RESEARCH ARTICLE OPEN ACCESS

Comparison of Analysis and Prediction Outcomes of ARIMA Model and LSTM Network Based on India's COVID-19

Premavathi T[1]
1(Department of Computer Engineering, Marwadi University, Rajkot
Email: prema.cse05@gmail.com)

Abstract:

The COVID-19 pandemic is a profound concern and an urgent issue that needs to be addressed first and foremost globally. It strongly affects all areas of life from public health, health, education, economy as well as human freedom of movement. The worrying thing is that there is no specific treatment as well as prevention in the early stages. This infectious disease could not be stopped anytime soon. With the rapid spread of the disease, the global health system collapsed because it did not anticipate the danger with its exponential rate of spread. The application of methods to predict the number of infections in machine learning can contribute to limiting the vulnerability to humanity. By predicting the number of cases, we can be better prepared, such as providing more hospital beds, producing more medical equipment, allocating more medical staff to areas with a sharp increase in the number of infections. This solution helps the government to have the necessary references to make the best and fastest decisions. This paper applies two models in machine learning that are Autoregressive-Integrated-Moving-Average (ARIMA) and Long Short Term Memory (LSTM) to choose the best solution to the problem [1].

***Keywords* — covid-19, ARIMA model, time series forecasting, auto regression, linear regression, random forest, SVM.**

I. INTRODUCTION

In December 2019, the world recorded the first cases of infectious disease caused by a new virus called SARS-CoV-2 in Wuhan city of China. After that, the disease quickly spread around the world at a rapid rate and became a pandemic for all of humanity. An infected person may have no symptoms at first, but can still spread the infection through direct contact or touching objects that may have been contaminated with the virus when they coughed. There is no specific treatment or vaccine for the pandemic in the early stages. It has challenged the global health system because of the exponential speed of its spread. The world health system has faced a shortage of hospital beds, emergency medical equipment, especially not enough medical staff or medical staff overload [2]. As a result, many patients did not receive timely care and many lives were unfortunately lost. The pandemic has also dealt a blow to the global economy [3] when everything has come to a standstill, education has also been suspended, and almost everything has to stop with the spread of the epidemic. The development of a vaccine is the fastest solution to limit the spread of the disease, and the early prediction of the number of infections also contributes to relieving the overwhelming pressure on the global system. Predicting the number of infections early helps us to build more field hospitals, urgently produce and supply equipment to areas with a high number of outbreaks. Help the government have the basis to make quick and right decisions. In this study, we want to provide one more reference in predicting the number of infections by applying machine learning methods. We apply 2 models that are ARIMA and Long Short Term Memory. Thereby comparing to find the best solution for prediction.

II. RELATED WORK

COVID-19 data is a time series data where the current date data is closely related to previous dates. Applying the methods that are used to deal with this kind of chronological data in machine learning is extremely useful for solving this problem.

The ARIMA model was applied by Sujeet Maurya et al., who applied the predictive model to COVID-19 time series data. The data includes 153 lines, of which 138 lines were used for training, and the remaining 15 lines were used for testing. The ARIMA model includes three main parameters p, d and q. The researchers took advantage of the autoregression function (PACF) and autocorrelation function (ACF) graphs to determine values for the p and q parameters. Sujeet Maurya et al. have found the optimal parameter for the model is (0,2,0) for 3 parameters p,d,q to apply to the data they are considering. The model gave an MSE result of 7240024855.066816 [4].

Hadeel I. Mustafa and colleagues collected data from three European countries that were most severely affected by COVID- 19. Different parameters were applied to the ARIMA model to compare and select the best results. The test results have shown that ARIMA(0,2,1), ARIMA(0,2,1), and ARIMA(1,2,0) models are the best choice for France, Italy and Spain with MAPE_France = 5,634, MAPE_Italy = 4,752, and MAPE_Spain = 5,849 values. Besides, they have given the predicted number of cases in the next 10 days for 3 countries including France from 140,320 to 159,619, Italy is 196,520 to 229,147 and Spain is 204,755 to 257,497 [5].

