iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37716

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Comparative Analysis of Deep Learning and Statistical methods for Covid-19 Cases Prediction

Shubhnesh Kumar Goyal
Associate Professor, Dharam Samaj (P.G.) College, Aligarh, India

Abstract: ***COVID-19 or novel coronavirus has affected the entire world causing widespread damages to human lives. Mathematicians and Statisticians have been trying to model the epidemiology of this disease to come with models that can predict its spread and outbreaks. Due to availability of time series like data in the case of this disease's spread, Deep Learning and Statistical models like ARIMA have been strong contenders for the prediction modelling. This paper presents an explorative and comparative approach to these methods comparing results of LSTM and CNN models for short term predictions in India during April 2020, and exploring the use of ARIMA models in the context. This study suggests that while LSTM outperforms CNN in simple time series modelling of cases registered, with inclusion of more volume of data features, CNN tend to outperform LSTM. Also, ARIMA model predictions done over Spain and Italy are found to be in very good agreement with the reported cases, suggesting an alternative method to Deep Learning approaches for short term predictions.***
Keywords: Deep Learning, CNN, ARIMA, COVID-19 prediction, Epidemiology, Optimisation

I. INTRODUCTION

First case of pneumonia like symptoms was found in Wuhan in 2019. This marked the first case of a novel coronavirus strain called COVID-19. Due to human-to-human transmission, this virus spread throughout the world and have caused widespread damages to human lives. To mitigate the effects of this virus on healthcare infrastructure, scientists have been trying to model the growth of number of cases using various models.

There have been statistical models like SIR modelling [3], deep learning models [1,2,8], models relying on global data [7] and exhaustive models build using agent and host simulation runs. All of these models have been suggested in varied capacities to predict how disease will spread and how the number of cases will grow. Perfecting the prediction of epidemiological spread of this disease will help governments prepare better for fighting COVID-19 as well as any future pandemic. Due to the availability of all these methods, there seem to be strong need of research in the direction of comparative analysis over pros and cons of each model along with best scenarios to apply them.

This research takes this approach and compare the LSTM model presented by S. Goyal [8] with a new CNN architecture based on model proposed by Huang C et. al. [2] in different states of data features. This study also explores the use of ARIMA based models and study them in case of Italy's and Spain's outbreak.

II. DATASET PREPARATION

International data was taken from John Hopkins maintained database [4,5] and national data was taken from the government-maintained dataset on the COVID-19 Tracker website [6]. Minmax scaling of the data is done in all of the models and the scaler is fitted only on the training data to prevent the forward bias. Two different sets of features are also used with details given afterward. Data till 12th April (from 14th March for India)
was used for reports.

III. NEURAL NETWORKS (CNN and LSTM)

Convolutional Neural Networks are a type of deep networks which are very often used for image classification. However, they can also be used to predict time series like videos and sounds. We predict a time series (the number of active cases of COVID-19 on a particular day) in which we have one-dimensional data, **we thus use one-dimensional convolution layer [2]**. Besides this LSTM model was imported [8] and benchmarked against CNN model for comparative analyses.

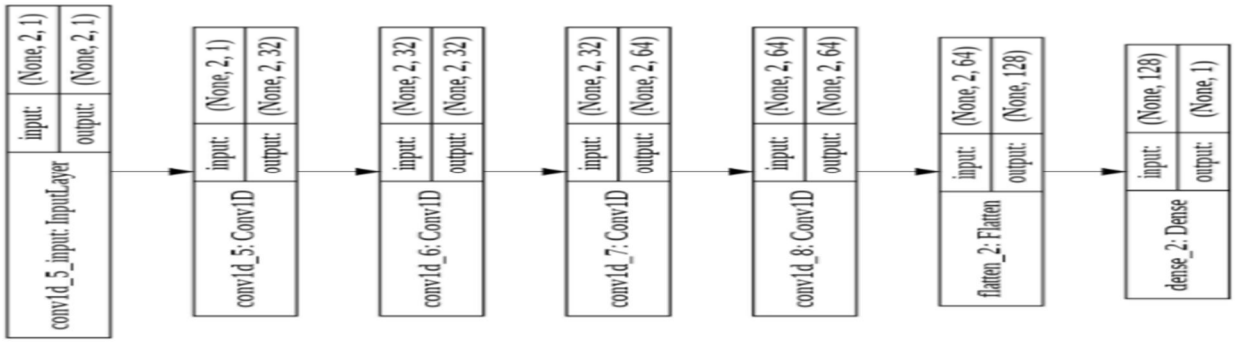


Figure 1 represents the network architecture of the CNN model used. All other parameters were fixed as in [8].

