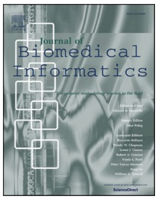
Journal of Biomedical Informatics 132 (2022) 104132

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin



A deep learning approach for Spatio-Temporal forecasting of new cases and new hospital admissions of COVID-19 spread in Reggio Emilia, Northern Italy

Veronica Sciannameo[a,1], Alessia Goffi[b,1], Giuseppe Maffeis[b], Roberta Gianfreda[b], Daniele Jahier Pagliari[c], Tommaso Filippini[d], Pamela Mancuso[e], Paolo Giorgi-Rossi[e], Leonardo Alberto Dal Zovo[f], Angela Corbari[f], Marco Vinceti[d,h], Paola Berchialla[g,*]

[a] *University of Padova, Unit of Biostatistics, Epidemiology and Public Health, Department of Cardiac, Thoracic, Vascular Sciences and Public Health, Italy*
[b] *TerrAria s.r.l, Via Melchiorre Gioia, 132, 20125 Milan, Italy*
[c] *Polytechnic of Turin, Department of Control and Computer Engineering, Italy*
[d] *Environmental, Genetic and Nutritional Epidemiology Research Center (CREAGEN), Section of Public Health, Department of Biomedical, Metabolic and Neural Sciences, University of Modena and Reggio Emilia, Via Campi 287, 41125 Modena, Italy*
[e] *Epidemiology Unit, Azienda Unità Sanitaria Locale-Istituto di Ricovero e Cura a Carattere Scientifico di Reggio Emilia, 42123 Reggio Emilia, Italy*
[f] *Studiomapp s.r.l., Via Pietro Alighieri, 43, 48121 Ravenna, Italy*
[g] *University of Torino, Department of Clinical and Biological Sciences, Italy*
[h] *Department of Epidemiology, Boston University School of Public Health, Boston, Massachusetts, USA*

ARTICLE INFO

Keywords:
COVID-19
Forecasting
ConvLSTM
Deep learning
Spatio-temporal
SARS-CoV-2

ABSTRACT

Background: Since February 2020, the COVID-19 epidemic has rapidly spread throughout Italy. Some studies showed an association of environmental factors, such as PM_{10}, $PM_{2.5,}$ NO_2, temperature, relative humidity, wind speed, solar radiation and mobility with the spread of the epidemic.

In this work, we aimed to predict via Deep Learning the real-time transmission of SARS-CoV-2 in the province of Reggio Emilia, Northern Italy, in a grid with a small resolution (12 km × 12 km), including satellite information.

Methods: We focused on the Province of Reggio Emilia, which was severely hit by the first wave of the epidemic. The outcomes included new SARS-CoV-2 infections and COVID-19 hospital admissions. Pollution, meteorological and mobility data were analyzed. The spatial simulation domain included the Province of Reggio Emilia in a grid of 40 cells of $(12\ \text{km})^2$. We implemented a ConvLSTM, ~~which is~~ a spatio-temporal deep learning approach, to perform a 7-day moving average to forecast the 7th day after. We used as training and validation set the new daily infections and hospital admissions from August 2020 to March 2021. Finally, we assessed the models in terms of Mean Absolute Error (MAE) compared with Mean Observed Value (MOV) and Root Mean Squared Error (RMSE) on data from April to September 2021. We tested the performance of different combinations of input variables to find the best forecast model.

Findings: Daily new cases of infection, mobility and wind speed resulted in being strongly predictive of new COVID-19 hospital admissions (MAE = 2.72 in the Province of Reggio Emilia; MAE = 0.62 in Reggio Emilia city), whereas daily new cases, mobility, solar radiation and $PM_{2.5}$ turned out to be the best predictors to forecast new infections, with appropriate time lags.

Interpretation: ConvLSTM achieved good performances in forecasting new SARS-CoV-2 infections and new COVID-19 hospital admissions. The spatio-temporal representation allows borrowing strength from data neighboring to forecast at the level of the square cell $(12\ \text{km})^2$, getting accurate predictions also at the county level, which is paramount to help optimise the real-time allocation of health care resources during an epidemic emergency.

[*] Corresponding author at: Centre for Biostatistics, Epidemiology and Public Health, Department of Clinical and Biological Sciences, University of Torino, Regione Gonzole 10, 10043 Orbassano (Turin), Italy.

E-mail address: paola.berchialla@unito.it (P. Berchialla).

[1] These authors equally contributed to the work.

https://doi.org/10.1016/j.jbi.2022.104132
Received 9 December 2021; Received in revised form 24 April 2022; Accepted 3 July 2022
Available online 11 July 2022

1. Introduction

Since the first case of SARS-CoV-2 infection, modelling the epidemic growth pattern has been considered crucial to understanding the pandemic's evolution and guiding the implementation of prevention and control measures [1].

The first case of SARS-CoV-2 infection in Italy was diagnosed in February 2020. Soon, the COVID-19 epidemic spread across Northern regions, including Lombardy, Veneto and Emilia-Romagna. In March, a lockdown, which limited individual mobility, and social distancing were imposed to slow down the diffusion of the SARS-CoV-2 virus. Based on mobility data, a few studies showed the effectiveness of the lockdown timing in reducing the daily number of infected people [2,3].

Soon after the epidemic's beginning, it was observed that the COVID-19 outbreaks were predominantly in the Northern Hemisphere wintertime, suggesting a potential transmission mechanism associated with cold temperature and stable humidity conditions [4].

Several studies have attempted to establish statistical relationships between meteorological variables and COVID-19 cases. Some found an association between a decreasing number of cases and an increase in temperature, humidity and solar radiation [5–7], although the overall variability of the virus spread could not be entirely explained by these factors [8].

Similarly, the observation that the pandemic had faster and wider spread in the most polluted areas contributed to formulating the hypothesis of positive interaction between pollution and pandemic spread. Many recent studies have linked air pollution to the increasing spread of COVID-19 [9–11]. Several have focused on Northern Italy, being the region with the first large outbreak in Europe and, at the same time, having high atmospheric pollution levels [12,13]. However, the statistical approaches applied in those studies cannot fully encompass the complexity of the epidemic dynamics, and recent literature [14] suggested that several factors besides climatic conditions and air pollution may play a pivotal role in the transmission of SARS-CoV-2.

In this study, we implemented a Convolutional Long-Short Term Memory (ConvLSTM) model [15], which is a Deep Learning (DL) algorithm, to predict the real-time transmission of SARS-CoV-2, building on the work of Paul et al. [16], which is among few studies that considered the spatial correlation in the diffusion of COVID-19.

DL algorithms are based on artificial neural networks (NN) with stacked layers composed of multiple neurons, which can learn increasingly complex data representations [17]. Kafieh et al. [18] conducted a comparative study concluding that LSTM models are the most promising DL approaches to forecast the SARS-CoV-2 epidemic.

