
PREDICTION AND ANALYSIS OF COVID-19 USING DEEP LEARNING MODELS

R. Indumathi*¹

*¹Student, Department Of MSC Computer Science, Sri Krishna Arts And Science College,
Coimbatore, Tamilnadu, India.

ABSTRACT

In this study, the number of novel coronavirus (COVID-19) positive reported cases for 32 Indian states and union territories is predicted using Deep Learning-based models. On an Indian dataset, the number of positive instances is predicted using recurrent neural network (RNN) based long-short term memory (LSTM) variations such as Deep LSTM, Convolutional LSTM, and Bi-directional LSTM. For forecasting daily and weekly cases, the LSTM model with the lowest error is used. It has been found that the suggested strategy produces short-term predictions with great accuracy, with error rates of less than 3% for predictions made each day and less than 8% for predictions made once a week. For the purpose of quickly identifying new coronavirus hotspots, Indian states are divided into various zones depending on the distribution of positive cases and daily growth rate.

Keywords: COVID-19, Prediction, Deep Learning, RNN, LSTM.

I. INTRODUCTION

2020 would go down in history as a disastrous year for humankind on this planet. The pandemic of pneumonia with unknown causes (new coronavirus), which was initially discovered in Wuhan, China, in December 2019 [1] and first fatalities reported on January 10, 2020, is rapidly engulfing the entire planet. The World Health Organization (WHO) has designated it as COVID-19 (Coronavirus illness 2019). John Hopkins University reports that as of 16-May-20, there were 4,563,458 confirmed cases of COVID-19. With 86,508 cases and a fatality rate of 3.2% and 0.2 fatalities per 100,000 people, India provides 1.9% of all cases. All nations are attempting to safeguard the lives of their citizens by putting in place policies like travel bans, quarantines, postponements and cancellations of events, social testing, separation, and both hard and soft lockdowns. The economic and societal effects of this virus are far more devastating than the lives it has claimed, particularly for developing and underprivileged nations. The potential devastation brought on by COVID-19 in India, where 32,303 people live per square kilometer in places like Mumbai, which is home to 18% of the world's population. The enormous population of India may therefore experience a rapid spread of this novel coronavirus. The Indian government is recommending a number of lockdowns to stop the virus's spread. With the exception of critical services, the whole country was initially placed under lockdown in lockdown 1.0 (March 25, 2020, to April 14, 2020), the entire nation was under complete lockdown except for essential services and in lockdown 2.0 (April 15, 2020, to May 3, 2020) was implemented with relaxation and lockdown 3.0 (May 4, 2020, to May 17, 2020), with additional relaxations in places where there were less coronavirus cases, was adopted with relaxation in areas where the virus was confined. The number of cases has dropped daily from 11.8% to 6.3% as a result of these lockdowns. However, the government cannot permanently impose a state of emergency since the economy could suffer greatly. So, a workable option might be to quarantine the most important areas, ensuring that those infected by the virus only stay there.

The World Health Organization reports that this virus is still without a vaccination and no antiviral medications are available for it, but medical organizations are working hard to develop a vaccine for this brand-new coronavirus. The vaccine may take at least 18 to 24 months to become ready, and it may take considerably longer to create enough of it to cover the majority of the population, even after expediting the typical vaccine term of 5 to 10 years. As the virus mutates, we are unsure of how long a vaccination might be effective. Every effort is taken to contain the coronavirus's spread, set up medical response systems to handle the influx of patients, and safeguard the front-line medical workers with sufficient supply of personal protective equipment (PPE), masks, and other necessities. Therefore, we can arrange our inventory appropriately if we know in advance the number of unique coronavirus cases, let's say over the upcoming few days.

II. RELATED WORK

There are minimal number of articles on the prediction of novel coronavirus cases in the literature, and few of them are examined below.

