IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 13, April 2025
Impact Factor: 7.67

Covid-19 Outbreak Forecasting based on Machine Learning Models

Mohit and Ritu Dagar
Department of Computer Science Engineering
Sat Kabir Institute of Technology & Management, Bahadurgarh, India
mohitbhardwajshal@gmail.com and ritudagarncce@gmail.com

Abstract: *One of the largest pandemic respiratory diseases in the world is the new coronavirus known as SARS-CoV-2 [1]. It has a high infection rate but a lower lethality rate. As per the worldometer site on 16 September 2021 in India, there are 33,380,438 COVID-19 cases, deaths are around 444,274, and people recovered are 32,590,504 in figures. However, USA has highest COVID-19 infected people with a value of approximately 42,504,484 and total death in figures 685,295. There are various vaccines [2] developed to halt the spread of COVID-19 like Pfizer-BioNTech, Moderna, Sputnik V, Covaxin, Covishield etc, but this virus is getting stronger in upcoming waves. This virus affects all age levels and infects animals too. Therefore, before it becomes difficult to stop the spread of COVID-19, we must properly predict the outbreak in any nation. The JHU CSSE COVID-19 Dataset [3] is used for this study, and we compare the accuracy of five different types of regression models i.e. Bayesian Ridge Regression, Polynomial Regression, Ridge Regression, Support Vector Regression, and Elastic Net algorithm. It has been found that Elastic Net outperforms among all models in terms of accuracy.*

Keywords: COVID-19, Data Science, Data Analytics, Bayesian Ridge, Elastic Net, Machine Learning Algorithm

I. INTRODUCTION

With the help of the latest data science analysis techniques, it is very easy to get faster and accurate predictions by utilizing the old dataset containing various important attributes that will predict outcome precisely. But it is not possible to get a very accurate prediction [4] due to the limited information or records available in the dataset. Regression is one most important machine learning techniques that will help to find out the relationship between the independent and dependent variables. Regression works well with continuous values. There are various variants available for regression, but for this research, we mainly used SVR, Ridge Regression, Elastic Net, Bayesian Ridge Regression, and Polynomial regression.

The main goal in this research is to compare different regression techniques for forecasting COVID-19 outbreak prediction [5] based on several parameters like mean Absolute error, mean squared error, etc. This early prediction of the COVID-19 outbreak can help government employees in various ways. For ex., they can make healthcare policy and take decisions accordingly. It will save people's lives and reduce the cost required for medicine.

II. LITERATURE SURVEY

By integrating supervised and unsupervised learning, the hybrid hierarchical ensemble described by Yakovyna et al. [6] enables us to improve the COVID-19 cases forecasting accuracy by 11% in terms of MSE, 29% in terms of area under the ROC, and 43% in terms of MPP metric. According to him, the most crucial aspect of COVID-19 spread for regression analysis and classification is virus pressure. To anticipate [7] the danger of COVID-19 and other diseases during a pandemic, a framework based on the Internet of Things and cloud computing has been presented that guarantees IoT sensor-based data collection and storage on the cloud system as well as self-assessment tests using COVID-19 websites and mobile apps. The hesitancy value has been taken into account in the patients, symptoms, and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 13, April 2025
Impact Factor: 7.67

disease tables in Intuitionistic Fuzzy Set. It has been discovered that calculating the distance between all diseases and all symptoms with hesitancy is a more effective method than determining the disease without the hesitancy value.

The optimised LSTM (popLSTM) [8] approach was put forward in deep learning to minimise the mistakes in calculating the number of verified COVID-19 cases using Basic LSTM. The statistics came from four countries South Korea, Hong Kong, Italy, and Indonesia, where the number of confirmed cases is rising significantly every day. According to the results, the popLSTM contributions of lowering the error value by considering the output gate's hidden state to enhance accuracy and integrating the output results to efficiently filter more specific information showed a noteworthy 4% improvement in accuracy in comparison to the prior model. The HBA-ANN [9] model is a novel hybrid intelligence model designed to forecast the daily COVID-19 cases in Russia, Brazil, India, and the US. Four statistical metrics R, RMSE, NSE, and SI, were used to compare its performance to those of independent ANN and GEP models. The findings showed that in each nation, the HBA-ANN model performed better than the GEP and ANN models.

