Content available at: https://www.ipinnovative.com/open-access-journals

Indian Journal of Microbiology Research

Journal homepage: https://www.ijmronline.org/

Original Research Article

A stochastic analysis and bibliometric analysis of COVID-19

Rakshita Chaudhary[1], Nisha Gaur[2*], Mohit Yadav[3], Siddharth Srivastava[4], Vanshika Chaudhary[5], Mohd Asif Shah[6]

[1] *Central Drugs Standard Control Organisation, MoHFW, New Delhi, India*
[2] *Dept. of Biotechnology, Gautam Buddha University, Greater Noida, Uttar Pradesh, India*
[3] *Dept. of Human Resource Management, Jindal Global Business School (JGBS), Sonipat, Haryana, India*
[4] *APL Apollo Tubes Ltd, Bulendsher, Uttar Pradesh, India*
[5] *Banasthali University, Radha Kishnpura, Rajasthan, India*
[6] *Kardan University, Kabul, Afghanistan*

ARTICLE INFO

Article history:
Received 03-07-2024
Accepted 09-11-2024
Available online 09-12-2024

Keywords:
Machine learning
COVID- 19
Neural network
Pearson's correlation
SARS- CoV- 2
Principal component analysis (PCA)

ABSTRACT

Background: COVID-19 a novel Corona Virus Disease which was caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Corona Virus 2) continues to pose a critical and urgent threat to global health. When an infected person comes in contact with a normal individual or when the infected person sneezes or coughs, the virus that triggers COVID-19 spreads.

Aim and Objective: The spread of novel SARS-CoV-2 was increasing, and the threats caused by it were becoming more severe in 2021. To counter the disease and save countless lives in danger, it is necessary to predict the trend of the number of cases and deaths and then implement the policies accordingly.

Materials and Methods: In this paper, the trends of the growth rate of worldwide cases and deaths were studied, and the future growth for 100 days was predicted using the Neural network model and Polynomial Regression model. For efficient planning, the countries were grouped using Principal Component Analysis and the predictions were made. The cases and deaths in different countries and states were related through the Pearson coefficient, and the heat maps were studied. Additionally, in this paper, a case study to predict the trend of cases, number of deaths and recoveries in India was also performed.

Result: The Indian states were grouped into four groups based on the Principal Component Analysis (PCA) results, and relevant remarks and trends were suggested. The growth of cases and deaths was studied, and the peaks were predicted for the next 200 days. In recent months, COVID-19 has generated a significant deal of anxiety as a global pandemic, and an increasing number of studies have been published in this area.

Conclusion: Consequently, a bibliometric examination of these papers may offer insight into current research hot subjects and trends. We are the first to join stochastic analysis of COVID-19 effects with bibliometric analysis of COVID-19. This prediction, if taken into consideration strategically during the planning of preventive measures of COVID-19 can help to reduce the cases to a great extent.

1. Introduction

COVID-19 a novel Corona Virus Disease which was caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Corona Virus 2) continues to pose a critical and urgent threat to global health. In December 2019, the first case of COVID-19 was discovered in the Hubei province of the People's Republic of China and spread worldwide (UMBERS 2020). Since its discovery, the disease has spread around the world and in March 2020 the World Health

* *Corresponding author.*
E-mail address: gaurnisha2007@gmail.com (N. Gaur).

https://doi.org/10.18231/j.ijmr.2024.058
2394-546X/© 2024 Author(s), Published by Innovative Publication.

Organization (WHO) declared it a pandemic as the overall number of patients confirmed to have the disease has exceeded 7,50,178 in 144 countries. However, the number of infected people was probably much higher, and more than 36398 people have died from COVID-19. In a short period, the epidemic has spread to over 200 countries. When an infected person comes in contact with a normal individual or when the infected person sneezes or coughs, the virus that triggers COVID-19 spreads. Due to these reasons, the overall number of patients has increased to 425,659,334 around the world and 351,561,123 people have died as of 21st February 2022 (UMBERS 2020).[1]

This virus affects the lower and upper respiratory tract along with cough, fever, weakness, shortness of breath, and lack of taste and scent. Based on MERS (Middle East Respiratory Syndrome) and SARS (severe acute respiratory syndrome) incubation, the infected person develops signs within 2-14 days, and as a result of this patients are at high risk of death. To confront COVID-19's rapid spread most countries resorted to complete lockdown to control the outbreak of COVID-19 but unfortunately at a high human and economic cost.[2]

With the massive loss of humans and destructive economic impact, the second wave of the pandemic presents a looming threat to society. Unlike India, due to a good testing and tracing system South Korea was able to control the situation. Additionally, European countries also have faced the second wave which was worse than the first wave.[3]Figure 1 showed the total active COVID-19 cases, total recovered and total death around the world.

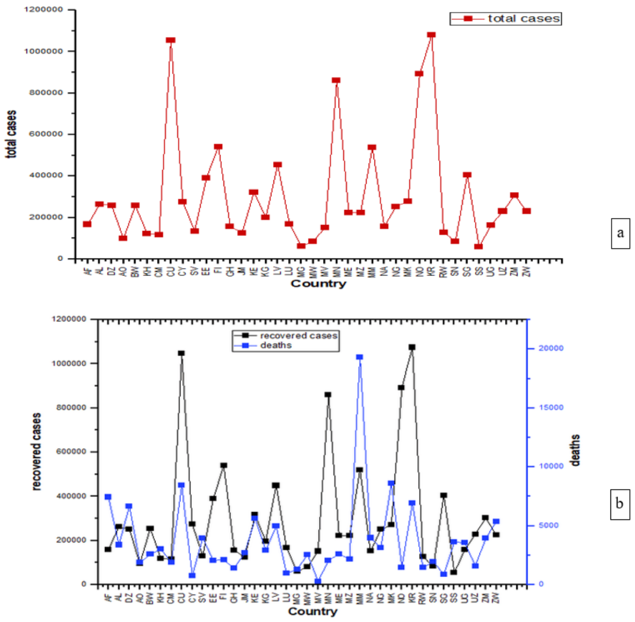

Figure 1: (**a**) Total number of cases, (**b**) deaths and recovered cases due to COVID-19 in the world

Despite all the protective measures taken by all countries around the world, as of the time of writing this article (February 12, 2022), 425,659,334 total cases have been reported with 351,561,123 deaths. Hence, it showed that COVID-19 is highly contagious and spreads very rapidly. Most countries were lifting lockdown restrictions slowly and travel restrictions so the risk of new infectious cases was very high. Though the development of vaccines has reduced the chances of infection overpopulation, reduced medical facilities and fear of vaccines are some factors that will increase the chances of infection during the next waves. Hence, this paper aims to simulate the possible wave outbreak in countries around the world. Mathematical modeling is one of the efficient tools to study contagious disease spread, its persistence, or when the world will return to its earlier situation.[4]

2. Materials and Methods

This section will cover the detailed data set of the study, principal component analysis, Pearson Coefficient, Neural Network, and Polynomial Regression Curve fitting. In addition, these mathematical tools were also used to predict the third wave in India. So, this present study involves the COVID-19 prediction around the world with 195 countries in consideration. In the Web of Science collection database, the global literature regarding COVID-19 published between 2020 and 2023 was searched. The search terms "COVID-19," "Novel Coronavirus," "2019-nCoV," and "SARS-CoV-2" were used to find the pertinent publications. The articles' bibliometric analysis was carried out using a VOS viewer.

