



OPEN

Spatio-temporal prediction of the COVID-19 pandemic in US counties: modeling with a deep LSTM neural network

Behnam Nikparvar^{ID1}, Md. Mokhlesur Rahman^{ID1,2}, Faizeh Hatami^{ID3} & Jean-Claude Thill^{ID3,4✉}

Prediction of complex epidemiological systems such as COVID-19 is challenging on many grounds. Commonly used compartmental models struggle to handle an epidemiological process that evolves rapidly and is spatially heterogeneous. On the other hand, machine learning methods are limited at the beginning of the pandemics due to small data size for training. We propose a deep learning approach to predict future COVID-19 infection cases and deaths 1 to 4 weeks ahead at the fine granularity of US counties. The multi-variate Long Short-term Memory (LSTM) recurrent neural network is trained on multiple time series samples at the same time, including a mobility series. Results show that adding mobility as a variable and using multiple samples to train the network improve predictive performance both in terms of bias and of variance of the forecasts. We also show that the predicted results have similar accuracy and spatial patterns with a standard ensemble model used as benchmark. The model is attractive in many respects, including the fine geographic granularity of predictions and great predictive performance several weeks ahead. Furthermore, data requirement and computational intensity are reduced by substituting a single model to multiple models folded in an ensemble model.

The highly infectious COVID-19 pandemic that has been with us since the early months of 2020 has been adversely affecting public health worldwide with about 155 million reported cases of infection and 32 million deaths as of May 5, 2021^{1–3}. Despite adopting non-pharmaceutical social distancing measures (e.g., travel ban, cancellation of flights, restrictions on gatherings, closure of schools and public transport) and pharmaceutical measures (i.e., vaccination), the number of coronavirus cases increased alarmingly fast in the United States and elsewhere throughout large periods of the pandemic^{4–8}. Since March 2021, the COVID-19 pandemic has jumped globally towards a new peak exceeding the previous peak of January 2021 due to uncontrolled outbursts in India, Europe, and South America^{3,5}. Considering the catastrophic consequences of this pandemic, this study envisioned to develop a purely data-driven model to forecast the number of infection cases and deaths 4 weeks ahead using a multi-variate deep long short-term memory (LSTM) network to guide policymakers to make timely appropriate decisions.

Although COVID-19 related infection cases and deaths have decreased by the time of this study, US citizens experienced the highest peak of the pandemic from November 2020 to February 2021 (Fig. 1). Despite strong pharmaceutical and non-pharmaceutical control measures, large cohorts of people have been affected daily in all states (Fig. 2). Well into this pandemic, the US has remained the most affected country in the world, with 21.39% and 18.23% of global confirmed cases and deaths, as of May 5, 2021². However, the severity of the pandemic has started to taper off nationally since late February thanks to the successes of recent vaccine administration (i.e., 32% of people was fully vaccinated and 45% had at least one dose as of May 5, 2021)³. Moreover, multiple new variants of the COVID-19 virus that transmit more readily from person to person and change the effectiveness of the vaccines are emerging and circulating in the US and around the world¹⁰. Additionally, people are hesitant to receive vaccine against COVID-19 infection in the US¹¹. Thus, a combination of evidence-based

¹The William States Lee College of Engineering, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA. ²Department of Urban and Regional Planning, Khulna University of Engineering & Technology (KUET), Khulna 9203, Bangladesh. ³Department of Geography and Earth Sciences, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA. ⁴School of Data Science, University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA. ✉email: jfthill@uncc.edu

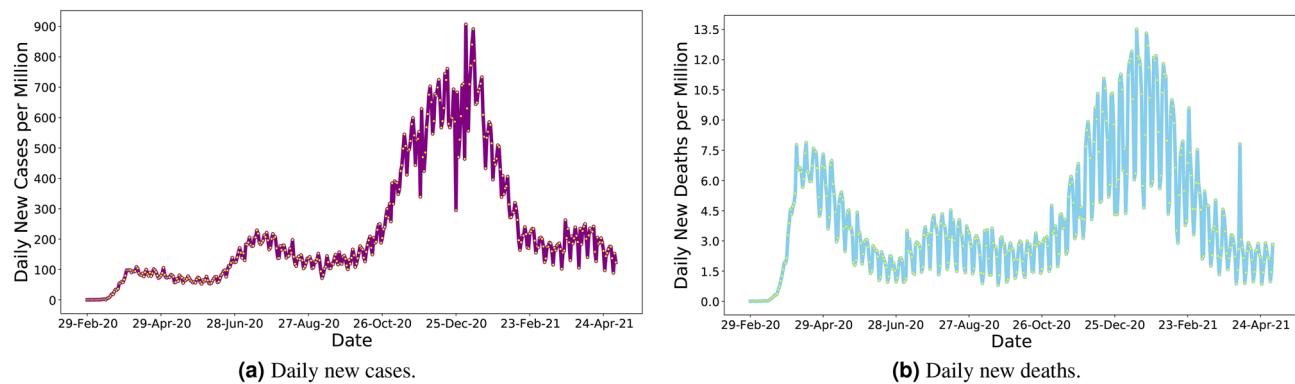


Figure 1. Daily new COVID-19 confirmed cases and deaths in the US, Data source⁹.

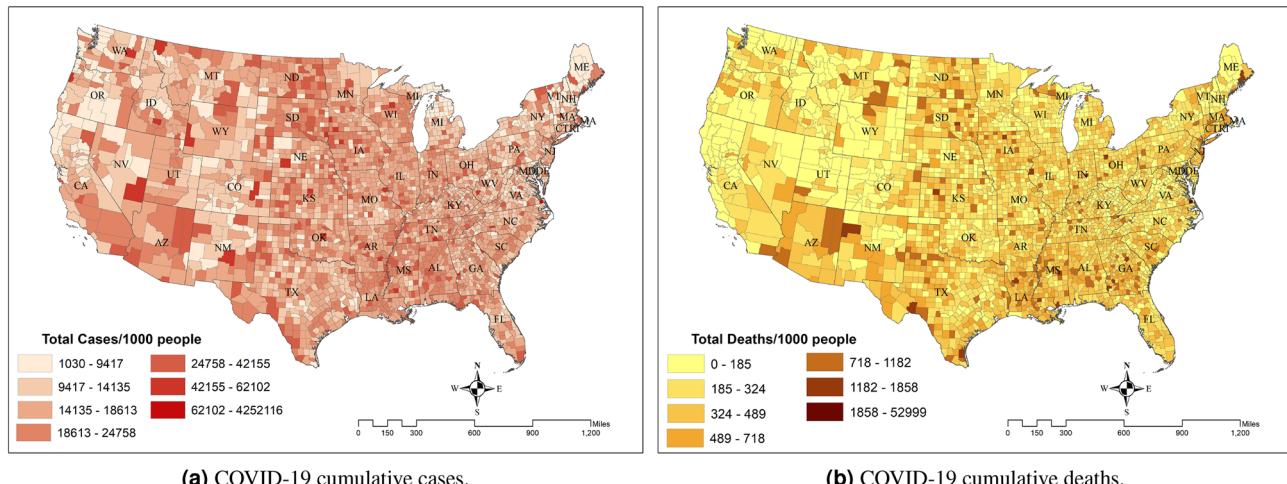


Figure 2. COVID-19 cases and deaths by county in the lower 48 United States, Data source¹⁹.

pharmaceutical and non-pharmaceutical public health mitigation measures remains important to curtail viral transmission locally and beyond¹²⁻¹⁸.

Literature review

The extant literature shows that, from the inception of the pandemic, researchers from around the world have been developing and implementing COVID-19 prediction models to understand the severity of the pandemic, delineate associated factors of virus infection, recovery, and death, and support the design of effective policies and operational measures to manage this unprecedented public health crisis^{1,15,20-22}. A wide range of methods and tools has been used in predictive models. Many studies have used machine learning and deep learning (e.g., random forest, support vector machine, decision tree, artificial neural network, ridge and lasso regression, nearest neighbor methods) to predict the transmission of the COVID-19 virus as well as individual and group responses to it²³⁻²⁶. A voluminous body of research has also been conducted with traditional epidemiological models to estimate viral diffusion and forecast the impact of policy interventions on the rate of infection (e.g., the Susceptible-exposed-infectious-removed (SEIR) model and variants of it)²⁷⁻³¹. However, compartmental models share several critical limitations. Most importantly, they are dependent on a considerable number of hypothesized input parameters, including the probabilities for transitions of population between the S, E, I and R states. Because of the strong sensitivity of SEIR models to changes in these input parameters, predictive accuracy can be substantially downgraded. In addition, SEIR models are based on oversimplifying assumptions, one of which being that the probability values or transition rates are homogeneous over population³² and constant over time. Instead, the rates of transition between the S, E, I and R population compartments for COVID-19 change over time and are very sensitive to socio-demographic conditions and mitigation policies. To alleviate some of these drawbacks, other classes of models have been advanced for estimating the probability of COVID-19 transmission and identifying compliance with social distancing measures. These include various hybrid models integrating neural networks and SEIR modeling^{33,34}, simulation systems (e.g., agent-based models)³⁵⁻³⁷ and video processing for object detection (e.g., You Only Look Once (YOLO))^{38,39}. Some other studies have used econometric models (e.g., linear regression, structural equation model) to identify the factors that influence the diffusion of COVID-19 and associated mitigation measures^{16,40}.

The convergence of recent developments in Information and Communication Technologies and in data analytics (i.e., production, access and storage of data, and analysis of information, cloud computing) has empowered people to apply Artificial Intelligence (AI) (e.g., machine and deep learning, text mining) to model and control complex issues in health care systems^{41–43}. Hence, AI-based techniques are widely used to monitor social distancing patterns of people and to assess scenarios on the transition of COVID-19 accurately¹⁵. They can solve multi-scalar, endogenous, non-linear, and ambiguous problems and extract insights from complex, unstructured and large data sets using their computational ability with reliable prediction compared to traditional approaches. Additionally, hybrid models (i.e., combining of multiple machine learning techniques) may enhance the robustness and generalization ability of machine learning models to handle reputedly wicked problems, such as a viral pandemic, quickly and effectively compared to single machine learning techniques and simulation models^{44,45}. One of the main motivations to predict the COVID-19 pandemic with machine learning is that it can more effectively estimate the effects of social distancing measures on the transmission of virus considering socio-economic control factors, which conventional econometric methods or compartmental models pain to capture¹⁵. However, high-quality data devoid of collection biases, robust methodologies, and validation with external data sets and models are necessary to develop a ubiquitous, reliable, and trustworthy model to provide consistent predictions across diverse systems (e.g., data, settlement contexts)^{15,43,46}.

Considering the rapid transmission of COVID-19 and the time lapse for the implementation of control measures, many studies have conducted spatio-temporal disease modeling to understand the geographic exposure and evolution patterns of associated risks of the pandemic^{47–51}. Spatio-temporal models are widely used by researchers to explain the temporal progress of disease over geographic regions and patterns of infection and mortality rates⁵². Recent studies have demonstrated that space and time are two critical factors for determining health risks during the COVID-19 pandemic^{53,54}. Spatio-temporal models consider infection rates of the surrounding neighbors to reduce the variability in the estimation that may exist locally⁵². These models also consider retrospective infection rates for short- and long-term prediction, controlling the irregular reporting of daily confirmed cases and deaths. Thus, spatio-temporal models yield better prediction, with a higher prediction accuracy and a lower tendency to over- or under-fit compared to traditional epidemiological and machine learning-based models⁴⁷. In this study, we used a deep learning-based space-time LSTM network to predict future COVID-19 cases and deaths, which can guide policymakers for timely interventions to reduce the severity of the pandemic and for more effective management of health care resources (e.g., nursing and medical staff, ICU beds, supplies, etc) under emergency conditions.

A wide range of studies have developed models for predicting COVID-19 cases and deaths and have investigated the factors that influence virus diffusion using machine learning and deep learning^{15,21,55}. They mentioned that lockdown and confinement measures, and socioeconomic factors significantly influence the outbreak of the pandemic and its dynamics. Specifically, the family of deep recurrent neural networks have proven to be an attractive approach for epidemic forecast⁵⁶ due to their acute capability to learn time series. Among these methods, uni-variate⁵⁷ and multi-variate^{58,59} LSTM models have been successful at predicting influenza and COVID-19, mainly due to their capability to memorize long-term dependencies. DeepGLEAM⁶⁰ uses a stochastic Diffusion Convolutional RNN (DCRNN) model, which considers short (commuting) and long range (air flight) mobility network connections, to forecast COVID-19 deaths at county and state granularities. Other versions of recurrent neural networks, such as attention networks⁶¹ and bidirectional LSTM⁶² networks, have also been reported to be successful. Taking stock of best practices in the extant literature, we propose a multi-variate and multi-time series long short-term memory (MTS-LSTM) network to simultaneously forecast confirmed cases, deaths, and mobility of all sub-populations at the county level using a single model. A study predicting the COVID-19 pandemic at this local level would provide unique insights for the policymakers to make targeted interventions to cope with the pandemic, given that many critical operational decisions and actions are made locally.

Previous models of epidemics have used a variety of covariates including mobility, underlying health, socio-demographic and socio-economic variables for predicting COVID-19 confirmed cases and death counts. In line with this precedence, we started with the implementation of models with various sets of independent variables. These variables included population density and measures at the sub-population level of race, educational attainment, age, family size and status, poverty, income, employment, housing, being in metropolitan or micropolitan area, health insurance status, percentage of deaths from other health conditions, percentage change in the time spent at work from Google mobility reports⁶³, and number of visits to Points of Interest from SafeGraph Places Schema data sets⁶⁴. After training a whole suite of model specifications and investigating the impact of these variables on pandemic severity by comparison of respective Root Mean Square Error (RMSE) statistics, it was found that COVID-19 is predicted with the highest accuracy with the SafeGraph mobility data as the sole covariate feature. The capability of the proposed deep learning model to predict COVID-19 dynamics with high performance with a single covariate is one of its distinctive strengths.

Data pre-processing

In this research, we proposed a multi-variate recurrent neural network with LSTM layers to predict the dynamics of a pandemic, which learns from multiple sample time-series simultaneously. To assess the performance of the proposed multi-time series long short-term memory (MTS-LSTM) method, we collected data on COVID-19 confirmed cases and deaths and foot traffic at the county level for a period of 33 weeks between January 26th and September 12th, 2020.

First, the data for deaths and confirmed cases of COVID-19 infection were downloaded from the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University⁶⁵ and pre-processed to correct inaccuracies. During data cleaning, for example, instances of decrease in cumulative counts were identified and replaced by the value of the previous day. We calculated the weekly sum of confirmed cases and deaths corresponding

to the Morbidity and Mortality Weekly Report (MMWR) weeks used by the Centers for Disease Control and Prevention (CDC) for all counties in lower 48 states of the US. MMWR weeks are based on the epidemiological calendar, with start on Sunday and end on Saturday. Daily reported cases and deaths were not selected since the new cases data are typically noisy. Instead, one may use the daily moving average. However, it is important to ensure no data leakage occurs between training, validation, and test data sets due to temporal dependency.

Next, the daily foot traffic patterns and points of interest (POI) for the top 5500 national brands were obtained from SafeGraph's Places Schema data set⁶⁴. SafeGraph provides visit data from around 18 million unique devices nationwide, which translates to 5–6% of the US population. These data are provided freely upon request for research purposes through the SafeGraph COVID-19 Data Consortium. We pre-processed and aggregated the foot traffic of the POIs into the county geographies and calculated their trailing average for each week.

Finally, to evaluate our model, we obtained predictions of an ensemble model implemented by COVID-19 Forecast Hub in collaboration with CDC for the period of 4 weeks ending on September 12th, 2020. Every week, the ensemble model accepts forecasts of COVID-19 confirmed cases and deaths from eligible models with a variety of assumptions and methodologies and takes the arithmetic mean and median of the forecasts for each geographic location. Evaluating the performance of the ensemble model and comparing it with the performance of the series of input models, the COVID-19 Forecast Hub recommends using the ensemble instead of individual models for policy considerations⁶⁶. In addition, since the ensemble model is an average of several input models, it has less variability in performance over time than a single input model. As a result, the ensemble model is a good criterion for assessing the performance of our model. However, only predictions of new cases are available for the ensemble model.

Model implementation

The codes were written in Python using the TensorFlow 2.5.0 package for deep learning. The implementation was conducted in the Google Colaboratory environment with GPU accelerators and high-RAM run-time shape. The model predicts all variables at the same time with a 4-week horizon ($T = 4$) and a time step of 1 week. We had confirmed cases, deaths, and mobility data for 33 weeks. The time series were divided into three sets. The last 4 weeks of the time series for all counties were kept aside for the test. We did not use data from these 4 weeks in any instances of the training process. Predictions of the Ensemble Model for the same 4 weeks were downloaded for evaluating the MTS-LSTM. Then, we used an out-of-sample validation (cross-validation) for training our model on the remaining 29 weeks ($n = 33 - 4 = 29$). After pre-processing of the time series, as explained in the model section, 70% of the feature vectors were used for training, and the rest (30%) for validation. This process was repeated ten times, each with a new set of randomly selected training and validation sets. All models in our experiments were trained using a batch size of 1024, an Adam optimizer with an initial learning rate of 10^{-3} and Huber loss function for 100 epochs. After a set of experiments, the window size was optimized with a length of three time steps ($l = 3$). This agrees with findings from other research that show the number of daily new cases of COVID-19 is related to population mobility of 3 weeks prior⁶⁷.

Results and discussion

Our research had three specific objectives. First, we aimed to study the impact of adding foot traffic (mobility) on prediction performance. Then, we wanted to find out how using multiple sample time series affects the performance of the predictions. Finally, we compared our predictions with the ensemble model.

To address the first question, we implemented the model with two specifications. The first model includes time series of two variables, deaths and new cases ($m = 2$), while the second model also includes the foot traffic as the third variable ($m = 3$). For the second objective, we trained the two models with different fractions of counties (1%–5%–10%–25%–50%–100%), henceforth called “experiments”, randomly selected among all counties in the data set and involved in the training and validation processes. We trained the model with ten “replications” to estimate the variation of predictions, each with a different set of randomly selected counties using a Monte Carlo sampling method.

Results are presented in Fig. 3, where the two panels on the left show the RMSE_{week} of new infection cases for all counties, and the two panels on the right show the same measure for deaths in all counties. Colors represent the portion of counties that have been used in an experiment (1 to 100%). The darker the color, the larger the fraction of counties that has been used in the experiment. Each observation point represents prediction of a single replication from the Monte Carlo process. In other words, points with the same color show all replications of a specific experiment. Notice that the Box-and-Whisker diagrams are plotted based on the forecasts from the experiment with 100% of counties.

What stands out in this figure is that Model 2 (with mobility as a predictor) predicts the new cases with lower RMSE median and variance than Model 1 (without mobility as a predictor) when 100% of counties are used in the training processes, and the same can also be said for experiments with smaller fractions of counties. Interestingly, Model 1 shows smaller RMSE values for the deaths compared to Model 2, while the differences are not significant. One possible explanation for this is that the window size is optimized with respect to the relationship between mobility and new cases. At the same time, the number of deaths has an extra delay in response to the number of new cases. That is, training the second model with a larger window size (possibly five or six for COVID-19) may lead to smaller death RMSE values for the second model compared to the first. From the figure, we can also find out that involving population mobility in the model improves the performance for predicting over a longer horizon window. Predictions of new cases for the third and fourth weeks have RMSEs less than the first-week prediction RMSE in the second model and their variance remains low. Median performances in the prediction of deaths are rather similar with the two model over a longer horizon, but Model 2 exhibits lower variances.

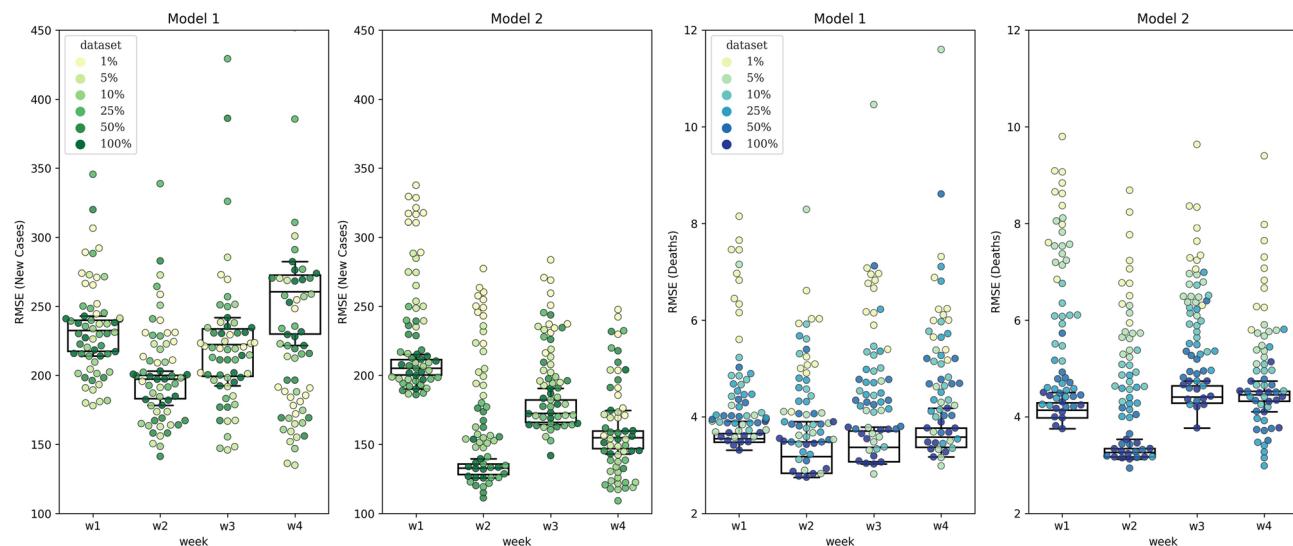


Figure 3. Comparison of predictions of Model 1 (without mobility as a predictor) and Model 2 (with mobility as a predictor). Each color represents a specific “experiment”, and each point represents a “replication” in the Monte Carlo process. The box-and-whisker diagrams are plotted based on the forecasts from the experiment with 100% of counties.

The color-coded RMSE values under different experiments show how increasing the number of counties in the training process affects the median and variance of the predictions (in Fig. 3). The pattern is more consistent with the death predictions mainly because the data on new infection cases are more subject to noise in comparison to the death data, which is more reliable.

The MTS-LSTM network may result in small negative predictions for some counties. For the confirmed cases for example the minimum value for 10 Monte Carlo repetitions is -9.076 , which is very small. The other two variables always remain positive in predictions. All negative values are replaced by zero before calculating RMSE values for evaluation. Alternatively, one could add an activation function to the last dense layer to make sure all predictions are non-negative.

Finally, we compared the prediction results of models 1 and 2 with the ensemble model from CDC. The ensemble model predictions for the 4 weeks between August 17th and September 12th were obtained⁶⁸. Figure 4 shows the dynamics of the new cases (left), deaths (middle), and foot traffic (right) over 33 weeks for several indicative counties. The predictions for the last 4 weeks from all three models are also reported in the panels of this figure. Models 1 and 2 are represented with 95% confidence intervals. Predictions based on the ensemble model are point predictions. Most counties have experienced the same mobility pattern over the study period, while their deaths and new cases series are shown to follow different trends and to be in different stages across the study period, including during the prediction horizon window. Additional results for counties with large, medium or small population size are available in the supplementary materials (see additional information section). These illustrative cases suggest that predictions from our MTS-LSTM model are overall acceptable, but that great variability exists across different population communities in the US.

A deeper comparison between predictions from Model 2 and the ensemble model is available in Fig. 5, where predicted values for each county are plotted against their actual values for each model by week of the prediction horizon. Rows one and two represent the new infection cases from Model 1 and new cases from the ensemble model, respectively. The RMSE_{week} values for the two models are in the same range, indicating very similar performance between the two models. This can also be seen in the two maps of Fig. 6, where the RMSE_{county} measure is depicted for both models. The third row in Fig. 5 shows the death prediction results against their actually reported values. As we observed in Fig. 3, the model’s performance to predict deaths is higher than for new cases, which is an indication that the deaths data are more reliable than data on new cases. RMSE_{total}^{new cases} is equal to 224.28, 169.84, and 159.92 for Model 1, Model 2, and the ensemble model, respectively. RMSE_{total}^{deaths} for Model 1 and Model 2 equals to 3.44 and 4.09, respectively.

Multi-time series long short-term memory model

We designed a multi-variate recurrent neural network with LSTM layers to predict three variables of COVID-19 confirmed cases and deaths, as well as foot traffic for multiple sub-populations using an iterated multi-step (IMS) estimation. LSTM is a gradient-based method developed by⁶⁹ to efficiently learn to store information over extended time intervals by recurrent back-propagation. This information is the long-term dependencies in a time series and can be transferred using a structure called memory cell. Memory cells have three gate structures named, input, forget, and output gates. These gates control the usage of historical information⁷⁰. Several versions of LSTM have been developed for applications in disease spread^{59,60}.

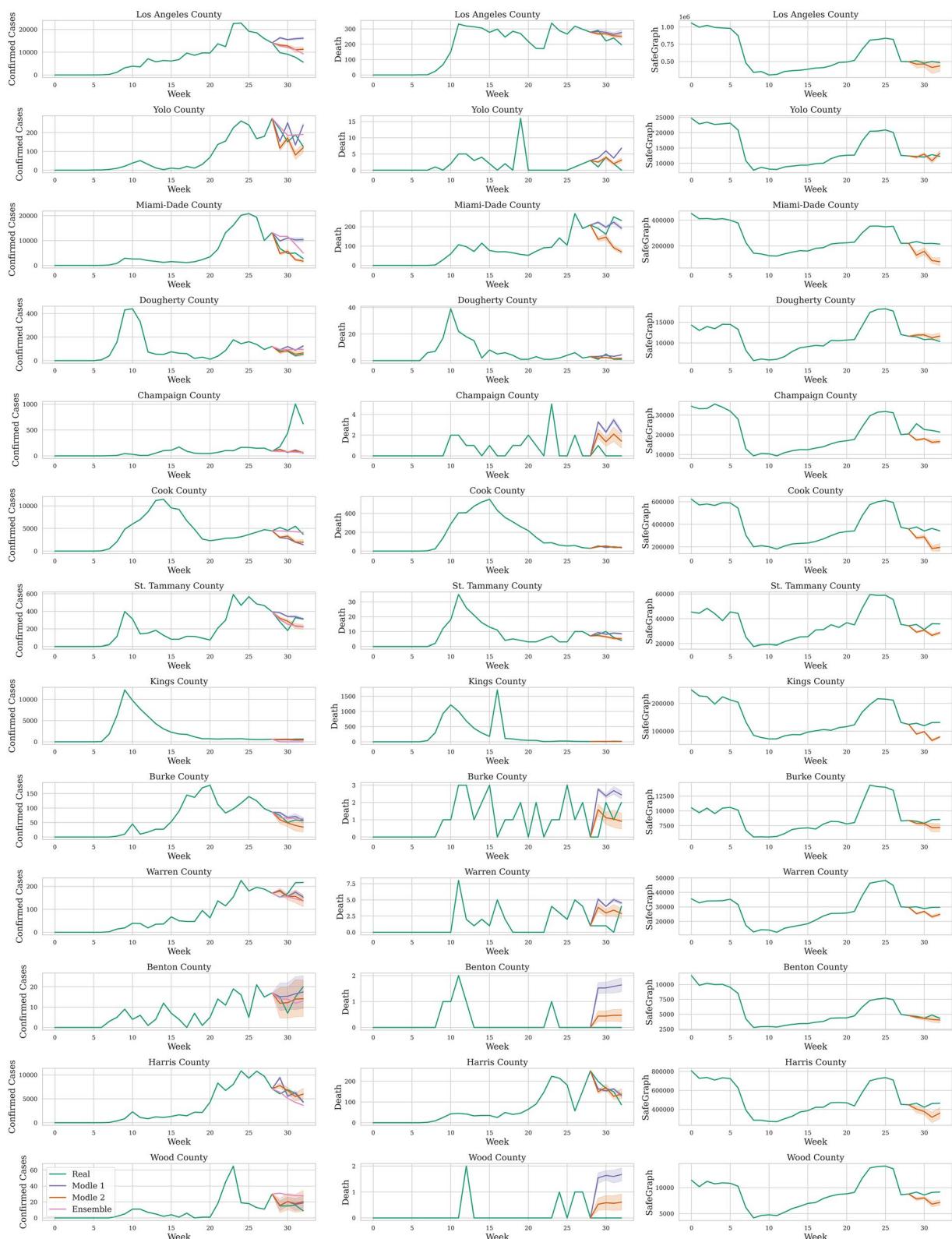


Figure 4. Dynamics of the new cases (left), deaths (middle), and foot traffic (right) over 33 weeks between January 26th and September 12th, 2020. Last 4 weeks include predictions from Model 1, Model 2, and the ensemble model and 95% intervals of confidence. Prediction of deaths is not available with the ensemble model.

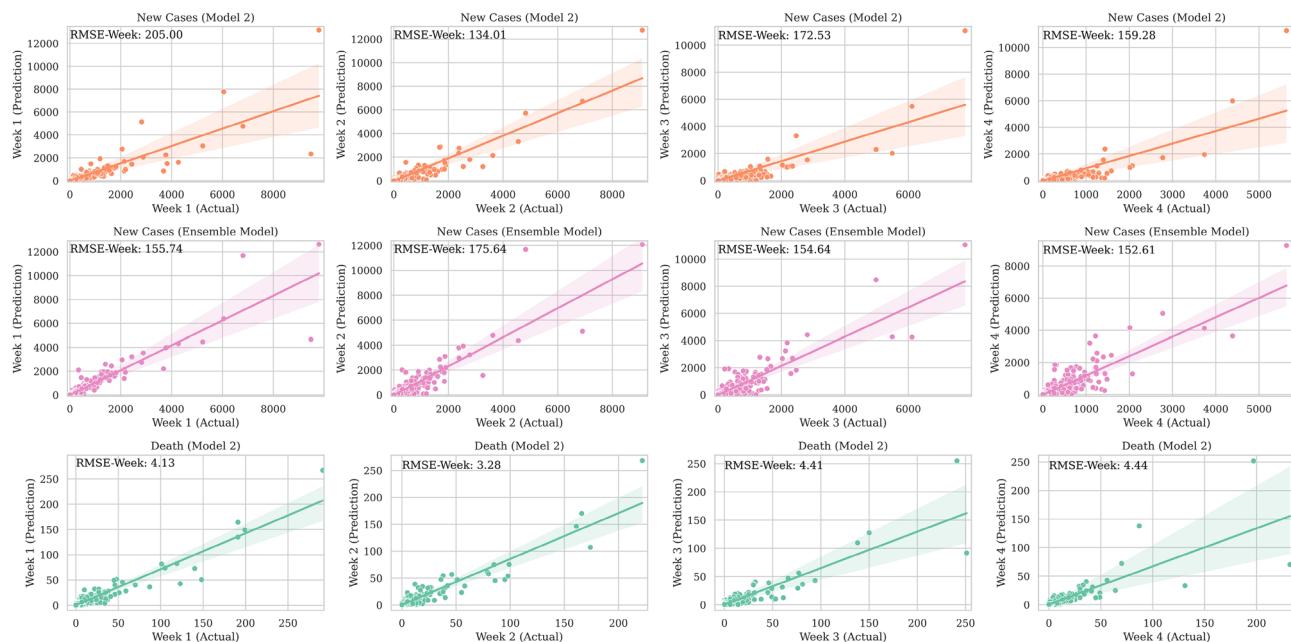


Figure 5. Predicted vs. real observations for new cases in Model 2 (top), new cases in ensemble model (middle), and deaths in Model 2.

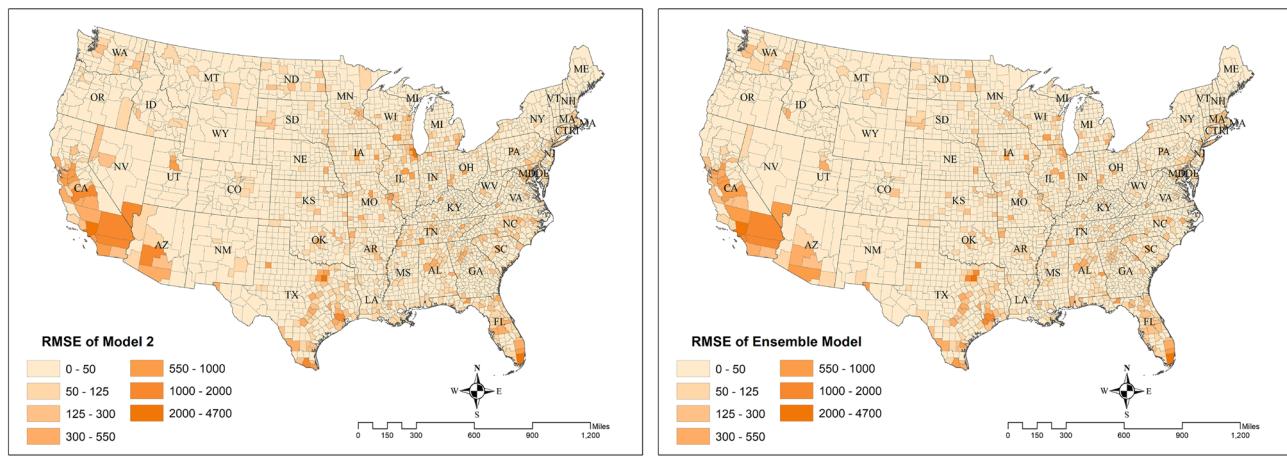


Figure 6. Comparison of predictions from Model 2 and Ensemble Model.

To aid the presentation of model results, we may refer to a uni-variate time series as a “feature”, and to a multi-variate time series as “features” from time-to-time. Also, a “feature vector” represents a sub-sequence of a multi-variate time series. Batches of feature vectors are the inputs to the network.

Data model preparation. Figure 7 demonstrates the data preparation and architecture of the network. x_i is a multi-variate time series for sub-population i with n time steps and m features. A window L of length l is selected to create the feature vectors and labels from each time series. $x_i^{(j)<t>}$ represents the time-step t of feature vector j from sub-population i . The sequence of feature values in each window (feature vector) is used to predict the next time step, where the actual values of the time series for those steps are used as the labels. In the process of creating feature vectors from each time series, the last sub-sequences with lengths less than l are dropped. This means that each time series has $n - l$ feature vectors of the same length.

We shuffled feature vectors from time series of all sub-populations to create a bag of feature vectors. Such an approach has three advantages compared to the single time series forecasting with LSTM layers. First and foremost, it provides more data for training compared to using the time series of each sub-population separately. Data are very limited at the early stages of the outbreaks, both in frequency and in quality. Therefore, data-driven approaches, especially deep learning methods with many parameters, are not applicable or end up with inaccurate predictions. Second, the spread of infectious diseases such as COVID-19 has complex diffusion and relocation

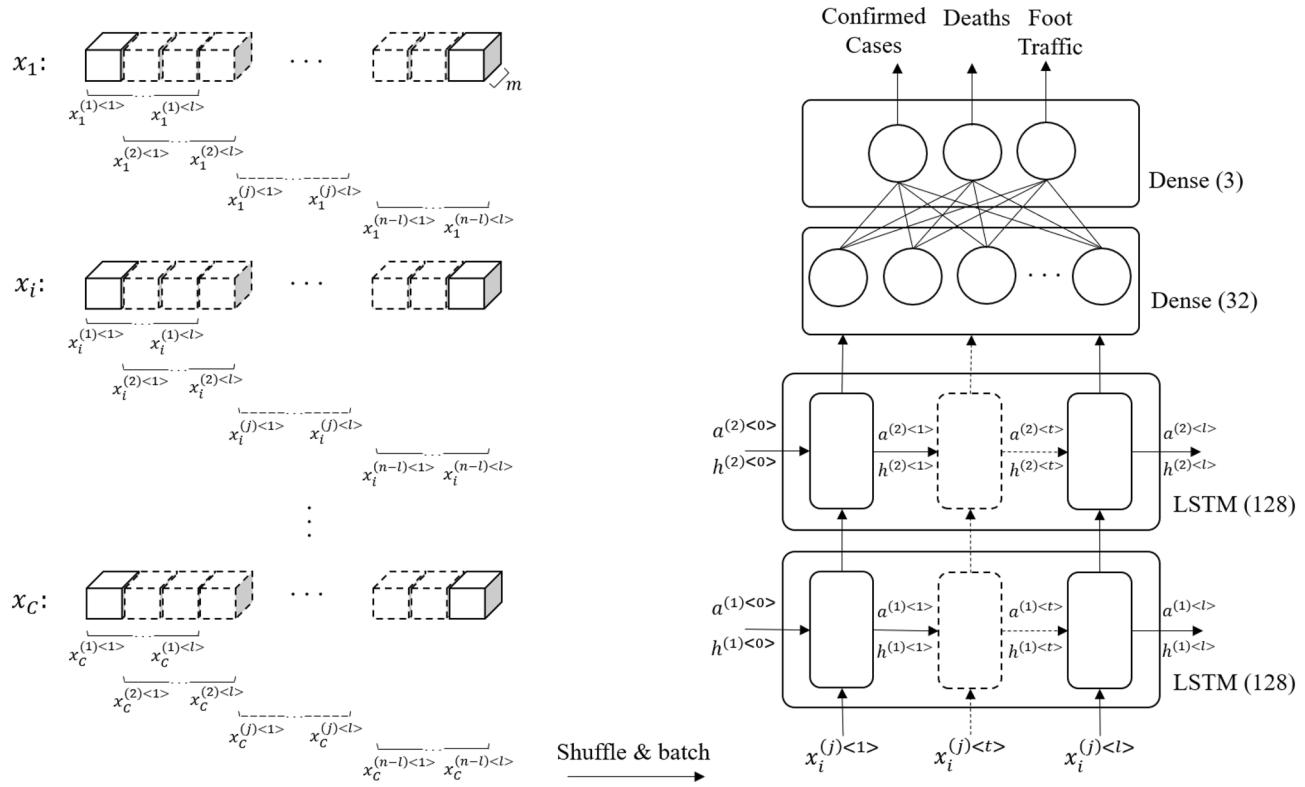


Figure 7. Proposed MTS-LSTM Model.

processes following mobility patterns across space and time⁷¹. That is, disease spread is not uniform, and outbreaks are in different stages in sub-populations across space. Such a method of creating a bag of feature vectors from all sub-populations provides an opportunity for the network to learn from later stages of an outbreak in sub-population A and uses this to predict outbreak dynamics in sub-population B. Third, training the machine learning model on feature vectors of multiple sub-populations simultaneously is expected to have more robust predictions when predicting disease dynamics spread in a new sub-population C. At the same time, intrinsic characteristics of different sub-populations such as mobility can affect the dynamics of disease in sub-populations. Thus, it is useful to have a multi-variate model that takes into account the mobility of sub-populations. In the next step, batches (bags) of feature vectors are created as inputs to the network. The size of each batch is klm, where k is the number of feature vectors within each batch.

MTS-LSTM architecture. The network has two LSTM layers with 128 neurons each, followed by a fully connected layer of 32 neurons, a drop out layer, and a fully connected regression layer with m neurons representing m features to be predicted. In our MTS-LSTM model, we have three features: new cases, deaths, and foot traffic. The structure of an LSTM memory cell is represented by the following equations:

$$\tilde{h}^{<t>} = \tanh(W_h[a^{<t-1>}, x_i^{(j)}] + b_h) \quad (1)$$

$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x_i^{(j)}] + b_u) \quad (2)$$

$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x_i^{(j)}] + b_f) \quad (3)$$

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x_i^{(j)}] + b_o) \quad (4)$$

$$h^{<t>} = \Gamma_u * \tilde{h}^{<t>} + \Gamma_f * h^{<t-1>} \quad (5)$$

$$a^{<t>} = \Gamma_o * \tanh(h^{<t>}) \quad (6)$$

where \$\tilde{h}^{<t>}\$ is the candidate value for updating memory cell parameter \$h^{<t>}\$. \$\Gamma_u\$, \$\Gamma_f\$, and \$\Gamma_o\$ are the update, forget, and the output gates. \$a^{<t-1>}\$ is the activation value from time step t - 1 of the feature vector. \$W_h\$, \$W_u\$, \$W_f\$, and \$W_o\$ are the network parameters and \$b_h\$, \$b_u\$, \$b_f\$, and \$b_o\$ are the biases.

The output features from the last LSTM layer, which provide representation for cases, deaths, and mobility, are concatenated before entering to the dense layer to provide a new feature representation. This new feature has information from previous time steps (time lags) of the feature being predicted as well as previous time steps (time lags) of secondary features. This way, all temporal dependencies, including various time lags between features, implicitly contributes to the predicted value of a feature in the back-propagation process. This is the simplest way to exploit temporal dependencies of various features. More explicit ways, with an added layer of complexity, exist to exploit dependencies of various features. For example, the cross-modal (cross-feature) LSTM⁷² separates the input features into a single time series and passes each of those through a separate three-layer LSTM stream and can be used for this purpose. A mechanism for information flow between the LSTM streams is proposed by authors to explicitly exploit multi-feature dependencies. The outputs from these LSTM layers are then concatenated into a new feature representation. This approach is more explicit because new hyperparameters are added to the model that can be controlled in a fine-tuning process. In this research study, we use the former approach. However, it is of interest to compare these two approaches of modeling between feature temporal dependencies in the future research.

Finally, one may use either of two approaches to forecast variables for multi-steps⁷³. The first approach is to optimize the multi-step forecasting objective function directly. This method is called direct multi-step (DMS) estimation, which is usually computationally expensive. Furthermore, the optimization of the objective function may fail if the model is highly non-linear or when the parameters are not well initialized⁷⁴. The other approach is to predict step by step in an iterative way. This method is called iterated multi-step (IMS) approach. The choice of method depends on the bias and variance of the prediction, on the number of prediction steps, and on the non-linearity of the model⁷⁵. IMS is easy to train, and forecasts of arbitrary length are possible. However, it is prone to accumulative errors⁷⁴. In this research, we used an IMS method since the length of the prediction horizon is not long.

Evaluation

We use three RMSE based metrics to evaluate predictions:

$$\text{RMSE}_{\text{county}} = \sqrt{\frac{\sum_{t=1}^T (P_t - O_t)^2}{T}} \quad (7)$$

$$\text{RMSE}_{\text{week}} = \sqrt{\frac{\sum_{j=1}^C (P_j - O_j)^2}{C}} \quad (8)$$

$$\text{RMSE}_{\text{total}} = \sqrt{\frac{\sum_{j=1}^C \sum_{t=1}^T (P_{jt} - O_{jt})^2}{CT}} \quad (9)$$

$\text{RMSE}_{\text{county}}$ calculates the RMSE of multi-step predictions for a single sample (county). T is the length of the prediction horizon and P_t and O_t are the prediction and observation for time step t. $\text{RMSE}_{\text{week}}$ represents the RMSE of predictions for a single time step (week) and all counties, where j is the county index and C is the number of counties. Finally, $\text{RMSE}_{\text{total}}$ presents a single-value metric based on RMSE with all predictions in different counties and weeks.

Conclusions and future work

We proposed a multi-variate LSTM-based recurrent neural network with mobility trained on multiple time series samples at the same time to predict the spatio-temporal spread of a pandemic. Our results show that adding mobility as a variable and using multiple samples to train the network improve the performance of the predictions both in terms of the bias and of the variance of forecasts. We also showed that the predicted results have similar accuracy and spatial patterns with the ensemble model used for benchmarking. While this may not add a new capability in modeling performance, our MTS-LSTM model has several attractive advantages. Compared to the single time-series LSTM, our model predicts the dynamics of disease within different sub-populations simultaneously. For a study area with n sub-populations, we used a single model instead of n different models. Given our experimental results, it may also be possible to replace the ensemble model which is based on several other models with a single model exhibiting the same level of performance, which would reduce data inputs and computational resources. From the point of view of training data size, this approach takes advantage of disease spread data from a sub-population to predict disease dynamics in other sub-populations. This is a very attractive feature, especially at the beginning of pandemics. At such crucial times, data availability is very limited, and machine learning and deep learning methods can be challenging to implement and use.

Recurrent neural networks are capable of being updated with new sets of observations as they become available. From this point of view, they are comparable with Bayesian learning. Besides, our model predicts all variables at the same time. This is important because of the causal relationship between variables that may impact the predictions. For example, staying at home when volunteering may be affected by the current number of new cases. Also, the model is highly flexible to add new variables such as hospital capacity or vaccination rates, as reliable data become available.

There is room to further improve our MTS-LSTM model. First, we optimized the size of the moving window based on the RMSE of new cases. It is better to optimize this hyperparameter by considering all variables (in our research, new infection cases and deaths). Second, we predicted the dynamics for a single period of time.

For the full evaluation of the model, it is useful to use the current model to predict dynamics for other periods as well. Third, after using a specific amount of data (about 10%), the model starts to improve slowly. This means, on the one hand that, if we have data from a portion of counties, we can get to an acceptable level of accuracy to predict dynamics of disease for all counties. On the other hand, this means there is room to use more complex models. In this respect, experimenting with different architectures such as adding bidirectional LSTM layers, 1D convolutional layers, encoder-decoder layers, and attention layers, presents a frontier for future work with opportunity to further enhance the predictive power of the MTS-LSTM model. Fourth, efficacy may also be improved by explicitly modeling temporal dependencies between features using a Cross-modal (cross-feature) LSTM⁷². Fifth, we did not account for spatial dependencies between counties in this research. All counties that participate in predictions of dynamics for a county have the same weights. However, we expect the counties closer to each other to have more similar values and, as a result, higher weights. Thus, it is useful to apply methods to include the spatial dependence. This can be conducted in a preliminary process using geo-statistical models or using the integration of LSTM with convolutional neural networks. The integration of LSTM with convolutional layers in graph based neural networks is also an interesting line of future research for disease spread prediction^{60,76,77}. Along a somewhat related line of thought, our analysis did not support using a series of covariates as predictors, which would better differentiate between local conditions. We would conjecture, however, that this requires further study, especially in the context of a model structure that captures spatial relationships more comprehensively.

Finally, the model can also be extended to series at different scales and granularities (e.g., state level and county level) to establish how performance changes or to experiment with policy-related scenarios (e.g., statewide lockdowns) that operate at various jurisdictional levels. Our MTS-LSTM model was trained with COVID-19 cases and deaths series in the United States and for a specific time period. It will be important to use this pre-trained model to predict disease dynamics in other geographic contexts, time periods, or even similar infectious diseases such as influenza in order to demonstrate the reproducibility of our methodology and results.

Data availability

Supplementary Information (e.g., codes, data, results) are available in GitHub at <https://github.com/behnamkp/Covid-19-Prediction.git>.

Received: 7 July 2021; Accepted: 20 October 2021

Published online: 05 November 2021

References

1. Gallo Marin, B. *et al.* Predictors of COVID-19 severity: A literature review. *Rev. Med. Virol.* **31**, 1–10 (2021).
2. Worldometers. COVID-19 Coronavirus Pandemic. <https://www.worldometers.info/coronavirus/> (Accessed 5 May 2021) (2021).
3. The New York Times. Coronavirus World Map: Tracking the Global Outbreak. <https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html> (Accessed 5 May) (2021).
4. Sartorius, B., Lawson, A. & Pullan, R. Modelling and predicting the spatio-temporal spread of COVID-19, associated deaths and impact of key risk factors in England. *Sci. Rep.* **11**, 5378 (2021).
5. Gamio, L. & Symonds, A. Global Virus Cases Reach New Peak, Driven by India and South America. <https://nyti.ms/3xYVO94> (Accessed on 5 May 2021) (2021).
6. Samuel, J. *et al.* COVID-19 public sentiment insights and machine learning for tweets classification. *Information* **11**, 314 (2020).
7. Borriello, A., Master, D., Pellegrini, A. & Rose, J. M. Preferences for a COVID-19 vaccine in Australia. *Vaccine* **39**, 473–479 (2021).
8. Samuel, J. *et al.* Feeling positive about reopening? New normal scenarios from COVID-19 US reopen sentiment analytics. *IEEE Access* **8**, 142173–142190 (2020).
9. Max Roser, E. O.-O., Ritchie, H. & Hasell, J. Coronavirus pandemic (COVID-19). <https://ourworldindata.org/coronavirus> (Accessed on 5 June 2021) (2020).
10. Centers for Disease Control and Prevention. About variants of the virus that causes COVID-19. <https://www.cdc.gov/coronavirus/2019-ncov/transmission/variant.html> (Accessed on 5 June 2021) (2021).
11. Ali, G. G. M. N. *et al.* Public perceptions of COVID-19 vaccines: Policy implications from US spatiotemporal sentiment analytics. *Healthcare* **9**, 1110 (2021).
12. Al Zobbi, M., Alsinglawi, B., Mubin, O. & Alnajjar, F. Measurement method for evaluating the lockdown policies during the COVID-19 pandemic. *Int. J. Environ. Res. Public Health* **17**, 5574 (2020).
13. Lu, X., Yuan, D., Chen, W. & Fung, J. A machine learning based forecast model for the COVID-19 pandemic and investigation of the impact of government intervention on COVID-19 transmission in China (2020). Preprint on webpage at www.researchsquare.com/article/rs-73671/v1.
14. Vinceti, M. *et al.* Lockdown timing and efficacy in controlling COVID-19 using mobile phone tracking. *EClinicalMedicine* **25**, 100457 (2020).
15. Rahman, M. *et al.* Machine learning on the COVID-19 pandemic, human mobility and air quality: A review. *IEEE Access* **9**, 72420–72450 (2021).
16. Rahman, M. *et al.* COVID-19 pandemic severity, lockdown regimes, and people's mobility: Early evidence from 88 countries. *Sustainability* **12**, 9101 (2020).
17. Silva, P. C. *et al.* COVID-ABS: An agent-based model of COVID-19 epidemic to simulate health and economic effects of social distancing interventions. *Chaos Solitons Fractals* **139**, 110088 (2020).
18. Rahman, M. M. *et al.* Socioeconomic factors analysis for COVID-19 US reopening sentiment with Twitter and census data. *Heliyon* **7**, e06200 (2021).
19. USAFacts. US COVID-19 cases and deaths by state. <https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/> (Accessed 5 June 2021) (2020).
20. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *BMJ* **369**(8242), m1328 (2020).
21. Swapnarekha, H., Behera, H. S., Nayak, J. & Naik, B. Role of intelligent computing in COVID-19 prognosis: A state-of-the-art review. *Chaos Solitons Fractals* **138**, 109947 (2020).
22. Xiang, Y. *et al.* COVID-19 epidemic prediction and the impact of public health interventions: A review of COVID-19 epidemic models. *Infect. Dis. Model.* **6**, 324–342 (2021).
23. Roy, A. & Kar, B. Characterizing the spread of COVID-19 from human mobility patterns and SocioDemographic indicators. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Advances in Resilient and Intelligent Cities*, 39–48 (2020).

24. Scarpone, C. *et al.* A multimethod approach for county-scale geospatial analysis of emerging infectious diseases: A cross-sectional case study of COVID-19 incidence in Germany. *Int. J. Health Geogr.* **19**, 32 (2020).
25. Polyzos, S., Samitas, A. & Spyridou, A. E. Tourism demand and the COVID-19 pandemic: An LSTM approach. *Tour. Recreat. Res.* **46**(2), 1777053 (2020).
26. Iwendi, C. *et al.* COVID-19 patient health prediction using boosted random forest algorithm. *Front. Public Health* **8**, 357 (2020).
27. Hou, X. *et al.* Intra-county modeling of COVID-19 infection with human mobility: assessing spatial heterogeneity with business traffic, age and race (2020). Preprint on webpage at <https://doi.org/10.1101/2020.10.04.20206763v1>.
28. Tang, B. *et al.* An updated estimation of the risk of transmission of the novel coronavirus (2019-nCoV). *Infect. Dis. Model.* **5**, 248–255 (2020).
29. Zhao, S. *et al.* Preliminary estimation of the basic reproduction number of novel coronavirus (2019-nCoV) in China, from 2019 to 2020: A data-driven analysis in the early phase of the outbreak. *Int. J. Infect. Dis.* **92**, 214–217 (2020).
30. Fanelli, D. & Piazza, F. Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos Solitons Fractals* **134**, 109761 (2020).
31. Choi, S. & Ki, M. Estimating the reproductive number and the outbreak size of COVID-19 in Korea. *Epidemiol. Health* **42**, e2020011 (2020).
32. Tolles, J. & Luong, T. Modeling epidemics with compartmental models. *JAMA* **323**, 2515–2516 (2020).
33. Soures, N. *et al.* SIRNet: Understanding social distancing measures with hybrid neural network model for COVID-19 infectious spread. *arXiv* (2020). preprint on webpage at [arXiv:2004.10376](https://arxiv.org/abs/2004.10376).
34. Chimmula, V. K. R. & Zhang, L. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos Solitons Fractals* **135**, 109864 (2020).
35. Wang, D. *et al.* Agent-based Simulation Model and Deep Learning Techniques to Evaluate and Predict Transportation Trends around COVID-19. *arXiv* (2020). Preprint on webpage at [arXiv:2010.09648](https://arxiv.org/abs/2010.09648).
36. Kai, D., Goldstein, G.-P., Morgunov, A., Nangalia, V. & Rotkirch, A. Universal masking is urgent in the covid-19 pandemic: Seir and agent based models, empirical validation, policy recommendations. *arXiv* (2020). Preprint on webpage at [arXiv:2004.13553](https://arxiv.org/abs/2004.13553).
37. Panovska-Griffiths, J. *et al.* Modelling the potential impact of mask use in schools and society on COVID-19 control in the UK. *Sci. Rep.* **11**, 8747 (2021).
38. Szczepanek, R. Analysis of pedestrian activity before and during COVID-19 lockdown, using webcam time-lapse from cracow and machine learning. *PeerJ* **8**, e10132 (2020).
39. Ahmed, I., Ahmad, M., Rodrigues, J. J., Jeon, G. & Din, S. A deep learning-based social distance monitoring framework for COVID-19. *Sustain. Cities Soc.* **65**, 102571 (2020).
40. Spada, A. *et al.* Structural equation modeling to shed light on the controversial role of climate on the spread of SARS-CoV-2. *Sci. Rep.* **11**, 8358 (2021).
41. Jiang, F. *et al.* Artificial intelligence in healthcare: Past, present and future. *Stroke Vasc. Neurol.* **2**, e000101 (2017).
42. Khalifa, N. E. M., Taha, M. H. N., Ali, D. E., Slowik, A. & Hassani, A. E. Artificial intelligence technique for gene expression by tumor RNA-Seq data: A novel optimized deep learning approach. *IEEE Access* **8**, 22874–22883 (2020).
43. Dexter, G. P., Grannis, S. J., Dixon, B. E. & Kasturirathne, S. N. Generalization of machine learning approaches to identify notifiable conditions from a statewide health information exchange. *AMIA Summits Transl. Sci. Proc.* **2020**, 152–161 (2020).
44. Wang, Y., Liao, Z., Mathieu, S., Bin, E. & Tu, X. Prediction and evaluation of plasma arc reforming of naphthalene using a hybrid machine learning model. *J. Hazard. Mater.* **404**, 123965 (2021).
45. Shao, Y. E., Hou, C.-D. & Chiu, C.-C. Hybrid intelligent modeling schemes for heart disease classification. *Appl. Soft Comput.* **14**, 47–52 (2014).
46. Roberts, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
47. Sun, J. *et al.* Forecasting the long-term trend of COVID-19 epidemic using a dynamic model. *Sci. Rep.* **10**, 1–10 (2020).
48. Chen, Y., Qin, R., Zhang, G. & Albanwan, H. Spatial temporal analysis of traffic patterns during the COVID-19 epidemic by vehicle detection using planet remote-sensing satellite images. *Remote Sens.* **13**, 208 (2021).
49. Briz-Redón, Á. & Serrano-Aroca, Á. A spatio-temporal analysis for exploring the effect of temperature on COVID-19 early evolution in Spain. *Sci. Total Environ.* **728**, 138811 (2020).
50. Liu, Q. *et al.* Spatiotemporal patterns of COVID-19 impact on human activities and environment in mainland China using night-time light and air quality data. *Remote Sens.* **12**, 1576 (2020).
51. Jarvis, K. F. & Kelley, J. B. Temporal dynamics of viral load and false negative rate influence the levels of testing necessary to combat COVID-19 spread. *Sci. Rep.* **11**, 1–12 (2021).
52. Ugarte, M. D., Adin, A., Goicoa, T. & Militino, A. F. On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Stat. Methods Med. Res.* **23**, 507–530 (2014).
53. Zhang, C. H. & Schwartz, G. G. Spatial disparities in coronavirus incidence and mortality in the United States: An ecological analysis as of May 2020. *J. Rural Health* **36**, 433–445 (2020).
54. Fitzpatrick, K. M., Harris, C. & Drawve, G. Fear of COVID-19 and the mental health consequences in America. *Psychol. Trauma Theory Res. Pract. Policy* **12**, S17–S21 (2020).
55. Lalmuhanawma, S., Hussain, J. & Chhakchhuak, L. Applications of machine learning and artificial intelligence for covid-19 (SARS-CoV-2) pandemic: A review. *Chaos Solitons Fractals* **139**, 110059 (2020).
56. Adhikari, B., Xu, X., Ramakrishnan, N. & Prakash, B. A. Epiddeep: Exploiting embeddings for epidemic forecasting. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 577–586 (2019).
57. Gautam, Y. Transfer learning for COVID-19 cases and deaths forecast using LSTM network. *ISA Transactions* (2021).
58. Chen, S. *et al.* Exploring feasibility of multivariate deep learning models in predicting covid-19 epidemic. *Front. Public Health* **9**, 661615 (2021).
59. Venna, S. R. *et al.* A novel data-driven model for real-time influenza forecasting. *IEEE Access* **7**, 7691–7701 (2018).
60. Wu, D. *et al.* Deepgleam: a hybrid mechanistic and deep learning model for covid-19 forecasting. *arXiv* (2021). Preprint on webpage at [arXiv:2102.06684](https://arxiv.org/abs/2102.06684).
61. Zhu, X. *et al.* Attention-based recurrent neural network for influenza epidemic prediction. *BMC Bioinform.* **20**, 575 (2019).
62. Ben Said, A., Erradi, A., Aly, H. & Mohamed, A. Predicting covid-19 cases using bidirectional lstm on multivariate time series. *arXiv* (2020). Preprint on webpage at [arXiv:2009.12325](https://arxiv.org/abs/2009.12325).
63. Aktay A., A. *et al.* Google COVID-19 Community Mobility Reports: anonymization process description (version 1.1). *arXiv* (2020). Preprint on webpage at [arXiv:2004.04145](https://arxiv.org/abs/2004.04145).
64. SafeGraph. SafeGraph Places Schema. <https://docs.safegraph.com/docs> (Accessed 15 Sept 2020) (2020).
65. Johns Hopkins University. COVID-19 dashboard by the center for systems science and engineering (CSSE) at Johns Hopkins University (JHU). <https://coronavirus.jhu.edu/us-map> (Accessed 15 Sept 2020) (2020).
66. Ray, E. L. *et al.* Ensemble forecasts of coronavirus disease 2019 (covid-19) in the us. *MedRxiv* (2020). Preprint on webpage at <https://www.medrxiv.org/content/10.1101/2020.08.19.20177493v1>.
67. Carteni, A., Di Francesco, L. & Martino, M. How mobility habits influenced the spread of the COVID-19 pandemic: Results from the Italian case study. *Sci. Total Environ.* **741**, 140489 (2020).

68. Reich, N. *et al.* har96, x. Zhang, jinghuichen, G. Espana, X. Xinyue, H. Biegel, L. Castro, Y. Wang, qjhong, E. Lee, A. Baxter, S. Bhatia, E. Ray, and abrennen, and ERDC CV19 Modeling Team (2020). Preprint on webpage at <https://github.com/reichlab/covid19-forecast-hub/tree/master/data-processed/COVIDhub-ensemble>.
69. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
70. Yao, W., Huang, P. & Jia, Z. Multidimensional lstm networks to predict wind speed. In *2018 37th Chinese Control Conference (CCC)*, 7493–7497 (IEEE, 2018).
71. Cromley, E. K. & McLafferty, S. L. *GIS and Public Health* (Guilford Press, XXX, 2011).
72. Velicković, P. *et al.* Cross-modal recurrent models for weight objective prediction from multimodal time-series data. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 178–186 (2018).
73. Shi, X. & Yeung, D.-Y. Machine learning for spatiotemporal sequence forecasting: A survey. *arXiv* (2018). Preprint on webpage at [arXiv:1808.06865](https://arxiv.org/abs/1808.06865).
74. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. *arXiv* (2015). Preprint on webpage at [arXiv:1506.03099](https://arxiv.org/abs/1506.03099).
75. Taieb, S. B. & Hyndman, R. Boosting multi-step autoregressive forecasts. In *International Conference on Machine Learning*, Vol. 32, 109–117 (PMLR, 2014). <http://proceedings.mlr.press/v32/taieb14.html>.
76. Velicković, P. *et al.* Graph attention networks. *arXiv* (2017). Preprint on webpage at [arXiv:1710.10903](https://arxiv.org/abs/1710.10903).
77. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. *arXiv* (2017). Preprint on webpage at [arXiv:1706.02216](https://arxiv.org/abs/1706.02216).

Author contributions

B.N., M.M.R., F.H., and J.C.T. conceived the study. B.N., M.M.R., F.H., and J.C.T. identified the deep learning modeling framework and designed the specific features of the final model. B.N. coded the model for implementation and ran all the experiments. B.N., M.M.R., F.H., and J.C.T. identified the extant literature. B.N., M.M.R., F.H. compiled the data used in the study. B.N., M.M.R., F.H., and J.C.T. analyzed the results. B.N., M.M.R., F.H., and J.C.T. contributed to the writing of the manuscript. J.C.T. supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.-C.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021