Using data from four developing nations namely Rwanda, Mozambique, Nepal, and Myanmar, the author [12] trained a BiLSTM with two hidden layers, one with 200 neurons and the other with 100 neurons. In terms of RMSE, the outcomes of the suggested multi-layer BiLSTM were contrasted with those of the multi-layer LSTM using identical parameters. Nepal, Mozambique, Rwanda, and Myanmar are the four developing nations where the multi-layer BiLSTM model performed better than the multi-layer LSTM model. To forecast the pandemic epidemic, Singh and Mittal [15] employed Linear Regression (LR), Support Vector Machine (SVM), and Polynomial Regression (PR). Using the recommended Ant Colony Optimisation technique (ACO), the author improves the parameters of the machine learning models that are presently in use. According to the outcome forecast, PR-ACO performs better than other ML methods in terms of final outcome. As an alternative to susceptible-infected-recovered (SIR) and susceptible-exposed-infectious-removed (SEIR) models, the author [16] in his paper compares machine learning and soft computing models to forecast the COVID-19 epidemic. Two machine learning models MLP and ANFIS shown a high potential for long-term

Data Analysis framework for Outbreak Analysis of COVID-19 prediction.

II. METHODOLOGY

Data Acquisition

This dataset is provided by the Centre for Systems Science and Engineering (CSSE) at Johns Hopkins University for analysis of COVID-19 outbreak. The dataset contains 279 rows with 613 columns. This dataset contains various attributes like latitude, longitude, confirmed COVID-19 cases with date, country, etc.

Data Cleaning

In this step, we remove unnecessary data by applying various techniques like replace null values with median or mode values, encoding categorical values, assigning useful names to columns, and removing outliers. Thus, we need to remove the noise from the data so that the algorithm can run faster and smoother.

Data Exploration

In this step, we will create a heatmap that shows the correlation between various input variables. Also, a pair plot, bar chart, and pie chart diagram are very useful for data visualization to understand the relationship between input and output variables.

Attribute Selection

Selecting the best feature will increase the accuracy of the model further for forecasting. Data visualization plays a very important role in selecting the features for our model. Also, the whole framework for COVID-19 outbreak prediction is shown in Fig. 1.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 13, April 2025
Impact Factor: 7.67

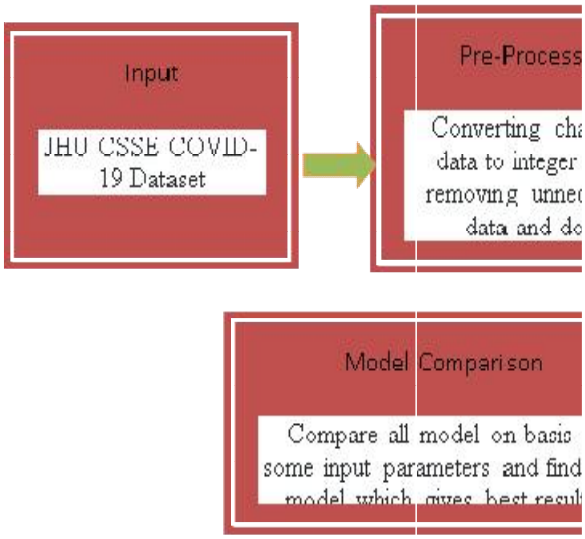

Figure 1: Data Analysis framework for Outbreak Analysis of COVID-19

Machine Learning Model

Elastic Net: It is a type of linear regression that uses both L1 & L2 regularisation. It eradicates the lasso regression limitation, and it is considered best when the number of samples used is less than the dimension data

Ridge Regression: LASSO regression is a form of linear regression useful when data suffers from multicollinearity. Unlike LASSO regression, it performs the L2 regularization technique. But the problem with this method it is not good for feature reduction because it does not decrease the number of variables, since it does not make the coefficient zero rather it minimizes it.

Bayesian Ridge: This method is useful for those datasets that have an insufficient number of data or are poorly distributed, and it is defined in probabilistic terms rather than a point estimate. The dependent variable is obtained via a probabilistic distribution. It calculates the posterior distribution for the model parameter.

Polynomial Regression: Polynomial regression is an extension of linear regression when the number of input variables is large and there is only one output variable. This method is not suitable for those datasets that contain outliers.

Support Vector Regression(SVR): Support vector machine solves the regression problem with the help of support vectors and a hyperplane known as support vector regression. It is a supervised method of learning that uses various types of kernels like linear kernel, RBF kernel, etc. It is sensitive to a noisy dataset and performs worse in the case of a huge dataset.

