, I designed the study, managed the data, conducted data analysis, and wrote the first draft of the manuscript up to its final version.

Funding

No funding was received for this research.

Availability of Data and Materials

The dataset used for this research, worldometer, was collected from Kaggle[11], containing comprehensive COVID-19 information for all nations. The data was applied to the seven machine learning algorithms mentioned and evaluated using the WEKA tool. The results were tabulated. Additionally, I found a graphical user interface website, worldometer.info, to generate graphical views of the data, providing pictorial representations for better understanding by the readers. Some of these graphical representations have been incorporated into this research paper.

References

1. a, b *World Health Organization. (2020). Report of the WHO-China joint mission on coronavirus disease (COVID-19). Retrieved from https://www.who.int/docs/default-source/*
2. ^ *Sender, R., Bar-On, Y. M., Gleizer, S., Bernshtein, B., Flamholz, A., Phillips, R., & Milo, R. (2021). The total number and mass of SARS-CoV-2 virions. Proceedings of the National Academy of Sciences of the United States of America, 118(25), e2024815118. https://doi.org/10.1073/pnas.2024815118. PMID: 34083352; PMCID: PMC8237675.*
3. a, b, c *Kumar, A., Narayan, R. K., Prasoon, P., Kumari, C., Kaur, G., Kumar, S., Kulandhasamy, M., Sesham, K., Pareek, V., Faiq, M. A., Pandey, S. N., Singh, H. N., Kant, K., Shekhawat, P. S., Raza, K., & Kumar, S. (2021). COVID-19 Mechanisms in the Human Body—What We Know So Far. Frontiers in Immunology, 12. https://doi.org/10.3389/fimmu.2021.693938*
4. a, b *Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., Hasell, J., Macdonald, B., Beltekian, D., & Roser, M. (2020). Coronavirus Pandemic (COVID-19). Published online at OurWorldInData.org. Retrieved from https://ourworldindata.org/coronavirus.*
5. ^ *Sharma, D. K., Chakravarthi, D. S., Boddu, R. S. K., Madduri, A., Ayyagari, M. R., & Khaja Mohiddin, M. (2023). Effectiveness of Machine Learning Technology in Detecting Patterns of Certain Diseases Within Patient Electronic Healthcare Records. In S. Yadav, A. Haleem, P. K. Arora, & H. Kumar (Eds.), Proceedings of Second International Conference in Mechanical and Energy Technology. Smart Innovation, Systems and Technologies (Vol. 290). Springer, Singapore. https://doi.org/10.1007/978-981-19-0108-9_8.*
6. ^ *Boddu, R. S. K., Santoki, A. A., Khurana, S., Koli, P. V., Rai, R., & Agrawal, A. (2022). An analysis to understand the role of machine learning, robotics and artificial intelligence in digital marketing. Materials Today: Proceedings, 56(4), 2288-2292. https://doi.org/10.1016/j.matpr.2021.11.637.*
7. ^ *Ramana, B. V., & Boddu, R. S. K. (2019). Performance Comparison of Classification Algorithms on Medical Datasets.*

In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC) (pp. 0140-0145). IEEE. https://doi.org/10.1109/CCWC.2019.8666497.

8. ^*Gulati, K., Kumar, S. S., Boddu, R. S. K., Sarvakar, K., Sharma, D. K., & Nomani, M. Z. M. (2021). Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to COVID-19 pandemic. Materials Today: Proceedings, 51, 38-41. https://doi.org/10.1016/j.matpr.2021.04.364.*
9. ^*Waikato University. (n.d.). Weka 3: Data Mining Software in Java. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/.*
10. ^*Omae, Y., Sasaki, M., Toyotani, J., Hara, K., & Takahashi, H. (2022). Theoretical Analysis of the SIRVVD Model for Insights Into the Target Rate of COVID-19/ SARS-CoV-2 Vaccination in Japan. IEEE Access. https://doi.org/10.1109/ACCESS.2022.3168985.*
11. a, b*Kaggle. (n.d.). COVID-19 Dataset. Retrieved from https://www.kaggle.com/datasets/imdevskp/corona-virus-report.*
12. ^*Karlinsky, A., & Kobak, D. (n.d.). The World Mortality Dataset: Tracking excess mortality across countries during the COVID-19 pandemic. medRxiv. https://doi.org/10.1101/2021.01.27.21250604.*
13. ^*Nguyen, N., Strnad, O., Klein, T., Luo, D., Alharbi, R., Wonka, P., Maritan, M., Mindek, P., Autin, L., Goodsell, D. S., & Viola, I. (2021). Modeling in the Time of COVID-19: Statistical and Rule-based Mesoscale Models. IEEE Transactions on Visualization and Computer Graphics, 27(2). https://doi.org/10.1109/TVCG.2020.3030415.*