2.1. *Dataset*

This study contains the data sets of the number of COVID-19 patients, number of deaths, and total number of recovered patients due to this disease around the world till 12th February 2022.

2.2. *Principal component analysis (PCA)*

PCA is used for the conversion of a large dataset that is in the form of a multidimensional matrix into a smaller matrix for better computation. It removes the redundancy in the data and makes the data smaller by transforming the entire matrix into linearly uncorrelated variables. These variables are called principal components and they explain the variation in data. This leads to the removal of redundancy and the similar components of the dataset are grouped which can also be used to get insights from a huge multidimensional data.

2.3. Principal component analysis of algorithm

Step 1: The entire dataset containing p countries having the values for q dates is converted to a $p \times q$ matrix M.

Step 2: Calculation of the Eigen values and the covariance matrix Ω

$\Omega = \frac{(\beta^T \ \beta)}{p}$

where β for the ith value is defined as $M_i \ - \ \frac{1}{p}\sum_1^p M_j$

Step 3: Calculate the Cumulative Explained Variance Ratio η(t) associated to the tth sample, where λ is the eigenvalue of the eigenvector e

$\eta(t) = \frac{(\sum_1^t \ \lambda_a)}{(\sum_1^p \ \lambda_a)}$

Step 4: To convert the p×q matrix as p×t matrix, choose the Eigenvectors whose η(t) > 0.95 where t is the number of Eigenvectors chosen

Step 5 : Hence, the final reduced dataset in terms of principal components is represented as

$\alpha = M_t \ E_t$

where E$_t$ is the set of t Eigenvectors

In PCA, the extraction of the features (data points) is done based on the variance and the newer dataset obtained has a higher variance than the original dataset. This leads to the development of a far more compact nonredundant feature matrix that is more useful and requires less computation power. This is done by finishing the eigenvalues and the eigenvectors of the covariance matrix. The largest eigenvalues have the strongest correlation with the dataset and they are called the principal component.[5]

2.4. Neural network

The neural network is a supervised machine learning model which is based on minimizing the cost functions to get the best-fit curve for regression models. Generally used for classification models, Neural networks can also be trained to predict the numerical values directly and hence can be used for regression. The data was extracted from https://www.worldometers.info/coronavirus/ using Beautiful Soup by text mining. This data was fed into the neural network to predict the trend of the growth of the curve. The implementation of the neural network is done using the MLP Regressor of the sci-kit learn package of Python.[5] To predict the trend of the COVID cases worldwide, a neural network of 500 hidden layers was used. The implementation was done using Anaconda in Python. For the neural network, the input data of cases from January 3, 2020, to February 12, 2022, was used.

2.5. Polynomial regression

Polynomial Regression is a technique of fitting the already available curve data with polynomials. A six-degree polynomial model is used in this paper and the trend of cases is fitted using the model. This gives us a prediction of how the data is going to change concerning time by properly optimizing the curve trend.[6] A sample polynomial as shown in equation 1 is used and the already available data is fitted to it to obtain the values of the parameters viz. a, b, c, d, e, f, g.

$f(x) = a + bx + cx^2 + dx^3 + ex^4 + fx^5 + gx^6 -$ (1)

2.6. Pearson correlation coefficient (PCC)

Correlation between sets of data is a measure of how well they are related. One of the most common measures of Correlation is the Pearson correlation coefficient which is the measure of linear correlation between two sets of data.[6] It is the covariance of two variables, divided by the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, where the results are between −1 and +1.

2.7. Pearson's correlation algorithm

Pearson's correlation coefficient, when applied to a population, Greek letter ρ (rho) is commonly used to represent the population correlation coefficient or the population Pearson correlation coefficient. Given a pair of random variables (A, B), the formula for ρ is

$\rho(a,b) = \frac{(cov(A,B))}{(\sigma_a \ \sigma_b \ )}$

Where:

cov is the covariance of the variables A and B

σ_a is the standard deviation of A

σ_b is the standard deviation of B

The formula for ρ can be expressed in terms of mean and expectation. Since

$cov(A,B) = E[(A - \mu_a)(B - \mu_b)$

The formula for ρ can also be written as

$\rho(a,b) = \frac{(E[(A-\mu_a)(B-\mu_b))}{(\sigma_a \ \sigma_b)}$

Where :

σ_a and σ_b are defined as above

μ_a is the mean of A

μ_b is the mean of B

E is the expectation

The formula for the ρ can be expressed in terms of the uncentered moment. Since

$\sigma_a = E(A)$

$\sigma_b = E(B)$

$\sigma_a^2 = E(A - E(A)^2) = E(A^2) - E(A)^2$

$\sigma_b^2 = E(B - E(B)^2) = E(B^2 \ ) - E(B)^2$

The formula for ρ can also written as z

$\rho(a,b) = E(AB) - E(A)E(B)\frac{(E(AB)-E(A)E(B))}{\sqrt{(}E(A^2)-E(A)^2 \ (E(B^2)-E(B)^2}$

2.8. A bibliometric analysis

A statistical technique known as bibliometrics uses mathematical techniques to quantitatively analyses research papers that are concerned with a particular topic. Additionally, it may evaluate the major research fields, assess the quality of the studies, and forecast the course of future research. Nearly all significant research publications are included in the Web of Science (WOS) online database, which also offers built-in analysis tools to provide representative results. The co-occurrence, co-authorship, citation, bibliographic coupling, co-citation, and themes were examined using VOS viewer (version 1.6.10).[7]

3. Result

3.1. *PCA for number of patients*

Based on the number of patients (Figure 2 a) provides the results of PCA methods for the classification of countries. It is a 3D figure as three principal components were analyzed for the countries present in the dataset. Table 1 shows that the first principal component accounts for 97.1% of the variation in the data.