Shreyansh Chordia et al. applied both ARIMA and PROPHET models to train for 224 days, from January 30, 2020 to September 8, 2020. The reviewed dataset recorded the data of many states of India day by day. After applying these two models, the best results obtained in Tamil Nadu based on the number of infections with R2 were 0.9984 and 0.9869 for ARIMA and PROPHET, respectively. And Karnataka gave the best results when predicting based on the number of deaths with R2 of 0.6697 and 0.9665 for ARIMA and PROPHET, respectively [6].

On May 30, 2020, Russia recorded 396,000 cases of COVID-19 infection, becoming the worst-affected country in the region. Lanlan Fang and colleagues studied data collected from this country. They have developed 3 ARIMA models including ARIMA(2,2,1), ARIMA(3,2,0), and ARIMA(0,2,1) to apply to the number of infections, deaths and recoveries. The researchers' experiments gave MAPE results of 0.6, 3.9, and 2.4 respectively for these 3 models [7].

The number of cases of COVID-19 in Italy of 105792 was recorded as of March 31, 2020. Nalini Chintalapudi and colleagues applied the ARIMA model to data during the 60-day lockdown in Italy. Results were obtained with an accuracy of 93.75 percent for infections and 84.4 percent for recoveries. The model also predicts that the number of infections will reach 182757 and the number of recoveries will be 81635 by May [8].

In this study, Raghavendra Kumar and colleagues applied all 3 types of models, which are AR, MA and ARIMA to the dataset taken from India. The RMSE results obtained for AR, MA, and ARIMA are 1083366, 1128500, and 1079058, respectively, and the average figure for all 3 models is 1096975 [9].

Sarbhan Singh et al used COVID-19 data collected from John Hopkins University and Malaysian Ministry of Health (MOH) websites. The researchers used data lines from January 22 to March 31, 2020 to train the ARIMA model. They extracted data from April 1 to April 17, 2020 for testing data, and used the data from April 18 to May 1, 2020 to predict the results. With the obtained results ARIMA(0,1,0) is the optimal model with MAPE and Bayesian Information Criteria (BIC) of 16.01 and 4.170 respectively [10].

In this study, Qiuying Yang et al. applied multiple models including: ARIMA(0,2,0) for the number of infections, ARIMA(2,2,1) for the number of deaths of the COVID- 19 of Hubei province. The training results obtained R2 are 0.956 and 0.823 respectively for ARIMA(0,2,0), ARIMA(2,2,1). On the other hand, they used the MAE calculation for validating data with the results obtained 18.1 for the number of infections and 5.2 for the number of deaths [11].

In this study, Naresh Kumar and colleagues used two models, ARIMA and PROPHET, to train COVID-19 datasets from many countries. MAPE results for 2 models ARIMA and PROPHET when applied to active cases are the best 0.586 and 1,481 for 2 countries US and UK, respectively. With the MAE assessment method, the best results for the two models ARIMA and PROPHET are 78.19 and 69.11 for the UK, respectively. Meanwhile, the results for 2 models ARIMA and PROPHET apply to the number of deaths with MAPE of 2,571 and 3,759 respectively for the US and Iran [12].

To train the dataset in India, Aishwarya Sen and colleagues employed a variety of machine learning algorithms, including Random Forest, Support Vector Machine, LASSO Regression, and Multilinear Regression.

According to the research results, the Random Forest method outperforms the other algorithms, with an R2 of 99.83 and an RMSE of 464196 [13].

III. METHODOLOGY

A. Autoregressive-Integrated-Moving-Average (ARIMA)

ARIMA is the acronym for "Auto Regressive Integrated Moving Average," which is made out of the initial letter of the sentence. The ARIMA model is an analytical method for determining the underlying meaning of time series data types (particularly non-stationarity time series) with an ordinal connection in terms of time [14]. It's ideal for forecasting future values based on historical time-sequential values. More specific:

Auto-Regressive represented by AR is a concept used to indicate the relationship between the observed value and prior lagged observed data values, all of which have a time-ordered relationship.

