Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Improving performance of deep learning predictive models for COVID-19 by incorporating environmental parameters

Roshan Wathore a,b, Samyak Rawlekar c, Saima Anjum a, Ankit Gupta a,b, Hemant Bherwani a,b,* , Nitin Labhasetwar a,b, Rakesh Kumar b,d

a CSIR-National Environmental Engineering Research Institute (CSIR-NEERI), Nehru Marg, Nagpur 440020, Maharashtra, India
b Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, Uttar Pradesh, India
c Indian Institute of Technology (IIT) Dharwad, Dharwad 580 011, Karnataka, India
d Council of Scientific and Industrial Research (CSIR), Anusandhan Bhawan, 2 Rafi Ahmed Kidwai Marg, New Delhi 110001, India

A R T I C L E   I N F O

Article history:
Received 31 August 2021
Revised 17 March 2022
Accepted 17 March 2022
Available online 8 April 2022

Keywords:
COVID-19
SARS-CoV-2
LSTM. Multivariate time series forecasting
Deep Learning

abstract

The Coronavirus disease 2019 (COVID-19) pandemic has severely crippled the economy on a global scale. Effective and accurate forecasting models are essential for proper management and preparedness of the healthcare system and resources, eventually aiding in preventing the rapid spread of the disease. With the intention to provide better forecasting tools for the management of the pandemic, the current research work analyzes the effect of the inclusion of environmental parameters in the forecasting of daily COVID-19 cases. Three univariate variants of the long short-term memory (LSTM) model (basic/vanilla, stacked, and bi-directional) were employed for the prediction of daily cases in 9 cities across 3 countries with varying climatic zones (tropical, sub-tropical, and frigid), namely India (New Delhi and Nagpur), USA (Yuma and Los Angeles) and Sweden (Stockholm, Skane, Uppsala and Vastra Gotaland). The results were compared to a basic multivariate LSTM model with environmental parameters (temperature (T) and relative humidity (RH)) as additional inputs. Periods with no or minimal lockdown were chosen specifically in these cities to observe the uninhibited spread of COVID-19 and explore its dependence on daily environmental parameters. The multivariate LSTM model showed the best overall performance; the mean absolute percentage error (MAPE) showed an average of 64% improvement from other univariate models upon the inclusion of the above environmental parameters. Correlation with temperature was generally positive for the cold regions and negative for the warm regions. RH showed mixed correlations, most likely driven by its temperature dependence and effect of allied local factors. The results suggest that the inclusion of environmental parameters could significantly improve the performance of LSTMs for predicting daily cases of COVID-19, although other positive and negative confounding factors can affect the forecasting power.

G A R N E R  R E S E A R C H  1 1 4 ( 2 0 2 3 ) 6 9 – 7 7

1. Introduction

The Coronavirus Disease 2019 (COVID-19) pandemic caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV 2) has severely impacted the social, economic and environmental aspects of human lives. Various sectors, industries and businesses have been harshly impacted by the restrictions due to this global crisis (Bherwani et al., 2021; Ranjbari et al., 2021). Worldwide response to COVID-19 included travel bans, social distancing, lockdown of non-essential services and working from home in an attempt to “flatten the curve” and reduce burden on the healthcare system. The extent and effectiveness of these restrictions varied greatly across administrative/political, demographic, economic, and environmental factors.