3.2. *PCA for number of deaths*

Table 2 shows the PCA analysis of the death data for 20 countries from 03/01/2020 to 12/02/2022. The eigenvalues are tabulated and the principal components are arranged in the hierarchical order based on the % variation of the data for which the variable accounts to.

3.3. *Neural network prediction*

The parameters of the neural network are shown in Table 3. Figure 2a shows the 3D plot of the prediction of cases worldwide using the neural network and the polynomial regression model used.[8]

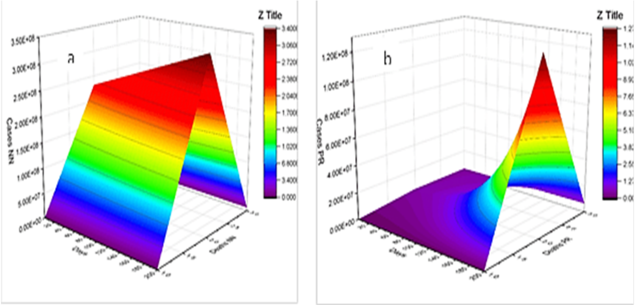

Figure 2: Prediction of trend of cases and death by (**a**) Neural network and (**b**) Polynomial Regression for 100 days.

The data is taken for 506 days and the neural network with the parameters shown is applied and the R^2 value is calculated.

3.4. *Polynomial regression*

The R^2 values as depicted in Table 4 are obtained which indicate an almost perfect fit. Figure 4 shows the box plot of the prediction of the cases and the deaths worldwide with the normal distribution of the data using Polynomial regression.

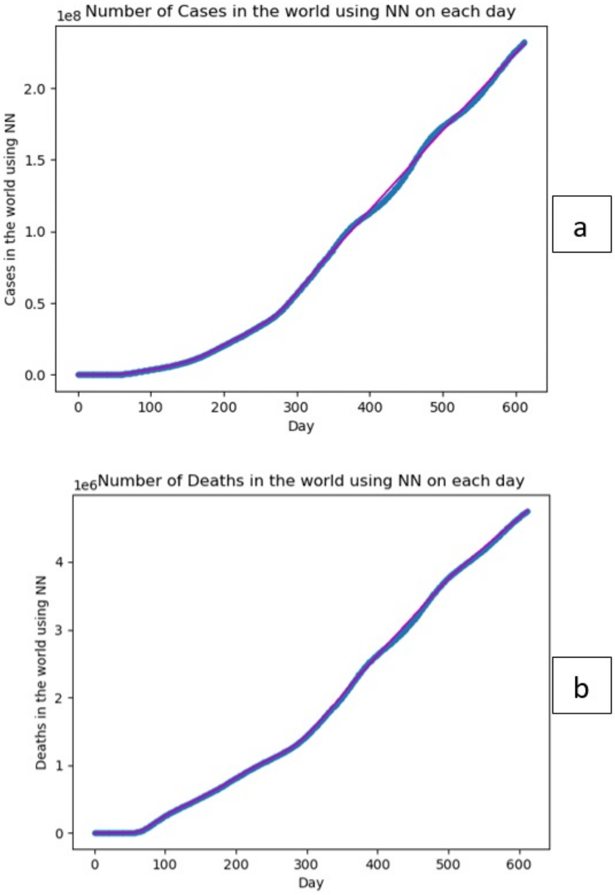

Figure 3: Curve fitting model by Neural Network for Worldwide (**a**) Cases and (**b**) Deaths

3.5. *Pearson's correlation*

Principal components analysis performed by doing transformations into some sets of variables values. The results were shown in the graphs where the countries and state axis were plotted in the first two component analysis and the variance was introduced. Principle component is shown by the variables such as total cases, total deaths and total recovered cases. The first component analysis showed the variability factor of nearly about 35%.[9]

4. Discussion

4.1. *PCA for number of patients*

According to the principal components, it can be seen that the countries can be divided into two groups. The first group consists of South Africa, Peru, Mexico, Brazil, India, Columbia, Russia and Argentina. The second group consists of the United States, Spain, Iran, Indonesia, France, Netherlands, United Kingdom, Germany, Poland, and Turkey.[10] So, the countries in the same group may experience similar growth patterns if other conditions

Table 1: Principal component analysis for number of patients in the countries

Principal Component Number	Total average Eigen value	Total average Percentage of Variance (%)	Total average Cumulative (%)
1-222 cases of different countries	0.995495554	0.450450502	99.38991608

Table 2: Principal component analysis for number of deaths of countries

Principal Component Number	Total average Eigenvalue	Total average Percentage of Variance (%)	Total average Cumulative (%)
1-222 cases of different countries	0.945945856	0.450450502	99.84036653

Table 3: Neural network parameters for COVID-19 prediction

Type of Framework	Neural Network
Alpha	1e-6
Lear Ning	0.0008
Number of hidden layers	500
Batch size	32
Number of Iterations	19958
Tolerance	1e-6
Activation Function	RELU
Optimizer	Adam

Table 4: R^2 value of the predictions

R^2 values	Cases by PR	Cases by NN	Deaths by PR	Deaths by NN
World	0.999999998	0.999999979	0.999966166	0.999959604
India	0.999999833	0.99999891	0.993258907	0.999963272

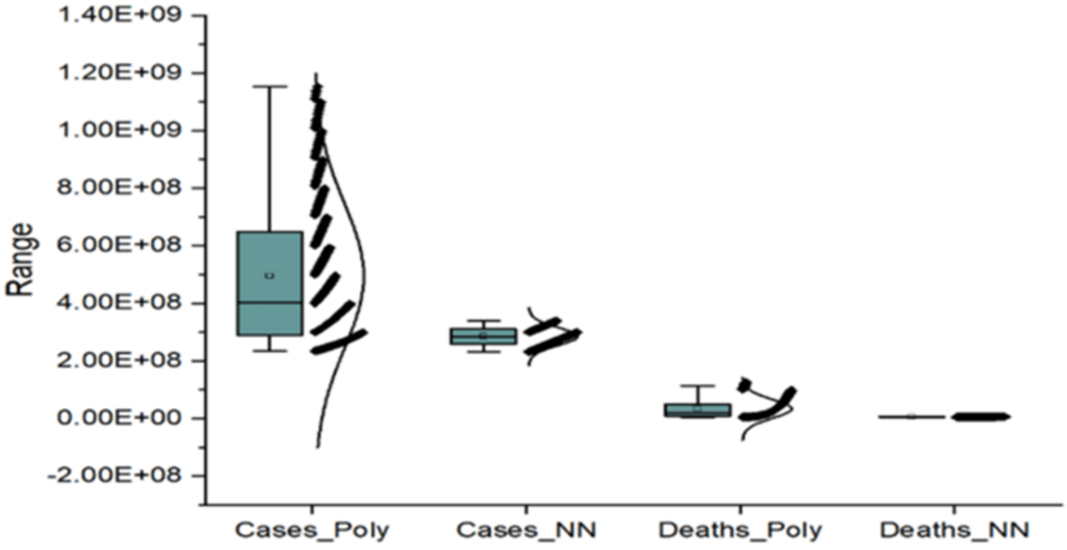

Figure 4: Box plot of the prediction of the cases and the deaths worldwide with the normal distribution of the data using Polynomial regression