With single feature (only confirmed cases as inputs) we have the following results for CNN

	Lookback=1	Lookback=2	Lookback=3	Lookback=4
R2 Train	0.92	0.99	0.8	0.99
R2 Test	0.19	0.89	0.16	-0.88

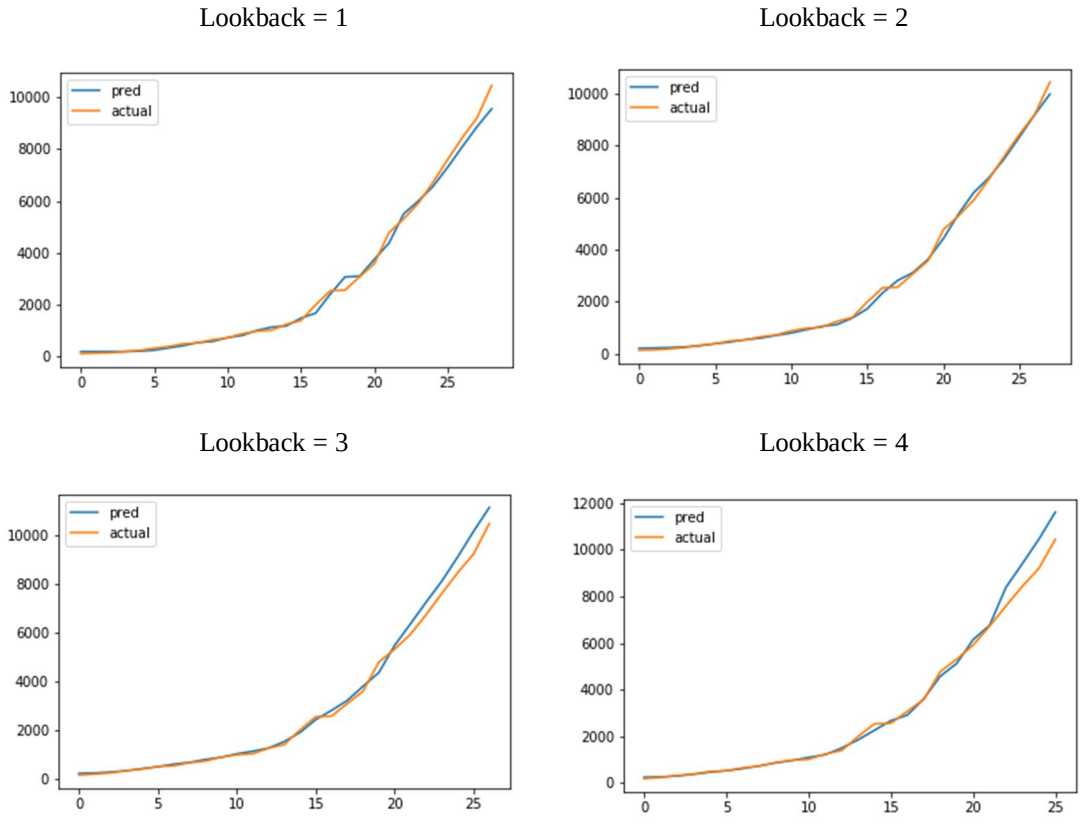


Figure 2 represents prediction vs actual cases using CNN model for different lookouts

The predictions using single features fall short when compared to the LSTM architecture suggested above, thus we went on to explore the benefits of using this architecture in high volume data i.e., with a greater number of features, because CNNs come with an advantage of reduced number of parameters due to pooling filters, and translational learning of features in the data.

