COMPARATIVE ANALYSIS OF EPIDEMIOLOGICAL, FORECASTING AND DEEP LEARNING MODELS FOR COVID19 SPREAD IN INDIA

APARNA VELLALA

AUGUST 2022

ABSTRACT

Covid19 is affecting across many nations and most population of the world. As per WHO there are 270million confirmed with about 5.3 million fatalities as on December 15th, 2021. Many governments, organizations and local bodies have been applying various models in order to estimate the disease spread and appliede varied strategies to curb the spread. There are many models proposed by mathematicians and statisticians for the same. In the current work a comparison is done with mathematical disease spread models SIR, SIRD, classic time series forecasting modelARIMA, and artificial neural network models RNN, LSTM with Covid19 India data. The study investigates the effect of disease containment policies and vaccination drives for Covid19 data in the context of India using SIR Model. All the models are built for multiple time prediction windows starting from 5 days up to 45 days. The models are evaluated with MAE, MAPE and RMSE for multiple states and India level data. It is inferred that the prediction time of 5 days has best results for SIR model. The ARIMA model can predict withacceptable performance up to 30 days. RNN and LSTM models can predict for 5 days within acceptable performance. The best model that can predict longer durations and has good performance is ARIMA model. A detailed report on the model details and performance is the outcome of this study.

9012

DOINumber:10.14704/nq.2022.20.8.NQ44921 **NeuroQuantology 2022; 20(8): 9012-9020**

LIST OF FIGURES AND TABLES

LIST OF ABBREVIATIONS

ADF............... Augmented Dickey–Fuller
ANN............... Artificial Neural Network
ARIMA............... Auto Regressive Integrated Moving Average
COVID19............Corona Virus Disease 2019
D...................... Dead/ Diseased

i

E…………………… Exposed
EDA………….. Exploratory Data Analysis
I…………………… Infectious
LSTM…………… Long Short-Term Memory
MAPE………….. Mean Absolute Percentage Error
MAE………….. Mean Absolute Error
R………………… Recovered
RMSE………….. Root Mean Squared Error
RNN………….. Recurrent Neural Network
S…………………… Susceptible
SARS Cov2………... Severe Acute Respiratory Syndrome Corona Virus 2
WHO .... World Health Organization
IN ………................ India

ACKNOWLEDGEMENTS

I would like to thank Mr. Praveen D. Chougale for mentoring and guiding me in completing this research. Also I would like to thank Liverpool John Moore's University, Liverpool, U.K. for providing me opportunity to do this research.

Table of Contents

1. INTRODUCTION

Covid19 is caused by SARS Cov2 Virus. This research is how to model the disease spread for Covid19 spread in India. Mathematical disease spreading models SIR and SIRD, time series advanced forecasting model ARIMA, deep learning models, RNN and LSTM are modeled for different predictive time windows. This research evaluates performance of models with error measurements techniques, MAPE, MAE and RMSE for the comparative study. Effect of policy decisions or disease control measures – lockdowns and immunization on the disease spread are also analyzed with SIR Model.

2. LITERATURE REVIEW

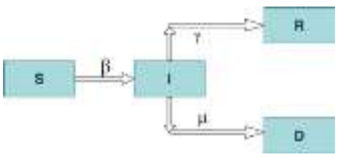
The literature is reviewed for the methods or architectures of the various models under three main categories, epidemiological models, statistical forecasting models and deep learning models.The data published in (Coronavirus Outbreak in India - covid19india.org, 2021) is picked from various state bulletins and is validated by the volunteers of the COVID19 India Org and is leveraged in the recent work (Arora et al., 2020).

Mathematical compartmental deterministic models are widely used for modelling disease spread. Typical infectious disease model, SIR Model is discussed in (Diekmann et al., 2012), (Ramanathan et al., 2012), (Hethcote, 2007)with 3 compartments, S for susceptible, I for infectious, R for Recovered in a closed population as shown in figure 1. In the model, β is the transmission rate that is constant, γ is the recovering rate that is fixed rate at which individualsget recovered or die.

SIR Model - Figure1

According to (Diekmann et al., 2012)R_0 is the number of infectious cases per typical infected person and is mathematically calculated in compartmental epidemic system.

(Chen et al., 2020) proposes time dependent SIR model by making transmission rate, $\beta$and recovery rate, γ as a variable that changes along with time. Another model, SIRDis created in the work (Fanelli and Piazza, 2020)and is as shown in figure 2



SIRD Model - Figure 2

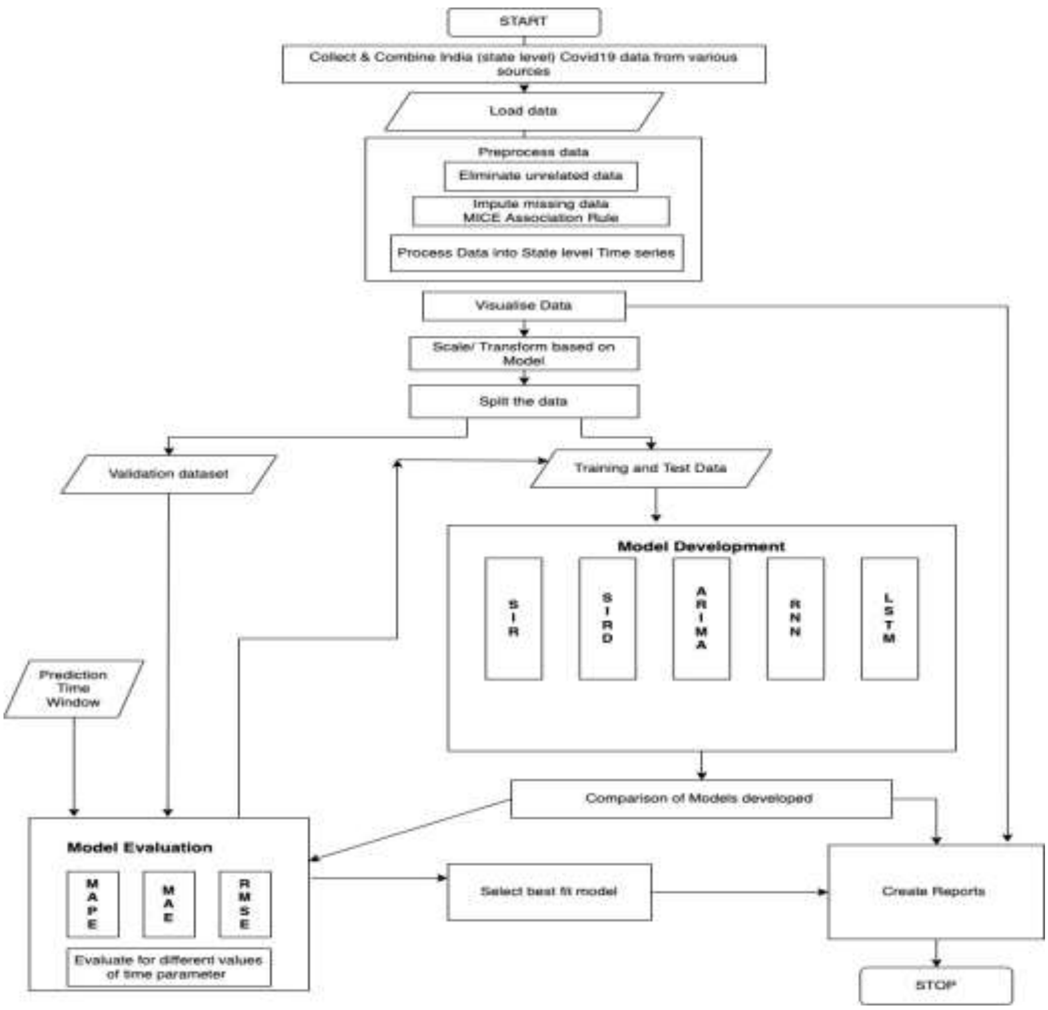
(Liao et al., 2020) proposed time window-based SIR model. The time window considered for SIR is from 3 to 30 and the hyperparameter can be tuned accordingly. The parameter evaluation is done based on the time window at the interval of time, t.The statistical forecasting models Auto Regressive (AR) model, Moving Average (MA) model, Auto-Regressive Moving Average (ARMA), and Auto-Regressive Integrated Moving Average (ARIMA)models are applied, and a comparison is published with the best model for the specific Covid19 data being ARIMA in (Alghamdi et al., 2019).

As per this study (Alzahrani et al., 2020), the parameters for ARIMA are chosen based on Akaike Information Criterion (AIC). Artificial neural networks (ANN) with feed forward are used for predictions and recurrent neural networks (RNN) architecture considers multiple entities present in a sequence. In the recent works(Arora et al., 2020; Bandara et al., 2020; Song et al., 2020; Zeroual et al., 2020), the basic structure of the vanilla RNN is leveraged and implemented for predicting time series. In the recent works (Siami-Namini et al., 2019; Bandara et al., 2020; Song et al., 2020; Zeroual et al., 2020)leverages LSTM that has stand-alone LSTM cell with gating mechanisms to address vanishing gradients or exploding gradients and long term short memory (LSTM).

The statistical forecasting models are evaluated using all or some of the accuracy measurement functions, RMSE, MAE, R^2, MAPE, RMSRE in the recent works (Alzahrani et al., 2020), (Owusu-sekyere and Harris, 2013; Alghamdi et al., 2019; Duan and Zhang, 2020; Petropoulos et al., 2020).

3. RESEARCH METHODOLOGY AND IMPLEMENTATION

For all the models, prediction is done for time windows, t = 5, 10, 15, 30, and 45 days. Model evaluation is done to do the comparative study with MAPE, MAE and RMSE.



Research Methodology Flow Diagram - Figure 3

The implementation is done in python. SIR and SIRD models are built for predicting the disease spread by solving ODEs respectively for the entire period of prediction i.e., 590 days. The models are also implemented for various time windows. Model building of ARIMA, Vanilla RNN and LSTM is done with train data of 70%, predictions are done for rest 30% for various time windows. Analysis is done by comparing the performance of each model with the time parameter and between the models.

The effectiveness of lockdowns is compared with the basic reproduction rate for time, t, R0(t) that are plotted from 0 to 120 days. The lockdowns or restrictions were mandated from 10 to 78 days i.e.,from 14th March 2020 to 31st May 2020. The plots with actual and predicted Infected, Susceptible and Recovered parameters for India and all the states for the initial 120 days is done to compare the impact of the lockdown restrictions.The immunization started in India on 16th January 2021 i.e., t=309. SIR Model is implemented with S0, I0, R0 initialized withthe actual values at time t=309. The effectiveness of immunizations is inferred by comparing the actual and the predicted S, I, R values.

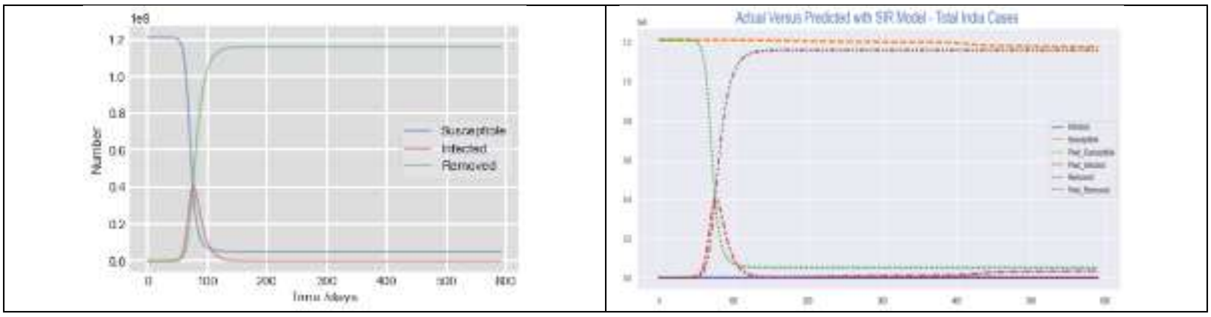
4. RESULTS AND INTERPRETATIONS

The results section consists of exploratory data analysis i.e., inferring trends for every state and India. The SIR model results are updated for to analyze the effect of lockdowns and immunizations. Intra model and inter model analysis is done on the performance metricsfor different prediction time windows.

4.1. Immunization and Disease Control Measures

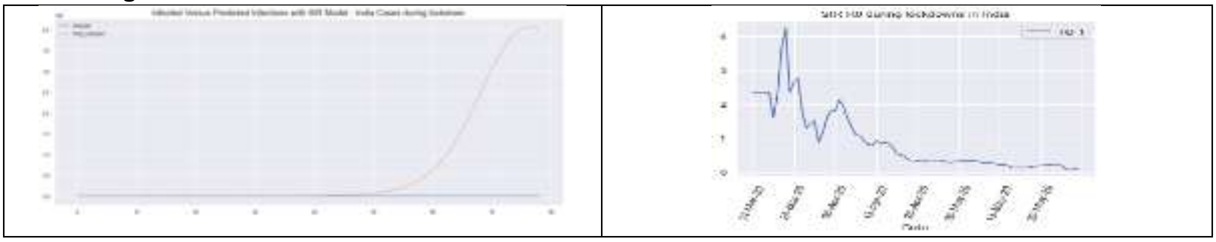
The SIR model results for India data are shown as in figure 4. As per the SIR Model, the maximum number of infections predicted without any restrictions is on 29th May 2020 and it is 405 million on day 78.The plots with comparison between actual S, I, R and predicted S, I, R time series is as shown in figure 8 for India.

The SIR models are created for 20 states of India. The plots and the predicted variables are published in GitHub,https://github.com/aparna2015/covid19ModellingResults.git.



SIR Model – Actual Versus Predicted - India - Figure 4

As part of the Covid19 disease containment, India imposed total lockdown from 14th March 2020 to 31st May 2020 i.e., from day 10 to day 78 of the timeseries. Figure 5 shows a comparison of actual infected to predicted infected for India during the lockdown. The actual number of infections is lower. Time based basic reproduction ratio is calculated and is plotted for lockdown duration as shown in figure 5. From these it can be deduced that the disease spread is lower than expected and lockdowns have been effective.



SIR parameters– India - Figure 5

4.2. Model Evaluations with prediction time windows

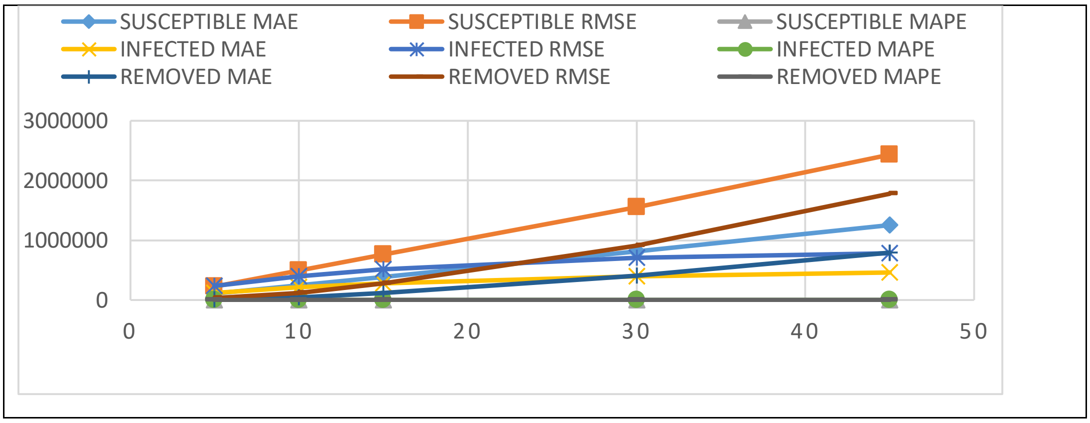
4.2.1. SIR and SIRD Model

The conventional SIR model is extended to predict shorter durations and supervised or initialized to the actual at interval, t days. The prediction is also done for t days i.e., t=5, 10, 15, 30 and 45. The models are created, and predictions are done for all the variables, S, I, and R (Removed). The performance measurement metrics, MAE, RMSE, and MAPE are calibrated for every parameter and various predicted time windows as shown in table 3 and figure 6for India.

SIR Model – India –performance metrics -Table 1

Predicted Time	SIR Model with various predicted time for India								
	SUSCEPTIBLE			INFECTED			REMOVED		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
5	110058	226397	0	121782	240271	19	16581	29123	9
10	249814	497414	0	211962	401117	34	45869	118596	13
15	389202	764072	0	279789	513686	45	114275	280853	13
30	811047	1553622	0	400099	706299	63	412970	915241	16
45	1255190	2437330	0	459426	784031	72	797209	1781121	23

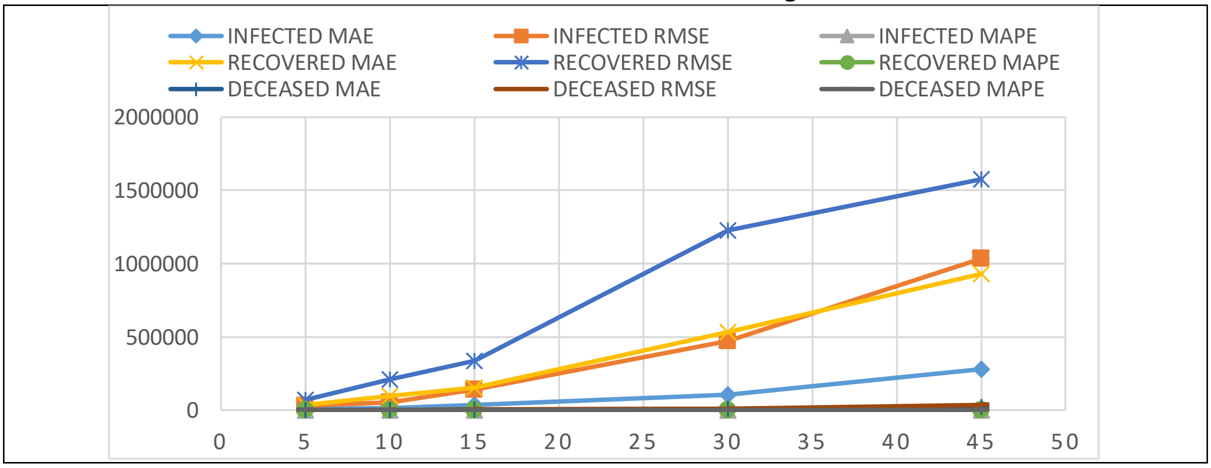
It can be inferred from the plots and metrics that the supervised learning of SIR model performance is better for lower prediction time windows than longer durations. The performance (considering MAPE) for the infected is acceptable with prediction time of 5 days. The performance of Infected is low for longer duration predictions and is having a clear increasing trend of errors as prediction time increases. Similar study id done on SIRD model as well.



SIR Model performance metrics comparison chart –Figure 6

4.2.2. ARIMA Model

The ARIMA model is implemented for Infected, Recovered and Deceased as univariate time series for India and selected four states. The stationarity is checked and the models are applied for predicting for different time windows. The performance metrics for the models for India as shown in figure 7.

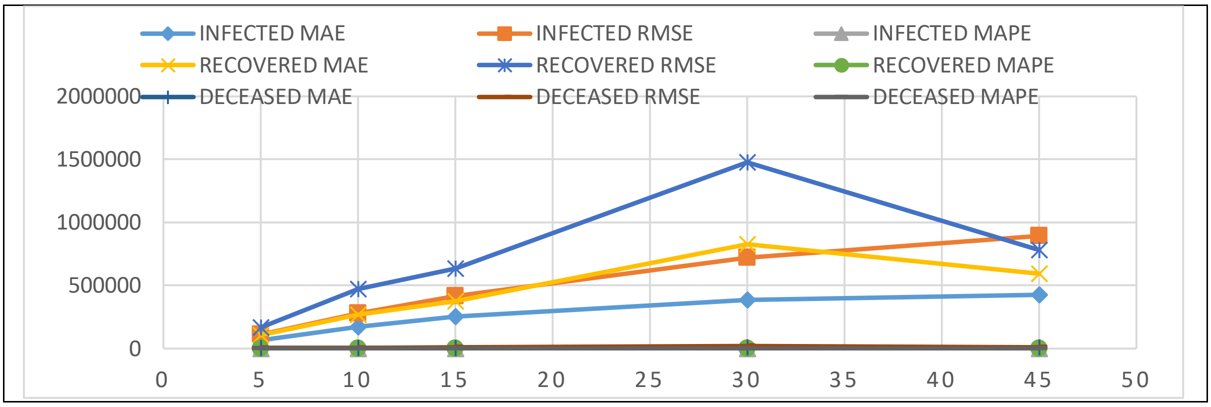


ARIMA Model performance metrics comparison chart - Figure 7

It can be inferred that the performance is reducing as the prediction window is increasing. However, the performance of this model is >70% (considering MAPE <30) for all the time series up to the prediction time of 30 days and for some time series the model is acceptable for 45 days.

4.2.3. RNN and LSTM Model

The performance metrics of RNN modelling for various time prediction windows are plotted in figure 8. The number of lags used for RNN modelling is 50. The performance India goes down when predicted time window is 5 to 30 and for 45, it is improving over the performance at 30. There is no direct correlation between the prediction time and the performance for RNN. Similar study is done for LSTM Model as well.

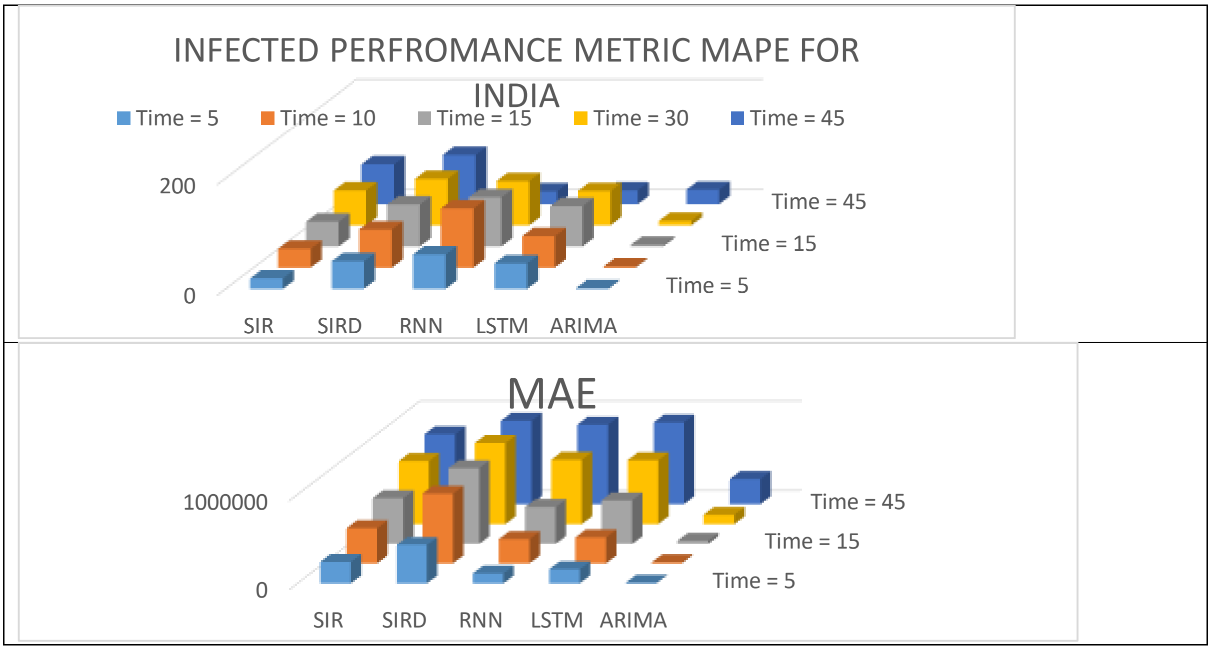


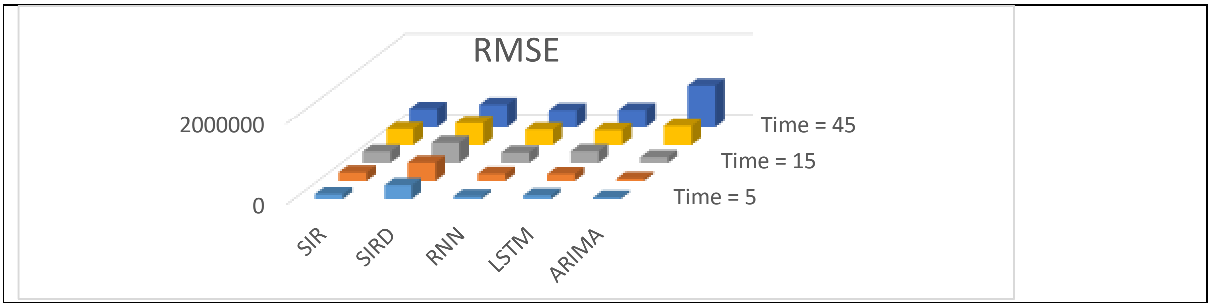
RNN Model performance metrics comparison chart - Figure 8

4.2.4. Inter Model Performance Evaluation

The performance of SIR Model is acceptable for prediction time window of 5 days. The number of datapoints required for this model is only 5 that are used for calibrating the basic reproduction ratio. The input given to SIR model is only one datapoint. The SIRD model also uses 5 days data for the basic reproduction ratio and is not having acceptable performance.All the model performance metrics are plotted in figure 9. As it is clearly indicating that the performance of ARIMA model is highest for predicted time windows of 5, 10, 15, and 30 days. It is highest for 45 days prediction when MAE is compared. The number of datapoints required for modelling ARIMA is high.

LSTM and RNN perform better for the shorter prediction windows when RMSE is considered. It is to be noted that LSTM and RNN training optimization is done considering MSE in the current study. The number of datapoints required for modelling ARIMA is high. The number of lags given as input to RNN and LSTM is 50. Hence, the number of datapoints required for every prediction are high as well.





India – Infected – Inter Model Performance Metrics Comparison Charts - Figure 9

5. CONCLUSIONS

The SIR model predictions are acceptable range for shorter durations i.e. 5 days while MAPE is considered. ARIMA has better performance for longer prediction time windows i.e. about 30 days with less than 20% MAPE. It is inferred that the lockdowns or the disease containment strategies have been effective and immunization has a negative effect on the disease spread in India.From the results, ARIMA model has best performance in this study and can predict longer durations up to 30 days with acceptable performance i.e., MAPE of 8. Followed by it is the SIR Modelfor prediction of 5 days. It has acceptable performance i.e., MAPE of 19.

The other method applied for lockdown effectiveness is the calibration of time-based basic reproduction rate that has a decreasing trend during lockdowns.SIR models are built similarly for the period of immunization. When the number of infected people predicted by the model and the actual number of infected people are compared, it can be concluded that immunization has played major role in containing the Covid19 disease spread within India.

5.1. Limitations and Future Work

This research is limited to certain location data. The compliance of the covid19 disease containment measures is different for every state/ location. The norms or regulations are different for every state. India data is cumulative sum of all the state level data. As future work,the effect of batch size and number of epochs for the training change the performance of the deep learning models. The performance of RNN, and LSTM models can be optimized with MAPE and/or MAE as well. In the LSTM model, if the number of lags is correlated to the performance of the predicted time windows.

REFERENCES

Anon (2021) *Coronavirus Outbreak in India - covid19india.org*. [online] Available at: https://www.covid19india.org/ [Accessed 1 Aug. 2021].

Arora, P., Kumar, H. and Panigrahi, B.K., (2020) Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons and Fractals*, [online] 139, p.110017. Available at: https://doi.org/10.1016/j.chaos.2020.110017.

Bandara, K., Bergmeir, C. and Smyl, S., (2020) Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. *Expert Systems with Applications*, 140.

Chen, Y.C., Lu, P.E., Chang, C.S. and Liu, T.H., (2020) A Time-Dependent SIR Model for COVID-19 with Undetectable Infected Persons. *IEEE Transactions on Network Science and Engineering*, 74, pp.3279–3294.

Diekmann, O., Heesterbeek, H. and Britton, T., (2012) *Mathematical tools for understanding infectious disease dynamics*. *Mathematical Tools for Understanding Infectious Disease Dynamics*.

Duan, X. and Zhang, X., (2020) ARIMA modelling and forecasting of irregularly patterned COVID-19 outbreaks using Japanese and South Korean data. *Data in Brief*, [online] 31, p.105779. Available at: https://doi.org/10.1016/j.dib.2020.105779.

Fanelli, D. and Piazza, F., (2020) Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos, Solitons and Fractals*, [online] 134, p.109761. Available at: https://doi.org/10.1016/j.chaos.2020.109761.

Hethcote, H.W., (2007) The Mathematics of Infectious Diseases The Mathematics of Infectious Diseases *. *Society for Industrial and Applied Mathematics*, [online] 424, pp.599–653. Available at: http://www.jstor.org/discover/10.2307/2653135?uid=3739736&uid=2&uid=4&uid=3739256&sid=21104838342357.

Owusu-sekyere, E. and Harris, E., (2013) Forecasting and Planning for Solid Waste

9019

Generation in the Kumasi Metropolitan Area of Ghana: An ARIMA Time Series Approach. *International Journal of Sciences*, 204, pp.69–83.

Petropoulos, F., Makridakis, S. and Stylianou, N., (2020) COVID-19: Forecasting confirmed cases and deaths with a simple time series model. *International Journal of Forecasting*, [online] xxxx. Available at: https://doi.org/10.1016/j.ijforecast.2020.11.010.

Ramanathan, A., Steed, C.A. and Pullum, L.L., (2012) Verification of compartmental epidemiological models using metamorphic testing, model checking and visual analytics. *Proceedings of the 2012 ASE International Conference on BioMedical Computing, BioMedCom 2012*, SocialInformatics, pp.68–73.

Siami-Namini, S., Tavakoli, N. and SiamiNamin, A., (2019) A Comparison of ARIMA and LSTM in Forecasting Time Series. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pp.1394–1401.

Song, X., Liu, Y., Xue, L., Wang, J., Zhang, J., Wang, J., Jiang, L. and Cheng, Z., (2020) Time-series well performance prediction based on Long Short-Term Memory (LSTM) neural network model. *Journal of Petroleum Science and Engineering*, [online] 186November 2019, p.106682. Available at: https://doi.org/10.1016/j.petrol.2019.106682.

Zeroual, A., Harrou, F., Dairi, A. and Sun, Y., (2020) Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, Solitons and Fractals*, [online] 140, p.110121. Available at: https://doi.org/10.1016/j.chaos.2020.110121.

Zhang, G., Eddy Patuwo, B. and Y. Hu, M., (1998) Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, 141, pp.35–62.