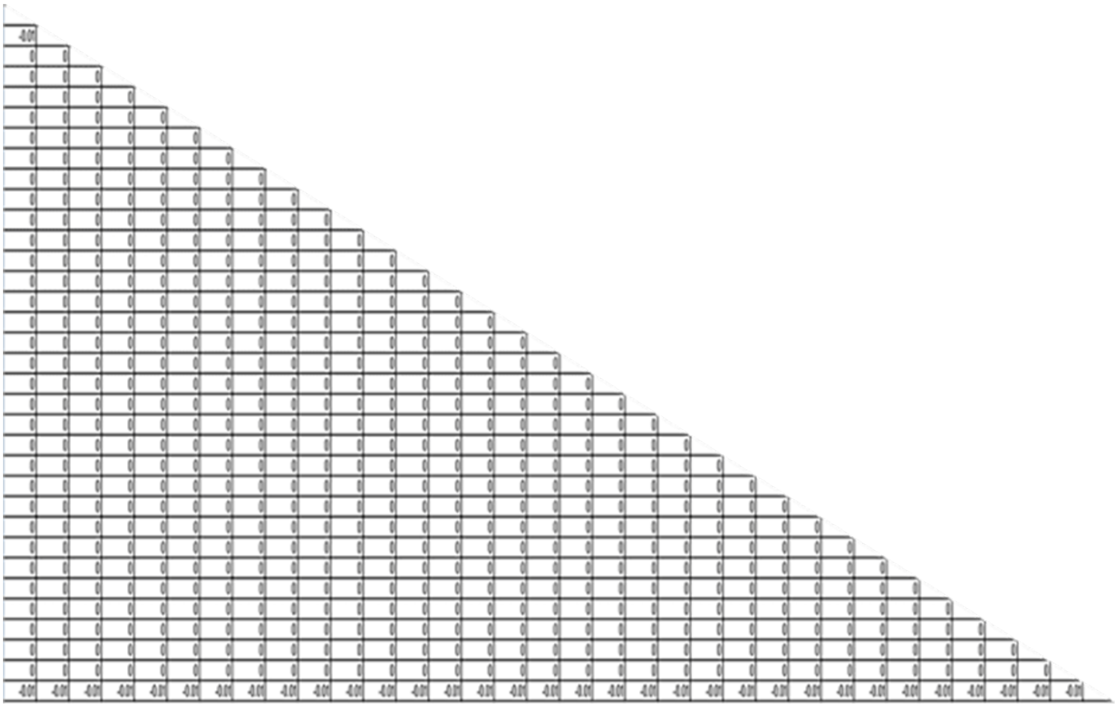

Figure 5: Pearson correlation of the world Total confirmed cases in number plot method. The number is showing how the countries are correlated with each other in terms of confirmed cases method

remain the same. Day-wise data has been taken and a 3-D graph has been drawn depending on the top three principal components. The more the magnitude of the eigenvalue, the more it accounts for the variation in the data. Chen C. et al. also studied the PCA for number of patient and concluded that there was significant difference between number of patients among the countries.[11,12]

4.2. PCA for number of deaths

Table 2 shows that when we divide the countries into groups by Principal Components based on the total number of deaths, the first group consists of Ukraine, Poland, Turkey, India, Brazil, Germany, and Italy. The second group consists of Russia, Indonesia, Columbia, Argentina, France, Spain, Peru, Netherlands, Iran, the United States, the United Kingdom, Mexico, and South Africa. A significant difference between the countries' groups can be seen in cases and deaths.

4.3. Neural network prediction

In the X axis of Figure 3, the days are present, and the Y and Z axis have the number of cases predicted by the neural network model and the polynomial regression model respectively. The Polynomial model predicts that the cases were increased for the next 100 days and the neural network model predicts a slow rise in the number of cases.

Figure 2b shows the 3D contour plot of the prediction of deaths worldwide for a period of 100 days. The deaths were predicted to increase with respect to rising days and reach to a maximum of 6888152 according to the Polynomial model and 4974181 according to the neural network model. So, it can be inferred that the neural network model predicts low deaths than the polynomial model for the worldwide deaths.[13]

Figure 3 show the trend of the curve-fitting data of the worldwide cases (Figure 3a) and deaths (Figure 3b) by neural network model. The Figure 4 shows the box plot of the prediction of the cases and the deaths worldwide with the normal distribution of the data using Neural network model as discussed above.

4.4. Polynomial regression

On implementing polynomial regression model using Python 3.8 and sci-kit learn module, satisfactory results were obtained. The R^2 values are shown in Table 4. For the world dataset, the equation is obtained to be f(X) = 461933.758 - 75264.227X^1 + 2132.323X^2 - 17.387X^3 + 0.083X^4 - 0.000X^5 + 0.000X^6 for the trend of cases with an R^2 value of 0.999. And for the trend of worldwide deaths, the equation comes out to be f(X) = 75895.500 - 9159.493X^1 + 211.012X^2 - 1.424X^3 + 0.005X^4 - 0.000X^5 + 0.000X^6 with an R^2of 0.999756857. Figure 3 show the trend of the curve-fitting data of the worldwide cases (Figure 3a) and deaths (Figure 3b) by the polynomial regression model.

4.5. Pearson's correlation

Figure 5 depicts the heat map of Pearson correlation coefficients amongst 20 countries in the world. It can be inferred from the graph that the countries having the value of the Pearson correlation coefficient as less than 0.90 are comparatively less correlated with each other than the other having a value greater than 0.95. The nature of the pandemic is that the cases were increased everywhere and a correlation can be established easily due to the same trend of growth everywhere.[14] So, the usual value of Pearson coefficient is greater than 0.50 cannot be applied here. As the Pearson correlation coefficient of United Kingdom and India is 0.86, it can be inferred that they are less correlated than the other counties that have above it 0.90 and even above 0.95 showing high correlation. This includes countries like France and India 0.93, Russia and France 0.98, etc. Some countries here even have a Pearson correlation coefficient of 1, for example, Italy and France and this depicts a great similarity in the trends in the number of cases in these countries suggesting adoption of similar measures.[15] With passage of the time it can be considered the correlation matrix and found that the countries and states having confirmed cases were not uniform in the targeted places. So, it was observed that the eigenvalues for death were greater than the confirmed cases.