To better capture the changing demographics, we tried to add different feature sets:

1) *Features Set 1:* Total and New confirmed cases, Total and New recovered cases, Total and New deaths.
2) *Features Set 2:* Total confirmed cases and top-five state totals based on the highest daily growth.

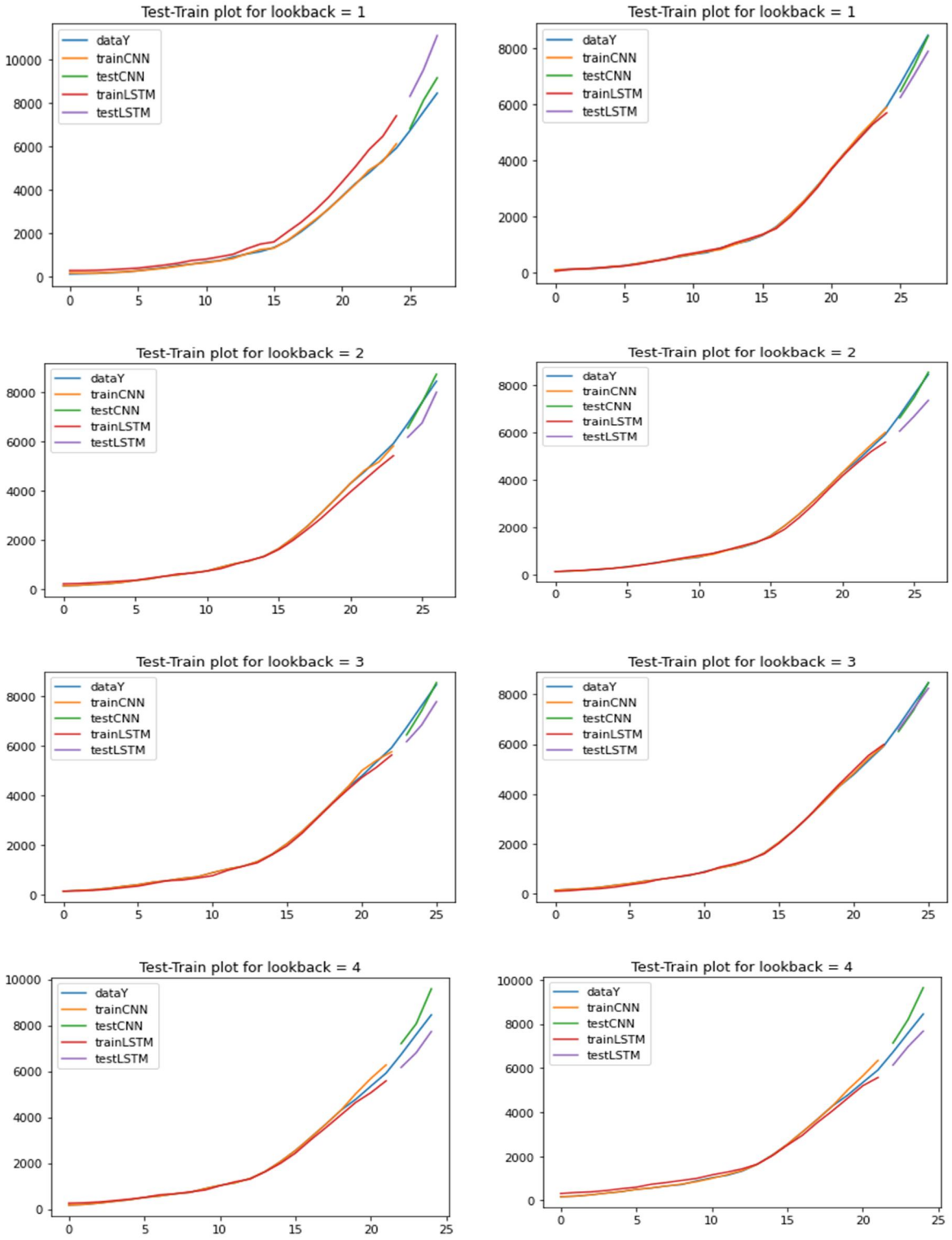


Figure 3 represents CNN vs LSTM predictions for First features set (Left) vs Second (Right) on 4 lookbacks

It can be inferred that as the amount of data in each training point i.e., timestep*features increase, CNN outperforms the LSTM on an average. It was shown that demographic and local factors can be used for prediction and they can help predict the uncertain and sudden changes in the next day and if these models are explored, then CNN would be a great option for training as it showed promising results in the feature-rich dataset and high-volume training point.