Our work aimed to predict the real-time transmission of SARS-CoV-2 in the province of Reggio Emilia in the Emilia-Romagna region, Northern Italy. Paul et al. [16] divided the map of Italy into relatively large grids and trained a ConvLSTM model with samples drawn from local distribution to predict new cases in an auto-regressive way. Differently from the work of Paul et al. [16], we dealt with the issue of relatively small grids, meaning sparse local distribution for training the model, and we also included meteorological, pollution and mobility data in the spatio-temporal model. Our model was then implemented into the web tool EPICO19 (EPIdemiological and logistic COvid19 model, https://www.epico19.eu/en/), which is used as support for public health professionals and decision-makers in managing outbreaks and assessing public health interventions.

2. Material and methods

2.1. Healthcare outcomes

The Local Health Authority (AUSL) of Reggio Emilia provided the number of newly diagnosed infections with SARS-CoV-2, corresponding to the number of new positive tests based on quantitative reverse transcription-polymerase chain reaction (RT-PCR), and the number of COVID-19 hospital admissions occurred in the province of Reggio Emilia (532,000 inhabitants, located in Northern Italy), from February 1, 2020.

Policies for testing, contact tracing and the diagnostic capacity changed from the first to the second wave [19]. In fact, in the first wave (i.e., from February 2020 until about July 31, 2020), tests were almost exclusively performed on suspect cases arriving at the emergency room and, in rare cases at home, only if severe symptoms were declared due to a lack of large availability of SARS-CoV-2 swab tests. In contrast, from the second wave (i.e., approximately from August 2020) onwards, availability of SARS-CoV-2 swab tests greatly increased, and tests were performed in outpatient dedicated facilities, the so-called "drive-through", in people referred to testing from contact tracing, even if asymptomatic or with mild symptoms [19]. This could suggest that the number of SARS-CoV-2 infections was underestimated during the first wave. Moreover, the exponential increase in SARS-CoV-2 infections in March and April 2020 determined an unprecedented demand for bed occupancy, which exceeded the existing capacities in several hospitals. This allowed for better planning of hospital beds needed for future waves [20,21]. Due to such differences between the first months of the epidemic and what happened afterwards, we decided to not use the data of the first wave and thus consider data from August 1, 2020, until September 20, 2021, the last available data at the time of writing.

2.2. Exposure input variable

Very-High Resolution (VHR) satellite and aerial images, with a ground resolution of 30–50 cm and 11 cm, respectively, were used to estimate the crowding index as a mobility indicator. We calculated the crowding index for each cell as the ratio between the daily density of light vehicles and the relative baseline value referred to in the pre-COVID-19 period. An Artificial Intelligence vision proprietary algorithm developed by Studiomapp has analyzed the VHR images to count the number of light vehicles, assessing the presence of people in the surroundings of hospitals, parking lots, supermarkets, working places, train stations, and logistics hubs (more details are available at STUDIOMAPP website [22].

Pollution data (PM_{10}, $PM_{2.5}$, NO_2 air concentration) were collected from the Urban Tool for Air Quality (UTAQ) (www.utaq.eu) developed by TerrAria s.r.l [23], which uses background concentrations data provided by the Copernicus Atmosphere Monitoring Service (CAMS) [24] along with local emissions and air quality measurements from ARPA Emilia-Romagna stations, for high-resolution pollutants concentration forecasting [20].

Meteorological data (air temperature, relative humidity, wind speed and solar radiation) were extrapolated from the COnsortium for Small-scale MOdeling (COSMO-5 M) model [25], available on the OpenData portal of the Regional Agency for the Protection of the Environment (ARPA) of Emilia-Romagna [26].

All the input data were collected daily from August 1, 2020, until September 20, 2021, i.e., for the entire observation period of the study.

2.3. Deep learning model

A ConvLSTM model consists of a Convolutional Neural Network (CNN) merged with an LSTM. CNNs are feed-forward NNs that combine convolution, Rectified Linear Unit (ReLU), pooling and fully connected layers to deal with the spatial correlations in data. The principle of CNNs is to exploit local connectivity among neurons to model spatial relations. Each neuron only processes inputs belonging to a spatially bounded area called receptive field and convolves them with weights called kernels or filters. Neurons in a convolutional layer are organized in a matrix called activation map or feature map. All neurons share the same kernels, which are applied to a differently centered receptive field, where the center position is slid across the rows and columns of the input matrix. Typically, multiple kernels are used in each layer to form multiple feature maps [27].

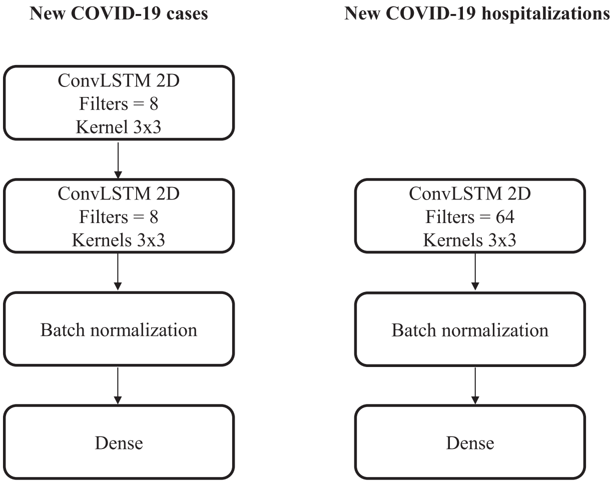


Fig. 1. ConvLSTM architecture for new COVID-19 cases forecasting (left) and for new COVID-19 hospitalization forecasting (right).

LSTM networks are instead DL models designed to deal with data sequences instead of single data points and therefore suited for processing time-series [28]. They are a particular type of Recurrent NN (RNN), able to handle long-term dependencies. The building block of a standard LSTM network is called LSTM unit or cell. Each unit comprises a memory part and three gates: an input gate, an output gate and a forget gate. An LSTM unit can remember values over arbitrary time intervals, and the three gates control the flow of information through the cell [27].

Using both CNN and LSTM network structures, ConvLSTMs can learn the dynamics of epidemic spread with high spatial resolution and a high degree of accuracy, thanks to their capability of building highly nonlinear representations [16]. In particular, ConvLSTMs are based on a recurrent layer, just like LSTMs, but where internal matrix multiplications are exchanged with convolution operations, like in CNNs [15]. The model can process a sequence of images, one slice at a time, similarly to how an LSTM goes through a series of data points one at a time.

Several experiments were conducted with different ConvLSTM structures and parameters to find the optimal model that can forecast a minimum validation error. More in detail, considering a certain redundant component present among the variables considered (such as temperature and solar radiation, PM_{10} and $PM_{2.5}$, NO_2 and mobility), we evaluated the correlation of the single variables with respect to the outcomes. We empirically tested the performance of the different models, individually adding input variables and varying the time lag and the structure of the model iteratively. The evaluation of the performance of the various tests determined the choice of the model in all its components.

We selected a DL model with two ConvLSTM layers and a final dense layer to predict the newly infected cases at the end of this process. We used eight kernels of size 3x3 in each ConvLSTM layer. We applied the dense layer, with a linear activation function, to the second ConvLSTM layer’s output produced at the last sequence element. We used a Mean Squared Error (MSE) loss function, optimizing it with the stochastic gradient descent (SGD) algorithm. We set the batch size to 10 and trained the model for 30 epochs, which means that the algorithm saw all the data 30 times.