Countries, where the Pearson correlation coefficient is less than 0.90, are approximately less correlated with each other when compared to the countries with higher values. It can be seen that the Pearson coefficient of United Kingdom and India is 0.84. And hence, they are less correlated than the ones which higher values like France and Russia. Generally, a higher correlation corresponds to a value of more than 0.95 in this case when the cases are increasing all around the world. This includes countries like Brazil and USA 0.96, Russia and France 0.99, etc. There are some countries where the Pearson correlation coefficient is 1 example Netherlands and Brazil that means the trend of deaths in these countries are similar and similar growth can be expected. The number is showing how the countries are correlated with each other in terms of deaths.[16]

5. Case Study of India

In India, when a student returned back from Wuhan, the province of China the COVID-19 was first confirmed in the state Kerala on January 27, 2020. To prevent the possibility of a stage 3 human to human transmission that can stimulate the spread of the coronavirus disease the Government of India has incorporated social distancing as a precautionary measure. In addition, to make aware the people about the peculiar epidemiological traits compared with previous two epidemics of SERS-CoV and MERS-CoV Indian Government also imposed a 14 hours voluntary public curfew (Janta Curfew) on March 22, 2020. However, a 21 days nationwide lockdown from March 25, 2020 to April 14, 2020 has declared by Government of India to prevent the spread of coronavirus disease among human (1.3 billion India population).[17]

To combat against COVID-19 pandemic in India in second wave, the lockdown has been extended up to May 03, 2020. This was further extended to 17th May 2020 by the Government of India and then NDMA finally extended this to 31st May 2020. Finally, from 8 June 2020, services began to resume in the garb of "Unlock 1". But as the restrictions were eased later due to the decrease in the number of cases, people came out in large numbers and the policies of the government supported them. In October 2020, a model suggested that the COVID had peaked in India and Indians had achieved herd immunity. Despite India's coronavirus numbers cross 1 crore-mark on December 19, people did not care because of rumors and being weary of staying inside their houses. This led to carelessness by the people only for the second wave to bounce back higher bringing the tally of the total cases to more than 30 million. Figure 7 shows the total number of cases, deaths and recovered cases due to COVID-19 in India states as on 27 June 2021. During the period of second wave, there was no proper vaccination or healthcare. But later, there were several drugs have been researched for example – by DRDO and other countries have also prepared medicines for COVID.[18] And as a result of proper healthcare and large-scale vaccinations, the death rate has decreased and the recovery rate is higher. But there were some gullible people who were afraid of the vaccines due to the rumors spread with respect to the side effects of the vaccines. Hence, due to the effect of several factors, the third wave was predicted to have a lower number of cases when compared to the second wave but still was higher than the first wave as the growth rate was increasing steadily.[19]

5.1. PCA

The PCA data of number of cases in India which implies that the first principal component accounts for 95.77% of the variation in the data. According to the principal components, it can be seen that the states can be divided into four groups as given the Table 5. Further, the states are divided into four groups shown in the Table 6. According to the principal components, the states can be divided into four groups as shown in Table 7.

5.2. Polynomial regression

The polynomial regression model was carried out on the data of daily Indian cases and daily Indian deaths.[19] The

Table 5: PCA result interpretation for Indian cases

S. No.	Groups	Conditions	Remarks	States
1	Group 1	PC2 > 0; PC3 > 0	In all the states cases graph goes from medium to down Except for goa in which there is an increase	Goa, Himachal Pradesh, Karnataka, Lakshadweep, Meghalaya, Mizoram, Puducherry, Sikkim, West Bengal
2	Group 2	PC2 < 0; PC3 > 0	All the states cases graph goes from medium to down except for Tamil Nadu case graph is going down	Arunachal Pradesh, Assam, Jammu, and Kashmir, Manipur, Nagaland, Odisha, Tamil Nadu, Tripura, Andhra Pradesh
3	Group 3	PC3 > 0; PC3 < 0	All the states cases graph is going down except for Maharashtra and Kerala the graph is not showing many changes	Bihar, Chandigarh, Chhattisgarh, Dadra and Nagar Haveli and Daman and Diu, Gujarat, Haryana, Jharkhand, Kerala, Madhya Pradesh, Maharashtra, Punjab, Rajasthan, Uttarakhand, Uttar Pradesh
4	Group 4	PC4 < 0; PC3 < 0	Except for Andaman and Nicobar Island where the graph is going down, all the states are not showing many changes	Andaman and Nicobar Islands, Delhi, Ladakh, Telangana

Table 6: PCA result interpretation for Indian deaths

S. No.	Groups	Conditions	Remarks	States
1	Group 1	PC2 > 0; PC3 > 0	All states the case is going down at some point and at some showing no changes	Chhattisgarh, Goa, Haryana, Jharkhand, Lakshadweep, Rajasthan, Uttarakhand
2	Group 2	PC2 < 0; PC3 > 0	All the states are showing cases going down except for Andaman and Nicobar Islands and Daman and Diu the cases are fluctuating highly going up and down	Andaman and Nicobar Islands, Dadra and Nagar Haveli and Daman and Diu, Delhi, Gujarat, Karnataka, Madhya Pradesh, Maharashtra, Punjab, Telangana, Uttar Pradesh
3	Group 3	PC3 > 0; PC3 < 0	All the states are showing cases graph going from steady to down Which is a good indicator for the future except for Nagaland going From up to down and again up	Arunachal Pradesh, Assam, Bihar, Himachal Pradesh, Kerala, Manipur, Meghalaya, Mizoram, Nagaland
4	Group 4	PC4 < 0; PC3 < 0	All states are showing cases graph doing down which is a good indicator for the future	Chandigarh, Jammu and Kashmir, Ladakh, Odisha, Puducherry, Sikkim, Tamil Nadu, Tripura, West Bengal, Andhra Pradesh

equations are $f(X) = -1622721.034 + 186954.462X^1 - 4354.908X^2 + 38.021X^3 - 0.144X^4 + 0.000X^5 - 0.000X^6$ for the number of cases and $f(X) = -11137.405 + 1355.512X^1 - 32.661X^2 + 0.294X^3 - 0.001X^4 + 0.000X^5 - 0.000X^6$ for the number of deaths.[20] Figure 6 shows the box plot of the prediction of the cases and the deaths in India with the normal distribution of the data using Polynomial regression and Neural Network model.

