Quantitative Biology 2022, 10(2): 125–138
https://doi.org/10.15302/J-QB-021-0276

RESEARCH ARTICLE

condLSTM-Q: A novel deep learning model for predicting COVID-19 mortality in fine geographical scale

HyeongChan Jo[1,†], Juhyun Kim[2,†], Tzu-Chen Huang[3], Yu-Li Ni[1,*]

[1] Division of Biology and Biological Engineering, Caltech, Pasadena CA 91125, USA
[2] The Division of Physics, Mathematics and Astronomy, Caltech, Pasadena CA 91125, USA
[3] Walter Burke Institute for Theoretical Physics, Caltech, Pasadena CA 91125, USA
* Correspondence: ynni@caltech.edu

Received April 17, 2021; Revised June 30, 2021; Accepted July 22, 2021

***Background*: Modern machine learning-based models have not been harnessed to their total capacity for disease trend predictions prior to the COVID-19 pandemic. This work is the first use of the conditional RNN model in predicting disease trends that we know of during development that complemented classical epidemiological approaches.**
***Methods*: We developed the long short-term memory networks with quantile output (condLSTM-Q) model for making quantile predictions on COVID-19 death tolls.**
***Results*: We verified that the condLSTM-Q was accurately predicting fine-scale, county-level daily deaths with a two-week window. The model's performance was robust and comparable to, if not slightly better than well-known, publicly available models. This provides unique opportunities for investigating trends within the states and interactions between counties along state borders. In addition, by analyzing the importance of the categorical data, one could learn which features are risk factors that affect the death trend and provide handles for officials to ameliorate the risks.**
***Conclusion*: The condLSTM-Q model performed robustly, provided fine-scale, county-level predictions of daily deaths with a two-week window. Given the scalability and generalizability of neural network models, this model could incorporate additional data sources with ease and could be further developed to generate other valuable predictions such as new cases or hospitalizations intuitively.**

Keywords: COVID-19; machine learning; deep learning; epidemiology; time series forecast

Author summary: Predictive models benefit governments and healthcare systems to combat the COVID-19 pandemic. Here we present the conditional long short-term memory networks with quantile output (condLSTM-Q), a well-performing model for quantile predictions on COVID-19 death tolls at the county level with a two-week forecast window. This fine geographical scale is a rare but valuable feature in publicly available predictive models, significantly benefit state-level officials to coordinate resources within the state. The quantile predictions from condLSTM-Q inform people about the distribution of the predicted death tolls, allowing better evaluation of the possible trajectories of the pandemic. Given the scalability and generalizability of neural network models, this RNN-based model could incorporate additional data sources with ease and could be further developed to generate other helpful predictions such as new cases or hospitalizations intuitively.

INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic has taken more than 1.1 million lives worldwide and more than 224,000 lives in the United States alone as of October 2020, when this manuscript was initially drafted. Predicting the pandemic trend in terms of deaths and cases precisely has been crucial, as it allows the

[†]These authors contributed equally to this work.

governments and healthcare systems to prioritize and distribute resources effectively [1].

Here we present the conditional long short-term memory networks with quantile output (condLSTM-Q) model, a novel deep learning model predicting the spatial-temporal distribution of COVID-19 death tolls. The condLSTM-Q was developed during the Caltech COVID-19 Initiative [2]; a campaign aimed to develop and explore novel models to complement the classic predictive epidemiological models for the pandemic. The optimization goal for participating models was to predict daily death tolls attributed to COVID-19 in each county across the United States with a two-week future prediction window. The county-level spatial resolution was deliberately chosen so that the predictions could inform state officials and local governments for better decision-making. In addition, the models were required to provide estimations of 10-quantiles (0.1 to 0.9 quantile with intervals of 0.1) as the outputs, as quantile prediction is useful when forecasting the extremes [3, 4].

The condLSTM-Q was among the best performers by the end of the campaign in early June 2020. To test its performance and robustness, the model was then deposited without further architectural modifications since mid-June 2020, and has continued to be trained and output predictions using updated data. Since its deposit, the model's predictions were comparable to the well-known, publicly available predictions by the Institute for Health Metrics and Evaluation (IHME) [5] throughout May−October when this manuscript was written, and was continued to be recorded during the publication process till June 2021.