For the prediction of new hospital admissions, we used a shallower model due to the lower daily numbers of subjects that were hospitalized due to COVID-19. In particular, the simpler model had one ConvLSTM layer and one dense layer with a linear activation function, and ConvLSTM layers included 64 filters of size 3x3. Also, in this case, we used the MSE loss function and the SGD optimizer; we set the batch size to 10 and trained for 30 epochs. The architectures of both ConvLSTM models (new COVID-19 cases and new COVID-19 hospitalizations) are reported in Fig. 1.

We evaluated the performance of the models in terms of Mean Absolute Error (MAE) compared with Mean Observed Value (MOV) and Root Mean Squared Error (RMSE).

All analyses were performed in R [29], taking advantage of Keras [30], a high-level NN API developed with a focus on enabling fast experimentation. The Keras R interface uses the TensorFlow [31] backend engine by default.

2.4. Data pre-processing

The spatial domain is a grid including the whole Province of Reggio Emilia with 40 square cells (8 rows × 5 columns), each with an extension

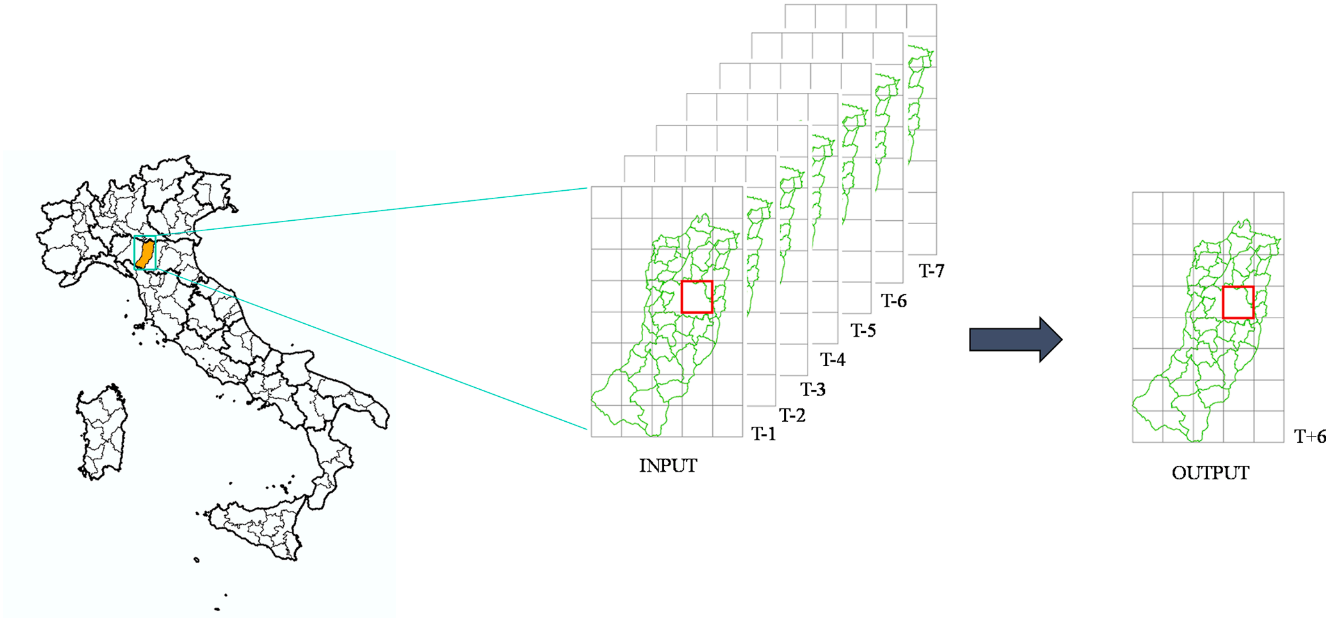


Fig. 2. On the left side, the Province of Reggio Emilia in Italy. On the right side, the area of the Reggio Emilia Province, divided in 40 cells. The red one is the cell that includes the Reggio Emilia city, i.e., the most populated area. Spatio-temporal representation of input and output data of ConvLSTM model (T-1 stands for yesterday, T-2 for two days ago, etc.).

Table 1
Descriptive values of environmental factors and healthcare outcomes in the training, validation and test periods. Mean, minimum and maximum values of the mean concentrations of environmental factors on the whole grid are reported.

	Training set (August 1, 2020 – January 31, 2021)	Validation set (February 1, 2021 – March 31, 2021)	Test set (April 1, 2021 – September 20, 2021)
New COVID-19 cases	22,895	11,064	9,386
New COVID-19 hospitalizations	1,680	826	604
Environmental factors, mean (range)			
NO_2 (μg/m^3)	17.5 (0.0, 220.3)	21.7 (0.0, 158.4)	9.6 (0.7, 150.2)
PM_{10} (μg/m^3)	23.2 (0.0, 104.7)	28.7 (1.7, 96.8)	13.4 (2.1, 52.9)
$PM_{2.5}$ (μg/m^3)	16.3 (0.0, 89)	17.7 (0.0, 80.8)	7.7 (0.8, 28.3)
Temperature (°C)	11.6 (−7.9, 34.2)	6.9 (−10.5, 17.7)	20.1 (−3.3, 34.0)
Solar radiation (W/m^2)	105.6 (2.1, 293.8)	134.4 (3.8, 226.3)	235.7 (7.2, 343)
Relative humidity (%)	72.3 (29.3, 100.0)	68.1 (31.7, 99.9)	54.7 (25.6, 99.8)
Wind speed (m/s)	1.8 (0.4, 11.3)	2.0 (0.4, 7.7)	2.4 (0.6, 7.6)
Crowding index (%)	68.2 (3.2, 98.7)	65.2 (6.3, 92.5)	79.3 (9.1, 98.9)

of 12 km × 12 km (Fig. 2). All the daily data (cases, hospital admissions, meteorological, air pollution, mobility) were mapped to this grid. Separate grid matrices were built for each variable considered, and they were concatenated as channels on a third axis.

Data from August 1, 2020, to January 31, 2021, were used as the training set, data from February 1, 2021, until March 31, 2021, as the validation set and data from April 1, 2021, until September 20, 2021, as the test set.

For each considered variable, mean and standard deviation (SD) were computed in the training test, and such values were used to standardize the respective variable both in training and in testing sets. The standardization was performed by subtracting the mean values and dividing by the SD.

Moreover, all model input and output data are considered as ‘7 days’ average to smooth the daily variation due to the daily variance of case detections and hospital admissions during the weekdays and the weekends.

Samples were obtained through a sliding window of 7 days, with a 1-day stride. Case and hospital admission models used a data window of 7 days for each input variable (e.g., from t-7 to t-1) to forecast the target variable at t + 6 (Fig. 2). Hence, the t + 6 forecast represents the weekly average from t0 to t + 6 of daily cases and hospital admissions, respectively.

