Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

Research on Various Time Forecasting Algorithms for Predicting Covid-19 Cases

K. Rajeswari
Department of Computer Engineering
Pimpri Chinchwad College of Engineering, Pune, India

S. R. Vispute
Department of Computer Engineering
Pimpri Chinchwad College of Engineering, Pune, India

Vighnesh Pathrikar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering, Pune, India

Tejas Podutwar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering, Pune, India

Akash Mandana
Department of Computer Engineering
Pimpri Chinchwad College of Engineering, Pune, India

Akshay Siddannavar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering, Pune, India

***Abstract*—The World Health Organization declared the Covid-19 outbreak a Public Health Emergency of International Concern on 30th January 2020, and a pandemic on 11 March 2020. Now one year after the outbreak of Covid-19 the virus has taken up new mutations and is becoming increasingly harder to predict in matters of its behavior as well as severity. In this study, various approaches for time series forecasting for coronavirus (Covid-19) cases have been compared. Statistical methods, i.e., Linear Regression, Susceptible Infectious Recovered (SIR), Auto-Regressive Integrated Moving Average (ARIMA), Generalized Autoregressive Conditional Heteroscedasticity (GARCH), Threshold GARCH (TGARCH), Exponential GARCH (EGARCH), and Seasonal Autoregressive Integrated Moving Average (SARIMA), have been compared to Deep learning method, Long Short-Term Memory (LSTM), using various performance parameters. Based on the predictions and forecasts of the better algorithm, health care workers could take proper decisions at the right time in providing kits to health centers and other aids to the population. As per our observations, as the number of days of forecast goes on increasing, the error rate of the model also increases. Forecasted trends also show that on average for fewer days, statistical methods tend to be better, whereas, for forecasts of greater days, Deep Learning methods tend to be better.**

Keywords*—Covid-19, Deep Learning, Statistical Modelling, Time-series forecast.

I. INTRODUCTION

In late December of 2019, many health centers reported a In late December of 2019, many health centers reported a bunch of patients with pneumonia of unknown cause. All the patients were linked to a seafood and wet animal wholesale market in Wuhan, China [9].

The World Health Organisation (WHO) then announced coronavirus (Covid-19), the virus responsible for the current pandemic. Currently, many countries are still struggling while some have controlled it to some extent with vaccination and preventive measures [8, 10, 14].

Now one year after the outbreak of Covid-19 the virus has taken up new mutations and is becoming increasingly harder to predict in matters of its behavior as well as severity.

In this survey, time-series models are compared to foretell the epidemiological trends of the Covid-19 pandemic. Based on the predictions and forecasts of the better algorithm, health care workers could take proper decisions at the right time in providing kits to health centers and other aids to the population. Present results give insights on the algorithms which are better than their counterparts at predicting the surge in cases thus highlighting the importance of social distancing and implementation of preventive measures of Covid-19.

II. LITERATURE SURVEY

Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average (SARIMA), are frequently used prediction models for univariate time series data forecasting.

Initially, the data is collected and converted to stationary data using differencing. Then using Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) the parameters for the ARIMA and SARIMA model are determined. Using the Kernel Density Estimate Plot (KDE) and Quantile–Quantile (Q–Q) plot the errors are realized. A sample plot is created, which is compared to the test data. After adjusting for errors, a forecast is made [1].

K.E. Arun Kumar et al. used ARIMA and SARIMA, statistical models, to develop a 60-day prediction of collective Covid-19 cases for top-16 nations in which the ARIMA model proves to be more accurate. They concluded that 9 countries showed an exponential increase in confirmed cases. 10 countries have shown exponential growth in the number of deceased people. And in 3 countries, the projections were stable [1].

Aykut Ekinci used a modified version of ARIMA, which incorporates the Generalized Autoregressive Conditional Heteroscedasticity Algorithm (GARCH), which is better for data containing volatility clustering, i.e., great changes tend to be followed by great changes, of either sign, and minor changes tend to be followed by minor changes [6]. Aykut Ekinci used one year's data to model one day ahead forecast of the days following the first-wave, second-wave, and third-wave, in which Autoregressive Moving Average - Generalized

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

Autoregressive Conditional Heteroscedasticity Algorithm (ARMA-GARCH) had the lowest RMSE on average, whereas ARMA - Threshold GARCH **(**ARMA-TGARCH) and ARMA - Exponential GARCH **(**ARMA-EGARCH) had mixed results. These models proved better when comparing sudden increases or decreases in the number of Covid-19 cases [3].

Deep learning has a significant contribution to the field of machine learning. Different time series forecasting algorithms like the Traditional Recurrent Neural Networks, Long Short Term Memory, Gated Recurrent Unit, fall under the category of deep learning which can be used for the forecasting of future Covid-19 cases.

Vinay Kumar Reddy Chimmula, Lei Zhang [12] had presented a model for the time series forecasting of Covid-19 cases in Canada with the help of deep learning where Long Short-Term Memory i.e LSTM was used to predict the real-time transmission of the new cases. The data for the research was taken from the dataset provided by Johns Hopkins University (JHU) and the Canadian Health authority. The dataset consisted of data of confirmed cases till the 31st of March 2020. The suggested model had no assumptions, unlike the statistical and machine learning models. Mean Square Error was used as the evaluation metrics. The results of the study were able to predict a certain exponential increase in the confirmed cases of Covid-19 for the coming months in Canada.

The research done by Nooshin Ayoobi et al. [15] consisted of three different methods i.e. LSTM, Convolutional LSTM, and GRU to forecast the new cases as well as the new deaths cases for the countries of Australia and Iran. The novel approach used the above-mentioned three models with their bidirectional extension which was done for the first time to determine the forecast of the new and the death cases. The neural networks of these models consisted of an input layer, output layer, and three hidden layers. The activation function used to catch the non-linearities in the data was the Rectified Linear Unit (ReLU). The forecast was done for the next one, three, and seven days. The results showed that the bidirectional extensions of these models proved to be better than the vanilla models.

Regression analysis is a forecasting model that estimates the relationship between two or more variables [3]. It is amongst the most commonly used Machine learning techniques for forecasting purposes. Linear Regression is a commonly used type of predictive analysis [4]. The overall idea is to find the dependency of one variable over the other using a best-fitting linear equation. For using the linear regression model, sklearn.linear_models package is imported from the library in Python [5]. Simple linear Regression and Multiple Linear Regression analysis help in understanding the various epidemic data points of India and the relationship between them. It is considered to be amongst the most significant predictive models. Limitations of the model can be due to the use of fewer independent variables or information and inaccuracy in the number of contact tracing cases. If the number of contact tracing cases is decreased, it will indirectly decrease the number of daily active cases [10].

The Susceptible-Infectious-Recovered (SIR) model belongs to basic epidemiological models which are considerably foretelling for diseases that are transmitted from person to person [2].

Mohammed N.Alenezi, Fawaz S. Al-Anzi, Haneen Alabdulrazzaq [7] have used a time-dependent SIR model that predicts infection rates as well as recovery rates for Kuwait. They experimentally studied the legitimacy of different R0 values over a time duration of 94 days and found that the SIR model is most accurate for R0 values in a range of 3 to 4.

PabelShahrear et.al [4], have used the Susceptible–Exposed–Quarantine–Infected–Recovered–Death (SEQIRP) model which is an altered version of the SIR model [11, 19]. The sensitivity of R0 to parameters has been studied. Sensitivity analysis demonstrates which parameter is more dominating for the system. Simulation is done using 4th order RK(Runge-Kutta) by the medium of MATLAB for the rate of susceptibility, infection, recovery and also for graphs of Number of Dead People vs Date, Number of Infected People vs Date for better understanding of the behavior of Covid-19 [4].

Weston C.Roda et al. [18] have demonstrated that the SIR model is considerably more accurate than the SEIR model for the outbreak of Covid-19 in Wuhan. They show that more intricate models need not be more accurate.

The dataset used by most of the researchers in their study was taken from JHU which is updated every single day. The general procedure for using this raw data for the prediction was illustrated. First, some preprocessing is done on the raw data where the data is converted into the time series data which is then divided into the testing and training set. The training data is then fit into the different models which are evaluated using the evaluation parameters like RMSE, MAE, RMSLE and MAPE, etc. Then the model is adjusted by looking at these performance parameters. Once training is done the model is tested with the testing dataset and then the appropriate model is chosen for forecasting. Here, as the number of days of forecast goes on increasing the error rate of the model is also increasing.

I. METHODOLOGY

A. Linear Regression

Simple linear regression analysis is a technique to find the relation between two variables. One of the variables is an independent variable while the other one is dependent and changes with the change in the independent variable. General representation of Linear regression is given by :

$$y = B_0 + B_1 x + e \quad (1)$$

The only limitation of using Linear regression is that it often uses the mean value of both input and output variables to define a relation between them. Just as the mean can't fully describe a single variable, the linear regression model is just not a reliable method for predicting variable relationships. Therefore, the various factors which hindered the use of SLR are addressed with the help of the Multiple Linear Regression (MLR) model.

While in Simple Linear Regression, only one independent variable is involved, Multiple linear regression uses two or more independent variables and one dependent variable to estimate the relationship between them.

General representation of Multiple linear regression is given by:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

$$y = B_0 + B_1X_1 + ...B_nX_n + e \quad (2)$$

where,

- y = the predicted or expected value of the dependent variable
- B_0 = value of y when all dependent variables are equal to zero
- B_1X_1= the regression coefficient (B_1) of the first independent variable (X_1)
- B_nX_n = the regression coefficient of the last independent variable
- e = model error (variation in our estimate of y) [21]

Initially, the dataset is cleaned by removing any missing and unwanted values. This process is called Data Cleaning. After the dataset is completely ready, it is split into two parts- Train and Test dataset (Mostly 80% for training and 20% for testing) The model is first trained using Train data and Test data to verify that the model fits the data well. Then the fitted regression equation with future dates is used for prediction [4].

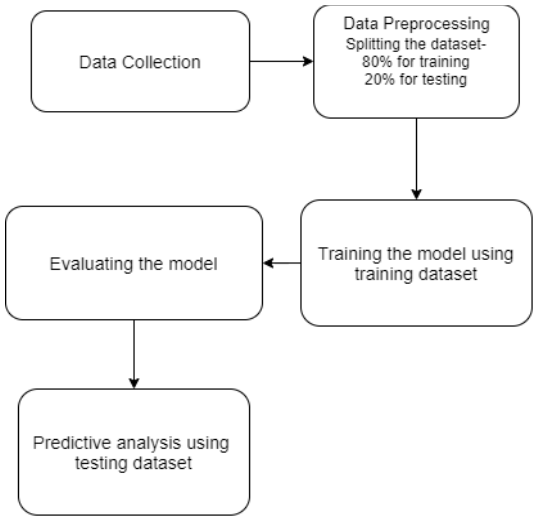

Fig. 1. [23] Process for Regression predictive modelling

B. ARIMA.

As opposed to classical regression, the main notion behind time series forecasting is that the measure of some variable at a certain period will depend on the measure of the same variable at previous periods [17].

ARIMA is a by-product of combining differencing with autoregressive and moving average models. The predictors include the lagged values as well as the lagged errors for the data. ARIMA model is expressed in terms of ARIMA(p, d, q), where,

- ***p*** is the order of the autoregressive section.
- ***d*** is the degree of first differencing.
- ***q*** is the order of the moving average section [16].

Thus, the following representation of ARIMA is used for predicting daily Covid-19 cases.

$$y_t = c + \phi_1 y_{t-1} + \cdots + \varphi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (3)$$

Where y_t is the number of cases for the t^{th} day, ε_t is the error between y_t and the actual value for the t^{th} day, and p and q are the parameters used for ARIMA, where p is calculated by considering the partially-correlated values with y_t and q is calculated by considering the correlated values with y_t [1].

C. SARIMA.

Seasonal-ARIMA (SARIMA) is formed by including further seasonal parameters $(P, D, Q)_m$, where m is the number of observations per year, to the ARIMA model.

The seasonal model contains both terms, from non-seasonal ARIMA, as well as the backshift of the seasonal period. The general representation is given by

$$\Phi_P(B^m)\varphi_p(B)(1-B^m)^D(1-B)^d y_t = \Theta_Q(B^m)\theta_q(B)w_t \quad (4)$$

Where y_t is non-stationary ARIMA, w_t is the gaussian white noise, $\phi(B)$ is the non-seasonal auto regressive component, $\theta(B)$ is the non-seasonal moving average component, D is the seasonal differencing, $\Phi_P(B^m)$ is the seasonal auto regressive component, and $\Theta_Q(B^m)$ is the seasonal moving average polynomial. B is the backshift operator where,

$$B^k y_t = y_{t-k} \quad (5)$$

The expressions for the non-seasonal autoregressive, moving average, seasonal terms for seasonal AR and seasonal MA model are [1]

$$\phi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - ... - \varphi_p B^p \quad (6)$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q \quad (7)$$

$$\Phi_P(B^m) = 1 - \Phi_1 B^m - \Phi_2 B^{2m} - ... - \Phi_P B^{Pm} \quad (8)$$

$$\Theta_Q(B^m) = 1 + \Theta_1 B^m + \Theta_2 B^{2m} + ... + \Theta_Q B^{Qm} \quad (9)$$

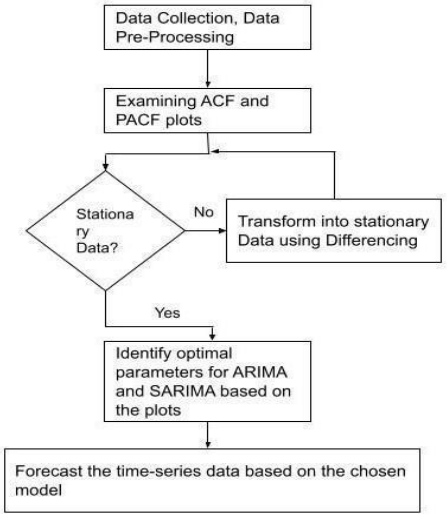

Fig. 2. [1] General Process followed for developing ARIMA and SARIMA models.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

D. ARMA-GARCH

The central assumption behind the ARMA model is that the data is stationary, hence the expected variance of all terms at a given point should be the same. This assumption is called homoskedasticity and it is an improbable assumption. To overcome this as well as volatility clustering, Engle [3] introduced the autoregressive conditional heteroscedasticity (ARCH) model as a new type of stochastic process. The general GARCH(1, 1) model is given by:

$$Y_{qt} = \mu + \alpha_1 Y_{q,t-1} + \alpha_2 Y_{q,t-2} + \ldots + \alpha_p Y_{q,t-p} \quad (10)$$
$$+ Y_{qt} + \varnothing_1 \varepsilon_{q,t-1} + \varnothing_2 \varepsilon_{q,t-2} + \ldots + \varnothing_r \varepsilon_{q,t-r} + \varepsilon_{qt}$$

$$\sigma^2_{qt} = \omega + \alpha \varepsilon^2_{q,t-1} + \beta \sigma^2_{q,t-1} \quad (11)$$

The mean equation presented in (11) known as the conditional variance equation, σ^2_{qt} displays one value in advance forecast variance built on past information. Conditional variance of the rate of growth for new cases for a particular country q, σ^2_{qt}, is a constant function, ω, the ARCH term, $\varepsilon^2_{q,t-1}$, and the GARCH term, $\sigma^2_{q,t-1}$. ARCH term is the previous data's consequence, measured as the interval of the squared error terms from the mean equation. The GARCH term specifies extended-term volatility, measured as last data's forecast variance.

D. SIR

It is a mathematical model belonging to basic epidemiological models which are considerably foretelling for diseases that are transmitted from person to person [2].

It is a compartmental model in which total population (*N*) is divided into:

- *S*: The number of susceptible individuals.
- *I*: The number of infectious individuals.
- *R*: The number of removed (and immune) or deceased individuals.

To represent the number of individuals in one compartment at a time '*t*':

- *S(t)*: The number of susceptible individuals.
- *I(t)*: The number of infectious individuals.
- *R(t)*: The number of removed (and immune) or deceased individuals.

Assumptions:

We need to make assumptions to simplify real-world situations into mathematical [19]:

1. Total population (sum of all compartments remains constant).
2. Birth rate and death rate are assumed to be negligible.
3. Everyone is susceptible.
4. Rate of infections is directly proportional to contacts.
5. Infectives recover/die at a constant rate.

Mathematical Representation:

Prior assumptions can be represented in form of mathematical equations:

$$S(t) + I(t) + R(t) = N \quad (12)$$

Equation (12) depicts the first assumption i.e, total population(sum of all compartments remains constant).[19]

Initial conditions: At time '$t = 0$'

$$S(0) = S_0 \quad (14)$$
$$I(0) = I_0 \quad (15)$$
$$R(0) = 0 \quad (16)$$

Equation (14) shows the initial amount of susceptible people represented by term S_0, at the start almost everyone will be susceptible. Equation (15) shows the initial number of infectious people represented by I_0. Initially, it is going to be some very small number of people, maybe just one person. Equation (16) shows that no people recovered initially.

We use a system of differential equations to represent the rate of change in each compartment:

$$dS/dt = -a(S \times I) \quad (17)$$
$$dI/dt = a(S \times I) - bI \quad (18)$$
$$dR/dt = bI \quad (19)$$

Equation (17) shows the rate of change of susceptibles concerning time, where '*a*' represents transmission rate, '*S*' is the number of susceptible people at that time '*t*' and '*I*' is the number of infectious people at a time '*t*' [19]. Thus the contacts between susceptibles and infectious are represented by '$S \times I$'. The sign is negative since people are leaving the Susceptible compartment [19]. Equation (18) shows the rate of change of infectious disease with respect to time, where '*b*' represents the recovery rate. In (18) it is shown that '$a(S \times I)$' people are added to the Infectious compartment and 'bI' people are removed from the Infectious compartment [19]. Equation (19) shows rate of change of recovered where '*b*' represents recovery rate and '*I*' is the number of Infectious people at a time '*t*' [19].

Fig. 3. [20] Compartments in the SIR model and transition rates between them.

The spread of disease depends on the rate of change of the Infectious compartment which is represented by Equation (18). If the value of 'dI/dt' is greater than zero, we can say the quantity of infectious individuals is increasing and we are moving towards an epidemic, otherwise, we can say the quantity of infectious individuals is decreasing and we are moving away from an epidemic (new people are being infected but they are recovering faster)

Around time '$t = 0$' :

Equation (18) would become

$$a(S_0 \times I_0) - bI_0 \quad (20)$$

If the value of (20) is greater than 0 then,

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

$$aS_0I_0 - bI_0 > 0 \quad (21)$$
$$aS_0I_0 > bI_0 \quad (22)$$
$$aS_0 > b \quad (23)$$
$$(aS_0/b) > 1 \quad (24)$$

(21) represents the situation when the value of (18) would be greater than zero near time '$t = 0$'. (22) is obtained by moving 'bI_0' to RHS. (23) is obtained by dividing both sides by I_0. (24) is obtained by dividing both sides by 'b'. Thus from (24) we can conclude that spread of the disease depends on the value of the term 'aS_0/b'. if 'aS_0/b' is greater than one then the disease is moving towards an epidemic and if its value is less than one then the disease isn't moving towards an epidemic.

Reproduction Number:
This term 'aS_0/b' From (24) is called 'basic reproduction number', represented by the symbol 'R_0'. it conveys the average number of secondary infections an infectious person will cause while he/she is infected.
Here the term 'a/b' is termed as contact ratio, symbol: q. It represents the fraction of the population that comes into contact with an infected individual during a period when they are infected.
Thus,

$$R = q \times S_0 \quad (25)$$

(25) represents the basic reproduction number in terms of contact ratio and the number of susceptible people around time $t = 0$.

E. DEEP LEARNING

The dataset where every data point is collected at periodic intervals of time is called the time series data [12]. So, Covid-19 data can be termed as the time-series data. The future values of the cases depend on the trends in the past. So, various mathematical and statistical methods have been suggested for modelling the time series data. But these models tend to have some limitations when used for modelling the time series data. These models fail to recognize the complex patterns in the time series data which is very useful for forecasting future values. These models work well for the few-step forecasts but not in the long-term forecasts. The missing values in the dataset affect their performance. So, to overcome these barriers of the statistical methods, deep learning models came into the picture to model the complex patterns in the time series data efficiently. The following section covers different deep learning models like Feed-Forward Neural Networks, RNN, and LSTM to be used for time series forecasting of the Covid-19 cases.

a) Feed Forward Neural Network: Multilayer perceptron is a type of feed-forward fully connected neural network consisting of layers of neurons. A neuron in these neural networks is a node that accepts some data as an input and performs some mathematical operation on it and then produces some output. The Covid-19 dataset is time-series data so this cannot be directly fed into the network as an input. So, some preprocessing needs to be done. The dataset is first converted into a regression dataset and then it is fed into the multilayer perceptron (MLP) for training purposes [27].

A feed-forward neural network consists of different layers, among which the first one is the input layer, the second set of layers consists of the hidden layers and the last layer is the output layer can be termed as a Multilayer Perceptron [25]. The MLP consists of at least three layers where the number of hidden layers will be one or more. Each neuron in a particular layer has some inputs where each input will be associated with a particular weight. Then the weighted sum of all the inputs is computed by the neuron which is then passed to an activation function. The data that we use in real-life examples are seldom going to be linear. So, most of the data that we deal with have non-linearities in them and to handle and analyze these non-linearities an activation function is used by the neuron. The different types of activation functions that are used can be sigmoid, tanh or ReLU. The weights associated with each input for a particular neuron are initialized with some random values. When the data is fed into the neural network it is propagated into the forward direction and the output is produced[24]. The error produced in the output is called the loss function. During the training process of the neural network, this loss function is minimized by adjusting the weights associated with the inputs and a better prediction is made.

$$y = g(w_0 + \sum_{i=0}^{m} x_i w_i) \quad (26)$$

Here, $y \rightarrow$ *output vector*

w_o = *bias*

$x_i w_i$ = *linear combination of inputs*

g = *non-linear activation function*

The linear combination of the weights and inputs $x_i w_i$) is added with a bias value (w_o). This weighted sum is then passed to a non-linear activation function (g), which gives the output for the perceptron.

b) RNN: RNN are special types of neural networks which are used to handle and analyze the sequential time series data. It is a type of neural network where the output at the particular time step is not only dependent on the inputs given to the network but also on the output of the previous time step [26]. This is an evolution of the previous feed-forward neural network discussed. The Time Series nature of the Covid-19 data has patterns in the past that need to be stored in a memory to generate better future predictions. This is where the RNNs perform better than the feed-forward networks where there is no relation of the output produced at a particular time step with the output produced at the previous time steps. So, the RNNs are said to have a memory where these outputs of the previous time steps are stored for future predictions [29]. So, the RNNS tends to perform better than the feed-forward neural networks in the time series forecasting of future values. RNN's suffer from the problem of vanishing gradients which leads to the incapability of the network to capture the long-term dependencies [28]. So, to tackle these limitations some modified models of the RNN were developed.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

Hidden Cell State:

$$h_t = f_w(x_t, h_{t-1}) \quad (27)$$

$$h_t = tanh(W^T_{hh} h_{t-1} + W_{xh}{}^T x_t) \quad (28)$$

Output Vector:

$$y_t = W_{hy}{}^T h_t \quad (29)$$

In RNNs, the hidden cell state (h_t) is dependent on the present inputs (x_t) as well as the previous cell state (h_{t-1}).

c) LSTM: LSTM is the modified version made from the RNN, which eliminates the problem of the long-term dependencies of the traditional RNN's. The output of each cell of the LSTM is passed to the cell at the next time step. The process is continued until we reach the final LSTM cell and generate the output. The design of the LSTM cells was made to tackle the problem of vanishing gradient by storing the information in the memory cell for a longer duration of time. The architecture of LSTM is built with a cell state and three different gates, each of which is used for handling different purposes [15].

Each LSTM block consists of three different gates, which are Forget gate, Input gate, and output gate. The forget gate is fed with the information from the previous hidden state and the current input. These are then passed to the sigmoid activation function. If the value comes out to be closer to zero then the gate discards this information and if it comes out to be near to one then it preserves the information. So, this helps the LSTM block to remove irrelevant information. The input gate helps to update the cell state of the LSTM block. This gate consists of two different activation functions. Among which the first is the sigmoid function which performs the same function as the forget gate and the second is the tanh i.e tan hyperbolic activation function to squish the values between -1 to 1 which helps the network work on the data uniformly. Then the sigmoid output is multiplied with the tanh output and then the cell state is updated accordingly. The output state will decide what information needs to be passed to the next cell as the previous cell state. In this way, the architecture of the LSTM helps eliminate the problem of long-term dependency.

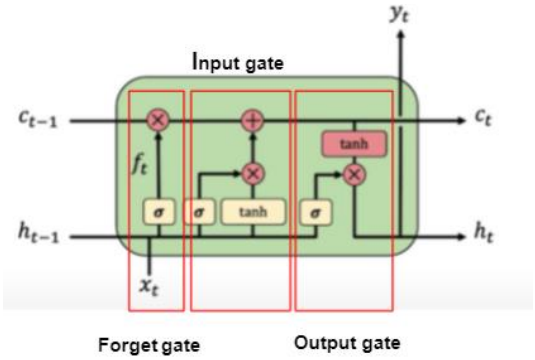

Fig. 4. LSTM cell with all the three gates

$$i_t = \sigma(w_i [h_{t-1}, x_t] + b_i) \quad (30)$$

The input gate helps to update the cell state of the LSTM block.

$$f_t = \sigma(w_f [h_{t-1}, x_t] + b_f) \quad (31)$$

The forget gate decides which information to forget from the h_{t-1} and x_t by passing it to the sigmoid (σ) activation function.

$$o_t = \sigma(w_o [h_{t-1}, x_t] + b_o) \quad (32)$$

The output (o_t) decides the value for the next hidden cell state.

where,
i_t = *represents input gate*
f_t = *represent forget gate*
o_t = *represent output gate*

σ = *sigmoid function*

w_x = *weight of the respective gate(x) neurons*

h_{t-1} = *output at previous time step (t-1)*

x_t = *input at the current time step*

b_x = *biases for the respective gates(x)*

TABLE I. Validation Metrics for time series forecasting of Covid-19 cases using statistical and deep learning methods

Ref No.	Algorithm(s)	Forecasting Range	Distribution	RMSE	Comment
[22]	LR	14 Days	Daily Active Cases	1.063962E+04	The number of daily active cases was predicted based on daily positive cases. It performed well even with just a single input variable
[22]	MLR	14 Days	Daily Active Cases	2.78134E+09	Multiple Linear Regression model has better accuracy as compared to the results obtained using Linear Regression model
[1]	ARIMA(p,d,q) (mean) SARIMA (p, d, q) (P, D, Q) $_m$(mean)	60 Days	Cumulative Cases	ARIMA 5.072E+03 SARIMA 8.119E+03	On average, ARIMA performed better than SARIMA. This is likely because of overfitting in the case of SARIMA
				ARMA	

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

[3]	ARMA (p, q) (mean) ARMA (p,q) -GARCH(1,1) (mean) ARMA (p,q) -TGARCH(1,1) (mean) ARMA (p,q) -EGARCH(1,1) (mean)	1 Day	Daily Case.	3.99E+02 ARMA-GARCH 3.457E+02 ARMA-TGARCH 3.464E+02 ARMA-EGARCH 3.944E+02	ARMA-GARCH has the lowest RMSE on average, whereas individually, ARMA-TGARCH and ARMA-EGARCH had mixed results
[7]	SIR	94 Days	Cumulative Cases	1.462E+03	The SIR model is almost fitted with the actual confirmed cases for R_0 values ranging from 3 to 4
[12]	LSTM	10 Days	Cumulative Cases	5.146E+01	LSTM was able to predict better even with limited data

IV. IMPLEMENTATION.

A. Introduction

In this section, we decided to compare and contrast different models on the same dataset (JHU Daily Cases for India [30]), by fixing the prediction period, and also by predicting for different time periods starting from 25th October, one day, seven days, and one month (up to 25th November).

In this section, the procedure followed for time series forecasting of daily Coivd-19 cases is discussed. Initially, the raw data consisting of cumulative values for Covid-19 cases were collected from the JHU dataset. In time series forecasting we require the past data of daily Covid-19 cases in order to forecast the future values hence these values are collected from the dataset and before using the data for training some preprocessing is applied.

In the preprocessing part, the data collected from the dataset consisted of the cumulative values of the Covid-19 cases so the first step was to convert these cumulative values into daily values. The dataset consisted of some negative and zero values which were removed. Fig. 5. depicts the daily Covid-19 cases in India as per JHU dataset.

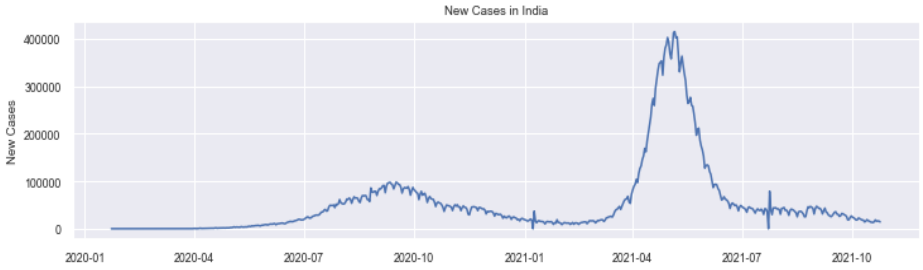

Fig. 5. Daily Covid-19 cases in India

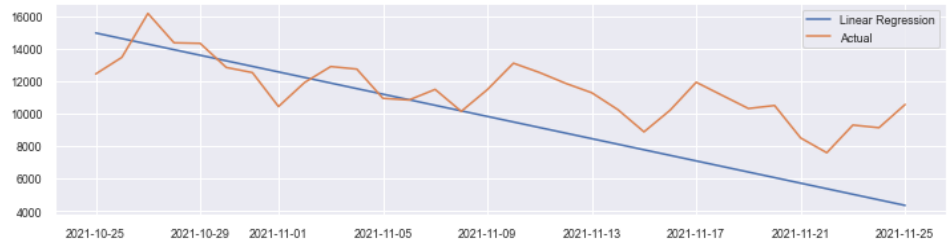

Fig. 6. .Linear Regression prediction with test data of daily active cases.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

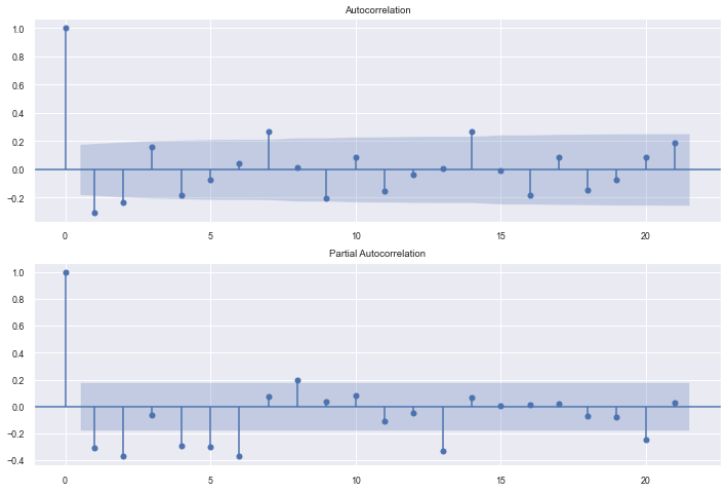

Fig. 7. ACF and PACF plots.

B.Linear Regression

On training the Linear Regression Model with an 80 days dataset, a straight line with a negative slope is formed. The Regression line can be seen coinciding with the actual cases trend line at a few points and then gradually moving away from it as the actual cases tend to rise, hence indicating a potential reversal. Thus, showing that Linear Regression works better on a linear dataset.

Fig. 6 depicts Linear Regression predicted vs Actual trend lines from 25 October 2021 to 25 November 2021. The root mean square error is calculated for one-day, one week, and one month. The one-day RMSE is 2.52E+03, seven-day RMSE is 1.45E+03 and the one-month RMSE is 2.76E+03.

C. ARIMA

On processing the initial data and checking for stationarity (by performing an Augmented Dickey-Fuller test), the data came out non-stationary. Hence by using first-order differencing, we got the following ACF and PACF plots. Fig. 7 depicts an ARIMA(2, 1, 2) model. After training ARIMA(2, 1, 2) on the recent subset of cases (from June) we got Fig. 8.

Fig. 8. depicts ARIMA(2, 1, 2) predictions from 25th October 2021 to 25th November 2021 with a one-day RMSE of 2.53E+03, seven-day RMSE of 2.47E+03, and a one-month RMSE of 4.32E+03.

D. SARIMA

For SARIMA, we tried a brute force approach to determine the parameters in a stepwise manner, by using a period of 3 months and minimizing AIC. This resulted in a SARIMA(2,1,2)(2, 0, 1, 3) with an AIC score of 2.36E+03. After training SARIMA(2,1,2)(2, 0, 1, 3) on the whole dataset we got Fig. 9.

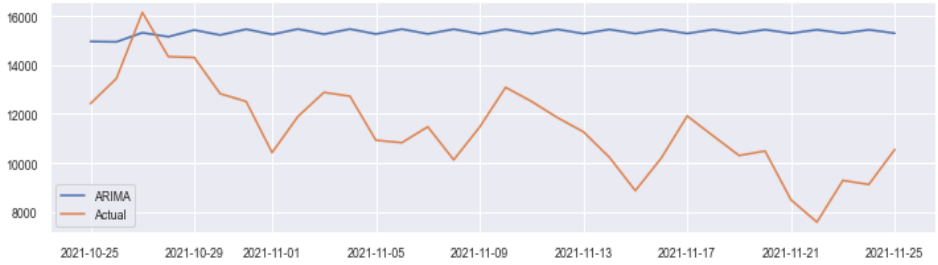

Fig. 8. ARIMA(2, 1, 2) predictions with test data of daily cases.

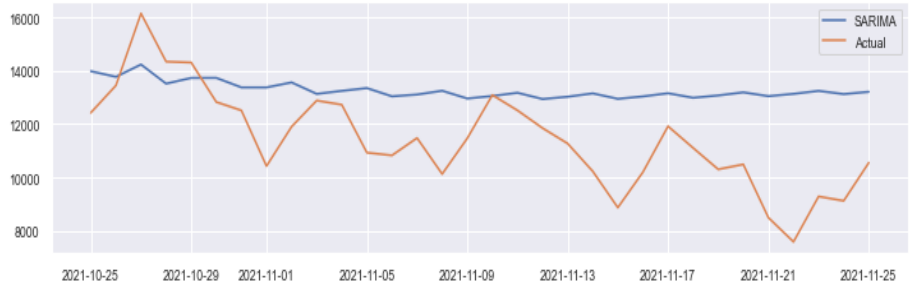

Fig. 9. SARIMA(2,1,2)(2, 0, 1, 3) predictions with test data of daily cases.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

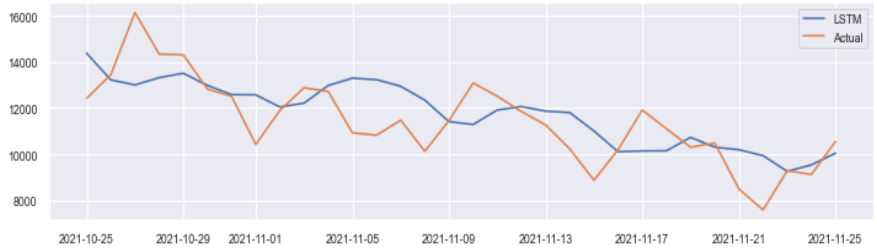

Fig. 10. Prediction of Covid-19 cases using LSTM

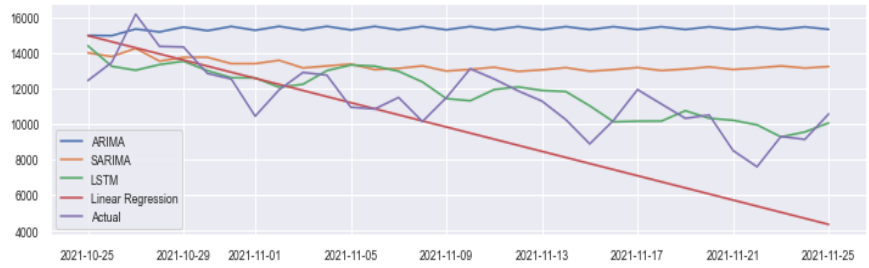

Fig. 11. Comparing the predictions for all the models

Fig. 9. depicts SARIMA(2,1,2)(2, 0, 1, 3) predictions from 25th October 2021 to 25th November 2021 with a one-day RMSE of 1.58E+03, seven-day RMSE of 1.48E+03, and a one-month RMSE of 2.46E+03.

E. LSTM

After the general preprocessing of data, we get the daily Covid-19 cases of the past data. Then some additional preprocessing was done on the dataset which includes the scaling of the values of cases in the range of 0 to 1. LSTM is sensitive to values of the data used for training hence the data needs to be scaled. The data used for training the model consisted of the daily Covid-19 cases from 9th of April 2020 till 24th of October 2021 which consists of the data of a total

of 563 days.

In order to feed the data into the LSTM model, we have to create sequences of the values from which the model produces the output. Before training the model we need to define the configuration of the model which consists of determining the number of layers and the number of units in each layer. So in this study, we have used the LSTM model with a total of three layers and one output layer where the first layer consists of 150 units, the second and the third layer consists of 64 units and the output layer consists of one unit

which corresponds to the value predicted by the model. Then the model was fit with the sequences of the data and was used for prediction of the future cases.

Fig. 10 shows the prediction of Covid 19 cases using LSTM from 25th of October 2021 to 25th November 2021. These predictions show the RMSE of 1.95E+03 for one-day prediction, 1.48E+03 for seven-days prediction and 1.41E+03 for one-month prediction.

TABLE II. Comparative analysis of various time series forecasting algorithms for daily cases

Algorithm	1 Day	7 Days	1 Month
Linear Regression	2.52E+03	1.45E+03	2.76E+ 03
ARIMA	2.54E+03	2.47E+03	4.32E+03
SARIMA	1.55E+03	1.48E+03	2.46E+03
LSTM	1.95E+03	1.48E+03	1.41E+03

V. CONCLUSION.

After having reviewed the related literature, we came across the following limitations. First, the time frame of the data-set used for the study and the predictions for the same don't align. Second, the forecasting range of the chosen literature varies, thus making it hard to conclude. Also, Linear Regression doesn't work optimally for non-linear data. The SIR model becomes inaccurate for long periods as one or more of the assumptions are violated.

In this survey, the results of various statistical and deep learning methods for predicting Covid-19 cases are compared. All of these included data extractions, data preprocessing, data visualization, prediction using machine learning algorithms, and error checking. As per our observations, as the number of days of forecast goes on increasing the error rate of all the models also increases.

As conventional knowledge states that, Deep Learning models tend to outperform Statistical Models in the long term. Our observations aligned with this statement, for forecasts of fewer days, Statistical methods, SARIMA for one-day and Linear Regression for seven-days, were better on average, whereas, for forecasts of greater days, Deep Learning method, LSTM for one month, tend to be more consistent.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 12, December-2021

REFERENCES

[1] K.E. ArunKumar, D.V. Kalaga, C.M. Sai Kumar et al., "Forecasting the dynamics of cumulative Covid-19 cases (confirmed, recovered and deaths) for top-16 countries using statistical machine learning models: Auto-Regressive Integrated Moving Average (ARIMA) and Seasonal Auto-Regressive Integrated Moving Average (SARIMA)", Applied Soft Computing Journal 103 (2021) 107161, 2021.

[2] https://en.wikipedia.org/wiki/Compartmental_models_in_epidemiology. Accessed 1/10/2021.

[3] Aykut Ekinci "Modelling and forecasting of growth rate of new Covid-19 cases in top nine affected countries: Considering conditional variance and asymmetric", Chaos, Solitons and Fractals 151 (2021) 111227.

[4] PabelShahrear, S. M. SaydurRahman, Md Mahadi Hasan Nahi "Prediction and mathematical analysis of the outbreak of coronavirus (Covid-19) in Bangladesh" Results in Applied Mathematics Volume 10, May 2021, 100145.

[5] reportCoronavirus disease (Covid-2019) situation reports. World Health Organization. [Online] Available at:: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[6] Nelson DB. Conditional heteroskedasticity in asset returns: a new approach. Econometrica 1991:347–70.

[7] Mohammed N. Alenezi, Fawaz S. Al-Anzi, Haneen Alabdulrazzaq, Building a sensible SIR estimation model for Covid-19 outspread in Kuwait, Alexandria Engineering Journal, Volume 60, Issue 3, 2021.

[8] [Su L., Ma X., Yu H., Zhang Z., Bian P., Han Y., Sun J., Liu Y., Yang C., Geng J., Zhang Z. "The different clinical characteristics of coronavirus disease cases between children and their families in China–the character of children with COVID-19 Emerg Microbes Infect", 9 (1) (2020), pp. 707-713

[9] Na Zhu, Ph.D., Dingyu Zhang, M.D., Wenling Wang, Ph.D., Xingwang Li, et al., for the China Novel Coronavirus Investigating and Research Team, 2020. A novel coronavirus from patients with pneumonia in China, 2019. The New England Journal of Medicine, vol. 382, no. 8, pp. 727-733.

[10] Cucinotta, Domenico, and Maurizio Vanelli. "WHO Declares Covid-19 a Pandemic." Acta bio-medica : Atenei Parmensis vol. 91,1 157-160. 19 Mar. 2020, doi:10.23750/abm.v91i1.9397

[11] May Robert M. "Infectious Diseases of Humans: Dynamics and Control" Oxford University Press, New York (1991)

[12] Vinay Kumar Reddy Chimmula, Lei Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks", Chaos, Solitons and Fractals 135 (2020) 109864

[13] Wang W., Tang J., Wei F. Updated understanding of the outbreak of 2019 novel coronavirus (2019-nCoV) in Wuhan, China. J Med Virol. 2020;92(4):441–44

[14] Muhammad Adnan Shereen, Suliman Khan Abeer Kazmi, Nadia Bashir, Rabeea Siddique, "Covid-19 infection: Origin, transmission, and characteristics of human coronaviruses" Journal of Advanced Research 24 (2020) 91-98.

[15] Nooshin Ayoobi, Danial Sharifrazi, Roohallah Alizadehsani, Afshin Shoeibi, Juan M. Gorriz, Hossein Moosaei, Abbas Khosravi, Saeid Nahavandi, Abdoul mohammad Gholamzadeh Chofreh, Feybi Ariani Goni, Ji˘rí Jaromír Kleme˘s, Amir Mosavi, "Time series forecasting of new cases and new deaths rate for Covid-19 using deep learning methods", Results in Physics 27(2021) 104495

[16] Forecasting: Principles and Practice (2nd ed) Rob J Hyndman and George Athanasopoulos, April 2018.

[17] ARIMA simplified. A simplistic explanation to the most popular Time Series Forecasting model out there. Abhishek Rajbhoj. September 2019.

[18] Weston C. Roda, Marie B. Varughese, Donglin Han, Michael Y. Li, Why is it difficult to accurately predict the Covid-19 epidemic?, Infectious Disease Modelling, Volume 5, 2020.

[19] Kermack, William Ogilvy, and Anderson G. McKendrick. "A contribution to the mathematical theory of epidemics." Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character 115.772 (1927): 700-721.

[20] https://en.wikipedia.org/wiki/File:Diagram_of_SIR_epidemic_model_states_and_transition_rates.svg

[21] Ghosal, Samit et al. "Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020)." Diabetes & metabolic syndrome vol. 14,4 (2020): 311-315. doi:10.1016/j.dsx.2020.03.017

[22] Rath, Smita et al. "Prediction of new active cases of coronavirus disease (Covid-19) pandemic using multiple linear regression model." Diabetes & metabolic syndrome vol. 14,5 (2020): 1467-1474. doi:10.1016/j.dsx.2020.07.045

[23] S. Shaikh, J. Gala, A. Jain, S. Advani, S. Jaidhara and M. Roja Edinburgh, "Analysis and Prediction of Covid-19 using Regression Models and Time Series Forecasting," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021.

[24] Sujath, R., Chatterjee, J.M. & Hassanien, A.E. A machine learning forecasting model for Covid-19 pandemic in India. Stoch Environ Res Risk Assess 34, 959–972 (2020). https://doi.org/10.1007/s00477-020-01827-8

[25] Junaid Farooq, Mohammad Abid Bazaz, A deep learning algorithm for modeling and forecasting of Covid-19 in five worst affected states of India, Alexandria Engineering Journal, Volume 60, Issue 1,2021, Pages 587-596, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2020.09.037.

[26] Abdelhafid Zeroual, Fouzi Harrouc, Abdelkader Dairi, Ying Sun "Deep learning methods for forecasting Covid-19 time-Series data: A Comparative study" Chaos, Solitons and Fractals 140(2020) 110121

[27] Car, Zlatan et al. "Modeling the Spread of Covid-19 Infection Using a Multilayer Perceptron." Computational and mathematical methods in medicine vol. 2020 5714714. 29 May. 2020, doi:10.1155/2020/5714714

[28] Safa Bahri, Moetez Kdayem, Nesrine Zoghlami, "Deep Learning for Covid-19 predictions", 2020 4th International Conference on Advanced Systems and Emergent Technologies

[29] Rauf, H.T., Lali, M.I.U., Khan, M.A. et al. Time series forecasting of Covid-19 transmission in Asia Pacific countries using deep neural networks. Pers Ubiquit Comput (2021). https://doi.org/10.1007/s00779-020-01494-0

[30] COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. https://github.com/CSSEGISandData/COVID-19, Accessed 6/12/2021.