This fact may be attributed to the grouping of weights that is done by CNN thus reducing the parameters to be tuned drastically, but this statement needs more research to follow.

IV. AUTO-REGRESSIVE INTEGRATED MOVING AVERAGE (ARIMA)

ARIMA, ('Auto Regressive Integrated Moving Average') is a class of models that explains a given time series based on its own past values and past errors. ARIMA models consist of both AR (Auto -Regressive) and MA (Moving Average Models) and can forecast values coming from a non-stationary time series.

AR model: It is a regressive model based on the past predictions

MA model: Regressive model based on past forecast errors

Lags of the stationarized series in the forecasting equation are called "autoregressive" terms, lags of the forecast errors are called "moving average" terms, and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series.

A stationary time series is one whose statistical properties such as mean, variance, auto-correlation are constant over time. Thus, Predicted value of Y = a constant and/or a weighted sum of one or more recent values of Y and/or a weighted sum of one or more recent values of the errors.

A nonseasonal ARIMA model is classified as an "ARIMA (p, d, q)" model, where:

1) p means the number of preceding ("lagged") Y values that have to be added/subtracted to Y in the model, so as to make better predictions based on local periods of growth/decline in our data. This captures the "autoregressive" nature of ARIMA.

2) d represents the number of times that the data have to be "differenced" to produce a stationary signal (i.e., a signal that has a constant mean over time). This captures the "integrated" nature of ARIMA. If d=0, this means that our data does not tend to go up/down in the long term (i.e., the model is already "stationary"). In this case, then technically you are performing just ARMA, not AR-I-MA. If p is 1, then it means that the data is going up/down linearly. If p is 2, then it means that the data is going up/down exponentially.

3) q represents the number of preceding/lagged values for the error term that are added/subtracted to Y. This captures the "moving average" part of ARIMA.

The forecasting equation is constructed as follows. First, let y denote the d^{th} difference of Y, which means:

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

If d=2: $y_t = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2}$

V. METHODOLOGY (USING STATSMODEL LIBRARY)

We create a time series object after pre-processing our dataset and get the time series plot with each observation recorded on a daily basis for the specified country. Next, we need to see if the series is stationary. We use the Augmented Dickey Fuller Test to test the stationarity of the time series observations. The null hypothesis (Ho) for the test is that the data is not stationary whereas the alternate hypothesis is that the data is stationary. The Level of Significance (LOS) is taken to be 0.05.

The p - value obtained doesn't lie in the level of significance, thus specifying that the series is not stationary. Rounds of differencing are performed to calculate the value of 'd' that makes our series stationary and it comes out to be 2 for Italy and Spain and 3 for India. Thus, parameter d is obtained.

Next, we calculate and plot the residuals along with their ACF and PACF by considering a simple ARIMA (1,2,1) model. The values of p and q are decided based on significance of coefficients, ACF and PACF plots of residuals, LLR test (Log Likelihood Ratio Test) LLF values and AIC (Information criteria)

The best fit for Italy comes out to be: ARIMA (2,2,1) and for Spain: ARIMA (3,2,2)

Statsmodel library doesn't support ARIMA with d>2. Thus, results for India couldn't be obtained.

VI. RESULTS

Predictions for the next 10 days (3rd April'20 to 12th April'20) were made and were compared to actual confirmed cases for Italy and Spain.

A. *Italy*

In Italy, the forecast done by the ARIMA (2,2,1) model turns out to be very close to the actual number of Confirmed Cases. The error in prediction ranged from 0.42% to 6.64%.