This pandemic has spurred tremendous research efforts to better understand the virus survival and transmission. One of the study areas that has been widely explored is the effect of environmental conditions on the virus transmission has been well reviewed (Wathore et al., 2020; Gautam et al., 2021a; Bherwani et al., 2020) with most studies reporting increased temperature, UV radiation and wind speed reduce the risk of COVID-19 spread. Wang et al., (2020) analyzed data from 100 Chinese cities and 1005 US counties and found that high temperature and high humidity reduce the transmission of COVID-19. In Indonesia,
Tosepu et al., (2020) determined that only average temperature was significantly correlated with the pandemic. Data from Italian province capitals showed that low wind speed, high moisture and occurrences of foggy days were associated with increased transmission of COVID-19 (Coccia, 2020). Wu et al., (2020) reported that with a unit increase in temperature, the COVID-19 cases reduced by about 3% whereas a unit increase in humidity caused the cases to shrink by about 1%. Lin et al., (2020) was concluded that low temperature with increased humidity causes increased spread of the virus and increased prevalence of the disease in Asian countries. Dbouk and Drikakis, (2020) explored energy and mass balance correlations with respect to the viability of the virus and concluded that high temperature and lower humidity significantly reduces the prevalence of the virus and reduces the transmission of disease. Some studies have suggested that the correlation with respect to humidity and temperature is not that linear as it is being projected and confounding variables such as physical distancing and restricted movements are an integral part of the rise and fall of cases and hence it may not be possible to pinpoint the exact relation of virus viability, transmissibility with the environmental attributes (Bherwani et al., 2020; Gupta et al., 2021). Bherwani et al., (2020a) incorporated environmental parameters (T and RH) into a Susceptible-Exposed-Infectious-Removed (SEIR) model and showed that the inclusion of environmental parameters is essential for improved model performance and systematic planning for handling the pandemic. However, allied parameters such as administrative restrictions, lockdowns, social and physical distancing have confounding effects and need detailed delineation in order to find out their impacts vis-a-vis environmental attributes.

Another study area which has been explored is the use of various modeling techniques to predict the COVID-19 cases in a country or region. Accurate forecasting is essential for managing a pandemic; it facilitates a better decision-making process and development of practical measures and strategic plans to enable preparedness and allocate (health) resources appropriately. Statistical models such as Auto Regressive Moving Average (ARIMA) have been used to predict COVID-19 cases (Bayyurt and Bayyurt, 2020; Benvenuto et al., 2020; Sahai et al., 2020; Singh et al., 2020). Other popular forecasting methods include epidemiological models such as the SEIR models and their variations (Bherwani et al., 2020; Gupta et al., 2020; Xu et al., 2020). Recently deep learning models such as Long Short-Term Memory (LSTM) and their variants have demonstrated improvements in the forecasting power compared to traditional forecasting approaches (Azarafza et al., 2020; Chimmula and Zhang, 2020). Devaraj et al., (2021) compared ARIMA and two LSTM variants (basic/vanilla and stacked) for daily COVID-19 cases in India and Chennai and determined that the stacked LSTM significantly outperformed the ARIMA model (46% reduction in mean absolute percentage error (MAPE)). Shaob et al., (2021) concluded that LSTM showed the best predictive performance when the cases in 4 countries (Pakistan, USA, India and Brazil) were predicted using varied techniques including LSTM, ARIMA, Artificial Neural Network Models (ANN), Exponential Smoothing/Error Trend Seasonality (ETS) and Gene Expression Programming (GEP). Srivastava et al., (2021) did a comparative study of LSTM, ARIMA, Holt’s Linear forecasting model, Exponential smoothing, and Moving-average model algorithms to forecast the number of new cases in six countries (Italy, Spain, France, USA, China, and Australia), and also concluded that LSTM gave the best performance. Kirbas et al., (2020) carried out a study in eight European countries using confirmed cases as the parameter of validation using ARIMA and Nonlinear Autoregression Neural Network (NARNN) and LSTM and again, LSTM was found the most accurate model with MAPE ranging from 0.16% to 2.5%. Moreover, modified versions of the LSTM’s such as stacked LSTM and bi-directional LSTM (Bi-LSTM) have also been compared and have shown better results than the basic LSTM at prediction of COVID-19 cases (Arora et al., 2020; Shastri et al., 2020; Zeroual et al., 2020).

Given the above insights, it is important to note that the above cited literature looking into forecasting using LSTMs and their variants for forecasting COVID-19 cases has at least one of the following shortcomings:

1. The models incorporated are univariate, implying the assumption of no external influence of other factors in the transmission of the virus. The input is the historical cases and the output is the predicted cases. Out of the above mentioned studies, only Devaraj et al., (2021) have considered multivariate LSTM’s by considering additional input parameters such as number of deaths, recoveries, latitude and longitude.

2. The duration considered for the training of forecasting models fail to capture various dynamic changes in the spread of the virus. This time period would likely capture only a monotonous increase or decrease in the daily COVID-19 cases without capturing the peaks. Out of the above-mentioned studies, only Devaraj et al., (2021), Chimmula and Zhang, (2020) and Shaob et al., (2021) have considered both rise and drop in cases.