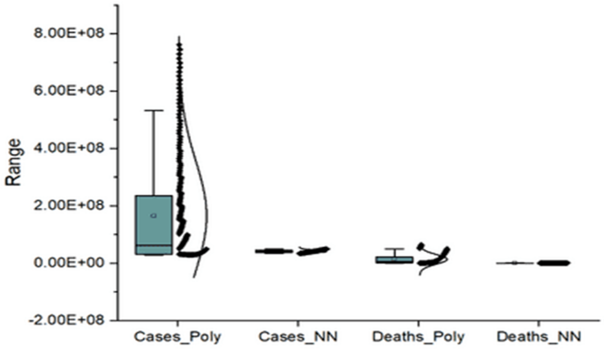

Figure 6: Box plot of the prediction of the cases and the deaths in India with the normal distribution of the data using Polynomial regression

5.3. *Pearson's correlation coefficient*

As depicted in Figure 7, the number of cases and treads in different Indian states are plotted as a numeric heat map of Pearson coefficients. In India, there has been a general increase in the number of cases because of the super spreader nature of the virus.[21] So, high correlation coefficients are expected because the cases

Table 7: PCA result interpretation for Indian recoveries

S. No.	Groups	Conditions	Remarks	Countries
1	Group 1	PC2 > 0; PC3 > 0	All the states are showing the cured person going down due to a decrease in the nod of confirmed cases except for Mizoram there is a higher recovery rate	Goa, Himachal Pradesh, Karnataka, Kerala, Lakshadweep, Meghalaya, Mizoram, Puducherry, Uttarakhand, West Bengal
2	Group 2	PC2 < 0; PC3 > 0	All the states are showing a decrease no of recovery which indicates decrease in no of confirmed cases except for Tamil Nadu which is showing a higher recovery rate	Arunachal Pradesh, Assam, Jammu and Kashmir, Manipur, Nagaland, Odisha, Sikkim, Tamil Nadu, Tripura, Andhra Pradesh
3	Group 3	PC3 > 0; PC3 < 0	All states are showing higher recovery rate as per the graph	Bihar, Chandigarh, Chhattisgarh, Dadra and Nagar Haveli and Daman and Diu, Gujarat, Haryana, Jharkhand, Madhya Pradesh, Maharashtra, Punjab, Rajasthan, Uttar Pradesh
4	Group 4	PC4 < 0; PC3 < 0	All states are showing higher recovery rate as per the graph except for Ladakh and Telangana which is showing a bit slower recovery rate compared to others	Andaman and Nicobar Islands, Delhi, Ladakh, Telangana

were increasing everywhere. But, even amongst the high correlation coefficient, some important correlations can be established by taking geographic and demographic data into consideration.[22] Usually, a coefficient is more than 0.5 considered having a high correlation,[23] but it can be seen here that all the coefficients are above 0.80. It can be said that a correlation of less than 0.90, for example Andaman and Nicobar Island and Punjab having their Pearson correlation coefficient as 0.87 have a comparatively low correlation. Generally, a coefficient above 0.95 can be considered as high similarity for this dataset because of its intrinsic nature of it being increasing everywhere.[23] Some states exhibit a high coefficient like Punjab and Odisha 0.97, Telangana and Puducherry 0.95, etc. And in some cases, the value is equal to 1 like in the case of Uttarakhand and Rajasthan indicating exactly same trend and high correlation.

It can be inferred that the states/union territories, where the Pearson correlation coefficient is less than 0.90, are comparatively less correlated with each other. As an example, if we consider the pair of Tripura and Uttarakhand, their Pearson correlation coefficient for the number of deaths is 0.81. And this means high correlation usually, but with the other state pairs having values more than 0.95 usually, this may seem a bit low. And the Pearson coefficient is meant to have a high value because there is usually migration between the states and the trend of increase and decrease of the cases among different states is similar. Some pairs like Tamil Nadu and Puducherry (0.99), Telangana and Odisha (0.96) have high correlation. And this can be geographically proved as well because these states are near each other and have similar demographics. Some states/union territories have Pearson correlation coefficient of 1, for example, Odisha and West Bengal.[10] That means they can have approximately the similar trend of deaths.

States/Union territories, where the Pearson correlation coefficient is less than 0.90, are comparatively less correlated with each other than the ones with a coefficient of more than 0.95. For example, states/union territories like the Andaman and Nicobar Island and Punjab their Pearson correlation coefficient as 0.87. So, they are less correlated than the other States/Union territories that have above 0.90 and even above 0.95 Pearson correlation coefficient which depicts that they are highly correlated to each other.[24]

This includes States/Union territories like Punjab and Odisha having 0.97, Telangana and Puducherry having a coefficient of 0.95, etc. Some states/union territories have Pearson correlation coefficient of 1, for example, Uttarakhand and Rajasthan. That means they are strongly correlated with similar increase/decrease in the number of confirmed cases. Here we observed the largest value eigenvalues and their corresponding eigenvectors. We observed that the largest eigenvalues of countries are same to the state of disease. The countries and states that were having largest eigenvalues from the start of pandemic (April 2020) decreased gradually (September 2020). After the peak time it gradually started increasing with little fluctuations in their values (January 2021). Again next largest eigenvalues