The foundation of the condLSTM-Q was based on the long short-term memory (LSTM) networks, a standard recurrent neural network (RNN) model well-suited for time-series predictions [6]. There were other works using LSTM and other neural networks for COVID-19 forecasting, such as risk assessment of countries [7], and predicting national COVID-19 mortality rates [8]. Nevertheless, to our knowledge, no county-level-quantile prediction model existed when our prototype was developed. We improved the classical LSTM model's performance by adapting and building on a "conditional" LSTM architecture which took in static data along with time series data [9, 10]. To implement the quantile outputs, we utilized the flexibility of neural networks to output a distribution of 10-quantiles that was required by the initiative.

Given the robust performance of condLSTM-Q and the fact that we have not exploited all possible variations of the model, this architecture from our pilot experiment is worthy of further exploration. Interdisciplinary collaboration between experts in machine learning and epidemiology would greatly facilitate building better variants that will provide longer prediction windows and clearer interpretability.

RESULTS

Predictions and performance

The condLSTM-Q provides a 14-day, county-level prediction with 10-quantiles. To illustrate this, we show predictions for several representative counties, identified by their federal information processing standards (FIPS) based county code, including Cook, Los Angeles, New York, Wayne, Philadelphia, Hennepin, Maricopa and Montgomery from June 12 to June 25, 2020 (Fig. 1). This prediction interval was after we finalized and froze the model architecture. The model was trained with the data up to June 11, 2020, and was agnostic to the data afterward (*i.e.*, post-June 12). The condLSTM-Q predicted the death trends in different phases in the pandemic with different dynamic ranges. For example, the number of daily deaths in Los Angeles fluctuated between 20 and 60 each day, whereas the counts in Philadelphia were roughly a third of that.

To validate the performance of the condLSTM-Q, we aggregated the sum of reported mortality cases of the counties and plotted the national trend with *The New York Times'* statistics (Fig. 2). At first, the precision of predictive values subject to limited data for model training. The predictions gradually improved around mid-April and successfully predicted the descending trend of COVID-19 mortality from May to July 2020 and the "second wave" of COVID-19 since late July 2020. To observe the changes in the prediction accuracy from day 1 to day 14 of the prediction (*i.e.*, across the two-week prediction window), we aligned the day 1, 3, 7, 10, and 14 of each of the prediction windows to their corresponding dates. The overall prediction accuracy improved as more training data became available. Before mid-April 2020, distal prediction (*i.e.*, day 14) varied much more than proximal predictions. However, after the model became stable with enough training data, the performance of both distal and proximal prediction converged with a smaller spread; also, the precision of distal prediction was not inferior to proximal predictions.

Visualization of the county-level prediction from August 6 to August 19, 2020, is shown in Fig. 3. This interval approximately covered periods of the COVID-19 "second wave" with the highest daily death attributed to COVID-19 in the United States, which impacted the Southwest and Southeast regions the most. For better visibility, the figure focuses on the Southwest regions only. With county-level resolution, condLSTM-Q demonstrated its ability to pick up inhomogeneous hot