3. Modeling is done focused on the number of cases in a larger region such as country or state with limited studies on city/district/county/province level data (Azarafza et al., 2020; Devaraj et al., 2021). Availability of accurate district/city level information facilitates better decision making than availability of state/country level information.

The current work attempts to 1) include the effects of environmental parameters, and 2) forecast daily cases using various LSTM variants due to their better forecasting power. To the best of the authors’ knowledge, only Bhimala et al., (2021) have proposed a weather integrated multivariate LSTM models to improve the model performance; however they assume a single parameter for every state in India, which is not practically applicable in such a scenario as environmental and meteorological parameters can exhibit hyperlocal variation even in the city scale.

In order to fill the gaps outlined above, this study looks into the performance of 3 univariate LSTM models (Basic LSTM, Stacked LSTM, and Bi-directional LSTM) to forecast daily cases in 8 cities across 3 countries – India (New Delhi and Nagpur), USA (Los Angeles and Yuma) and Sweden (Stockholm, Skane, Uppsala, and Vastra Gotaland) with varying climatic zones (tropical, sub-tropical and near-frigid respectively). The methods section describes the dataset, preprocessing methods, brief introduction on the different LSTM variants used and the metrics for evaluating model performance. Subsequent to the methodology, the results are explained and the conclusions derived therein are discussed. The paper is unique in its way that it discusses the incorporation of environmental parameters in LSTM models used for forecasting COVID-19 daily cases, which is clearly delineated in the results and discussion sections. Finally, the potential for future work and limitations of this study are outlined.

2. Methodology

2.1. Dataset

A total of 8 study cities across 3 countries (India, Sweden, and the USA) with varying climate zones (tropical, temperate, and frigid) were considered. The study periods were also chosen so as to capture periods with low or minimal lockdown to observe the
near-uninhibited spread of the virus. Additionally, locations across different climates were considered to explore the effect of environmental parameters (daily average RH and T) on the daily cases. Table 1 summarizes the study locations, date ranges for analysis, and the environmental parameters observed during the study period. Daily cases and environmental parameters were taken from publicly available datasets. The duration of the analysis ranged from 164 to 204 days.

Swedish strategy during the onset of the COVID-19 pandemic was exploring options of voluntary measures with no specific and strict lockdowns in force (Ludvigsson, 2020), making the country an ideal case to explore the spread of COVID-19 with minimal interference of external factors. India had gone into complete lockdown from 24th March 2020, which resulted in significant reductions in pollution levels across the country. The lockdowns consisted of 4 phases, with the fourth phase ranging from 18 to 31 May 2020, during which lockdown restrictions gradually started lifting. This was followed by 6 unlocking phases from June to November (Ambade et al., 2021a, 2021b, 2021c, 2021d; Chelani and Gautam, 2021; Gautam et al., 2021). In Arizona, USA, the statewide lockdown order expired in May 2020, which eventually led to a sharp rise in cases that eventually declined by October 2020. Similarly, in California, restrictions gradually started relaxations starting from May 2020.

2.2. Data preprocessing and model preparation

The daily cases and daily averaged environmental data for the selected cities were further passed for preprocessing. Missing environmental data was imputed by interpolating over the missing values using the pandas interpolate module with a linear method which assumes that the missing values are equally spaced. (https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.interpolate.html). A lag of 6 days was incorporated for the environmental parameters to account for the virus incubation period (Cheng et al., 2021; WHO, 2020). An appropriate running average of 5–7 days was applied on the daily cases time series depending on the location to account for the sharp rise and drops in cases due to various infrastructural lags in testing time and lack of testing which was observed on the weekends and removal of within week variations (Adiga et al., 2021). A split of 80% and 20% were considered for the training and test data respectively. The inputs were normalized and reshaped before passing it through the models. Models were prepared in Python using the Keras library (https://keras.io). Univariate LSTM variants considered for this study were the basic LSTM, Bidirectional LSTM (Bi-LSTM), and Stacked LSTM. A basic multivariate LSTM was applied by incorporating two environmental parameters – daily averaged T and RH. These models are briefly explained below:

**LSTM:** First introduced by Hochreiter and Schmidhuber, (1997), a Long Short-Term Memory (LSTM) network is a variant of a Recurrent Neural Network (RNN). Traditional RNN’s are capable of storing short term past information i.e. the previous time step. RNN’s are not suitable for longer term predictions as the gradients are prone to vanishing (i.e. the solution does not converge) or exploding gradient (i.e. the solution diverges). LSTM’s on the other hand are capable of retaining past information over a longer period of time, thus tackling the problem of long-term dependencies and give more accurate predictions. LSTM’s are hence well suited for forecasting of time-series (Gers et al., 2000).

LSTM uses three gates as indicated in Fig. 1 and the subsequent equations. The forget gate (\(f_t\)) is responsible for forgetting unnecessary information, while the input gate (\(i_t\)) is used for adding new or useful information. The output gate (\(o_t\)) controls the flow of the information updates the hidden states at every time step (Arora et al., 2020).

\[
\begin{align*}
    f_t &= \sigma(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f) \\
    i_t &= \sigma(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i) \\
    o_t &= \sigma(W_o \cdot [C_{t-1}, h_{t-1}, x_t] + b_o) \\
    C_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\
    C_t &= f_t \cdot C_{t-1} + i_t \cdot C_t
\end{align*}
\]

Here,

1. \(f_t\) represents the equation for the forget gate
2. \(i_t\) represents the equation for the input gate
3. \(o_t\) represents the equation for the output gate
4. \(x_t\) is the input vector
5. \(h_t\) and \(C_t\) are hidden layer vectors
The LSTM model consists of the input layer, a single hidden layer, and the dense layer.

The Stacked LSTM consists of the input layer followed by two LSTM layers and the dense layer.

The Bi-Directional LSTM consists of the input layer, the Bi-LSTM layer, and the dense layer.

The Multivariate LSTM consists of the input layer, a single hidden layer, and the dense layer.

The model parameters are summarized in Table 2.

2.3. Metrics Used

For this work, the accuracy of the above-indicated models was evaluated using the mean absolute percentage error (MAPE) calculated by the following equation:

\[
\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|
\]  

(6)

Other metrics used were the Pearson’s correlation coefficient (\(r\)), coefficient of determination \((R^2)\), and the root mean square error (RMSE).

\[
r = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{N} (y_i - \bar{y})^2}}
\]  

(7)

\[
R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}
\]  

(8)

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}
\]  

(9)

where, \(y_i\)'s represents the actual values, \(\hat{y}_i\)'s represents the predicted values from the model, \(y\) represent the mean values of the y-variable and \(x\) represent the mean values of the x-variable. All metrics were calculated in Python using the sklearn.metrics module (https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics).

3. Results & Discussion

Table 3 shows the model \(R^2\), RMSE and MAPE values for all the LSTM variants for the locations considered. For the colder regions, the MAPE is higher compared to other locations due to the relatively lower number of cases observed in this region and lack of explicit peaks. In general, the multivariate LSTM model significantly outperformed the other models, displaying an average improvement of 61 – 71% in the MAPE compared to the univariate models (Table 4). The improvement is similar to a study done by Shetty and Pai, (2021) who observed a 66% improvement in the MAPE (from 20.73% to 7.03%) after implementing a Cookoo search algorithm for better forecasting of COVID-19 cases in the state of Karnataka, India.

MAPE values for the multivariate LSTM ranged from 0.6% (Skane) to 8.9% (Vastra Gotaland). Average model error across the 8 locations is 4.1%. MAPE Results are comparable to Shastri et al., (2020), where MAPE's ranged from 2.17% to 4% and 2.00% to 10.00% for various LSTM variants in India and USA respectively. MAPE values observed by Kırbaş et al., (2020) for 8 European countries ranged from 0.16% to 2.5%. Abbasmehr and Paki, (2021) obtained MAPE values of 0.81% and 0.77% for USA and India respectively with a Beysian Optimized LSTM (Mean MAPE of 2.6% for 10
Chowdhury et al., (2021) predicted daily cases in Bangladesh using LSTM and achieved a MAPE of 4.51%. In New Delhi, Arora et al., (2020) observed MAPE of 6.17% for weekly predictions for Bi-LSTM; whereas in this work, the MAPE is 4.5% for Bi-LSTM and 3% for Multivariate LSTM, although the time period of analysis is different.