ConvLSTM forecasted, separately for each cell in the grid, the average number of new infections (i.e., cases with RT-PCR positive SARS-CoV-2 test) and hospital admissions due to COVID-19 (i.e., a hospital admission occurring in a confirmed case from 2 days before diagnosis up to 21 after). Summing up these values, we obtained the total new infected and the total number of hospitalized COVID-19 cases at the province level.

An anchored walk-forward approach has been implemented in the test set to mimic the realistic setting in which the web tool EPICO19 is used. In this approach, the predictions are on a time horizon of one week, and once a week is over, the actual values are added to the training set before the model is trained again, and the forecasts for the next week are made.

More in detail, each Monday, data are updated, and each week a new training with all the available data from August 1, 2020, is performed to forecast the following week. In this way, all the available information improves the performance.

3. Results

In the entire Province of Reggio Emilia, the new COVID-19 cases in the train, validation and test sets were 22,895, 11,064 and 9,386, respectively; meanwhile, the new COVID-19 hospital admissions were respectively 1,680, 826 and 604 (Table 1). The mean, minimum and maximum of the mean daily values of meteorological, pollution and mobility indicators in the three periods are reported in Table 1.

The trained model allowed us to predict the outcomes of the upcoming week based on the data we collected in the previous 7 days, i.e., feeding the model with the data of the last week, up to Sunday, on Monday, the model forecasts daily outcomes up to next Sunday.

Daily new cases, mobility index and meteorological variables with different lag times were tested. We conducted ablation studies on the validation set to select the best combination of variables and optimize lag times to achieve the best model accuracy [32].

The validated models allowed us to predict the number of new COVID-19 cases/hospital admissions in the upcoming week. The forecasting model of new COVID-19 cases included the daily new cases collected in the previous 7 days, the crowding index with a lag time of 14 days, the solar radiation (lag time = 28 days) and the $PM_{2.5}$ air concentration (lag time = 14 days). The forecasting model of new

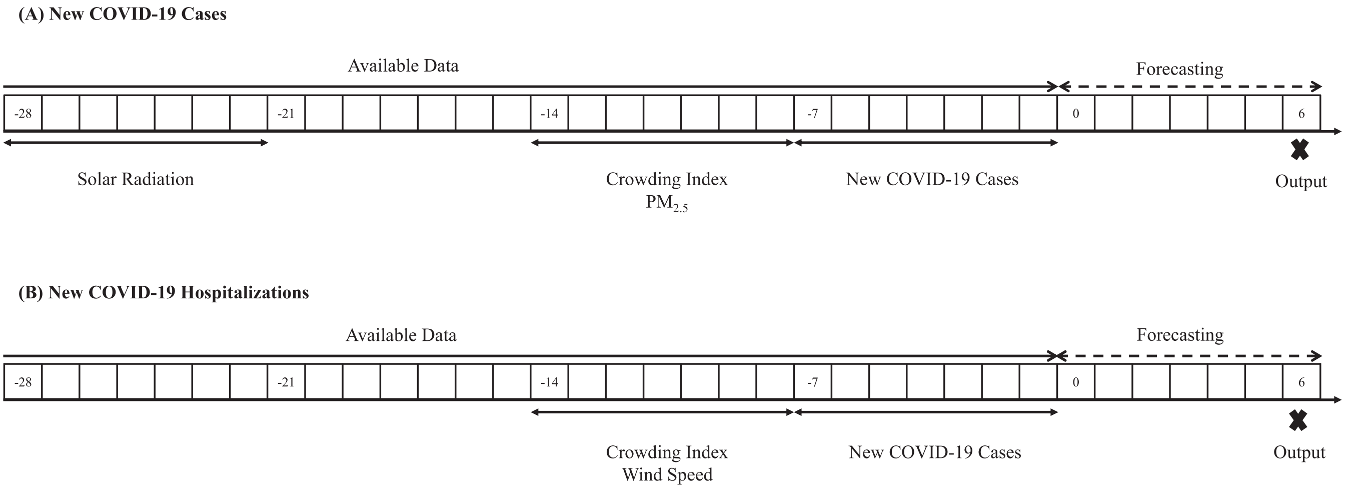


Fig. 3. Extraction from the data of temporal intervals of observations and predictions.