Performance Metrics and Loss Function

$$\text{Mean Absolute Error (MAE)} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (1)$$

$$\text{Mean Squared Error (MSE)} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad (2)$$

$$\text{Loss Function} = \text{MSE} \quad (3)$$

Where

y_i = Actual value

$\hat{y}_i$ = Predicted value

N= Total number of observations

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 13, April 2025
Impact Factor: 7.67

Heat map Chart

Fig. 2. represents the heatmap that shows the relationship between various parameters. It helps for finding the correct attribute that affects the life expectancy of human being.

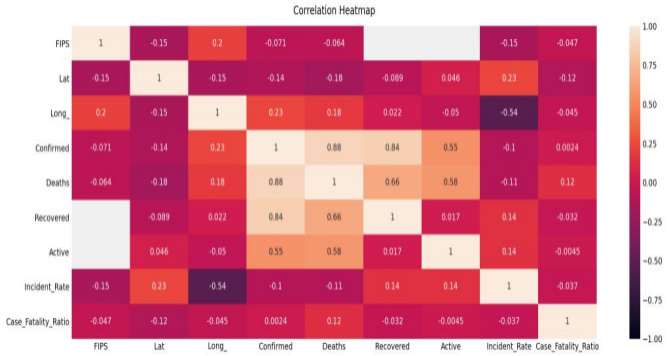

Figure 2: Heatmap diagram

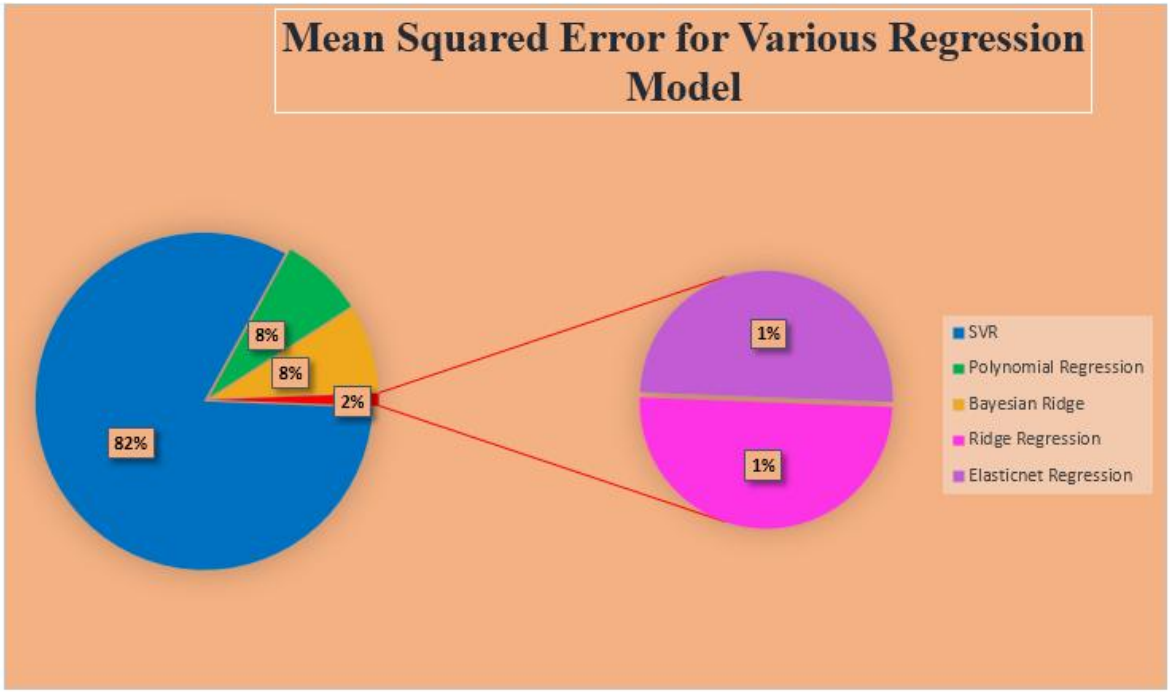

Figure 3: Mean Squared Error of Various Regression ML Models

IV. RESULT AND DISCUSSION

In our research, we did a comparative analysis of various types of regression models. In this comparison, our SVR model performs very badly but Elastic Net and Ridge Regression beat all models in terms of accuracy. Fig. 3. shows Mean Squared Error values of various Regression machine learning models used in our research and Fig. 4. compares various Regression machine learning models used in our project in terms of Mean Absolute Error. Elastic Net and Ridge regression give the best accuracy among all models, with mean absolute error values of 0.078709, 0.07871, respectively. Their share in mean squared error percentage is 1%,1%, respectively. SVR gives very low accuracy with a value of 0.974 as shown in Fig. 4. Its share in the mean squared error is 82%. Bayesian Ridge and Polynomial Regression accuracy are very similar, with values 0.313702 and 0.296034 respectively.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 13, April 2025
Impact Factor: 7.67