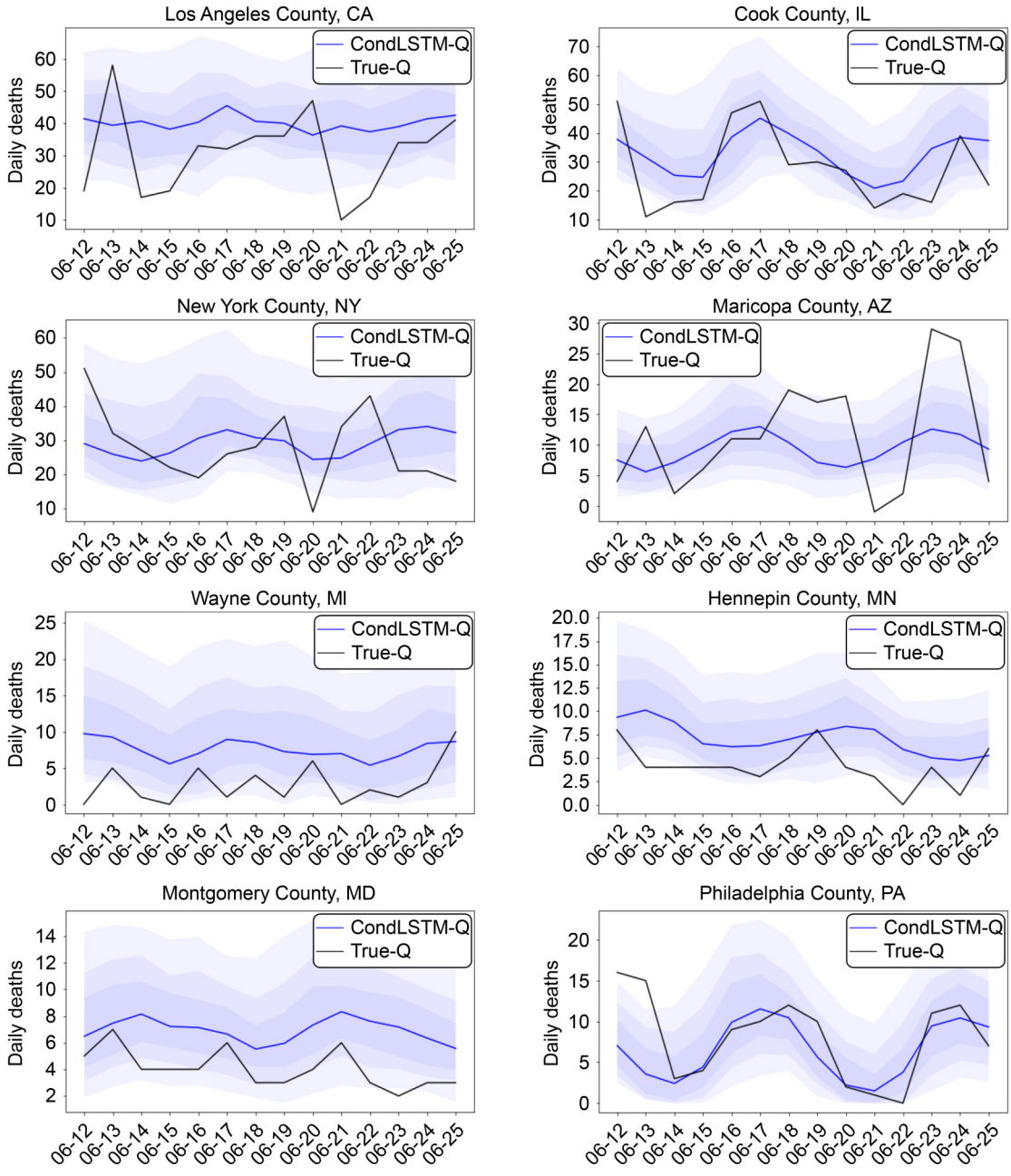

Figure 1. Predictions in representative locations. Representative counties including Cook, Los Angeles, New York, Wayne, Philadelphia, Hennepin, Maricopa, and Montgomery. The condLSTM-Q was able to keep track of the death trends in different phases in the pandemic with different dynamic ranges. For instance, Los Angeles had daily deaths fluctuating between 20−60 per day whereas Philadelphia counts were roughly a third of that.

spots within specific states. In contrast, other models with coarser geographical resolution could only observe an average trend of the whole state. In addition, at this resolution, one can observe state-state interactions of spread in the state border. For instance, quite a few counties in Arizona and Nevada had trends more similar to their neighboring California hot spots than to other counties within the states.

Comparison with other models

To evaluate the performance of condLSTM-Q, we compared its predictions with the IHME model's prediction [5]. The IHME model generates state-wise predictions. Thus we aggregated our predictions to corresponding states for comparison, and compared root mean square error (RMSE) of the state-wise forecasts on two-week intervals. We averaged across the nine