Patient Information Based Algorithm (PIBA) has been reported by Wang et al. for predicting the number of deaths brought on by this COVID-19 in China. According to predictions, 13% of deaths would occur altogether in Hubei and Wuhan, while 0.75–3% would occur elsewhere in China. They also stated that varied climates and temperatures would affect the death rate. Based on the United States spread analysis, a case was provided in that demonstrated a direct correlation between temperature and COVID-19 instances. It predicted a sharp decline in the number of cases in India throughout the summer, but this didn't materialize. In order to forecast COVID-19 instances and the Spanish stock market over the short term, Ahmar and Val employed ARIMA and Sutte ARIMA. As of April 16, 2020, they have provided their predictions with a MAPE of 3.6%. For estimating the number of positive cases in Italy, Spain, and France, Ceylan employed ARIMA models. He recorded a MAPE of between 4% and 6%. In Fanelli and Piazza conducted COVID-19 forecasting and analysis in China, France, and Italy. They have predicted how many ventilation units Italy will need based on their findings. They classified the population into four groups—vulnerable, healthy, infected, and dead—and used that information to forecast the number of cases. Deep learning model (LSTM) was used by Reddy and Zhang to forecast when the outbreak in Canada would end. Their long-term model accuracy is 92.67%, compared to 93.4% for the short term.

We provide a deep learning-based methodology for estimating the potential patient population for COVID-19 infection. For several Indian states and union territories, we have forecasted the number of new coronavirus positive cases from one day in advance to one week in advance. For the prediction, we have employed recurrent neural networks and LSTM-based models. On the other hand, we tested a variety of LSTM models on an Indian dataset and discovered that more complex LSTM models, such as stacked LSTM, convolutional LSTM, and bi-directional LSTM models, provide better accuracy than straightforward LSTM models.

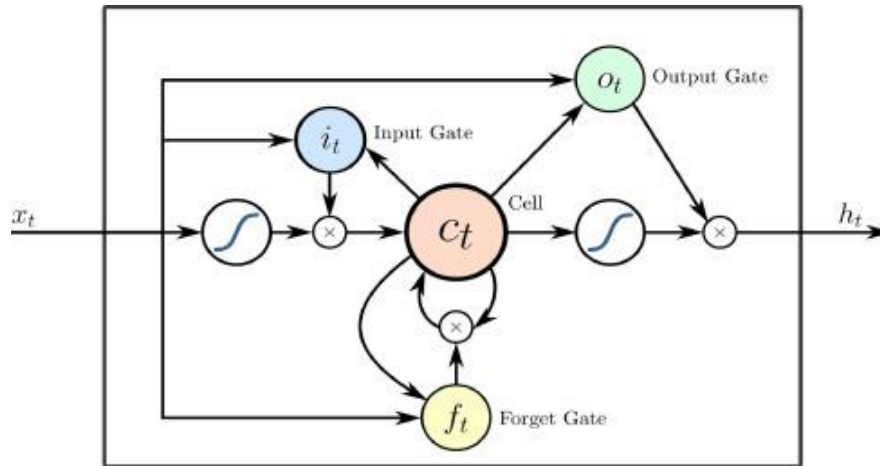
III. METHODOLOGY

Background: Recurrent neural networks (RNN)

According to deep learning theory, a deep sequential or hierarchical model performs classification or regression tasks more effectively than shallow ones. Recurrent neural networks have hidden states that are spread out throughout time, which enables them to store a lot of historical data. Since they can handle variable length sequential data, forecasting applications are where they are most frequently utilized. Because they only use the hidden layer activation functions of the previous time step, recurrent neural networks have the significant drawback of being unable to solve the vanishing gradient or exploding gradient problems and being able to store only short-term memories.

Long-short-term memory (LSTM) & its variants:

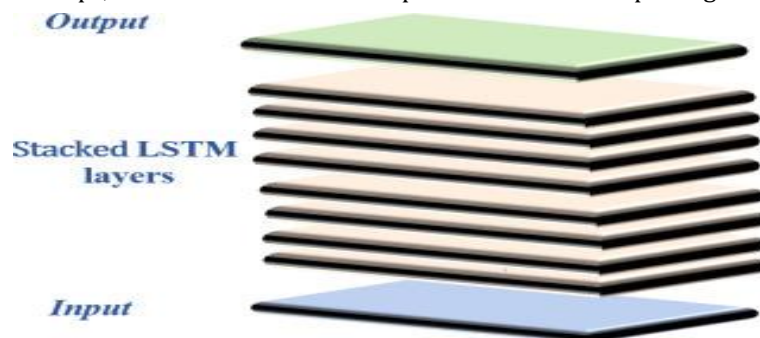
LSTMs are among the most practical solutions for prediction problems since they estimate future forecasts based on different highlighted aspects available in the dataset. The data in LSTMs passes through elements referred as cell states. LSTMs are capable of accurate recollection or forgetfulness. Time series data are information that has been accumulated over a period of time, and LSTMs are often suggested as a reliable approach to forecast using these data values. In this kind of design, the model moves from the previous state of being shrouded to the next step of the arrangement. Long short-term memory cells (LSTM) are used in conjunction with RNNs for long-term memory storage because RNNs can only store a certain amount of data. RNN's problems with exploding and vanishing gradients are solved by LSTMs. Similar to RNN, LSTM cells have hidden units that can be changed out for memory blocks.