V. CONCLUSION

We simply conducted a comparison analysis for regression approaches in this study report. Elastic Net, Bayesian Ridge Polynomial, Ridge Regression, Polynomial Regression, and Support Vector Machine Regression were the five machine learning techniques that we employed. We had compared the accuracies of various machine learning models for finding which is best for forecasting the COVID-19 outbreak in the country. After this comparative analysis, it is found that Elastic Net gives more accurate and reliable results as compared to other regression techniques utilized in our research.

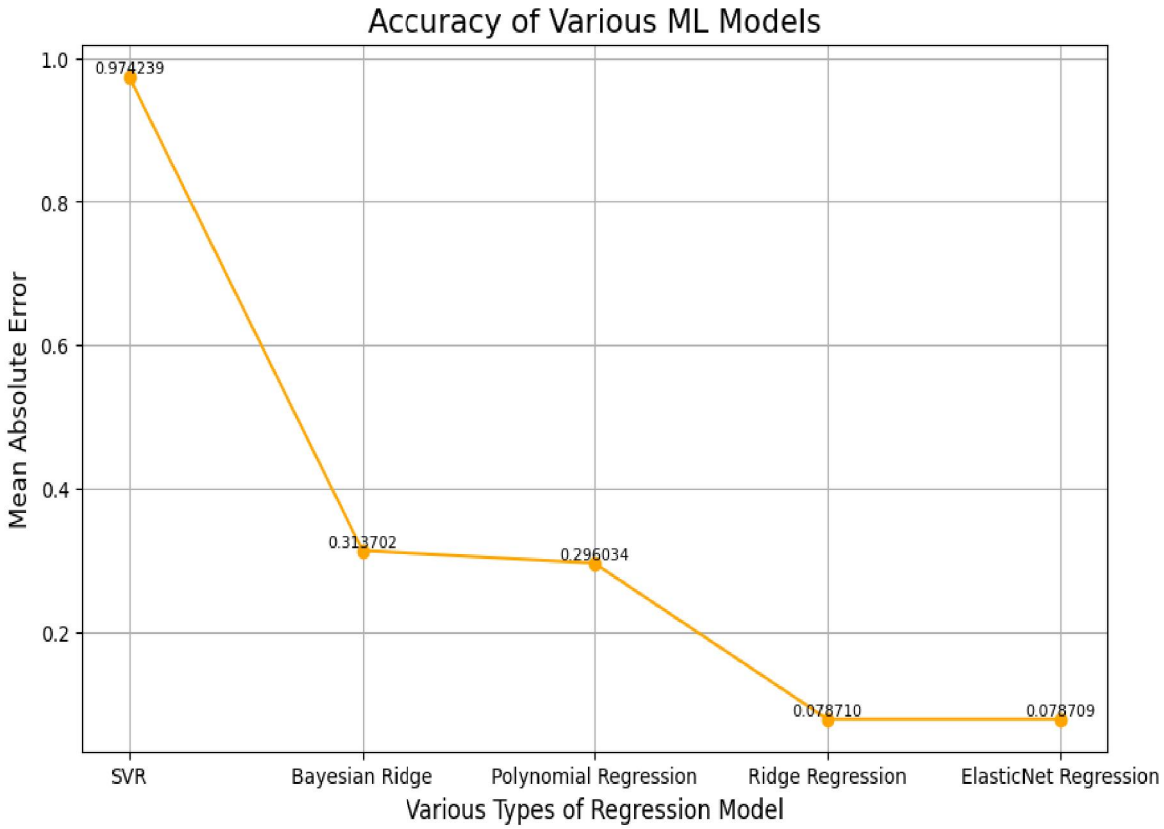

Fig. 4. Mean Absolute Error of Various Regression ML Models

Elastic Net and Ridge regression accuracy differ by only a small amount. If we use Elastic Net for forecasting it provides the best consultancy service and accurate prediction for COVID-19 suffering patients. This research will also become helpful for healthcare policy makers and government employees in making quick decisions regarding COVID-19. This research will also become helpful in those countries where very less healthcare facilities are available.