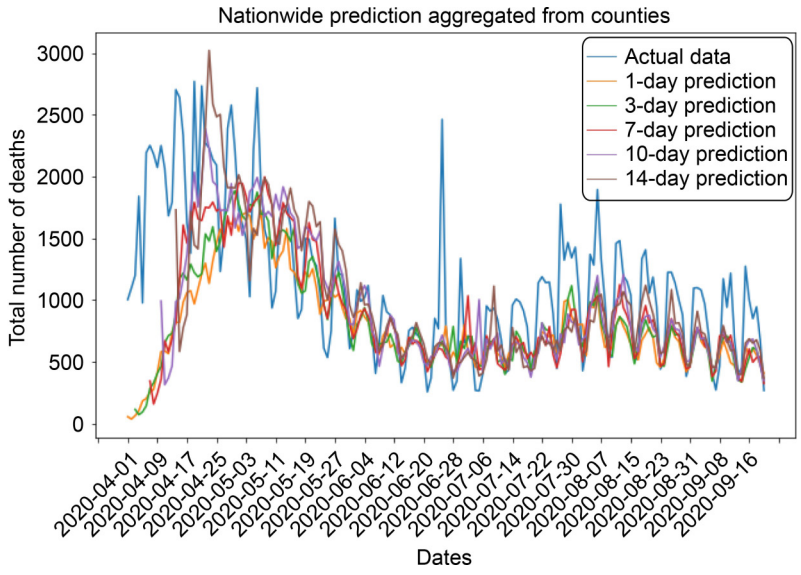

Figure 2. Nationwide prediction on the total number of deaths, aggregated from counties. At each time point, the actual data from that day is shown in blue, and the predicted values returned from models trained until 1, 3, 7, 10, and 14 days ago are shown in different colors. The model's overall accuracy is relatively low in the beginning when there was not enough data for the model to be trained on, but the prediction started to follow the trend well since late April.

quantile predictions from condLSTM-Q to estimate the mean of the distribution, in order to allow direct comparison against single value predictions from IHME.

As shown in Fig. 4, the state-wise predictions of condLSTM-Q were consistently comparable to the predictions from the IHME model. The condLSTM-Q initially had higher RMSE in early May 2020, but showed better performance over most intervals since then after training data became ample. Even though the condLSTM-Q was trained using county-level data with the pinball loss function, it demonstrated a robust performance on a different geographical scale under a different metric (*i.e.*, RMSE). We also measured the RMSE by setting zeros as control (*i.e.*, placing zeros on the full 14-days interval across all 50 states in the United States). Our results revealed that both IHME model and our condLSTM-Q presented a similar pattern with the control, indicating that there could be an uncaptured variance in both models. We also noticed a sharp peak of the epidemic curve (approximately 20,000 deaths) on June 24, 2020, when *The New York Times* made bulk adjustments in its data due to changes in tallying criteria.

The usefulness of conditional layer for categorical data

With a classical LSTM network that does not have a conditional layer for initializing the hidden states based on categorical data as in condLSTM-Q, the categorical data should be stacked to the exact dimension of the time series data to be fed into the model. As mentioned in the Introduction, this may lead to suboptimal performance of the model because it introduces constant values in time series features which actually are not sequential data. To test this, we compared the prediction accuracy of two models: the model with the aforementioned stacking method, which we refer to as the "pseudo-categorical" LSTM model, and condLSTM-Q. The pseudo-categorical LSTM model had a pinball loss of 0.115 during the prediction period of May 22−June 4, 2020, and 0.0994 during June 5−June 18, 2020. The pinball loss of condLSTM-Q, on the other hand, was 0.0959 and 0.0744, respectively, over the same prediction periods.

To investigate whether such a decrease in pinball loss was observed in every county, we looked at the difference in pinball loss between condLSTM-Q and the pseudo-categorical LSTM model. Figure 5A shows a distribution of such differences in counties with a high number of deaths (> 50). The difference was obtained by subtracting the pinball loss of a pseudo-categorical model from condLSTM-Q's loss, so its bias to the left means condLSTM-Q's loss is significantly lower than the pseudo-categorical model ($p < 0.0005$, one-sided Wilcoxon signed-rank test)

The difference between these two models is well-demonstrated in Fig. 5B and C, which shows example predictions of May 22−June 4 and June 5−June 18, 2020, in New York—the county with the most significant difference in pinball loss between the models. When the condLSTM-Q was already capturing a down-turned trend, the pseudo-categorical LSTM was still predicting an upward trend. We also found that the