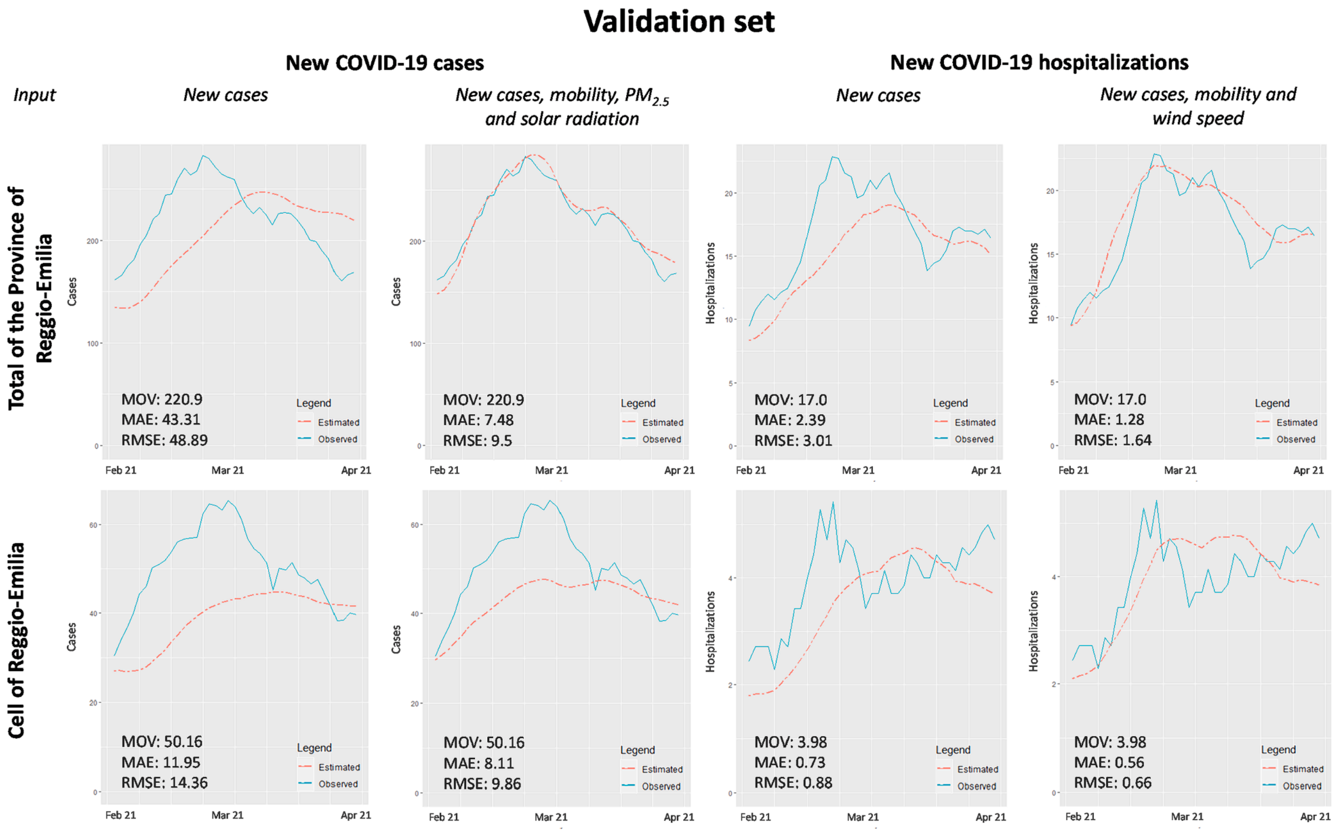


Fig. 4. Observed and predicted distributions in the validation set (February 1, 2021, until March 31, 2021) of new COVID-19 cases and new COVID-19 hospital admissions, in the total Province of Reggio Emilia and in the cell of Reggio Emilia. MOV = Mean Observed Value, MAE = Mean Absolute Error, RMSE = Root Mean Squared Error.

COVID-19 hospital admissions included the daily new cases collected in the previous 7 days, the crowding index (lag times = 14 days) and wind speed (lag times = 14 days), as shown in Fig. 3.

To assess the contribution of meteorological and mobility variables in the forecasting model of COVID-19 new cases and new hospital admissions, we have reported in Fig. 4 MOV, MAE and RMSE computed on the validation set for the final selected model and a simpler model with only the new cases collected in the previous 7 days (without meteorological and mobility variables).

When considering the prediction of new COVID-19 cases in the entire Province, the MAE decreased from 43.31 to 7.48, corresponding to an 83% reduction. We observed a similar result for RMSE (Fig. 4).

When considering the prediction of new hospital admissions, the reduction of the validation error is lower, going from 2.39 to 1.28 for MAE and 3.01 to 1.64 for RMSE (Fig. 4). Looking at the cell of Reggio Emilia only (red square in Fig. 2), the error reduction between the model with meteorological and pollution variables, crowding index and those with only new infections in the previous week is lower.

In Fig. 4 and Fig. 5, the number of predicted and observed cases of infections/hospital admissions over the validation and test periods are superimposed at the Province level and the cell of Reggio Emilia only. Models that include meteorological and pollution variables and crowding index have a smaller gap between the predicted and observed values than a simpler model with only new infections in the previous week as input data.

Finally, we reported the distributions of absolute errors and the observed values on the test set (Table 2). A mean observed value (MOV) of 52.35 new cases with a mean absolute error (MAE) of 22.27 was observed (Province level). In the Reggio Emilia cell, there was a MOV of 13.79 with an MAE of 5.78. Looking at the forecasting of hospital admissions over the province, there was a MOV of 3.31 with a MAE of 2.72. In the cell of Reggio Emilia, a MOV of 1.2 with an MAE of 0.62 were computed. Furthermore, in Table 2, more details about the distribution of the observed values and the absolute errors on the test set are reported in terms of mean, SD, median, first and third quartile (IQR), minimum (Min) and maximum (Max).

To illustrate the ConvLSTM performance on the whole grid in the test set, we report in Fig. 6 the MOV and MAE separately for each cell, showing that the models reach good performance, both in forecasting new cases and new hospital admissions. MOVs and MAEs for each grid in the validation period are reported in Supplementary Fig. 1.

4. Discussion

In an emergency setting, accuracy in the short-term forecasting of the number of new COVID-19 cases and hospital admissions is of fundamental importance to optimize the resources, i.e., shifting hospital wards from COVID-19 to non-COVID-19 dedicated. This study used a DL approach based on ConvLSTM to forecast new COVID-19 cases and new hospital admissions.

Other ML/DL methods have been proposed to address spatio-temporal forecasting. Among them, Wen and colleagues [33] compared several DL methods for air pollution prediction (a non-COVID-19 context), and they found out that combining CNN and LSTM was the most promising approach when dealing with spatio-temporal forecasting in comparison with simple LSTM, ARMA, support vector machine or logistic regression. Yelsilkanat [34] forecasted the daily number of COVID-19 cases in a spatio-temporal setting using Random Forest (RF). However, this work made the spatial forecasting at the country level in this work, without considering adjacencies on a grid map. Indeed, to our knowledge, a RF cannot deal with data adjacencies on cells in a grid as CNN does.