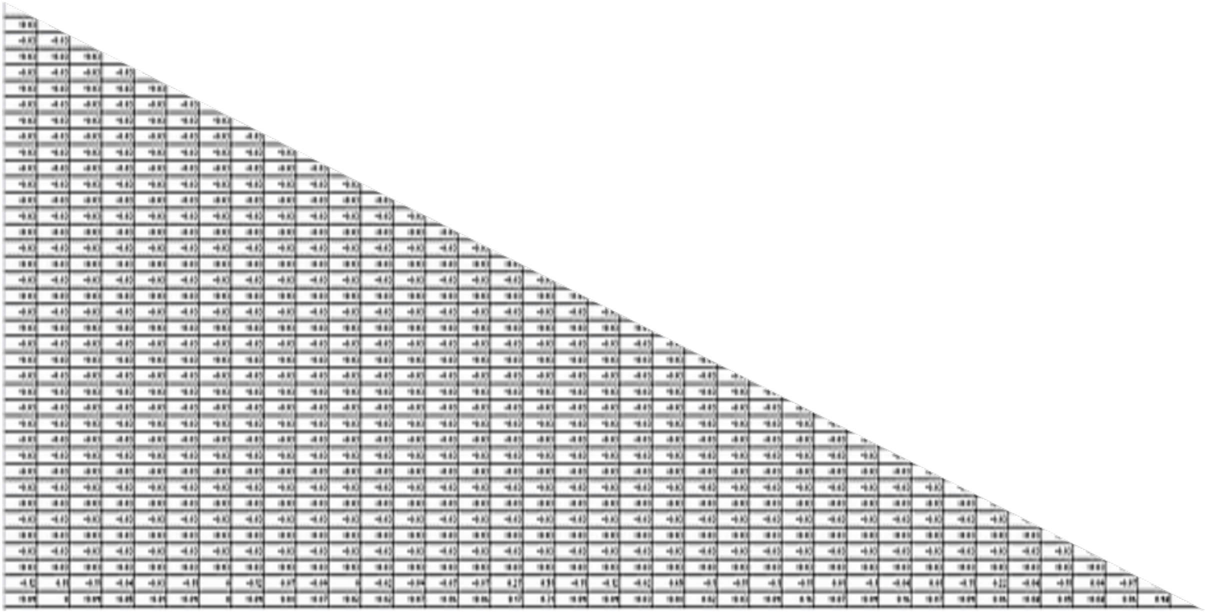

Figure 7: Pearson correlation of the India Total confirmed cases in number plot method. The number is showing how the states are correlated with each other in terms of confirmed cases

were increased from September 2021 to October 2021, by following the trend of the first largest eigenvalues. After the peak time it gradually increases. Death cases were reliable till January 2022.[25]

5.4. Bibliometric analysis of the keywords

The final analysis included keywords that were submitted by the paper's authors and appeared more than five times in the WOS core database. 4348 of the 9,886 keywords met the requirement. The keywords with the highest frequency were "COVID-19" and "coronavirus", which were strongly associated with "pneumonia" and "epidemiology". To display the frequency of the keywords that appeared more than ten times, a word cloud was also constructed.[26] The most prevalent condition was listed as "COVID-19," followed by "pneumonia," "outbreak," and “infection. Figure 8 shows the bibliometric analysis of the keyword used and annual scientific Production for Covid. By using bibliometric analysis, six clusters of the cited references were discovered. Six clusters were displayed in various color schemes. The most often referenced author in the red cluster is Bierman I. The same information was utilized to analyses co-authorship by country. Data on a strong network of international collaboration was filtered and collected using simple criteria. Just 93 of the 195 nations met the requirements as a result.[27]Figure 8 also displays relationship of the co- citations of authors and inferred that the top three clusters, which are colored red, green, and blue, stand in for the research areas of clinical characteristic, disease transmission, and treatment and also depicts the pattern of the countries' corresponding authors, with USA having the strongest influence on the research with the greatest number of publications in a single nation, followed by the China and then India.

6. Conclusion

In this paper, Principal component analysis, Polynomial Regression, and Neural Networks were used to predict the trend of COVID-19 cases both worldwide and for India. Considering the rate of deaths and cases in different countries, similar policies can be adopted for the countries in the same group. A case study pertaining to India has been carried out and the modelling algorithms have been implementing leading to significant results. R^2 values > 0.999 have been obtained for the curve fitting data and the trends for the next 100 days have been predicted. With consideration of 222 countries, the number of cases and deaths in the world is predicted to increase and taking immediate preventive measures are suggested. The Pearson correlation study shows high correlation between the data as the average value of the coefficients is more than 0.99. This study has been performed both for Indian states and the worldwide countries which show a similar result, saying that the reasons for increase in the cases in one state are linked with the other. A special case study on India has been performed and the data from 36 Indian states and Union Territories have been studied. The cases and deaths are predicted to steadily increase in the forthcoming 100 day according to the Neural network model and it can be noted that the polynomial regression model predicts a