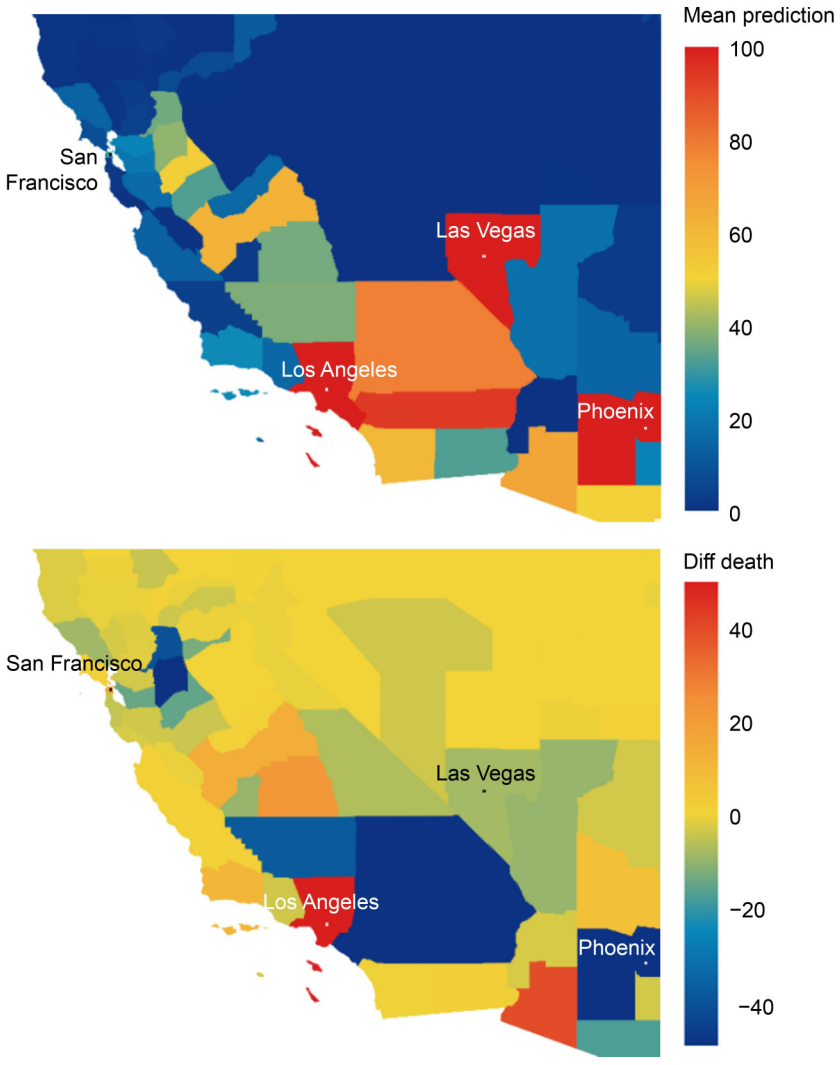

Figure 3. County-wise prediction by condLSTM-Q. County-wise absolute death counts (Top) and the differences with the ground truth (Bottom), summed from August 6 to August 19, 2020. This interval was roughly the peak of the "second wave" of deaths in the United States. For better visibility, the figure only shows the west coast of the United States. Note that the top and the bottom figures have different color scales.

condLSTM-Q had a much tighter spread of prediction, allowing planning based on the predictions much more possible.

Explainability of the model

To quantify the relative importance of each feature for the prediction, we did a series of permutation tests as suggested in previous studies [11, 12]. We quantified the feature importance based on how much the model's validation loss increased after permuting in a validation set. An increased validation loss after shuffling the feature shows that the model relied on that feature for the prediction and vice versa.

Time-series features were shuffled across counties first and then permuted across time. As each county can have drastically different values in time-series features, shuffling across counties and the temporal sequence ensures more realistic input data for LSTM than without. In contrast, categorical features were only shuffled across counties as they did not have a temporal component. Such permutation was executed ten times for each feature, and the resulting ten input data sets were provided to the trained model. Feature importance was measured at three different time frames (May 15−28, July 15−28, and October 8−21, 2020) with the same procedure.

The result is shown in Figs. 6 and 7. For categorical data (Fig. 6), its importance to the prediction was generally higher during the earlier phase, when there were insufficient time-series data for the model to be trained on. The top three features with the greatest

Dates	14-day prediction RMSE: State level		
	condLSTM-Q	IHME	Control
2020-05-04	54.814	**38.189**	67.881
2020-05-19	**18.396**	20.812	38.046
2020-06-03	**13.709**	14.152	27.152
2020-06-24	**74.154**	74.295	77.597
2020-07-04	16.291	**14.790**	28.945
2020-07-18	48.690	**47.098**	64.760
2020-08-06	**32.927**	33.749	55.354
2020-08-21	**14.276**	21.912	37.628
2020-09-02	**16.325**	18.915	34.328
2020-09-18	**13.123**	14.421	28.974