While the MAPE values in this study are higher than those observed by Kırbas et al., (2020) and Abbasimehr and Paki, (2021), this can be partly attributed to different model architectures, parameters, regions considered (countries vs cities) and significantly lesser number of daily cases observed in our study, especially in Sweden which further amplifies the errors.

R² values for the multivariate LSTM ranged from 0.925 to 0.995 which is comparable to the LSTM results from Shoaib et al., (2021) who looked into country-level daily cases. The correlations for the multivariate LSTM in this study are significantly higher than other LSTM variants likely due to ample environmental data available over the past week for predictions (since the lookback period is 7 for the models). The Stacked LSTM was on average the worst performing model, with MAPE ranging from 4.9% (New Delhi) to 25.5% (Stockholm); with an average of 14.2% across the 8 locations considered. Bi-LSTM showed the highest MAPE for 3 out of 4 locations.
in Sweden (except for Uppsala) but improved performance in India; the overall performance was similar to the Basic LSTM, with average MAPE of 10.3% (Bi-LSTM) and 10.4% (Basic LSTM).

While a majority of studies have seen better performance from LSTM variants, this was not the case in our study. Stacked LSTM and Bidirectional LSTM, in general, allow for greater model complexity and are suited for more complex input patterns. Passing univariate inputs into these LSTM variants makes the models prone to overfitting, which in turn deteriorates the model performance on the test dataset. For instance, Said et al., (2021) also observed a reduced performance of Bi-LSTM against Basic LSTM in Qatar, where they looked into multivariate time series data enriched with data related to lockdown measures. While there are several LSTM results available, a direct one-to-one comparison with this study would be fallacious due to the varying locations, model architectures and time period considered. Metrics of absolute errors (eg. RMSE, MAE, etc.) rather than relative errors such as MAPE prevents comparisons (ArunKumar et al., 2021). A comprehensive summary of COVID-19 prediction models and their results has been compiled elsewhere (Ghafouri-Fard et al., 2021).

Table 3
Summary of RMSE and MAPE values obtained for the various LSTM models.

| Cities       | Model       | R²   | MAPE (%) | RMSE   |
|--------------|-------------|------|----------|--------|
| Vastra Gotaland | Basic       | 0.716 | 16.8     | 9.828  |
|              | Stacked     | 0.526 | 20.7     | 12.385 |
|              | Bidirectional | 0.64  | 17.7     | 12.002 |
|              | Multivariate | 0.925 | 8.9      | 6.685  |
| Stockholm    | Basic       | 0.881 | 13.2     | 11.852 |
|              | Stacked     | 0.553 | 25.5     | 22.311 |
|              | Bidirectional | 0.804 | 18.9     | 14.769 |
|              | Multivariate | 0.969 | 8.7      | 7.944  |
| Skane        | Basic       | 0.811 | 6        | 3.997  |
|              | Stacked     | 0.678 | 8        | 5.217  |
|              | Bidirectional | 0.673 | 6.2     | 5.179  |
|              | Multivariate | 0.995 | 0.6     | 0.486  |
| Uppsala      | Basic       | 0.842 | 8.5      | 0.959  |
|              | Stacked     | 0.596 | 7        | 4.889  |
|              | Bidirectional | 0.931 | 8.1     | 0.632  |
|              | Multivariate | 0.993 | 2.1     | 0.175  |
| Yuma         | Basic       | 0.841 | 10.9     | 5.659  |
|              | Stacked     | 0.63  | 18.4     | 8.528  |
|              | Bidirectional | 0.859 | 9.3     | 5.325  |
|              | Multivariate | 0.99  | 3       | 0.892  |
| Los Angeles  | Basic       | 0.568 | 4.7      | 57.703 |
|              | Stacked     | 0.11  | 5.5      | 82.798 |
|              | Bidirectional | 0.336 | 6.3     | 71.487 |
|              | Multivariate | 0.978 | 0.8     | 9.325  |
| New Delhi    | Basic       | 0.885 | 4.7      | 171.525|
|              | Stacked     | 0.866 | 4.9      | 184.932|
|              | Bidirectional | 0.896 | 4.5     | 163.258|
|              | Multivariate | 0.794 | 3       | 142.112|
| Nagpur       | Basic       | 0.473 | 18.1     | 208.935|
|              | Stacked     | 0.29  | 23.3     | 326.86 |
|              | Bidirectional | 0.83  | 11.5    | 118.522|
|              | Multivariate | 0.964 | 5.4     | 71.77  |