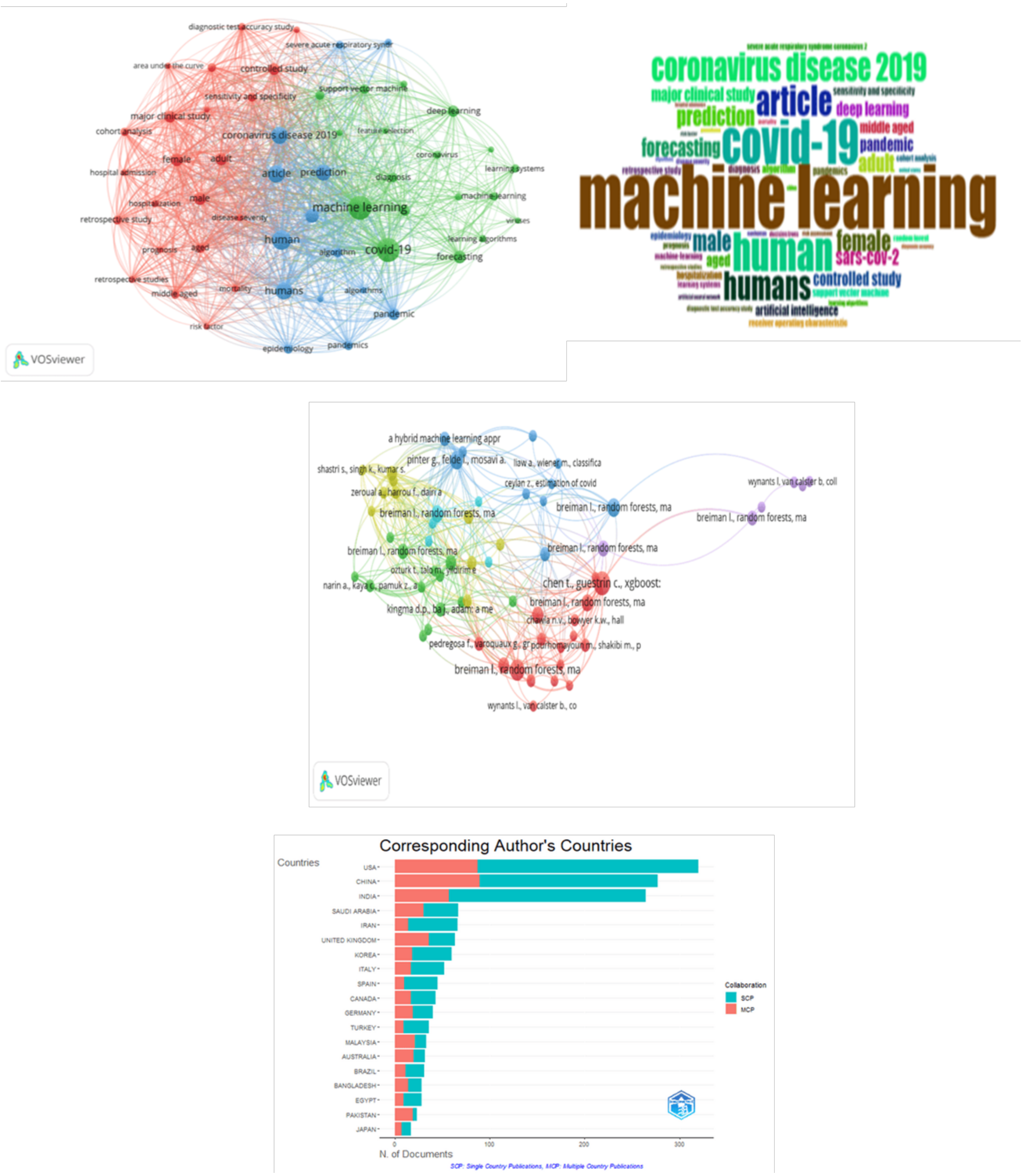

Figure 8: Bibliometric analysis of COVID-19

peak on the number of deaths in India on 13th January, 2022. Hence, preventive measures are recommended with strengthening of medical facilities. Bibliometric analysis concluded that more and more scholarly papers are being produced as the pandemic spreads. Understanding COVID-19 and developing strategies to stop its spread are both made possible by scientific and medical research. Future directions still include the development of vaccines and effective pharmacological therapies.

7. Source of Funding

None.

8. Conflict of Interest

All authors mutually declare that they do not have any conflict of interest.

References

1. MBERS, S I & ASSESSMENT, W R. Coronavirus disease 2019 (COVID-19); 2020.
2. Khedher NB, Kolsi L, Alsaif H. A multi-stage SEIR model to predict thepotential of a new COVID-19 wave in KSA after lifting all travel restrictions. *Alexandria Eng J*. 2021;60(4):3965–74.
3. Gothai E, Thamilselvan R, Rajalaxmi RR, Sadana RM, Ragavi A, Sakthivel R, et al. Prediction of COVID-19 growth and trend using machine learning approach. *Mater Today Proc*. 2023;81(2):597–601.
4. Raj V, Renjini A, Swapna MS, Sreejyothi S, Sankararaman S. Nonlinear time series and principal component analyses: Potential diagnostic tools for COVID-19 auscultation. *Chaos Solitons Fractals*. 2020;140:110246.
5. Islam SMD, Bodrud-Doza M, Khan RM, Haque MA, Mamun MA. ExploringCOVID-19 stress and its factors in Bangladesh: a perception-based study. *Heliyon*. 2020;6(7):e04399.
6. Pal R, Sekh AA, Kar S, Prasad DK. Neural network-based country wiserisk prediction of COVID-19. *Appl Sci*. 2020;10(18):6448.
7. Stephanp, Veugelers R, Wang J. Reviewers are blinkered by bibliometrics. *Nature*. 2017;544(7651):411–2.
8. WHO Coronavirus Disease (COVID-19) Dashboard. Available from: https://covid19.who.int/.
9. Singh A, Dey J, Bhardwaj S. Is this the beginning or the end of COVID-19 outbreak in India? A data driven mathematical model-based analysis. *medRxiv*. 2020;doi:10.1101/2020.04.27.20081422.
10. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497–506.
11. World Health Organization Coronavirus Disease (COVID-19) Dashboard. Available from: https://who.sprinklr.com.
12. Chen C, Dubin R, Kim MC. Emerging trends and new developments in regenerative medicine: a scientometric update (2000 - 2014). *Expert Opin Biol Ther*. 2014;14(9):1295–1317.
13. Government of India COVID-19 Dashboard. Available from: https://www.mygov.in/covid-19.
14. Ambikapathy B, Krishnamurthy K. Mathematical modelling to assess the impact of lockdown on COVID-19 transmission in India: model development and validation. *JMIR Public Health Surveill*. 2020;6(2):e19368.
15. Sarkar K, Khajanchi S. Modeling and forecasting of the COVID-19 pandemic in India; 2020. Available from: https://arxiv.org/abs/2005.07071.
16. Menon A, Rajendran NK, Chandrachud A. Modelling and simulation of COVID-19 propagation in a large population with specific reference to India . medrxiv; 2020. doi:10.1101/2020.04.30.20086306.
17. Bhola J, Venkateswaran VR, Koul M. Corona Epidemic in Indian context: Predictive Mathematical Modelling. medrxiv; 2020. doi:10.1101/2020.04.03.20047175.
18. Gupta R, Pal SK. Trend Analysis and Forecasting of COVID-19 outbreak in India. MedRxiv; 2020. doi:10.1101/2020.03.26.20044511.
19. Brown A, Horton R. A planetary health perspective on COVID-19: a call for papers. *Lancet*. 2020;395(10230):1099.
20. Mahase E. Covid-19: WHO declares pandemic because of "alarming levels" of spread, severity, and inaction. *BMJ*. 2020;368:m1036.
21. Zhong H, Wang Y, Zhang ZL, Liu YX, Le KJ, Cui M, et al. Efficacy and safety of current therapeutic options for COVID-19 - lessons to be learnt from SARS and MERS epidemic: A systematic review and meta-analysis. *Pharmacol Res*. 2020;157:104872.
22. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):727–33.
23. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet*. 2020;395(10223):507–13.
24. Eynde JJV. COVID-19: A Brief Overview of the Discovery Clinical Trial. *Pharmaceuticals (Basel)*. 2020;13(4):65.
25. Mukherjee R. Global efforts on vaccines for COVID-19: Since, sooner or later, we all will catch the coronavirus. *J Biosci*. 2020;45(1):68.
26. Chan JFW, Yuan S, Kok KH, Kai-Wang K, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the2019 novel coronavirus indicating person-to-person transmission: a study of afamily cluster. *Lancet*. 2020;395(10223):514–23.
27. Yousefi B, Valizadeh S, Ghaffari H, Vahedi A, Karbalaei M, Eslami M. A global treatments for coronaviruses including COVID-19. *J Cell Physiol*. 2020;235(12):9133–42.

Author's biography

Rakshita Chaudhary, Senior Technical Data Associate https://orcid.org/0000-0001-8357-7259

Nisha Gaur, Assistant Professor https://orcid.org/0000-0002-8699-6659

Mohit Yadav, Associate Professor

Siddharth Srivastava, Production Engineer

Vanshika Chaudhary, PhD Student

Mohd Asif Shah, Dean of Faculty

Cite this article: Chaudhary R, Gaur N, Yadav M, Srivastava S, Chaudhary V, Shah MA. A stochastic analysis and bibliometric analysis of COVID-19. *Indian J Microbiol Res* 2024;11(4):342-353.