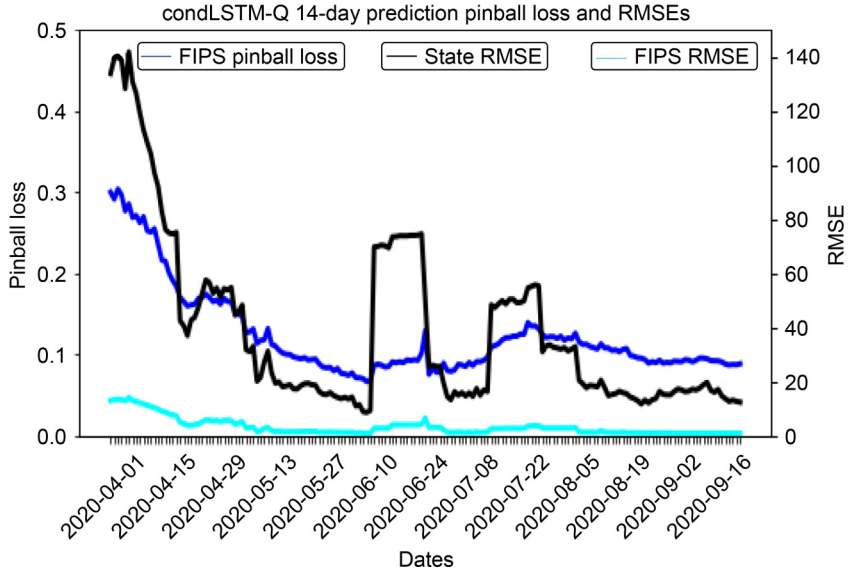

Figure 4. Performance analysis. State-level prediction (Top). State-wise, two-week prediction RMSE of condLSTM-Q and IHME model. We matched the starting dates of our predictions to the dates when the IHME model was updated, where each model would have access to the training data up to the day before the onset of the two-week prediction. We also tried placing zeros in all the predictions and calculated RMSE as a control. Both IHME and condLSTM-Q showed a similar pattern with the control, indicating that there is an uncaptured variance for both models. Bottom: County(FIPS)-level performance measured by pinball loss and RMSE, along with state level RMSE for better comparison with other models such as the IHME predictions. The error for both metrics showed a steady decline since April with several jumps from June to August, which is from the abnormality in *The New York Times*' data due to their bulk adjustments following the changes in tallying criteria.

importance during the earlier phase in May were gross domestic product (GDP) in 2015, GDP in 2016, and heart disease mortality. GDP was also one of the most important features in the later phase in July, but it was not as important as the population estimate (2018) and estimated mortality (2015−2017). These two features were the most important categorical features in October as well, followed by the number of eligible people for medicare (2018).

In contrast, the ranking of the importance of the time series features (Fig. 7) was relatively stable compared to categorical features. The number of deaths and cases were the two most essential features in all three time frames. In the beginning, the number of cases was more prominent but was surpassed by the number of deaths later on. Seasonality also had high importance, especially during the earlier phase in May and July 2020, when there were fewer data for the model to be trained on.

DISCUSSION

Contribution and novelty

This work is the first use of the conditional LSTM model in predicting the COVID-19 trend that we know of. The model has the merit of providing fine-scale, county-level predictions of daily deaths with a two-week

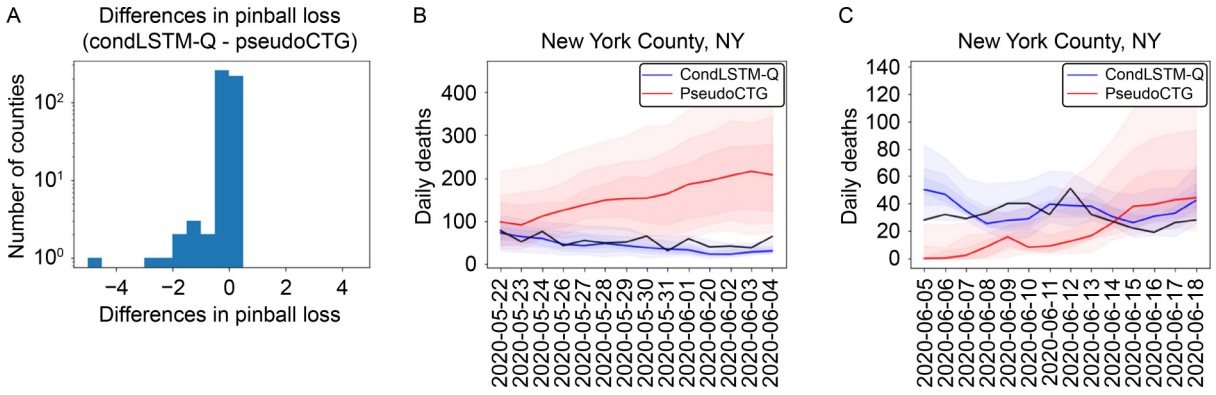

Figure 5. Effectiveness of conditional architecture. (A) A histogram of differences in pinball loss between condLSTM-Q and the pseudo-categorical model (pseudoCTG), in counties with a high number of total death (>50). The loss from the pseudo-categorical model was subtracted from condLSTM-Q's loss, so negative values mean condLSTM-Q has lower pinball loss. (B, C) Representative case study of New York, in two different time frames. The condLSTM-Q not only captures the trend better, but also has a tighter distribution when compared to the pseudo-categorical model trained on the same training data with categorical data stacked to match the time series.

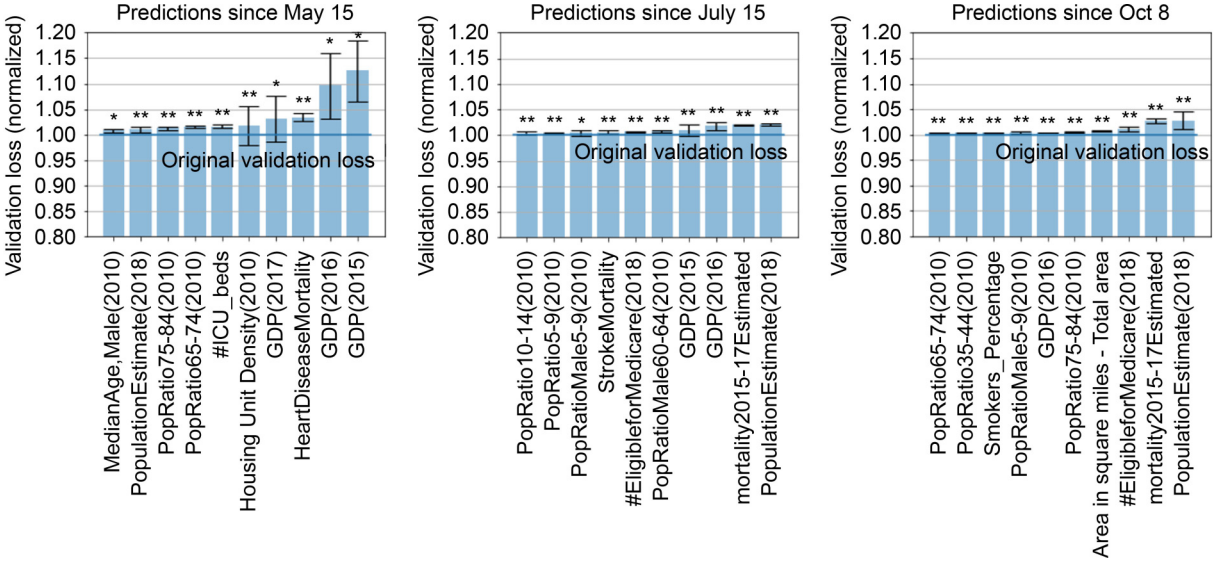

Figure 6. Importance of each categorical feature. The permutation feature importance of the 10 most important categorical features in three prediction windows (May 15−28, July 15−28, and October 8−21, 2020) is shown. The categorical features were more important during the earlier phase rather than in the later phase. The list of important features also changed over time: GDP and heart disease mortality for the earlier phase; population, mortality, and the number of eligible people for medicare for the later phase. Asterisks indicate statistical significance based on one-sample sign test to original validation loss (*p < 0.05, **p < 0.01).

window. The model's performance was robust and comparable to, if not slightly better than, the well-known IHME predictions. This model type is new for epidemiology modeling, and our work is far from testing all possible model architectures. One could foresee that the model's performance in terms of stability, precision, and length of the forecast window will enhance with further optimization. We also see that conditional neural networks could be generalized to trend forecasting of other infectious diseases in the future and are not limited to COVID-19.