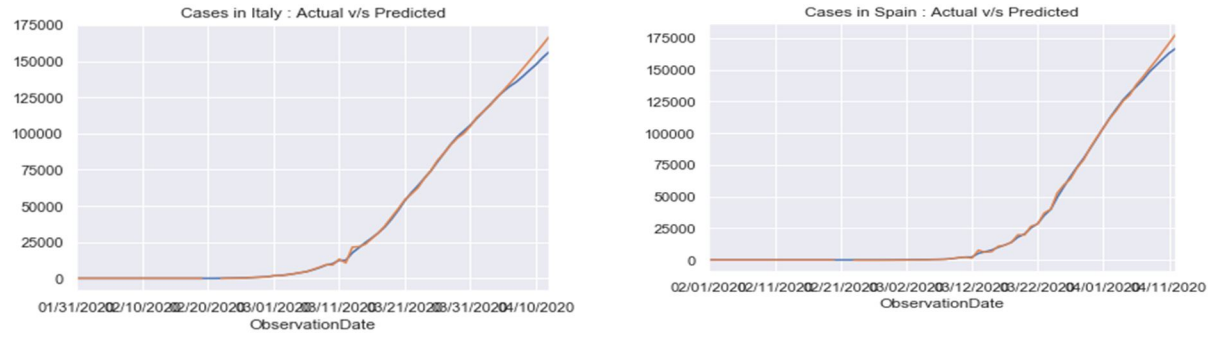


Figure 4 represents ARIMA (2,2,1) model prediction for Italy and ARIMA (3,2,2) model prediction for Spain

B. *Spain*

Predictions for the next 10 days in Spain from the ARIMA (3, 2, 2) model come out as approximately true indicators for the total number of confirmed cases with the error in prediction ranging from 1.09 % to 6.59 %. This remains a promising candidate for modelling the cases without using deep learning models, but due to resource constraints, for now we have presented results only for Spain and Italy.