Integrated in the word ARIMA is represented by the letter I which indicates the difference between successively observed values over time.

Moving-Average is also represented by MA is a method of analyzing data points by calculating the averages of distinct subsets of the entire data set [15].

There are 3 parameters p, d and q to build an ARIMA model, specifically as follows:

The parameter-p is used to specify the number of lag observations.

The parameter-d is used to specify the number of observations that must be computed differently.

The parameter-q is used to specify the size of the moving average window .

B. Long-Short-Term-Memory (LSTM)

The LSTM is a type of recurrent neural network (RNN) in which the results of previous stages are used as input for the following step. LSTM is utilized to handle time-series data when the time gap between two consecutive events is not predefined. A unit in LSTM is made up of the following main components: a cell, a forget-gate, an input-gate, an output-gate in which the gates are responsible for controlling the flow of information in and out of the cell, and cells take on the role of remembering values for a certain period of time [16].

Forget-gate deletes the information in the cell when it is not in use. The cell's state is 0 information is no longer needed to store, otherwise the information will be kept if the state is 1.

The input-gate controls whether or not information is saved in the cell.

The output-gate is in charge of determining what information to retrieve from the cell [17].

IV. REGRESSION METRICS FOR EVALUATING REGRESSION MODELS

1. Coefficient-of-Determination (R2)

R^2 is used to determine the strength of a linear connection between two variables, and it is frequently employed by academics while doing trend analysis [18].

2. Mean-Absolute-Error (MAE)

Mean Absolute Error computes the average difference between computed and real data [19].

3. Root-Mean-Squared Error (RMSE)

The RMSE is determined by taking the square root of the average square of the difference between the actual and estimated values [20].

4. Mean-Absolute-Percentage-Error (MAPE)

MAPE is determined as the average of the absolute value of dividing the difference between the real and calculated values by the actual value [21].

V. DATA VISUALIZATION

A. India COVID-19 Dataset

The dataset used in this study was gathered over time in several Indian states. The number of infections increases and decreases differently in each location. The data utilized in this investigation are summarized in Table 1 below [22].

TABLE I. INDIA'S STATEWIDE COVID-19 DATA

Sno	Date	State/Union Territory	…	Cured	Deaths	Confirmed
1	2020-01-30	Kerala	…	0	0	1
2	2020-01-31	Kerala	…	0	0	1
…	…	…	…	…	…	…
18109	2021-08-11	Uttar Pradesh	…	1685492	22775	1708812
18110	2021-08-11	West Bengal		1506532	18252	1534999

B. COVID-19 Data Visualization Across India

The line graph depicts the upward and downward trend of COVID-19 over time, from January 30, 2020, to August 11, 2021, entire India in Figure 1.

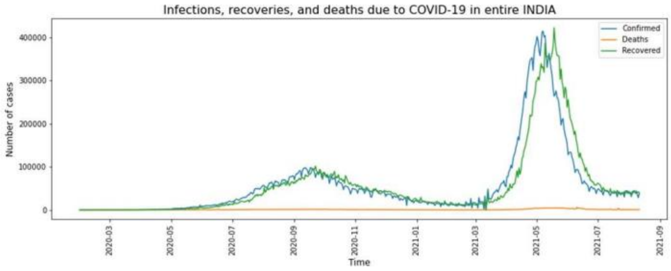

Fig. 1. The trend of COVID-19 infections, recoveries, and fatalities across India

C. Each state's COVID-19 data visualization

In Figures 2 to 9, the graphs clearly illustrate an increase or lower trend in the number of infections, recoveries, and deaths over time in the selected Indian states in descending order of decreasing population.

Uttar Pradesh

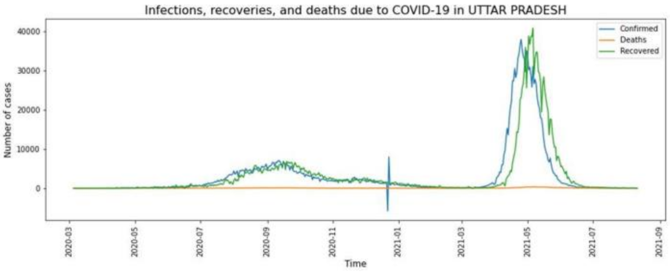

Fig. 2. The trend of COVID-19 infections, recoveries, and fatalities in Uttar Pradesh.