LSTM cell

LSTM deep/LSTM stacked:

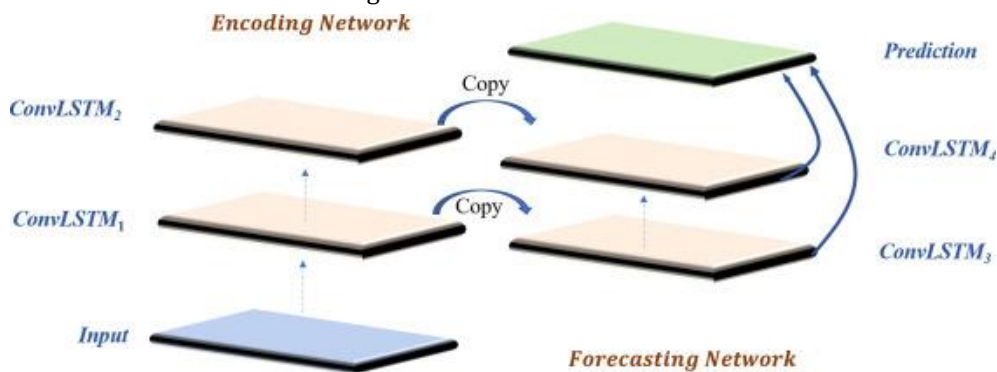
The above-described basic LSTM is extended by stacked LSTM, also referred to as deep LSTM. There are numerous hidden layers and numerous memory cells in layered LSTM. The depth of neural networks, where each layer holds some information and transmits it to the next, is increased by stacking a number of layers. Sequence data is provided by the top LSTM layer to the layer below it, and so forth. Instead of delivering a single output for all time steps, it offers an individual output for each time step using a stacked LSTM structure.



LSTM deep/LSTM stacked

LSTM with convolutions (Conv-LSTM):

In a convolutional LSTM, the last two dimensions of the input x vector, cell output y vector, hidden state vector h , and the gates (i_t, f_t, o_t) are all 3D tensors. The inputs of Conv-LSTM determine both the past states of linked cells and the future states of each cell in the grid.



Convolutional LSTM network for forecasting

LSTM in both directions (Bi-LSTM):

Standard RNNs simply consider the inputs in one direction and ignore any future knowledge. By using the LSTM's bidirectional topology, this problem is solved. By taking into account both the past and the future, the bidirectional LSTM (Bi-LSTM) recovers the entire temporal information of time t . This divides the standard

RNN's hidden neurons into forward and backward states, with the forward state neurons not coupled to the backward state neurons and vice versa. This structure is comparable to the conventional unidirectional RNN without the backward states. This structure eliminates the requirement for additional time delays that are used in traditional RNN. The following is a summary of the Bi-LSTM training procedure over time:

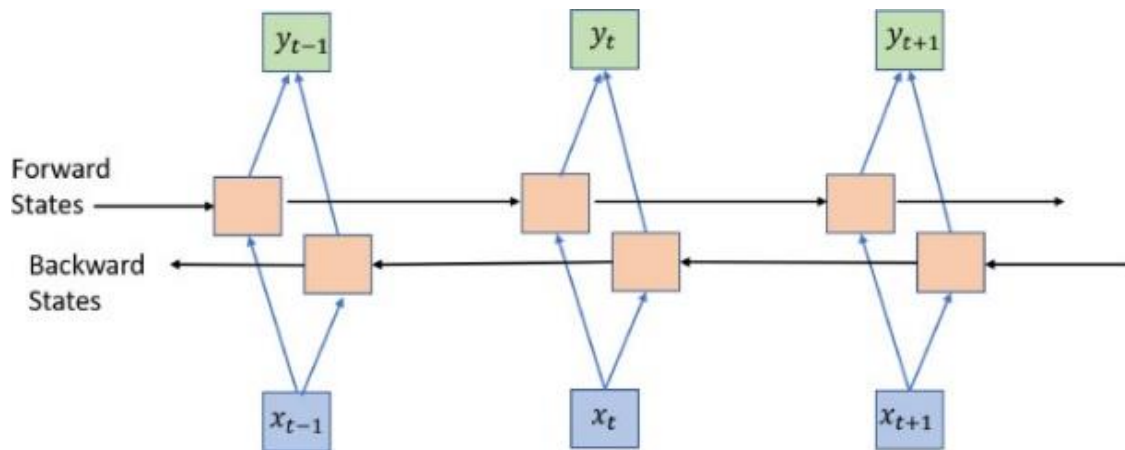
1. Forward pass:

- Run input data for time $1 \leq t \leq T$ through Bi-LSTM and evaluate the predicted outputs as calculated in standard RNN.
- Run forward pass for forward states from $t=1$ to $t=T$ and backward states from $t=T$ to $t=1$.
- Run the forward pass for output neurons.

2. Backward Pass:

- Evaluate the objective function derivative for time $1 \leq t \leq T$ calculated in forward pass.
- Run backward pass for output neurons.
- Run backward pass for forward states from $t=T$ to $t=1$ and backward states from $t=1$ to $t=T$

3. Update weights:

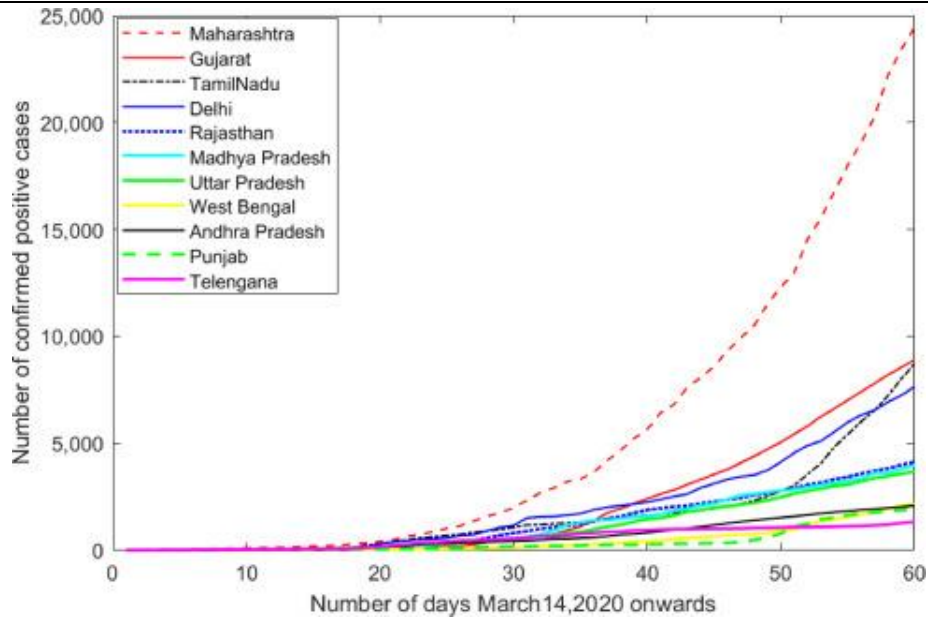


Bidirectional LSTM

IV. DATA ANALYSIS

Description of a data set:

The Ministry of Health and Family Welfare provided the data set for this paper (Government of India). As the number of occurrences increases or decreases depending on other environmental/physical variables, this data is very stochastic in nature. It is made up of 32 distinct time-series data of COVID-19 instances that have been officially confirmed in each of the states and union territories since March 14, 2020 (4). The linear weighted moving average missing data statistics technique is used to impute the missing values in each of the individual series so that the model can continue to learn sequentially and make reliable predictions for the future. We collected information for our research (from March 14, 2020 to May 14, 2020). Data has been divided into training data (from March 14, 2020 to May 8, 2020) and testing data (from May 9, 2020 to May 14, 2020). The 11 Indian states where the number of COVID-19 positive cases exceeds 1000 in 60 days are shown in Fig. 5. This graph demonstrates that while the number of cases is increasing linearly in states like Rajasthan and Madhya Pradesh, and at a very high rate in a small number of states like Maharashtra (14.3%), Tamil Nadu (18.3%), Gujarat (19.3%), and Delhi (13.7%), as well as states like Telangana, West Bengal, and Punjab. Except for these 11 states, the growth rate is quite low in all other states and union territories.

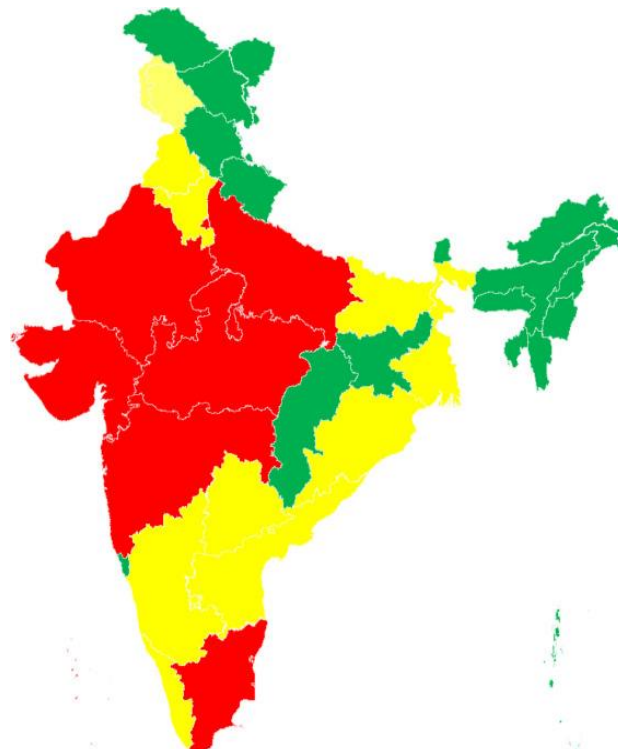


Indian states with number of COVID-19 positive cases above 1000 from March 14, 2020 to May 14, 2020

Dot plot analysis:

Based on the positive COVID-19 instances and daily rise depicted on the India map, we split India into three groups.

- Mild Zone—All states are retained in the mild zone if the overall number of positive COVID-19 cases is less than 200 and the daily increase is less than 2%. It is highlighted in green.
- Moderate Zone—States with 200–2000 COVID-19 positive patients and a daily increment of less than 5% fall into this category. It is highlighted in yellow.
- States that have more than 2000 COVID-19 positive patients and a daily increment of more than 5% are considered to be in the severe zone. It is highlighted in red.

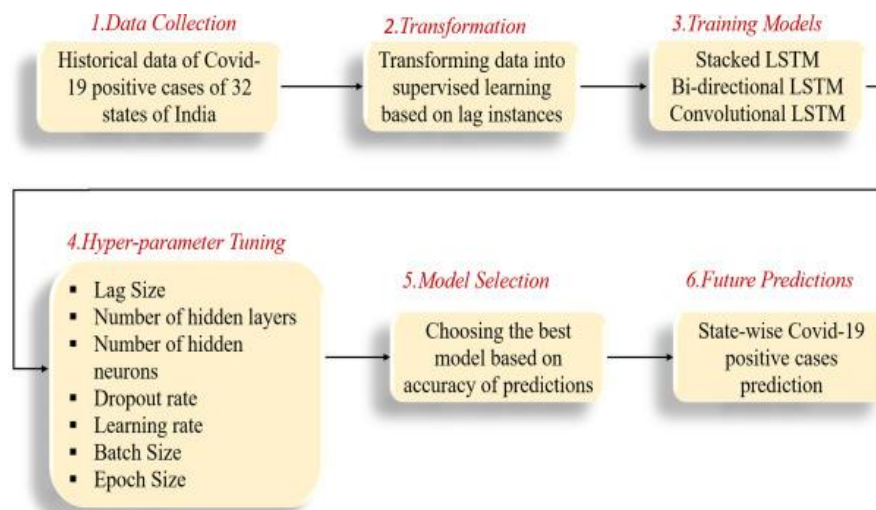


Division of India in the severe (red), moderate (yellow) and mild (green) zones depending upon the number of confirmed COVID-19 positive cases and daily rise based on the data till May14,2020.

V. RESULTS AND DISCUSSION

Model creation:

On open source libraries like Numpy, Pandas, Tensorflow (Google), and Keras, the tests are carried out. Python, a high-level general-purpose programming language, is utilised as the application programme interface for deep learning frameworks (APIs). The resulting APIs are used to create the current model structures for Deep LSTMs, Bi-LSTMs, and Conv-LSTMs, among other recurrent neural network versions. In order to create future projections of the number of confirmed cases present in any given region, these models are used to learn the dynamic dependant structure existing in the data as well as mapping the learning sequence present. Due to the dynamically changing dataset structure, we used history data from March 14, 2020 to May 14, 2020 for training and testing our prediction models. These models are given region-based historical data on the number of instances emerging daily. Every one of these models has undergone intensive hyper-parameter tuning. The mean squared error loss in these models is optimised using the Adam optimiser. The testing data set is used to evaluate these models' errors, and the best model with the highest degree of accuracy is chosen to serve as the prediction model for COVID-19 data. As the number of confirmed cases cannot be expressed in decimal numbers, output values are then rounded to the nearest integer value. Below is a diagram of our suggested methodology.



Layout of proposed method

Design of the deep LSTM/stacked LSTM models:

A stacked LSTM model is an LSTM model that consists of numerous LSTM layers. Another part of hyperparameter optimization is choosing how deep the model should be; this can normally range from a single layer to a three- to four-layer deep model architecture, with a three-layer architecture being most commonly employed for difficult learning tasks. A two-layer deep LSTM system with 100 hidden neurons units per layer makes up the model employed in this experiment. The lag structure is discovered to be the input form for the model, with 3 steps and 1 feature. Additionally, the model makes use of the ReLu activation function to get around the vanishing gradient problem, which is a common issue with recurrent neural networks.

Design of the Convolutional LSTM (Conv-LSTM) model:

Conv-LSTMs are a type of LSTM that replaces matrix multiplication with convolutional operations at each of the current cell gates inside the LSTM cell. This allows the model to map long-term sequential dependencies in the data because convolutional input and recurrent transformations take place inside each cell. A single convLSTM2D layer and a heavily coupled output layer made up the Conv-LSTM model employed in this experiment. The lag structure is the input shape that is fed into the model, and the number of subsequences and input characteristics are discovered to be 4, 2, and 1, respectively. The data input is reorganised into the input shape described above, together with the kernel size and 64 filters. The integer value (1,2) serves as the convolution window and is utilised to create the feature map. Using the test data with three time lags, the trained model is then utilised to produce day-ahead forecasts for each of the states/UTs.

Design of a bi-directional (Bi-LSTM) LSTM model:

Bi-LSTM differs from unidirectional LSTM in that it can at any time protect data from both the past and the future by combining the two concealed states. The LSTM that runs backwards protects data from the future. As they consider both the past and the future, Bi-LSTMs better understand the settings. A single hidden layer of 100 hidden neuron units is employed in the Bi-LSTM model used in this experiment, which is wrapped in a bidirectional wrapper class structure. This entails duplicating the system's main intermittent layer so that there are now two layers sitting on top of one another. The method of combining the two layers is chosen to be 'concat' from a variety of possibilities available as 'sum' or 'mul' value. The number of steps, often referred to as the lag structure behind in space, is chosen as the input shape provided into the model. Finally, a dense layer structure is sent through the bidirectional wrapper classes to generate the prediction values. The model is trained using the mean squared error as a loss parameter and trained through a predetermined number of epochs, which helped to resolve gradient problems.

Mistake comparison:

Following is a comparison of the effectiveness of our suggested prediction approaches using performance measure indices like mean absolute percentage error (MAPE).

MAPE measures accuracy as a rate, which may be calculated as the ratio of the real values to the projected values divided by the actual values for each time frame. In Table 1, we have used convolutional, stacked, and bi-directional LSTM models to assess the MAPE of the 32 Indian states. The average MAPE for stacked LSTM is 4.81%, for bi-directional LSTM it is 3.22%, and for conv-LSTM it is 5.05%. In a few states, MAPE of 0% (ideal) shows that our model correctly predicted the number of occurrences. MAPE ranges from 15.35% in bi-directional LSTM to 15.37% in convolutional LSTM and up to 30.67% in stacked LSTM. On a 15-day testing dataset, these errors are estimated in order to choose the optimal model. A model with a wide error range may fluctuate more with little changes and be more susceptible to large departures from expected results. Additionally, Bi-LSTM is better suited for prediction purposes than convolutional and stacked LSTM since it has a smaller error range and the lowest average error across all models. So, in order to evaluate predictions of COVID-19 positive cases, we employed the Bi-LSTM model.

Table 5.1: Mean Absolute Percentage Error (MAPE) of states and union territories (UTs) of India by convolutional, stacked and bi-directional LSTM models

S.No.	States/UTs	Convolutional LSTM	Stacked LSTM	Bi-directional LSTM
1	Andaman and Nicobar	0	0.2	0
2	Andhra Pradesh	3.2	1.6	1.24
3	Arunachal Pradesh	0	0	0
4	Assam	7.28	6.3	5.49
5	Bihar	7.03	4.95	5.3
6	Chandigarh	8.76	8.3	6.64
7	Chhattisgarh	12.94	11.05	10.9
8	Delhi	2.86	3.4	2.13
9	Goa	0	0	0
10	Gujarat	2.78	2.02	0.99

S.No.	States/UTs	Convolutional LSTM	Stacked LSTM	Bi-directional LSTM
11	Haryana	5.94	5.23	4.35
12	Himachal Pradesh	5.57	3.81	2.68
13	Jammu and Kashmir	2.36	1.82	1.53
14	Jharkhand	5.46	3.53	2.95
15	Karnataka	3.06	2.31	1.71
16	Kerala	2.04	0.74	0.63
17	Ladakh	12.23	11.19	7.63
18	Madhya Pradesh	4.38	4.44	1.9
19	Maharashtra	2.43	2.23	1.29
20	Manipur	0	0	0
21	Meghalaya	1.1	0.55	0.55
22	Mizoram	0	0	0
23	Odisha	7.79	6.4	5.88
24	Puducherry	3.13	12.65	3.13
25	Punjab	18.02	12.07	7.95
26	Rajasthan	1.3	2.35	1.35
27	Tamil Nadu	7.17	5.33	3.53
28	Telangana	1.83	1.39	0.97
29	Tripura	21.16	30.67	15.35
30	Uttar Pradesh	3.37	2.32	1.11
31	Uttarakhand	2.03	2.26	1.8
32	West Bengal	6.25	4.95	4.16
Average	5.05	4.81	3.22	

Prediction:

As evaluated on the India dataset for 15 days, the anticipated number of COVID-19 positive cases closely matches the actual number of cases (April 30, 2020, to May 14, 2020). In Fig. 8, the error percentage is indicated in red; for instance, this model's error for May 14, 2020 is 0.279%. For data from May 9 to May 15, this model is also tested state-wise for daily and weekly forecasts. Utilizing data up to May 8, 2020, we used the

Bi-LSTM model to construct a 1-week prediction from May 9, 2020 to May 15, 2020. The following day's prediction was then assessed using the previous day's actual values. For four Indian states, daily and weekly prediction errors are determined. (Maharashtra, Tamil Nadu, Delhi and Rajasthan). On our website, which was created for novel coronavirus forecasts, you can also find predictions for other states. It is clear from this table that the weekly prediction MAPE ranges from 4% to 8%, and the daily MAPE is within 3%. For weekly predictions, we can see that the inaccuracy starts to rise significantly after the fourth day. As a result, we may conclude that this model is very accurate for predictions made within a short period of time (between one and three days).



15 days comparison of predicted and actual Covid-19 positive cases by bi-directional LSTM model for India for the year 2020

Table 5.2: Daily and weekly error percentages for one-week testing data using Bi-directional LSTM model

Empty Cell	Maharashtra		Tamil Nadu		Delhi		Rajasthan	
Empty Cell	Daily Error %	Weekly Error%	Daily Error %	Weekly Error %	Daily Error %	Weekly Error %	Daily Error %	Weekly Error %
09-May	2.70	2.20	5.31	5.00	2.74	1.57	0.65	0.30
10-May	1.35	0.41	1.01	7.61	0.58	3.18	6.82	4.30
11-May	0.96	2.17	0.30	12.30	1.11	5.43	0.15	3.64
12-May	0.87	6.45	0.78	12.50	0.14	6.28	0.73	5.28
13-May	2.63	8.96	6.55	9.80	3.63	7.58	0.85	6.56
14-May	1.18	10.51	1.20	5.93	0.61	8.58	2.14	5.82
15-May	0.00	12.66	1.88	0.89	0.15	10.55	3.12	6.00
MAPE	1.39	6.20	2.43	7.72	1.28	6.17	2.06	4.56

The state authorities will be assisted in balancing the load that the medical infrastructure can support by the extremely accurate short-term estimates made at the state level. Predictions may also lead to the imposition or removal of lockdowns, which would ensure that economic activity can resume even though doing so might otherwise pose a threat to millions of people's ability to support themselves.

VI. DISCUSSION

Our suggested model's output can aid authorities and planners in making lockdown decisions. The state-wise forecasts will assist the state authorities in balancing the load that the medical infrastructure can support. This will also ensure that economic activity can continue, which could otherwise pose a challenge to millions of people's ability to support themselves. We categorise the zones into mild, severe, and moderate categories and explain how we did it. To stop or limit the rise in the number of new coronavirus patients, distinct preventative measures should be implemented for each zone.

The following is a list of preventive measures by zone:

- **Warm Zone:**

Despite the fact that these states are all under quarantine zones, they can nonetheless begin economic activity. To ensure that the situation does not spiral out of control, a COVID-19 test should be made compulsory for all incoming visitors from neighboring states.

- **Moderate Zone:**

Depending on the precautions it takes, this zone could transition into a mild or severe zone over the next few days. To stop the transmission of viruses, containment zones must be sealed off and identified. These locations can withstand a soft lockdown in non-contained zones with limited economic activity. To stop the spread of unknown and asymptomatic cases, extensive testing is necessary. Moving towards the mild zone should be the focus of all efforts.

- **Severe Zone :**

In this area, things have already spiraled out of control and put tremendous strain on healthcare providers and infrastructure. The key to controlling virus propagation is a strict lockdown with no economic activity, identification, and sealing of containment zones. Before a zone is declared virus-free, it must pass several tests. The mild zone should be reached first, followed by the moderate zone, using all available effort.

VII. CONCLUSION

The number of COVID-19 positive cases in Indian states can be predicted using the deep learning algorithms we have proposed in this paper. The rise in the number of positive cases in India has been the subject of an exploratory data analysis. States are divided into mild, moderate, and severe zones based on the number of cases and daily growth rate in order to implement more practical lockdown measures at the state level as opposed to a national lockdown, which could have negative socioeconomic effects. As prediction models, long short-term memory (LSTM) cells based on recurrent neural networks (RNN) are employed. The most accurate LSTM variation is selected after testing it on 32 states and union territories using deep LSTM, convolutional LSTM, and bi-directional LSTM models. Bi-directional LSTM performs the best and convolutional LSTM performs the poorest based on prediction errors. For all states, daily and weekly forecasts are generated, and it is discovered that bi-LSTM provides incredibly accurate results (error less than 3%) for short-term forecasting (1–3 days). On a website created for the general public, predictions are readily available to the public. These forecasts will be useful for researchers, planners, and state and federal government officials as they manage services and set up medical infrastructure. Other nations may adopt the suggested approach and preventive measure.

VIII. CONFLICT OF INTEREST STATEMENT

The authors affirm that they have no known financial or interpersonal conflicts that would have appeared to have an impact on the research presented in this study.

IX. REFERENCES

- [1] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, and others Clinical characteristics of individuals with a 2019 new coronavirus infection in Wuhan, China, according to the Lancet 395 (10223) (2020)
- [2] Z. Alsafi, M. Khan, A. Kerwan, A. Al-Jabir, C. Sohrabi, N. OfiNeill, et al. Global emergency declared by World Health Organization: analysis of the 2019 new coronavirus (COVID-19) IJ Surgery (2020)
- [3] The group W.H., et al. identifying the virus that causes the coronavirus disease (COVID-19). 2020a
- [4] <https://coronavirus.jhu.edu/> Johns Hopkins Coronavirus Resource Center 2020

- [5] Coronavirus resource centre at Johns Hopkins, 2020. Mortality analysis.
<https://coronavirus.jhu.edu/data/mortality>
- [6] Indian people (2020) worldometer, 2020. Available at: <https://www.worldometers.info/world-population>
- [7] Cities listed in order of population density List of cities proper by population density - Wikipedia, 2020.
<https://en.wikipedia.org/wiki/>