Table 4
MAPE statistics for the various LSTM variants used in this study.

| Location         | Multivariate LSTM | Basic LSTM | Stacked LSTM | Bi-directional LSTM |
|------------------|-------------------|------------|--------------|---------------------|
| Vastra Gotaland  | 8.9               | 16.8       | 20.7         | 17.7                |
| Stockholm        | 8.7               | 13.2       | 25.5         | 18.9                |
| Skane            | 0.6               | 6          | 8            | 6.2                 |
| Uppsala          | 2.1               | 8.5        | 7            | 8.1                 |
| Yuma             | 3                 | 10.9       | 18.4         | 9.3                 |
| Los Angeles      | 0.8               | 4.7        | 5.5          | 6.3                 |
| New Delhi        | 3                 | 4.7        | 4.9          | 4.5                 |
| Nagpur           | 5.4               | 18.1       | 23.3         | 11.5                |
| Average (Std Dev)| 4.1 (3.3)         | 10.4 (5.3) | 14.2 (8.6)   | 10.3 (5.4)          |
| Average Improvement (%) of Multivariate LSTM | – | 60.8 | 71.3 | 60.6 |
Correlations of smoothed daily cases with environmental parameters after considering a lag period of 6 days.

Moreover, since the considered period captures peaks, the correlations are subject to both rising and fall of cases. From these observations, it can be inferred that there is a range of temperature which is ideal for the virus transmission and survival, with colder temperatures generally favoring virus spread and vice versa. RH showed mixed correlations with the daily cases. These correlations however can also be driven by the inherent dependence of the RH with temperature, hence making it difficult to effectively determine the relationship with the virus survival and transmission.

Fig. 4A-H shows the time series of the actual and predicted daily cases for the case of multivariate LSTM in the 8 cities considered. In general, the multivariate model performs well in predicting and following the trend of the daily cases, the model accounts for the clearly observable peaks in daily cases as observed in the warmer regions of LA, Nagpur, New Delhi, and Yuma and the cold region of VG. In Delhi, the model captures double peaks as well. For the colder regions, where there is a lack of explicit peaks, the MAPE is higher due to the relatively lower number of cases observed in this region.

4. Conclusion

This work is specifically aimed at forecasting the uninhibited spread of COVID-19, with minimal interference of other parameters so as to confirm the hypothesis of improved model prediction by integrating environmental parameters into conventional models such as LSTM. The datasets considered ranged from 5 to 7 months which included a period of uncertainty when researchers were still characterizing the virus transmission and survival. In conclusion, the research presents the improved potential for deep learning models incorporated with environmental parameters as inputs for better and improved prediction of the daily COVID-19 cases in the selected locations, consisting of 8 cities across the globe with varying climatic zones. The multivariate LSTM model significantly outperformed the other univariate models. The proposed T and RH integrated multivariate LSTM model can help the decision-makers and the authorities to effectively manage lockdown measures, resources and available infrastructure (Das et al., 2021; Lemaitre et al., 2021; Tomar and Gupta, 2020).

5. Limitations and future work

It is important to note that the work done here can be further improved. While additional data was readily available, their inclusion would have been subject to various additional influences such as lockdowns, festivals and social distancing parameters, which would have potentially introduced bias in the models. The training set data is still small in the context of deep learning, hence limiting this work to a simple model architecture; with more data, the predictive power of LSTM’s is expected to increase, which would further enable the development of deeper and complex models and their optimization. The model does not account for many other factors such as demographic, socio-economic, political, infrastructural, asymptomatic individuals, pollution levels, lockdown status, vaccination status, and many other non-linear factors which can significantly affect the transmission of the disease. Additional information on the characterization of the virus and virus-laden particles (bioaerosols) and its influence on the local microenvironment may also provide additional insights on the pandemic (Gollakota et al., 2021). It is expected that the inclusion of such parameters would further improve the predictive power of these forecasting models.