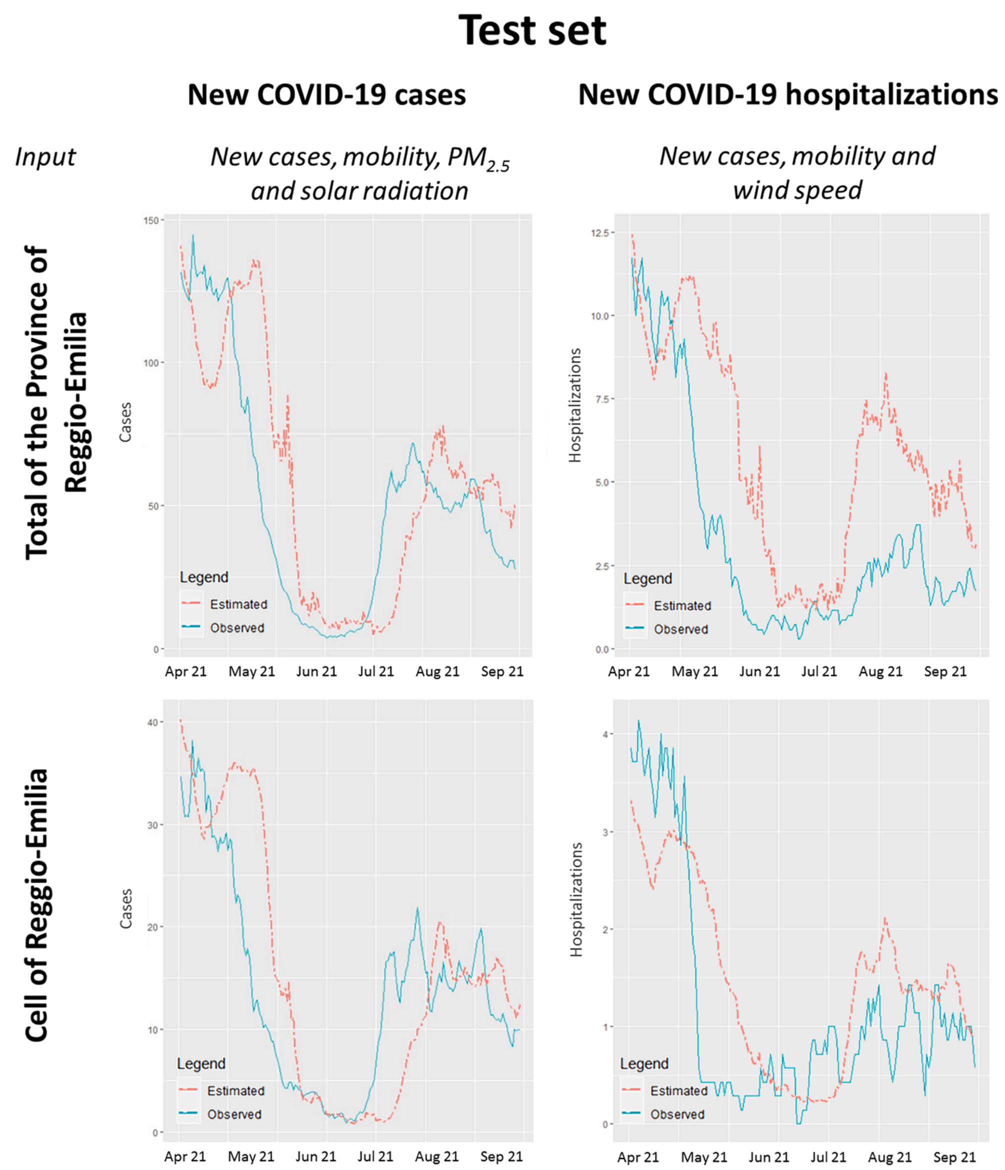


Fig. 5. Observed and predicted distributions in the test set (from April 1, 2021, until September 20, 2021) of new COVID-19 cases and new COVID-19 hospital admissions, in the total Province of Reggio Emilia and in the cell of Reggio Emilia.

Furthermore, our aim was not only to forecast at a very fine-grain spatial resolution, but also to deal with a set of covariates, developing a single architecture for the implementation in a web tool.

Given these considerations, ConvLSTM was selected for this work since it meets all the requirements mentioned above and provides acceptable errors.

Our ConvLSTM models achieved good performances, particularly at the level of the entire Province and allowed us to manage the spatio-temporal correlation of data. Indeed, the forecasting in an area likely depends on past data (temporal dimension) and data from neighborhoods due to mobility from/to an area, and meteorological or environmental conditions, spanning regions.

We observed in our study that, even at a local level (single grid cells), the predictions are good (Fig. 6), with a maximum MAE of 5.78 cases compared to a MOV of 13.79 cases in the Reggio Emilia cell.

However, the model tended to overestimate the number of new COVID-19 hospital admissions. A possible explanation could be the effect of vaccination, which began in January 2021. In September 2021, 74% of the population was fully vaccinated. In other words, we tested a model trained during the period August 2020-January 2021, when the vaccination campaign had not yet started, and tested it in a period with more than 70% of people immunized. Moreover, in Italy, in September 2021, the “green pass” was introduced to allow access to public indoor places only to individuals with vaccination, negative test, or evidence of recent recovery from infection. Such data, which affect the epidemic trend, are not available for the current study, and this is a limitation that leads to overestimation, particularly in the forecasting of new COVID-19 hospitalizations.

The ablation studies performed showed that mobility information, solar radiation, $PM_{2.5}$ and wind speed are helpful in improving the performance in the forecasting of new COVID-19 infections/hospital admissions, confirming results already found in the literature [2,35,36].

In addition, other variables, such as temperature, relative humidity, PM_{10}, NO_2, did not improve the forecasting, probably, due to high correlations between humidity and other meteorological variables (temperature, relative humidity, wind speed and solar radiation), and between PM_{10}, $PM_{2.5}$ and NO_2 and mobility. The inclusion of all these highly correlated variables can produce noise instead of improving the forecasting. Furthermore, we cannot exclude that the observed associations with meteorological and environmental variables and the number

Table 2
Results of predictions of the 7th day after of new infected cases and the 7th day after of new hospital admissions, in the testing periods (from April 1, 2021, until September 20, 2021). RMSE = Root Mean Squared Error, Min = Minimum, Max = Maximum, SD = Standard Deviation, IQR = Inter-Quantile Range.

	Test set **(From April 1, 2021, until September 20, 2021)**			
	New COVID-19 cases *New cases, PM$_{2.5}$, mobility, solar radiation*		**New COVID-19 hospitalizations** *New cases, mobility and wind speed*	
Province of Reggio Emilia (whole map)				
	Observed Values	*Absolute Errors*	*Observed Values*	*Absolute Errors*
Mean (SD)	52.35 (40.14)	22.27 (18.74)	3.31 (3.21)	2.72 (1.80)
Median (IQR)	49.14 (18.43–65.57)	17.85 (6.52–31.92)	2.14 (1.00–3.71)	2.76 (1.11–4.15)
Min-Max	3.57–144.71	0.41–79.17	0.29–11.71	0.00–6.51
RMSE		29.07		3.23
Reggio Emilia Cell				
	Observed Values	*Absolute Errors*	*Observed Values*	*Absolute Errors*
Mean (SD)	13.79 (9.65)	5.78 (5.69)	1.2 (1.14)	0.62 (0.48)
Median (IQR)	12.86 (4.71–18.14)	4.10 (0.99–8.42)	0.86 (0.43–1.29)	0.51 (0.27–0.87)
Min-Max	0.86–38.14	0.00–23.8	0.00–4.14	0.01–2.08
RMSE		8.09		0.78

of new cases are due to a causal link, but simply due to confounding being associated with other variables that are the actual causative factors [37].

Our study presents several limitations. The first one is intrinsic to the DL approach, which is challenging to apply in the initial stages of epidemics because it needs enough data for the training process. Furthermore, DL algorithms focus mainly on making predictions at the expense of interpretability. For example, it is challenging to figure out how solar