REFERENCES

[1] T.-H. Song, L. Clemente, X. Pan, J. Jang, M. Santillana, and K. Lee, "Fine-grained forecasting of COVID-19 trends at the county level in the United States," npj Digit. Med., vol. 8, no. 1, p. 204, April 2025, doi: 10.1038/s41746-025-01606-1.

[2] Z. M. Nia et al., "Leveraging deep-learning and unconventional data for real-time surveillance, forecasting, and early warning of respiratory pathogens outbreak," Artificial Intelligence in Medicine, vol. 161, p. 103076, Mar. 2025, doi: 10.1016/j.artmed.2025.103076.

[3] T. H. Kim, R. Chinthaginjala, A. Srinivasulu, S. P. Tera, and S. O. Rab, "COVID-19 health data prediction: a critical evaluation of CNN-based approaches," Scientific Reports, vol. 15, no. 1, p. 9121, Mar. 2025, doi: 10.1038/s41598-025-92464-0.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 13, April 2025
Impact Factor: 7.67

[4] G. Battineni, N. Chintalapudi, and F. Amenta, "Forecasting of COVID-19 epidemic size in four high hitting nations (USA, Brazil, India and Russia) by Fb-Prophet machine learning model," Applied Computing and Informatics, vol. 21, no. 1/2, pp. 2–11, Jan. 2025, doi: 10.1108/ACI-09-2020-0059.
[5] M. Z. Abedin, M. H. Moon, M. K. Hassan, and P. Hajek, "Deep learning-based exchange rate prediction during the COVID-19 pandemic," Annals of Operations Research, vol. 345, no. 2–3, pp. 1335–1386, Feb. 2025, doi: 10.1007/s10479-021-04420-6.
[6] V. Yakovyna, N. Shakhovska, and A. Szpakowska, "A novel hybrid supervised and unsupervised hierarchical ensemble for COVID-19 cases and mortality prediction," Scientific Reports, vol. 14, no. 1, p. 9782, April 2024, doi: 10.1038/s41598-024-60637-y.
[7] N. K. Tyagi and K. Tyagi, "IoT and cloud-based COVID-19 risk of infection prediction using hesitant intuitionistic fuzzy set," Soft Computing, vol. 28, no. 5, pp. 3743–3755, Mar. 2024, doi: 10.1007/s00500-023-09548-0.
[8] I. Sembiring, S. N. Wahyuni, and E. Sediyono, "LSTM algorithm optimization for COVID-19 prediction model," Heliyon, vol. 10, no. 4, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26158.
[9] S. N. Qasem, "A novel honey badger algorithm with multilayer perceptron for predicting COVID-19 time series data," The Journal of Supercomputing, vol. 80, no. 3, pp. 3943–3969, Feb. 2024, doi: 10.1007/s11227-023-05560-1.
[10] R. Li, Y. Song, H. Qu, M. Li, and G.-P. Jiang, "A data-driven epidemic model with human mobility and vaccination protection for COVID-19 prediction," Journal of Biomedical Informatics, vol. 149, p. 104571, Jan. 2024, doi: 10.1016/j.jbi.2023.104571.
[11] O. Gaidai, J. Sheng, Y. Cao, Y. Zhu, and S. Loginov, "Generic COVID-19 epidemic forecast for Estonia by Gaidai multivariate reliability method," Franklin Open, vol. 6, p. 100075, March 2024, doi: 10.1016/j.fraope.2024.100075.
[12] S. P. Cumbane and G. Gidófalvi, "Deep learning-based approach for COVID-19 spread prediction," International Journal of Data Science and Analytics, June 2024, doi: 10.1007/s41060-024-00558-1.
[13] B. Chen et al., "High-resolution short-term prediction of the COVID-19 epidemic based on spatial-temporal model modified by historical meteorological data," Fundamental Research, vol. 4, no. 3, pp. 527–539, May 2024, doi: 10.1016/j.fmre.2024.02.006.
[14] P. Bodapati, E. Zhang, S. Padmanabhan, A. Das, M. Bhattacharya, and S. Jahanikia, "A Global Network Analysis of COVID-19 Vaccine Distribution to Predict Breakthrough Cases among the Vaccinated Population," COVID, vol. 4, no. 10, pp. 1546–1560, Sep. 2024, doi: 10.3390/covid4100107.
[15] S. Singh and S. Mittal, "Pandemic Outbreak Prediction using Optimization-based Machine Learning Model," 2023 3rd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS), Kalady, Ernakulam, India, 2023, pp. 154-159, doi: 10.1109/ACCESS57397.2023.10199872.
[16] T. Sher, A. Rehman, and D. Kim, "COVID-19 Outbreak Prediction by Using Machine Learning Algorithms," Computers, Materials & Continua, vol. 74, no. 1, pp. 1561–1574, 2023, doi: 10.32604/cmc.2023.032020.
[17] S. A. Rakhshan, M. S. Nejad, M. Zaj, and F. H. Ghane, "Global analysis and prediction scenario of infectious outbreaks by recurrent dynamic model and machine learning models: A case study on COVID-19," Computers in Biology and Medicine, vol. 158, p. 106817, May 2023, doi: 10.1016/j.compbiomed.2023.106817.
[18] S. Raheja, S. Kasturia, X. Cheng, and M. Kumar, "Machine learning-based diffusion model for prediction of coronavirus-19 outbreak," Neural Computing and Applications, vol. 35, no. 19, pp. 13755–13774, Jul. 2023, doi: 10.1007/s00521-021-06376-x.
[19] S. Natarajan, M. Kumar, S. K. K. Gadde, and V. Venugopal, "Outbreak prediction of COVID-19 using Recurrent neural network with Gated Recurrent Units," Materials Today: Proceedings, vol. 80, pp. 3433–3437, 2023, doi: 10.1016/j.matpr.2021.07.266.
[20] S. Namasudra, S. Dhamodharavadhani, and R. Rathipriya, "Nonlinear Neural Network Based Forecasting Model for Predicting COVID-19 Cases," Neural Processing Letters, vol. 55, no. 1, pp. 171–191, Feb. 2023, doi: 10.1007/s11063-021-10495-w.
[21] P. Martínez-Fernández, Z. Fernández-Muñiz, A. Cernea, J. L. Fernández-Martínez, and A. Kloczkowski, "Three Mathematical Models for COVID-19 Prediction," Mathematics, vol. 11, no. 3, p. 506, Jan. 2023, doi: 10.3390/math11030506.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 13, April 2025
Impact Factor: 7.67