Maharashtra

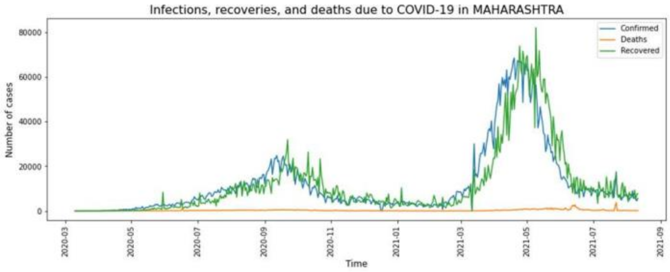

Fig. 3. The trend of COVID-19 infections, recoveries, and fatalities in Maharashtra.

West Bengal

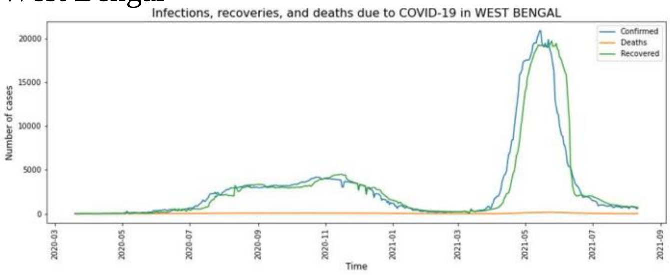

Fig. 4. The trend of COVID-19 infections, recoveries, and fatalities in West Bengal.

Tamil Nadu

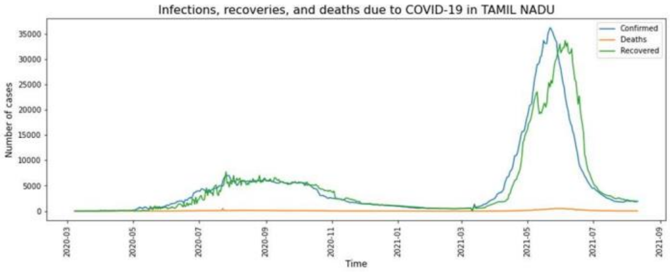

Fig. 5. The trend of COVID-19 infections, recoveries, and fatalities in Tamil Nadu

Rajasthan

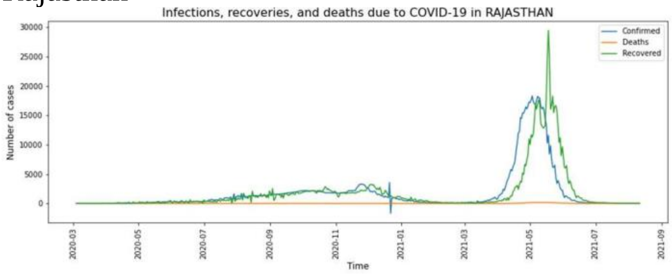

Fig. 6. The trend of COVID-19 infections, recoveries, and fatalities in Rajasthan.

Karnataka

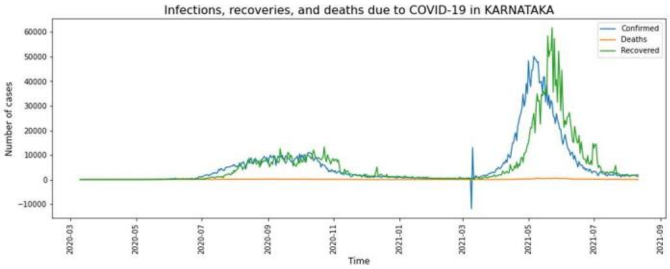

Fig. 7. The trend of COVID-19 infections, recoveries, and fatalities in Karnataka.