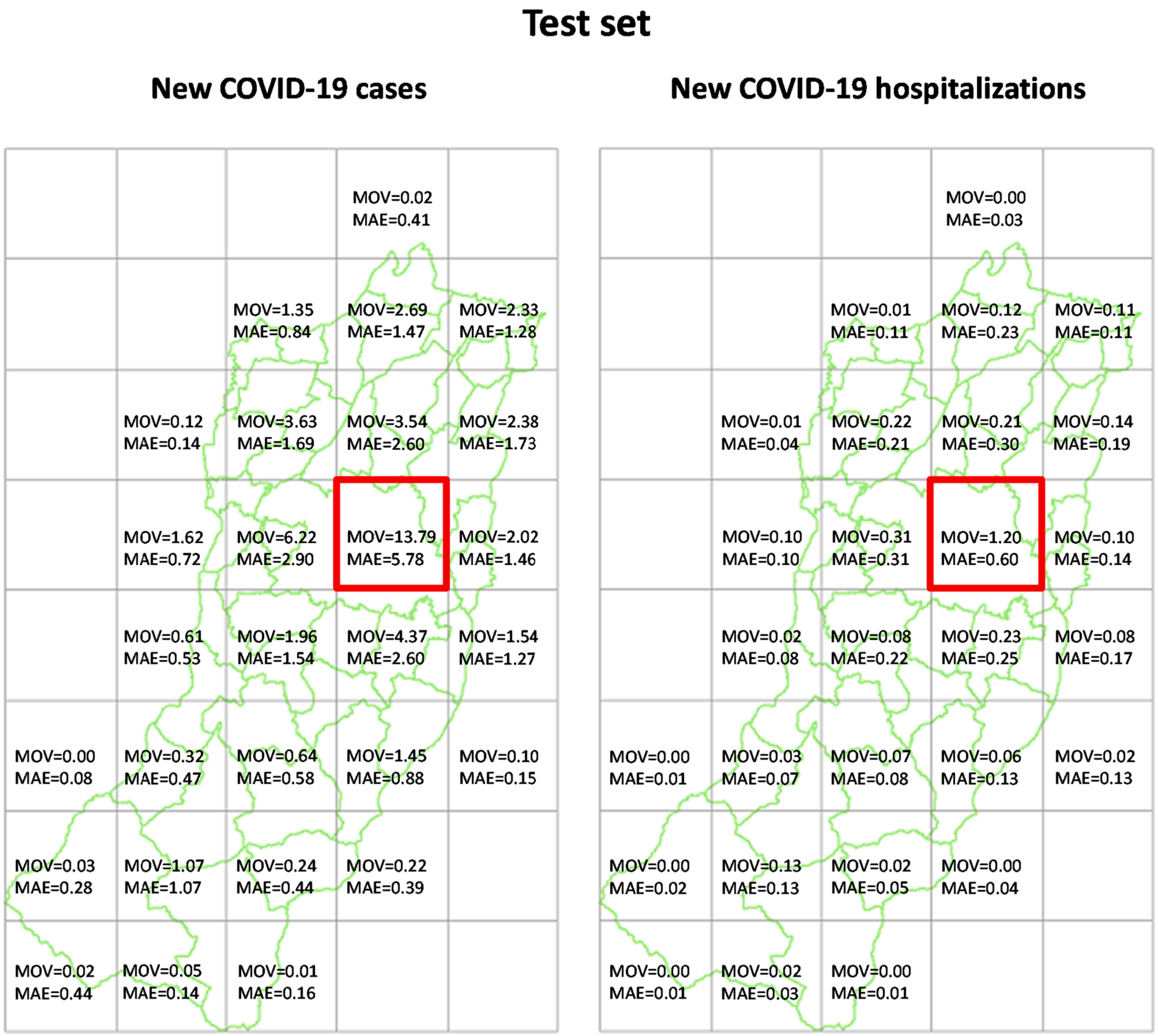


Fig. 6. Mean Observed values (MOV) and Mean Absolute Errors (MAE) in the test period (April 1, 2021, until September 20, 2021) for each cell in the grid. The red cell is the one containing most of Reggio Emilia city, the most populated cell.

radiation with lag times of 28 days behaves in the forecasting mechanism. DL methods are “black-boxes”, in which understanding the mechanisms used to forecast and evaluate the contribution of single input variables is still an open challenge.

5. Conclusions

In conclusion, we showed that ConvLSTMs, a model embedded into the web application EPICO19 (EPIdemiological and logistic COvid19 model, https://www.epico19.eu/en/), might perform well in forecasting new cases and hospital admissions due to COVID-19, taking advantage of a spatio-temporal DL representation. Information about mobility, meteorology and air pollution can be mapped to a fine-resolution spatial grid. This approach also allows for accurate predictions at a local level (small areas with a limited extension of 12 km × 12 km). The capability of the ConvLSTM to forecast at a local level could be helpful in optimizing the real-time allocation of health resources to support public health professionals and decision-makers in managing outbreaks and assessing public health interventions.

Funding

This study was funded by the European Spatial Agency (ESA, ESA Contract No. 4000131356/20/NL/AF) with the support of the Ministry for technological innovation and digitization and the Italian Space Agency.

7. Ethics committee

Area Vasta Emilia Nord Ethics Committee, approval no. 2020/0135719.

CRediT authorship contribution statement

Veronica Sciannameo: Methodology, Formal analysis, Writing – original draft. **Alessia Goffi:** Formal analysis, Software, Writing – original draft. **Giuseppe Maffeis:** Conceptualization, Project administration, Software. **Roberta Gianfreda:** Project administration, Software. **Daniele Jahier Pagliari:** Methodology, Writing – original draft. **Tommaso Filippini:** Writing – review & editing. **Pamela Mancuso:** Data curation. **Paolo Giorgi-Rossi:** Conceptualization. **Leonardo Alberto Dal Zovo:** Data curation. **Angela Corbari:** Data curation. **Marco Vinceti:** Conceptualization, Writing – review & editing, Supervision. **Paola Berchialla:** Conceptualization, Methodology, Writing – original draft, Supervision.

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.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2022.104132.

References

[1] S. Du, J. Wang, H. Zhang, W. Cui, Z. Kang, T. Yang, B. Lou, Y. Chi, H. Long, M. Ma et al., Predicting COVID-19 using hybrid AI model. 2020.

[2] M. Vinceti, T. Filippini, K.J. Rothman, F. Ferrari, A. Goffi, G. Maffeis, N. Orsini, Lockdown timing and efficacy in controlling COVID-19 using mobile phone tracking, EClinicalMedicine 1 (25) (2020 Aug), 100457. https://www.sciencedirect.com/science/article/pii/S2589537020302017.

[3] N. Kishore, R. Kahn, P.P. Martinez, P.M. De Salazar, A.S. Mahmud, C.O. Buckee, Lockdowns result in changes in human mobility which may impact the epidemiologic dynamics of SARS-CoV-2, Sci. Rep. 11 (1) (2021) 6995, https://doi.org/10.1038/s41598-021-86297-w.

[4] M.M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, A. Amoroso, Temperature, humidity, and latitude analysis to estimate potential spread and seasonality of Coronavirus Disease 2019 (COVID-19), JAMA Netw Open 3 (6) (2020) e2011834, https://doi.org/10.1001/jamanetworkopen.2020.11834.

[5] F. Benedetti, M. Pachetti, B. Marini, R. Ippodrino, R.C. Gallo, M. Ciccozzi, D. Zella, Inverse correlation between average monthly high temperatures and COVID-19-related death rates in different geographical areas, J. Translational Med. 18 (1) (2020) 251, https://doi.org/10.1186/s12967-020-02418-5.

[6] C. Merow, M.C. Urban, Seasonality and uncertainty in global COVID-19 growth rates, Proc. Natl. Acad. Sci. U.S.A. 117 (44) (2020) 27456–27464.

[7] G.F. Ficetola, D. Rubolini, Containment measures limit environmental effects on COVID-19 early outbreak dynamics, Sci. Total Environ. 761 (2021). https://www.sciencedirect.com/science/article/pii/S0048969720379638.

[8] P. Mecenas, R.T.d.R.M. Bastos, A.C.R. Vallinoto, D. Normando, A.M. Samy, Effects of temperature and humidity on the spread of COVID-19: a systematic review, PLOS ONE 1515 (99) (2020), https://doi.org/10.1371/journal.pone.0238339 e0238339e0238339.