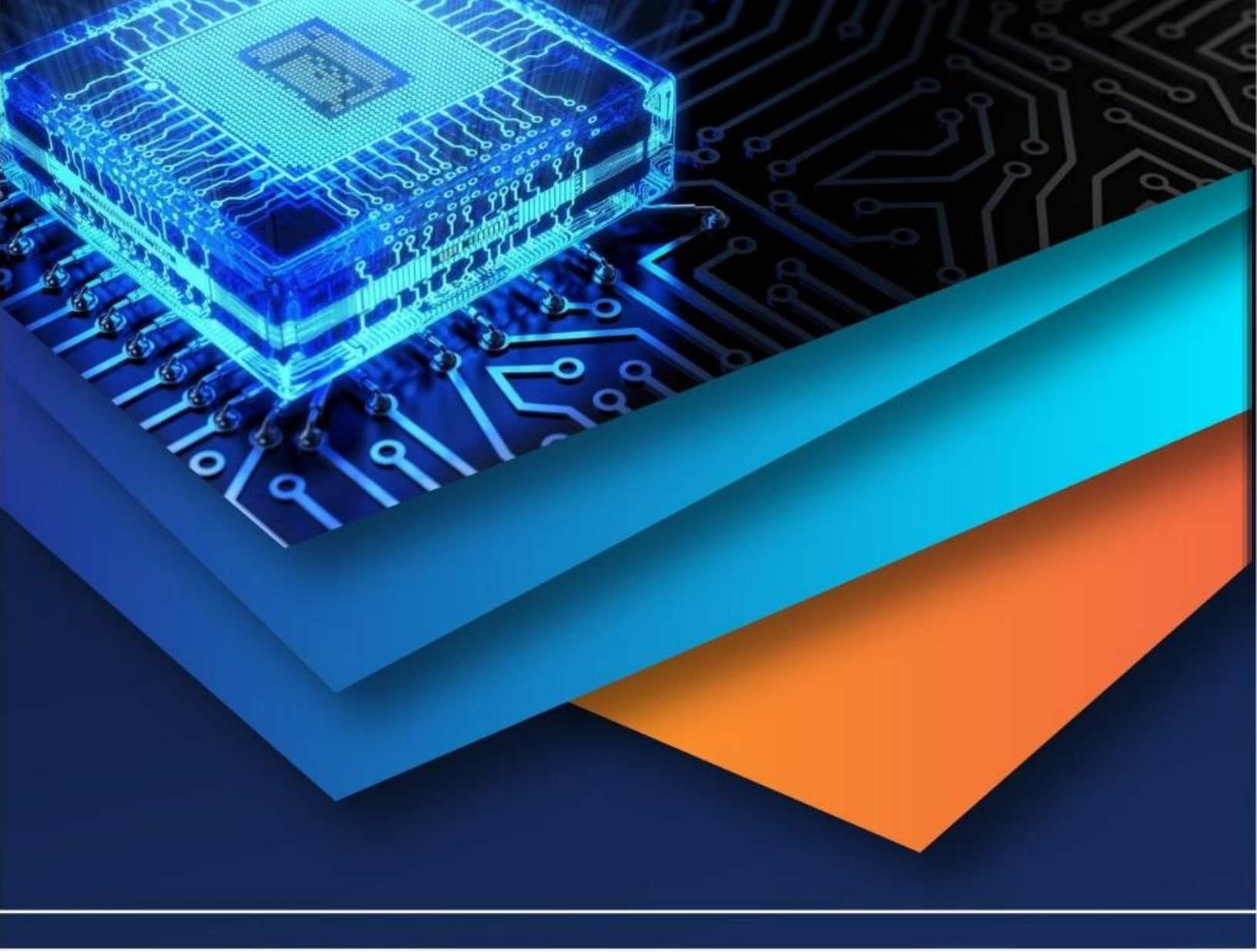
VII. DISCUSSION

This study showed that for single feature datasets, LSTMs outperform CNNs for short term predictions but as features increase and training data becomes voluminous, CNNs start outperforming LSTMs. This study also shows the effects of changing lookback on prediction accuracy and for lookbacks 1, 2, and 3, excellent results were obtained but with lookback increased to 4, predictions tend to deviate for the data in the scope of this paper. This study also shares detailed architecture guidelines for this CNN model which has performed really well and we think that can be used for preparation of prediction reports by relevant authorities.

We then explore the use of comparatively less computationally expensive methods like ARIMA, which don't rely on many variables and can thus help in prediction in small datasets. ARIMA (2,2,1) was found to be in excellent agreement with Italy while ARIMA (3,2,2) brilliantly captured the Spain's outbreak. This shows this method's potential in small window prediction of COVID-19 cases. We have finally concluded the study with sharing the detailed procedure of arriving at the ARIMA models as well and hope that this can be used to better mitigate any upcoming challenges. Future work can be extended in fine tuning of presented models, applying them into segregated areas and localised lockdown time periods, extending ARIMA method to India, and integrating all this analysis into an automated pipeline for better industrial inclusion and scaling.

REFERENCES

[1] Chae, Sangwon et al. "Predicting Infectious Disease Using Deep Learning and Big Data." International journal of environmental research and public health vol. 15,8 1596. 27 Jul. 2018, doi:10.3390/ijerph15081596
[2] Huang C, Chen Y, Ma Y, Kuo P. Multiple-Input Deep Convolutional Neural Network Model for COVID-19 Forecasting in China. medRxiv; 2020. DOI: 10.1101/2020.03.23.20041608.
[3] Joceline Lega, Heidi E. Brown, Data-driven outbreak forecasting with a simple nonlinear growth model, Epidemics, Volume 17, 2016, Pages 19-26, ISSN 1755-4365, https://doi.org/10.1016/j.epidem.2016.10.002.
[4] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis., 20 (2020), pp. 533-534
[5] https://github.com/datasets/covid-19
[6] https://www.covid19india.org/
[7] Shubhnesh Kumar Goyal, "Predicting COVID-19 Cases in India using Global Data", International Journal of Science and Research (IJSR), https://www.ijsr.net/search_index_results_paperid.php?id=SR21720164321, Volume 10 Issue 7, July 2021, 1387 – 1394
[8] Shubhnesh Kumar Goyal, "Time Series Analysis using Deep LSTM Networks for predicting COVID-19 Cases in India", International Journal of Science and Research (IJSR), https://www.ijsr.net/search_index_results_paperid.php?id=SR21806211654, Volume 10 Issue 8, August 2021, 301 - 305



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY