Predictions for COVID-19 transmission in India using LSTM,SVR and Random forest networks

Akash Anil Patil
Student
SCOPE
VIT Chennai
Pune, India
akashanil.patil.2020@vitstudent.ac.
in

***Abstract*—Coronavirus announced as a worldwide pandemic by WHO, has arisen as the most forceful sickness, affecting over 90% nations of the world. The infection began from a solitary person in China, is presently expanding worldwide at a pace of 3% to 5% day by day and has become a ceaseless cycle. A few examinations even foresee that the infection will remain with us until the end of time. India being the second most populated nation on the planet is likewise not spared, and the infection is spreading as a network level transmitter. Along these lines, it turns out to be truly critical to examine the conceivable effect of COVID-19 in India and figure how it will act in the days to come. In present work, expectation models dependent on hereditary programming have been created for affirmed cases (Confirmed Cases), passing cases (Death Cases) and recovered cases across three most influenced states specifically Maharashtra, Gujarat and Delhi just as entire India. The proposed forecast models comprising autoregressive integrated support vector regression (SVR), Long short term memory (LSTM) and Random Forest (RF) are assessed for time series prediction of confirmed cases, deaths and recoveries in India based on data stored by Kaggle in between 1 March 2020 to 20 November 2020 The performance of models is measured by mean absolute error, root mean square error. The forecast of different boundaries (number of positive cases, number of recuperated cases, and so forth) acquired by the proposed technique is exact inside a specific reach and will be a helpful apparatus for directors and wellbeing authorities. What's more, the models could be profoundly dependable for time arrangement expectation of COVID-19 cases in India. And the models could be highly reliable for time series prediction of COVID-19 cases in India.**

Keywords—COVID-19, Coronavirus, Time series forecasting, Prediction, LSTM, SVR

I. INTRODUCTION

Extreme Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) or Covid sickness 2019 (COVID-19) is a worldwide pandemic flare-up, the world is confronting today. The infection, which relocated from Bats to human, beginning from Wuhan, the capital of Hubei territory of China, has impacted in excess of 180 nations of the world. The main patient was accounted for on 8 December 2019, in Wuhan by the Chinese organization [1]. After a month, the primary demise from the infection was accounted for on 9 January 2020. Around the same time, world wellbeing association (WHO) proclaimed that a novel Covid infections has been recognized and is growing multi-along the side ordinary [2]. The infection which began from China, moved to the third world and first case in Thailand was accounted for on 13 January 2020 [3]. The Chinese specialists attempted to contain the infection by forcing certain exacting activities including air terminal terminations, interstate terminations, railroad interferences, public get-together suspension, stopping public vehicle, conclusion of shops, mass exercises and whatever other movement which represents get-togethers were brought to a prompt end. These measures where utilized to limit the impact of network level transmission of the infection [4]. The Chinese specialists assumed control over the control of the circumstance and gathered information from 2018 International Air Transport Association (IATA) to distinguish and check the irresistible illness weakness records (IDVIs) in new nations where the infection may have communicated outside China [5]. It should be noticed that IDVI has a scope of [0, 1], higher the estimation of IDVI, lower is the danger of illness transmission and weakness. The virus influenced in excess of 85,000 Chinese populace and the underlying objections influenced were Hong Kong, Bangkok, Tokyo and Taipei, all having an IDVI above 0.65 [6].

Despite the fact that various endeavours have been brought into place, yet the infection was not controlled and by 19 January 2020, various cases over the world were accounted for [3].WHO proclaimed the infection as a crisis circumstance for the entire world on 31 January 2020 and by 11 March 2020, it was pronounced as another undermining worldwide pandemic [4].Starting at 12 May 2020, just about 4,006,257 have been accounted for over the globe with an absolute demise check (DC) of 278,892 measuring for an everyday increment of 26% and 28% expansion in affirmed cases (CC) to passings every day [6].The most exceedingly awful influenced nation being USA, the second most influenced nation is Russian Federation, trailed by United Kingdom, Spain, Italy, Germany, France and Turkey.

India the second most crowded nation of the world with 1.3 billion individuals to serve, having a normal family pay positioned at 112th out of 164 nations by the world bank and with a 150th position in worldwide medical care by world financial gathering. This basic condition was under the scanner of the entire world, when the COVID-19 pandemic previously set foot on Indian soil [9]. The primary case was accounted for on 30 January 2020 and it was normal that India won't have the option to endure the warmth and because of absence of fundamental administrations, life of millions of individuals will be in question. The significant purpose behind such little effect of COVID-19 on India, is because of the opportune reaction from the separate focal and state governments.

Since the first day of the unleashing of the virus in India, the Indian govt. has been examining every single individual going to the nation from China and individuals who had any

Chinese travel history in the previous few days. The principal cross country lockdown for 21 days was declared by the public authority on 23 March 2020 and was additionally stretched out by an additional 156 days till 27 August 2020, which has been additionally reached out till 15 Sept 2020 (based on the current situation we can have another lockdown out till 30 Dec 2020). Significant measures included ideal reaction to give medical services offices, contact following, broad testing, network preparation and others have assisted with containing infection and keep a low death rate. Odisha, Kerala and Tamil Nadu has a long history of catastrophic events and prudent steps have just been taken by the public authority. Maharashtra on an entire uses robots to screen social separating and lockdown. A group control procedure has likewise been utilized to analyse and contain the infection. This has been finished by studying, identifying and contact following of around 3 km of region where multiple patients are analysed [9].

In present work, another hereditary programming based model (GP) [18] for times arrangement expectation of the COVID-19 situations in India has been proposed to appraise the conceivable spread of the infection. The dataset for assessment is taken from [10].The GEP model has been utilized to foresee the all-out number of cases in India dependent on two significant boundaries, these incorporate affirmed cases (CC), and demise cases (DC).

In this paper, we put forth an attempt to foresee the flare-up of COVID-19 dependent on past transmission information. The proposed estimate models including autoregressive incorporated integrated support vector regression (SVR), Long short term memory (LSTM) and Random Forest (RF) are assessed for time series prediction of affirmed cases, deaths and recuperations (recoveries).Our outcomes are relied upon to alarm the general medical services suppliers of India to set themselves up for the emergency against COVID-19

II. LITRATURE REVIEW

Every infectious disease shows certain example patterns and such examples should have been recognized dependent on transmission elements of such flare-ups. Interceding measures to kill such irresistible infections depend on the strategies used to assess the flare-up when it happens. Any flare-up in a nation or area normally happens at various degrees of extent regarding time for example occasional changes, transformation of infection after some time.

In [1]. [2].a transmission model for jungle fever and in [3]. a numerical model for dissecting elements of tuberculosis has been created to contemplate the transmission utilizing numerical models. In [4].a laplacian based decay is utilized to explain the non-straight boundaries in a Pine Witt infection. A changed SIRS model in [5].effectively assisted with controlling the syncytial infection in new-born children.

Computerized reasoning(AI) and portable figuring(Mobile Computing) are one of the vital variables for the achievement of innovation in medical care frameworks [16]. In the realm of savvy gadgets, information is being produced in the uncommon path than at any other time and advanced the part of AI in medical care [16].The present reality is more associated than any other time this assisted with sharing share the real time infectious data /information between the nations. Considering flow pandemic circumstance numerous scientists concocted a few numerical models to foresee the transmission of novel Covid [17]. [18].. The significant disadvantage of the current models are linear, non-worldly and a few suppositions while demonstrating the organization. The Coronavirus is a period time series dataset and it is strongly prescribed to utilize the Network Algorithms to remove the examples from it.

In this paper, we put forth an attempt to foresee the flare-up of COVID-19 dependent on past transmission information. Past examinations on COVID-19 expectations did not consider a time series data set a while building up the model we looked into this issue and tried to resolve this by utilizing LSTM organizations, SVR and Random Forest Predictive calculations.

In Nov-Dec the circumstance is again going to deteriorate and the demise rate could increment by 30%.And in this way we hope to alarm the general medical services suppliers of India to set them up for the emergency against COVID-19 once more

III. METHODS AND DATASETS

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

A. DataSets

A The COVID-19 information utilized in this research is gathered from COVID-19 in India Dataset and Novel Corona Virus 2019 Dataset which is accessible on kaggle.[19], furnished with number of affirmed cases until November 20, 2020. The informational collection likewise incorporates number of fatalities and recuperated patients before every day's over. The dataset is accessible in the time arrangement design with date, month and year. The COVID-19 dataset is separated into preparing set (80%) on which our models are prepared and testing set (20%) to test the presentation of the model.

B. LSTM Network for modelling time series

LSTM is termed as Long Short Term Memory algorithm which was created as the solution for short-term memory. It was created by Hoch Reiter and Schmidhuber in 1997. It has mechanisms named as gates which regulate the flow of information. These gates basically learn which data in the given set is important to keep and which is not to keep. With this it helps to pass on only relevant and necessary data rather than all data to further process to make predictions .Most of the prediction results are based on this Algorithm. Long Short Term Memory algorithm various applications are voice recognition, text generation and speech synthesis. You can even use LSTM to create great captions for videos.

In LSTM we have 3 different types of gates which regulate the data flow. The Gates termed as Forget Gate, Input Gate, and Output Gate.

1. Forget gate

To start with we have the forget gate of LSMT. Basically this gate decides on which or what information should be retained and which information should be thrown away. The Data or the Information which is passed on from the previous hidden state and the current data which is inputted through sigmoid function is processed. Further the values from the forget gate comes out between 0 and 1. The values which are closer to 0 means they are to be removed (forget) and the values closer to 1 means to insert (keep).

$$f_t = \sigma(w_f[h_{t-1}, x_t] + b_f)$$

2. Input Gate

Now second comes the Input Gate of LSTM. It is basically used to update the cell state. First step we do is to pass the hidden state and the current data input into the sigmoid function. Further that function decides which values or data(information) should be updated by transforming the values to be between 0 and 1.Here the values marked as 0 means not important data, and the values which are marked as 1 means the data is important. Then you also pass the similar data i.e. the hidden state data and the current input into tanh function to get values between -1 and 1.It helps to regulate the network. Further we multiply the tanh output data values with the sigmoid output values. Then the sigmoid output decides which information is more important to keep from the tanh output data values.

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i)$$

3. Cell State

Presently we ought to have enough data to ascertain the cell state. To begin with, the cell state gets point wise increased by the fail to remember vector. This has a chance of dropping qualities in the cell state in the event that it gets increased by values almost 0. At that point we take the yield from the information entryway and do a point wise expansion which refreshes the cell state to new qualities that the neural organization finds applicable. That gives us our new cell state.

$$\tilde{c_t} = tanh(w_c[h_{t-1}, x_t] + b_c)$$
$$c_t = f_t * c_{t-1} + i_t * \tilde{c_t}$$
$$h_t = o_t * tanh(c^t)$$

4. Output Gate

Last we have the yield gate. The yield gate chooses what the following concealed state ought to be. It Recollect that the hidden state contains data on past sources of info. The hidden state is likewise utilized for expectations (predictions).To starts with, we pass the past hidden state and the current input data to a sigmoid function. At that point we pass the recently changed cell state to the tanh function. We duplicate the tanh output with the sigmoid output to choose what data the hidden state should convey. The yield is the shrouded state. The new cell state and the new covered up is then continued to whenever step.

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o)$$

$i_t \rightarrow represents\ input\ gate.$
$f_t \rightarrow represents\ forget\ gate.$
$o_t \rightarrow represents\ output\ gate.$
$\sigma \rightarrow represents\ sigmoid\ function.$
$w_x \rightarrow weight\ for\ the\ respective\ gate(x)\ neurons.$
$h_{t-1} \rightarrow output\ of\ the\ previous\ lstm\ block(at\ timestamp\ t-1).$
$x_t \rightarrow input\ at\ current\ timestamp.$
$b_x \rightarrow biases\ for\ the\ respective\ gates(x).$

C. Support Vector Regression Model(SVR)

a) Positioning Figures and Tables: Place Support Vector Machine may be used as a regression technique, maintaining all the most options that characterize the rule (maximal margin). The Support Vector Regression (SVR) uses constant principles because the SVM for classification, with solely a couple of minor variations. First of all, as a result of output may be a real it becomes terribly tough to predict the data at hand that has infinite potentialities. Within the case of regression, a margin of tolerance is ready in approximation to the SVM which might have already requested from the matter.

b) However besides this truth, there's additionally a additional sophisticated reason, the rule is additional sophisticated so to be taken in thought. However, the most

plan is usually the same: to attenuate error, individualizing the hyperplane that maximizes the margin, keeping in mind that a part of the error is tolerated.Support Vector Machines (SVMs) are documented in classification issues. The utilization of SVMs in regression isn't in addition documented, however. These forms of models ar called Support Vector Regression (SVR).In most regression models, the target is to attenuate the add of square errors

c) Support Vector Regression. Support Vector Regression basically offers the flexibleness to outline what quantity error which can be applicable in our model and can realize associate appropriate line (or hyperplane in higher dimensions) to suit the info. In distinction to OLS, the target perform of Support Vector Regression is to attenuate the coefficients — additional specifically, the l2-norm of the constant vector — not the square error. The error term is instead handled within the constraints, wherever we tend to set absolutely the error but or capable such a that margin, referred to as the most error, ϵ (epsilon). We are able to tune alphabetic character to realize the specified accuracy of our model. Our new objective performs and constraints are as follows:

Minimize:

$$MIN\ \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2$$

Constraints:

$$\left|y_i - w_i x_i\right| \leq \varepsilon$$

Minimize:

$$MIN\ \frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 + C\sum_{i=1}^{n}\left|\xi_i\right|$$

Constraints:

$$\left|y_i - w_i x_i\right| \leq \varepsilon + \left|\xi_i\right|$$

D. Random Forest(RF)

Random forest, as it name suggests, it consists of an oversized variety of individual call trees that operate as ensemble. Every individual tree within the random forest spits out category class} prediction and therefore the class with the foremost votes becomes our model's prediction.In the Machine Learning world, Random Forest models area unit a sort of non-constant quantity models that may be used each for regression and classification. They're one among the foremost common ensemble strategies which belong to the specific category of Bagging methods.A large variety of comparatively unrelated models (trees) in operation as a committee can outperform any of the individual constituent models.Ensemble strategies involve victimization several learners to reinforce the performance of any single one among them singly. These strategies will be delineating as techniques that use a gaggle of weak learners (those WHO on the average slightly higher results than a random model) along, so as to form a stronger, mass one.

One of the most drawbacks of decision Trees is that they're terribly susceptible to over-fitting: they are doing well on training information, however don't seem to be therefore versatile for creating predictions on unseen samples. Whereas there are a unit workarounds for this, like pruning the trees, this reduces their prognosticative power. Typically they're models with medium bias and high variance; however they're straightforward and simple to interpret.

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

•ni sub(j)= the importance of node j
•w sub(j) = weighted number of samples reaching
•C sub(j)= the impurity value of node j
•left(j) = child node from left split on node j
•right(j) = child node from right split on node j

. In a very traditional decision tree, once it's time to separate a node, we have a tendency to contemplate each attainable feature and decide the one that produces the foremost separation between the observed data within the left node vs. those within the observed data of right node. In distinction, every tree in a very random forest will decide solely from a random set of options. This forces even a lot of variation amongst the trees within the model and ultimately ends up in lower correlation across trees and a lot of diversification.

Random Forest models sum up and combine the simplicity level of Decision Trees with the flexibility and power of an ensemble model. In a forest of trees, we forget about the high variance of an specific tree, and are less concerned about each individual element, so we can grow nicer, larger trees that have more predictive power than a pruned one.

IV. Results and Experimentations

The techniques utilized in this study depend on information guided methodologies and are totally unique in relation to past investigations. Our methodologies and prescient results will give help to limiting the contaminations and conceivable end of current COVID-19 pandemic. We prepared our organization with information until November 20, 2020. In this investigation we found that strategies or choices taken by government will enormously influence the current outbreak. Several concentrates on determining of COID-19 transmission depend on the R0 technique notwithstanding; they did exclude the affectability examination to locate the significant highlights. We analysed our model forecasts utilizing mean square mistake (MSE).

In Fig. 1 we plotted the total number of confirmed cases till the date on the daily bases. From the figure we can see that, India didn't witness its peak yet and its expected number of cases will soon increase exponentially despite the social distancing.

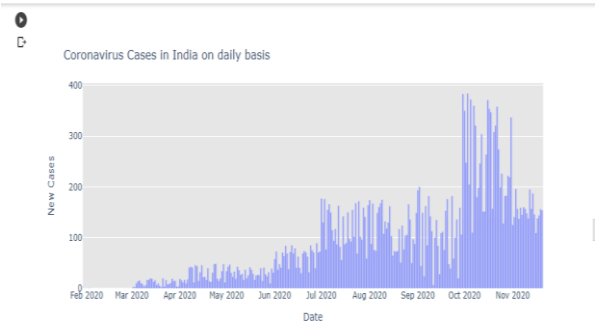

Fig. 1. ExampleCorona Cases in India on daily bases

In Fig. 2 Based on the previous data from the covid-19 India dataset from kaggle ,we tried to prdict the total number of patients who are recovered , dead , active and confirmed.The modelled faded color line is the predcited line while the dark color link is based on the actuall data.

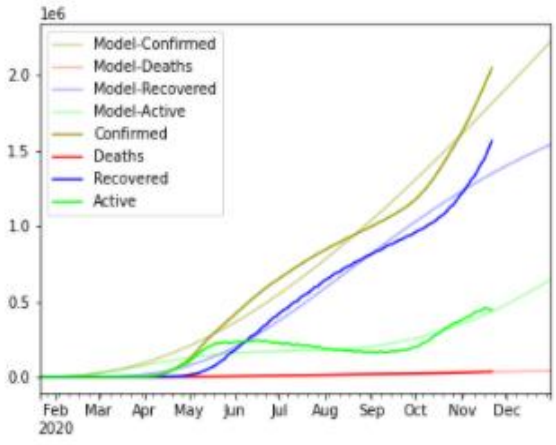

Fig. 2. Actual Data and Modelled predicted Data

In Fig 3 The actual data and the prdiced data is compared with the help of LSTM modelling and the Mean Square Error (MSE) method and Root Mean Square (RMSE) Error meothod till the 20 November which is the current data set state.

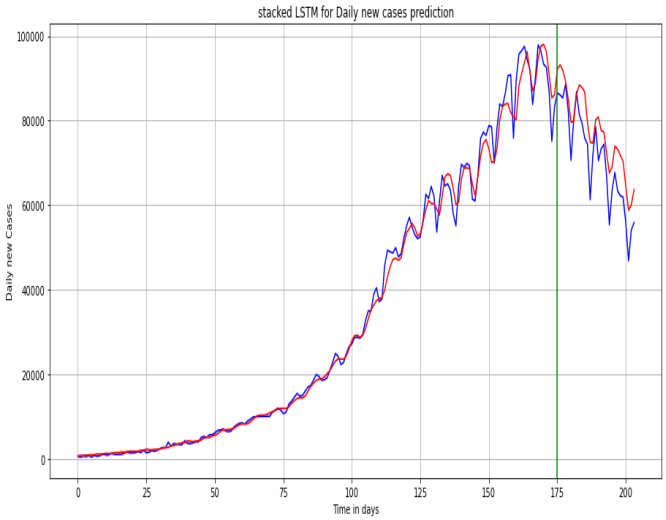

Fig. 3. ExampleCorona Cases in India on daily bases

Fig. 4. In Fig 4 , 5 and 6 the data is assessed for time series prediction of confirmed cases, deaths and recoveries in India based on data stored by Kaggle in between 1 March 2020 to 20 November 2020.The performance of models is measured by mean absolute error, root mean square error.

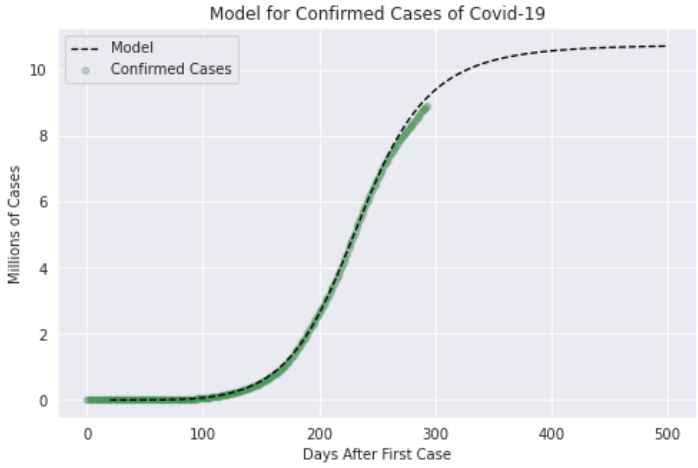

Fig. 5. Total number of confirmed cases till 20 Nov with predicted model(dotted line), X-axis=No of days after first case and Y-axis=Thousands of cases.

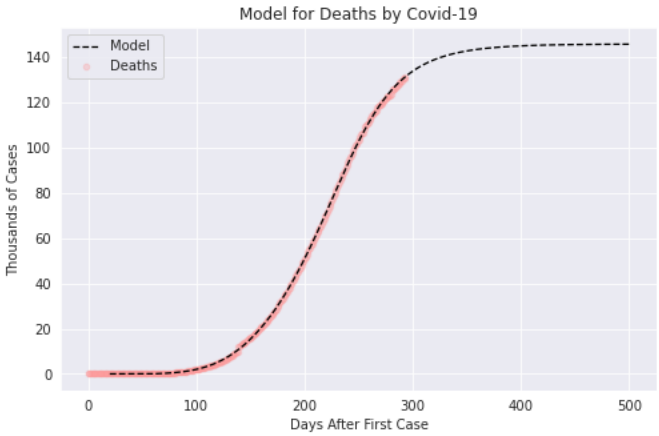

Fig. 6. Total number of deaths till 20 Nov with predicted model(dotted line) X-axis=No of days after first case and Y-axis=Thousands of cases.

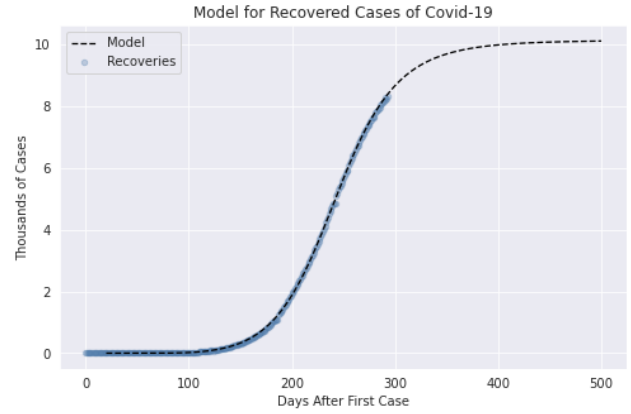

Fig. 7. Total number of confirmed cases till 20 Nov with predicted model(dotted lines),X-axis=No of days after first case and Y-axis=Thousands of cases

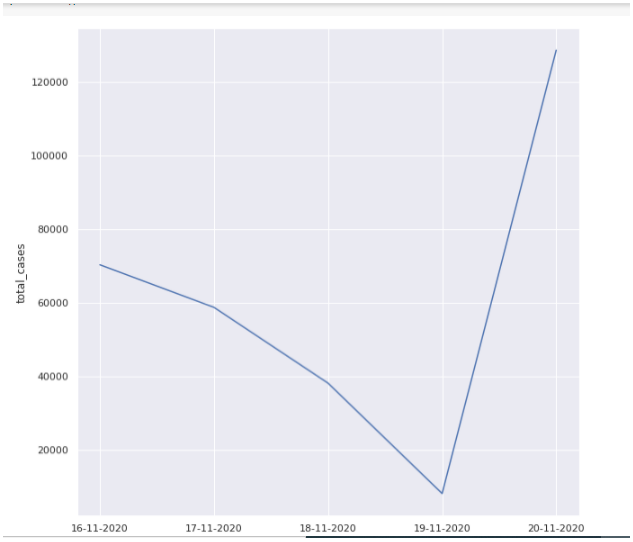

Fig. 8. Total number of cases past 5 days with SVR. X-axis=Dates and Y-axis=Total Cases.

In Fig 8 ,As you can see there is a sudden increase and spike in the month of November caused by the excess spreading of Covid 19 after the lockdown.Now government should soon plan on keeping another lockdown.

.

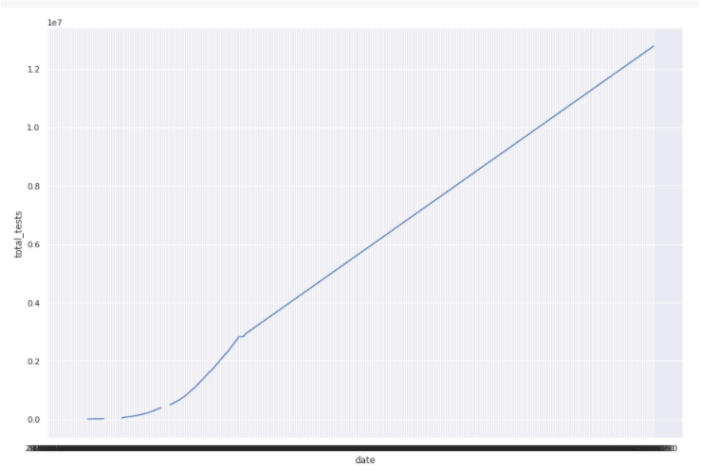

Fig. 9. Total number of tests per day till 20 Nov using SVR X asis = number series of Days till 20November. And Y ais is total tests data.

In Fig 9,As you can see that the total number of tests are increasing gradually throughout the time which tells us that the Covid 19 is spreading again.

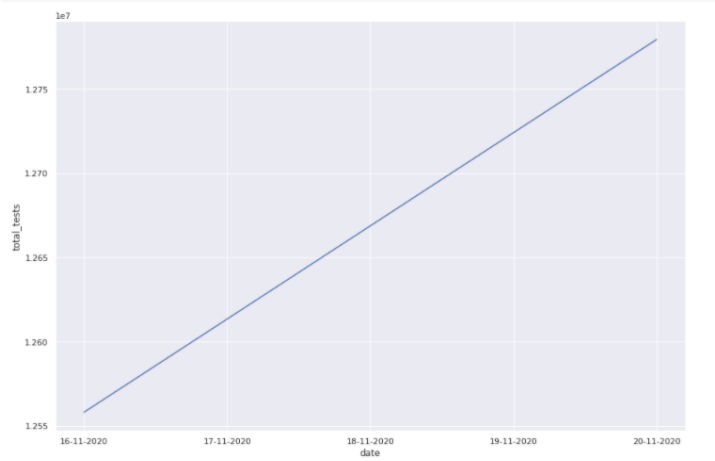

Fig. 10. Total number of tests till past 5 days Where X-axis is the 5 days and the Y axis is number of test cases till 20 Nov.

In Fig 10,As you can observe that there is a straight line which you can also call as a dangerous situation , the line implies that the number of tests are gradually increasing with respect to the days and there is no stopping it.This test is done for only 5 days.

TABLE I. MSE AND RMSE

Sr No	***Algorithm***	***MSE***	***RMSE***
1	LSTM	0.00046 0.00589	0.02156 0.07674
2	SVR	0.82081	0.13543
3	Random Forest	0.02219	0.14897

TABLE II. MSE AND RMSE

	ForecastId	ConfirmedCases	Fatalities
count	12212.000000	12212.000000	12212.000000
mean	6106.500000	1208.889125	53.222486
std	3525.445078	6234.287452	417.608734
min	1.000000	0.000000	0.000000
25%	3053.750000	6.000000	0.000000
50%	6106.500000	81.000000	0.000000
75%	9159.250000	367.000000	3.000000

V. CONCLUSION

In this paper, we have proposed profound learning models for foreseeing the quantity of COVID-19 positive cases, recovered and passing in Indian states. An exploratory information examination of the expansion in the quantity of positive cases ,cases, recovered and passing in India has been finished. In light of the quantity of cases and day by day development rate and test rate, states are ordered into mellow, moderate and extreme zones for practical lockdown estimates state-wise in contrast with securing the entire country, which may cause some monetary issues. Recurrent neural organization (RNN) based long momentary memory (LSTM) Support vector relapse (SVR) and Random Forest (RD) cells are utilized as forecast models. Daily and week after week expectations are determined for all India, and it is discovered that among all different calculations utilized LSTM gives extremely precise results(error under 3%) for transient expectation (1–3 days). Forecasts are freely accessible at a site created for the overall population. These forecasts will be useful for state and public government specialists, scientists and organizers for overseeing administrations and orchestrating clinical framework likewise. The proposed model and preventive methodology can be trailed by different countries too.Using predictive deep learning algorithms we were able to predict the growth in next month I.e. December.

ACKNOWLEDGMENT

The preferred The author thank Professor A Swaminathan for his comments and support and simultaneously help from our working group to complete this paper. Expressed are those of the author only, no other representation

REFERENCES

[1] *C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, L. Zhang, G. Fan, J. Xu, X. Gu, et al.*

[2] *WHO.*
Statement regarding cluster of pneumonia cases in wuhan, china;
World Health Organization: Geneva, Switzerland (2020)
China Lancet, 395 (2020), pp. 497-506
Available online: https://www.who.int/china/news/detail/09-01-2020-who-statementregarding-cluster-of-pneumonia-cases-in-wuhan-china(accessed on 17 February 2020)

[3] *WHO.*
Novel coronavirus–thailand (ex-china)
World Health Organization: Geneva, Switzerland (2020)
Available online: https://www.who.int/csr/don/14-january-2020-novel-coronavirus-thailand-ex-china/en(accessed on 17 February 2020)

[4] *WHO director-general's opening remarks at the media briefing on COVID-19 – 11 march 2020. 2020. [Online; accessed 21-March-2020]*

[5] *Moore M., Gelfeld B., Okunogbe A.T., Christopher P.. Identifying future disease hot spots: Infectious disease vulnerability index; RAND corporation: Santa monica, CA, USA.2016.*

[6] *Availableonline: https://www.rand.org/pubs/research-reports/RR1605.html (accessed on 17 February 2020).*

[7] *WHO. Situation report; world health organization: Geneva, switzerland. 2020.*
Available online: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/.

[8] *J. Riou, C.L. Althaus*
Pattern of early human-to-human transmission of wuhan 2019-ncov
bioRciv (2020)

[9] *J.A. Backer, D. Klinkenberg, J. Wallinga*

[10] *The incubation period of 2019-ncov infections among travellers from wuhan*
China medRciv (2020)

[11] *T. Lancet*
India under COVID-19 lockdown
Lancet (London, England), 395 (10233) (2020), p. 1315

[12] *Eubank S., Guclu H., Kumar V.A., Marathe M.V., Srinivasan A., Toroczkai Z., Wang N.. Modelling disease outbreaks in realistic urban social networks. 2004. Nature, 429, 6988, 180–184*

[13] *J. Riou M. Mandal, S. Jana, S.K. Nandi, A. Khatua, S. Adak, T.K. Kar, C.L. Althaus*

[14] *I.E. Frank, R. Todeschini*
The data analysis handbook
Elsevier, Amsterdam (1994)

[15] *T. Liu, J. Hu, M. Kang, L. Lin, H. Zhong, J. Xiao, A. Deng*
Transmission dynamics of 2019 novel coronavirus
2019-nCoV (2020)

[16] *C.-J. Huang, Y.-H. Chen, Y. Ma, P.-H. Kuo*
Multiple-input deep convolutional neural network model for covid-19 forecasting in china

[17] *medRxiv (2020)*

[18] *N.M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A.R. Akhmetzhanov, S.-M. Jung, B. Yuan, R. Kinoshita, H. Nishiura*

[19] *Epidemiological characteristics of novel coronavirus infection: A statistical analysis of publicly available case data*

[20] *Analysis and forecast of COVID-19 spreading in china*
Italy and France Chaos, Solitons & Fractals, 134 (2020), p. 109761I.E. Frank, R. Todeschini