[9] Y. Zhu, J. Xie, F. Huang, L. Cao, Association between short-term exposure to air pollution and COVID-19 infection: evidence from China, Sci. Total Environ. 20 (727) (2020), 138704. https://www.sciencedirect.com/science/article/pii/S004896972032221X.

[10] T. Filippini, K.J. Rothman, S. Cocchio, E. Narne, D. Mantoan, M. Saia, A. Goffi, F. Ferrari, G. Maffeis, N. Orsini, V. Baldo, M. Vinceti, Associations between mortality from COVID-19 in two Italian regions and outdoor air pollution as assessed through tropospheric nitrogen dioxide, Sci. Total Environ. 15 (760) (2021), 143355. https://www.sciencedirect.com/science/article/pii/S0048969720368868.

[11] T. Filippini, K.J. Rothman, A. Goffi, F. Ferrari, G. Maffeis, N. Orsini, M. Vinceti, Satellite-detected tropospheric nitrogen dioxide and spread of SARS-CoV-2 infection in Northern Italy, Sci. Total Environ. 15 (739) (2020), 140278. https://www.sciencedirect.com/science/article/pii/S0048969720337992.

[12] E. Conticini, B. Frediani, D. Caro, Can atmospheric pollution be considered a co-factor in extremely high level of SARS-CoV-2 lethality in Northern Italy? Environ. Pollut. 1 (261) (2020), 114465. https://www.sciencedirect.com/science/article/pii/S0269749120320601.

[13] C. Copat, A. Cristaldi, M. Fiore, A. Grasso, P. Zuccarello, S.S. Signorelli, G.O. Conti, M. Ferrante, The role of air pollution (PM and NO2) in COVID-19 spread and lethality: a systematic review, Environ. Res. 1 (191) (2020), 110129. https://www.sciencedirect.com/science/article/pii/S0013935120310264.

[14] L. Hu, W.-J. Deng, G.-G. Ying, H. Hong, Environmental perspective of COVID-19: atmospheric and wastewater environment in relation to pandemic, Ecotoxicol Environ. Saf. 219 (2021) 112297. https://pubmed.ncbi.nlm.nih.gov/33991934.

[15] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, W. Woo, Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.) Advances in Neural Information Processing Systems. Curran Associates, Inc.; 20Available from: https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.

[16] S.K. Paul, S. Jana, P. Bhaumik, A multivariate spatiotemporal model of COVID-19 epidemic using ensemble of ConvLSTM networks, J. Institution Engineers (India): Ser. B. (2020), https://doi.org/10.1007/s40031-020-00517-x.

[17] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[18] R. Kafieh, R. Arian, N. Saeedizadeh, Z. Amini, N.D. Serej, S. Minaee, S.K. Yadav, A. Vaezi, N. Rezaei, S. Haghjooy Javanmard, K. Blyuss, COVID-19 in Iran: forecasting pandemic using deep learning, Computational Math. Methods Med. 2021 (2021) 1–16.

[19] T. Filippini, F. Zagnoli, M. Bosi, M.E. Giannone, C. Marchesi, M. Vinceti, An assessment of case-fatality and infection-fatality rates of first and second COVID-19 waves in Italy: COVID-19 fatality rate in Italy, Acta Biomed. 92 (S6) (2021) e2021420. https://www.mattioli1885journals.com/index.php/actabiomedica/article/view/12241.

[20] F. Verelst, E. Kuylen, P. Beutels, Indications for healthcare surge capacity in European countries facing an exponential increase in coronavirus disease (COVID-19) cases, Euro Surveill. (2020), https://doi.org/10.2807/1560-7917.ES.2020.25.13.2000323. March 2020.

[21] F.M. Grosso, A.M. Presanis, K. Kunzmann, C. Jackson, A. Corbella, G. Grasselli, A. Andreassi, A. Bodina, M. Gramegna, S. Castaldi, D. Cereda, D.D. Angelis, A. Castrofino, G. Del Castillo, L. Crottogini, M. Tirani, A. Zanella, M. Salmoiraghi, Decreasing hospital burden of COVID-19 during the first wave in Regione Lombardia: an emergency measures context, BMC Public Health 21 (1) (2021), https://doi.org/10.1186/s12889-021-11669-w.

[22] STUDIOMAPP. Available from: https://www.studiomapp.com/.

[23] F. Ferrari, G. Maffeis, J. Flemming, R. Gianfreda, Utaq, a tool to manage the severe air pollution episodes, Environ. Eng. Manage. J. 19 (10) (2020). http://eemj.eu/index.php/EEMJ/article/view/4209.

[24] Copernicus. Available from: https://atmosphere.copernicus.eu/.

[25] COSMO model. Available from: http://www.cosmo-model.org/.

[26] ARPA. Available from: https://datacatalog.regione.emilia-romagna.it/catalogCTA/group/open-data-arpae.

[27] F. Emmert-Streib, Z. Yang, H. Feng, S. Tripathi, M. Dehmer, An introductory review of deep learning for prediction models with big data, Front. Artif. Intell. 3 (2020) 4. https://www.frontiersin.org/article/10.3389/frai.2020.00004.

[28] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780.

[29] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing [Internet]. Vienna, Austria; 2019. Available from: https://www.R-project.org.

[30] Chollet F, others. Keras [Internet]. 2015. Available from: https://keras.io.

[31] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y.U. Yuan, X. Zheng, Large-scale machine learning on heterogeneous systems, TensorFlow (2015).

[32] Newell Allen. A Tutorial on Speech Understanding Systems, in: Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium. New York Academic; 1975.

[33] C. Wen, S. Liu, X. Yao, L. Peng, X. Li, Y. Hu, T. Chi, A novel spatiotemporal convolutional long short-term neural network for air pollution prediction, Sci. Total Environ. 1 (654) (2019 Mar) 1091–1099. https://www.sciencedirect.com/science/article/pii/S0048969718344413.

[34] C.M. Yeşilkanat, Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm, Chaos, Solitons Fractals 1 (140) (2020 Nov), 110210. https://www.sciencedirect.com/science/article/pii/S0960077920306068.

[35] E. De Angelis, S. Renzetti, M. Volta, F. Donato, S. Calza, D. Placidi, R.G. Lucchini, M. Rota, COVID-19 incidence and mortality in Lombardy, Italy: An ecological study on the role of air pollution, meteorological factors, demographic and socioeconomic variables, Environ. Res. 1 (195) (2021 Apr), 110777. https://www.sciencedirect.com/science/article/pii/S0013935121000712.

[36] A. Paez, F.A. Lopez, T. Menezes, R. Cavalcanti, M.G.da.R. Pitta, A Spatio-Temporal Analysis of the Environmental Correlates of COVID-19 Incidence in Spain. Geogr. Anal. n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/gean.12241.

[37] A.A. Chudnovsky, Letter to editor regarding Ogen Y 2020 paper: “Assessing nitrogen dioxide (NO_2) levels as a contributing factor to coronavirus (COVID-19) fatality”, Available from: Sci. Total Environ. 20 (740) (2020 Oct), 139236 https://www.sciencedirect.com/science/article/pii/S0048969720327534.