Gujarat

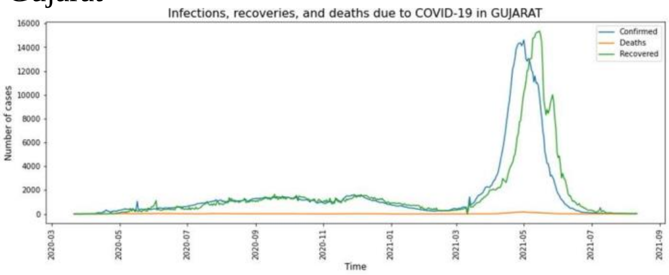

Fig. 8. The trend of COVID-19 infections, recoveries, and fatalities in Gujarat.

Andhra Pradesh

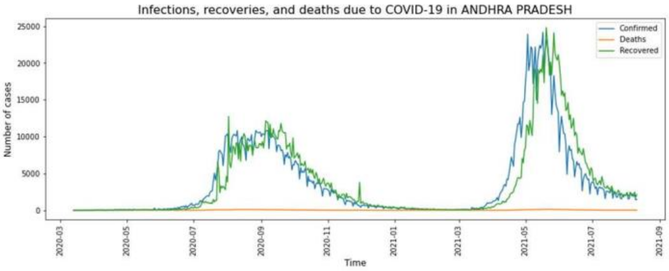

Fig. 9. The trend of COVID-19 infections, recoveries, and fatalities in Andhra Pradesh.

VI RESULT

Two models, ARIMA and LSTM, were utilized to train the COVID-19 dataset in this study, one for the entire country and the other for each particular state. Data is extracted in two parts: 80 percent for training data from 30-Jan-20 to 21-Apr-21, and 20% for testing data from 22-Apr-21 to 11-Aug-21. The two models' prediction outputs are then compared and assessed using common methods like R^2, MAE, RMSE, and MAPE.

A. Compare data-driven forecast outcomes across india

Comparative results between the two models based on data from all over India are presented in Table II.

TABLE II. COMPARE PREDICTION OUTCOMES BETWEEN TWO MODELSBASED ON DATA ACROSS INDIA.

	ARIMA(1,1,1)				LSTM			
	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE	R^2
Confirmed cases	9360.887	0.083	13232.094	0.990	10505.922	0.093	15189.436	0.987
Deaths	344.217	0.238	639.582	0.812	877.051	0.334	1238.633	0.309
Confirmed cases	9888.926	0.070	14863.169	0.986	20629.958	0.106	32861.329	0.932

B. Compare data-driven forecast outcomes by state

The comparison outcomes between the two models are based on the number of cases, fatalities, and recoveries in selected Indian states, as shown in Tables III, IV, and V.

1. Based on confirmed cases by state

TABLE III. COMPARE PREDICTION OUTCOMES BETWEEN TWO MODELS BASED ON CONFIRMED CASES BY STATE

State	ARIMA(1,1,1)				LSTM			
	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE	R^2
Uttar Pradesh	640.209	0.403	1306.150	0.983	696.190	2.596	1294.985	0.979
Maharashtra	2041.494	0.133	3282.161	0.963	2158.124	0.145	3368.953	0.955
West Bengal	242.287	0.066	422.581	0.997	363.127	0.080	636.514	0.993
Tamil Nadu	294.751	0.020	480.655	0.998	1262.297	0.067	2053.158	0.972
Rajasthan	312.409	0.216	673.569	0.989	354.187	1.285	757.457	0.985
Karnataka	1659.114	0.153	2801.363	0.969	2062.789	0.190	3387.304	0.953
Gujarat	110.864	0.092	235.030	0.997	278.253	0.412	532.380	0.982
Andhra Pradesh	1192.854	0.173	1883.666	0.938	1497.447	0.193	2310.988	0.906