[22] Y. Kim, C.-R. Park, J.-P. Ahn, and B. Jang, "COVID-19 outbreak prediction using Seq2Seq + Attention and Word2Vec keyword time series data," PLoS ONE, vol. 18, no. 4, p. e0284298, Apr. 2023, doi: 10.1371/journal.pone.0284298.
[23] S. Khalilpourazari and H. Hashemi Doulabi, "Robust modelling and prediction of the COVID-19 pandemic in Canada," International Journal of Production Research, vol. 61, no. 24, pp. 8367–8383, Dec. 2023, doi: 10.1080/00207543.2021.1936261.
[24] H. Verma, S. Mandal, and A. Gupta, "Temporal deep learning architecture for prediction of COVID-19 cases in India," Expert Systems with Applications, vol. 195, p. 116611, Jun. 2022, doi: 10.1016/j.eswa.2022.116611.
[25] O. Sharif, M. Z. Hasan, and A. Rahman, "Determining an effective short term COVID-19 prediction model in ASEAN countries," Scientific Reports, vol. 12, no. 1, p. 5083, Mar. 2022, doi: 10.1038/s41598-022-08486-5.
[26] P. Pham, W. Pedrycz, and B. Vo, "Dual attention-based sequential auto-encoder for Covid-19 outbreak forecasting: A case study in Vietnam," Expert Systems with Applications, vol. 203, p. 117514, Oct. 2022, doi: 10.1016/j.eswa.2022.117514.
[27] M. O. Alassafi, M. Jarrah, and R. Alotaibi, "Time series predicting of COVID-19 based on deep learning," Neurocomputing, vol. 468, pp. 335–344, Jan. 2022, doi: 10.1016/j.neucom.2021.10.035.
[28] S. Ardabili et al., "COVID-19 Outbreak Prediction with Machine Learning," Algorithms, vol. 13, no. 10, p. 249, Oct. 2020, doi: 10.3390/a13100249.