Nevertheless, the universal availability of city-level weather information, including weather forecasts, enables quick and easy integration of these parameters into forecasting models. It is also expected that the other more complex univariate LSTM variants used in this study (stacked and bidirectional LSTM’s) would further improve upon the integration of environmental parameters and additional training data. Additionally, there are ways to improve the accuracy of these forecasting models such as data augmentation, generative adversarial networks and transfer learning by using some of the earlier epidemiological models as a pre-trained network.

The outcome of this work suggests that the inclusion of daily averaged environmental parameters could significantly improve the prediction capability of deep learning forecasting model for COVID-19. Hence, it is recommended to integrate publicly available weather data (historical and forecast) for enhanced accuracy in the forecasting of city-level COVID-19 cases, although other positive and negative confounding factors can affect the forecasting power.

CRediT authorship contribution statement

Roshan Wathore: Investigation, Method, Resources, Software, Visualization, Writing – original draft. Samyak Rawlekar: Resources, Software, Writing – original draft. Saima Anjum: Investigation, Resources, Software. Ankit Gupta: Formal Analysis, Visualization, Validation. Hemant Bherwani: Conceptualization, Formal Analysis, Project Administration, Resources, Supervision, Writing – original draft, Writing – review & editing. Nitin Labhasetwar: Supervision, Validation, Writing – review & editing. Rakesh Kumar: Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Table 5

| Cities        | Correlation with Temperature | Correlation with RH | Temperature (Avg, Std) | RH (Avg, Std) |
|---------------|-----------------------------|---------------------|------------------------|--------------|
| Vastra Gotaland | 0.561                       | -0.287              | 10.5, 5.9              | 72.4, 13.6   |
| Stockholm     | 0.435                       | -0.377              | 11.1, 6.3              | 66.8, 13.5   |
| Skane         | 0.572                       | 0.303               | 12.5                   | 72.5, 10.6   |
| Uppsala       | 0.006                       | -0.207              | 11.2, 6.4              | 68.2, 13.2   |
| Yuma          | 0.200                       | -0.248              | 32.5, 3.3              | 26.3, 9.0    |
| Los Angeles   | -0.157                      | 0.293               | 23.0, 3.6              | 53.8, 15.3   |
| New Delhi     | -0.174                      | -0.014              | 31.2, 2.5              | 69.2, 13.7   |
| Nagpur        | -0.632                      | -0.004              | 28.3, 3.0              | 53.4, 15.0   |
High Temperature and High Humidity Reduce the Transmission of COVID-19. SSRN Electron. J. https://doi.org/10.2139/ssrn.3551767.

Wathore, R., Gupta, A., Bherwani, H., Labhasetwar, N., 2020. Understanding air and water borne transmission and survival of coronavirus: Insights and way forward for SARS-CoV-2. Sci. Total Environ. 749. https://doi.org/10.1016/j.scitotenv.2020.141486.141486.

WHO, 2020. Transmission of SARS-CoV-2: implications for infection prevention precautions [WWW Document]. URL https://www.who.int/news-room/commentaries/detail/transmission-of-sars-cov-2-implications-for-infection-prevention-precautions (accessed 3.6.21).

Wu, Y., Jing, W., Liu, J., Ma, Q., Yuan, J., Wang, Y., Du, M., Liu, M., 2020. Effects of temperature and humidity on the daily new cases and new deaths of COVID-19 in 166 countries. Sci. Total Environ. 729,. https://doi.org/10.1016/j.scitotenv.2020.139051 139051.

Xu, C., Yu, Y., Chen, Y., Lu, Z., 2020. Forecast analysis of the epidemics trend of COVID-19 in the USA by a generalized fractional-order SEIR model. Nonlinear Dyn. 101, 1–14. https://doi.org/10.1007/s11071-020-05946-3.

Zeroual, A., Harrou, F., Dairi, A., Sun, Y., 2020. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. Chaos Solitons Fractals 140,. https://doi.org/10.1016/j.chaos.2020.110121 110121.