2. Based on deaths by state

TABLE IV. COMPARE PREDICTION OUTCOMES BETWEEN TWO MODELS BASED ON FATALITIES BY STATE

State	ARIMA(1,1,1)				LSTM			
	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE	R^2
Uttar Pradesh	15.230	0.415	22.623	0.963	14.661	0.437	22.684	0.9619
Maharashtra	269.604	0.750	568.405	0.068	271.955	0.427	557.040	0.118
West Bengal	5.213	0.150	6.927	0.984	13.460	0.215	20.154	0.866
Tamil Nadu	23.792	0.153	34.571	0.954	88.940	0.309	140.471	0.243
Rajasthan	4.197	0.355	7.824	0.985	10.408	0.380	16.528	0.931
Karnataka	33.945	0.183	57.118	0.903	65.709	0.251	108.432	0.658
Gujarat	2.319	0.219	4.244	0.991	2.683	0.224	5.141	0.984
Andhra Pradesh	5.332	0.128	7.065	0.955	5.947	0.150	7.809	0.946

3. Based on recoveries by state

COMPARE PREDICTION OUTCOMES BETWEEN TWO MODELS BASED ON RECOVERIES BY STATE

State	ARIMA(1,1,1)				LSTM			
	MAE	MAPE	RMSE	R^2	MAE	MAPE	RMSE	R^2
Uttar Pradesh	1198.773	0.244	2623.052	0.956	1404.901	0.291	2978.352	0.939
Maharashtra	4271.786	0.267	6194.404	0.913	4980.193	0.294	7523.154	0.864
West Bengal	237.579	0.051	573.934	0.995	390.680	0.066	903.629	0.987
Tamil Nadu	615.279	0.041	985.347	0.992	1288.514	0.074	1995.730	0.969
Rajasthan	817.007	0.251	1760.427	0.939	1024.712	0.813	2302.970	0.897
Karnataka	3487.678	0.232	5738.305	0.867	8588.158	0.367	14124.737	0.211
Gujarat	269.286	0.194	500.464	0.991	331.635	0.339	644.928	0.984
Andhra Pradesh	1058.993	0.118	1604.884	0.947	1849.396	0.168	2898.631	0.832

VII CONCLUSION

The goal of this research is to use models to forecast the type of time-series data that COVID-19 data contains. Furthermore, using a variety of models allows us to gain a more objective perspective, allowing us to select the best model. ARIMA and LSTM are two time-series data prediction models that are used to train a dataset that spans 559 days from January 30, 2020, to August 11, 2021. This dataset covers multiple Indian states by date. The ARIMA model has the most accuracy when forecasting the number of infections in Tamil Nadu state (0.998), while the LSTM has the highest accuracy when predicting the number of infections in West Bengal (0.993). In general, both models produce excellent results, although ARIMA outperforms the LSTM model. As a result, it is clear that ARIMA is well suited to the prediction of time-series data.

REFERENCES

[1] Wikipedia contributors, “COVID-19 - Wikipedia,” Wikipedia, The Free Encyclopedia., 2021. https://en.wikipedia.org/wiki/COVID-19 (accessed Apr. 07, 2022).

[2] Khanh, H. Q., Damodharan, P., & Kumar, D. (2022, March). Data acquisition based COVID-19 Spread Prediction Analysis. In 2022 International Conference on Electronics and Renewable Systems (ICEARS) (pp. 1651-1655). IEEE.

[3] Mridha, K., Kumbhani, S., Pandey, A. P., & Damodharan, P. (2021, December). Automatically Detect the coronavirus (COVID-19) disease using Chest X-ray and CT images. In 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA) (pp. 150-156). IEEE.

[4] S. Maurya and S. Singh, “Time Series Analysis of the Covid-19 Datasets,” 2020 IEEE Int. Conf. Innov. Technol. INOCON 2020, pp. 1–6, 2020, doi: 10.1109/INOCON50539.2020.9298390.

[5] Damodharan, P., & Ravichandran, C. S. (2019). Inclusive strategic techno-economic framework to incorporate essential aspects of web mining for the perspective of business success. International Journal of Enterprise Network Management, 10(3-4), 329-349.

[6] Pandian, M. T., Damodharan, P., Bhavya, K. R., Singh, S., Anitha, K., & Aggarwal, A. K. (2022). Clustering time series for automatic similarity measurement selection of Database. International Journal of Human Computations & Intelligence, 1(3), 1-7.

[7] Damodharan, P., & Ravichandran, C. S. (2019). Applicability evaluation of web mining in healthcare E-commerce towards business success and a derived cournot model. Journal of medical systems, 43, 1-10.

[8] Singh, S., Aggarwal, A. K., Ramesh, P., Nelson, L., Damodharan, P., & Pandian, M. T. (2022, August). COVID 19: Identification of Masked Face using CNN Architecture. In 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC) (pp. 1045-1051). IEEE.

[9] Sahana, Sudipta, Damodharan Palaniappan, Sunil Devidas Bobade, Shaik Mohammad Rafi, B. Kannadasan, and N. Jayapandian. "Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data." Journal of Pharmaceutical Negative Results (2022): 485-495.

[10] S. Singh et al., “Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models,” J. Infect. Dev. Ctries., vol. 14, no. 9, pp. 971–976, 2020, doi: 10.3855/JIDC.13116.

[11] Mridha, K., Jha, S., Shah, B., Damodharan, P., Ghosh, A., & Shaw, R. N. (2022, January). Machine learning algorithms for predicting the graduation admission. In International Conference on Electrical and Electronics Engineering (pp. 618-637). Singapore: Springer Singapore.

[12] Kavitha, M. S., & Damodharan, P. (2013, July). Pcloud implementing saas in distributed system. In 2013 International Conference on Current Trends in Engineering and Technology (ICCTET) (pp. 416-417). IEEE.

[13] A. Sen, U. Kala, and A. Manchanda, “Analysis and prognosis of COVID- 19 pandemic in India - A machine learning approach,” Proc. 2021 1st Int. Conf. Adv. Electr. Comput. Commun. Sustain. Technol. ICAECT 2021, pp. 836–841, 2021, doi: 10.1109/ICAECT49130.2021.9392449.

[14] Jayashree, M. M., & Damodharan, P. (2018). An improved multilevel resource handling strategy for cloud based video streaming. International Journal of Scientific Research in Science and Technology, 4(8), 344-351.

[15] Damodharan, P., Aravind, P., Gomathi, K., Keerthana, R., & ManishaSamrin, K. (2017). Controlling input device based on Iris movement detection using artificial neural network. int j sci, 2(2), 634-642.

[16] Inthumathi, M. S., & Damodharan, P. (2016). PPDM and Data Mining Technique Ensures Privacy and Security for Medical Text and Image Feature Extraction in E-Health Care System. International Journal of Computer Science and Information Technologies, 6(6), 5126-5129.

[17] Kavitha, M. S., & Damodharan, P. (2013, July). Pcloud implementing saas in distributed system. In 2013 International Conference on Current Trends in Engineering and Technology (ICCTET) (pp. 416-417). IEEE.

[18] “Coefficient of Determination (R Squared): Definition, Calculation - Statistics How To,” 2018. https://www.statisticshowto.com/probability-and-statistics/coefficient-of-determination-r-squared/ (accessed Apr. 07, 2022).

[19] Veena, K., P. Damodharan, and N. Suguna. "Intrusion Detection System using Intelligent Deep Boltzmann Machine." (2019).

[20] Pandian, M. T., & Damodharan, P. (2023). Forming the Cluster in the RFID Network for Improving the Efficiency and Investigation of the Unkind Attacks. Computational Intelligence in Analytics and Information Systems: Volume 2: Advances in Digital Transformation, Selected Papers from CIAIS-2021, 295.

[21] Institute of Business Forecasting & Planning, “Mean Absolute Percentage Error (Mape),” SpringerReference, 2011. https://ibf.org/knowledge/glossary/mape-mean-absolute-percentage-error-174#:~:text=The MAPE calculation is as,bars stand for absolute values (accessed Feb. 22, 2022).

[22] Sudalairajkumar, “COVID-19 in India | Kaggle,” Kaggle, 2020. https://www.kaggle.com/sudalairajkumar/covid19-in-india (accessed Jan. 27, 2022).