One contribution unique in our model is the fine geographical resolution. County-level spatial resolution is a feature not available in other famous and publicly available forecasts such as IHME, DELPHI, and Los Alamos National Laboratory [13, 14], which provide state-wise predictions. This provides unparalleled opportunities for investigating trends within the states and interactions between counties along state borders. In addition, by analyzing the importance of the categorical

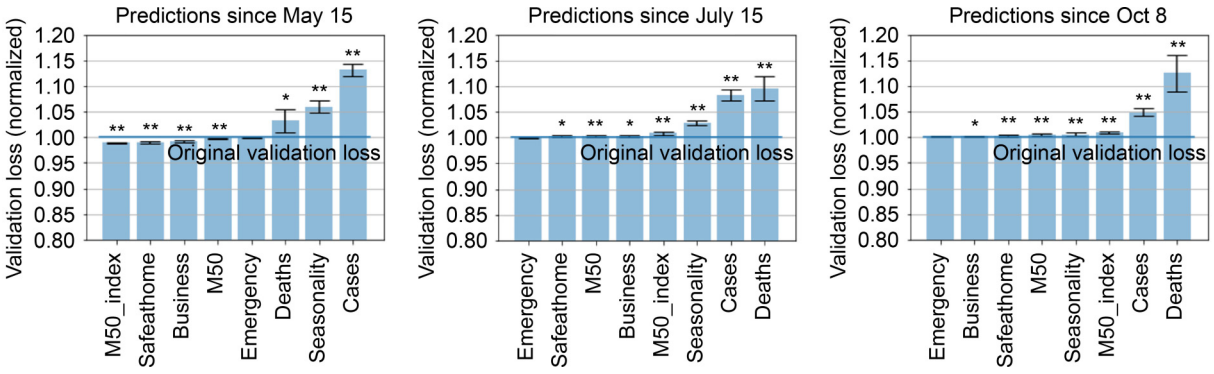

Figure 7. Importance of each time series feature. The permutation feature importance of the time series features in three prediction windows (May 15−28, July 15−28, and October 8−21, 2020) is shown. The number of cases, deaths, and flu seasonality are the three most important time series features in the model. The importance of seasonality, however, went down nearly to zero in the later phase. Asterisks indicate statistical significance from one-sample sign test to original validation loss (*p < 0.05, **p < 0.01).

data, which will be discussed below, one could learn which features are risk factors that affect the death trend and provide handles for officials to ameliorate the risks.

Effect of categorical features on predictions

The other special advantage of using conditional LSTM is that it provides insight into the effect of each categorical and time series feature on the predictions. Figure 8 shows the total number of deaths predicted by each model after altering categorical features with the highest permutation feature importance. As mentioned in the result section, GDP in 2015 and 2016 and heart disease mortality were the most important features in the beginning, but later on, the dominant factors shifted to the population estimate, estimated mortality, and the number of eligible people for medicare; an increase in any of these features led to an increase in predicted counts and vice versa.

Such a pattern in the earlier phase shows the possible vulnerability of people with heart disease to COVID-19, which is consistent with previous studies reporting elevated risk of poorer outcomes with COVID-19 in people with congenital heart disease [15] and coronary heart disease [16−18]. The positive correlation between GDP and the predicted number of deaths in the earlier phase is also consistent with a study claiming a positive correlation between GDP and the number of confirmed cases of COVID-19 in China [19]—nevertheless, the GDP estimates the value of goods and services produced by each county [20], which inevitably correlates with its population and other variables. Therefore, it would require further investigation to remove the dependencies of each factor.

On the other hand, in the later phase, the critical features have shifted to population estimates, mortality rate, and the number of eligible people for medicare. This may indicate that the number of deaths due to COVID-19 in the later phase is more related to the overall healthcare resources and general health condition of the population in each county.

Additional benefits

The LSTM-based structure is scalable and flexible. We were limited to a few available data sources during the initiative and have not expanded actively to other data sources that possibly provide additional information for predictions. Testing the robustness of this architecture was the original goal of this pilot study. Once there are additional data sources, the condLSTM-Q could adapt and be retrained to extract information from the new data source without adjusting the architecture. For instance, if one hypothesizes that the temperature affects the death trend, the model can be easily retrained and tested with an additional time-series feature of temperature added to the input data stream. These untapped variations of parameters would be interesting follow-ups for future works.

Long-term robustness of the model and its potential limitations

We were interested in the robustness of the original model's long-term performance. Many factors changed over time that were not our original inputs that later became relevant to the pandemic—for instance, new effective treatment and updated guidelines for in-hospital patients [21]. Thus, we had the model updated with a daily data stream without changing the structure or input variables until mid-June 2021. As shown in the aggregated national prediction (Fig. 9), the model's

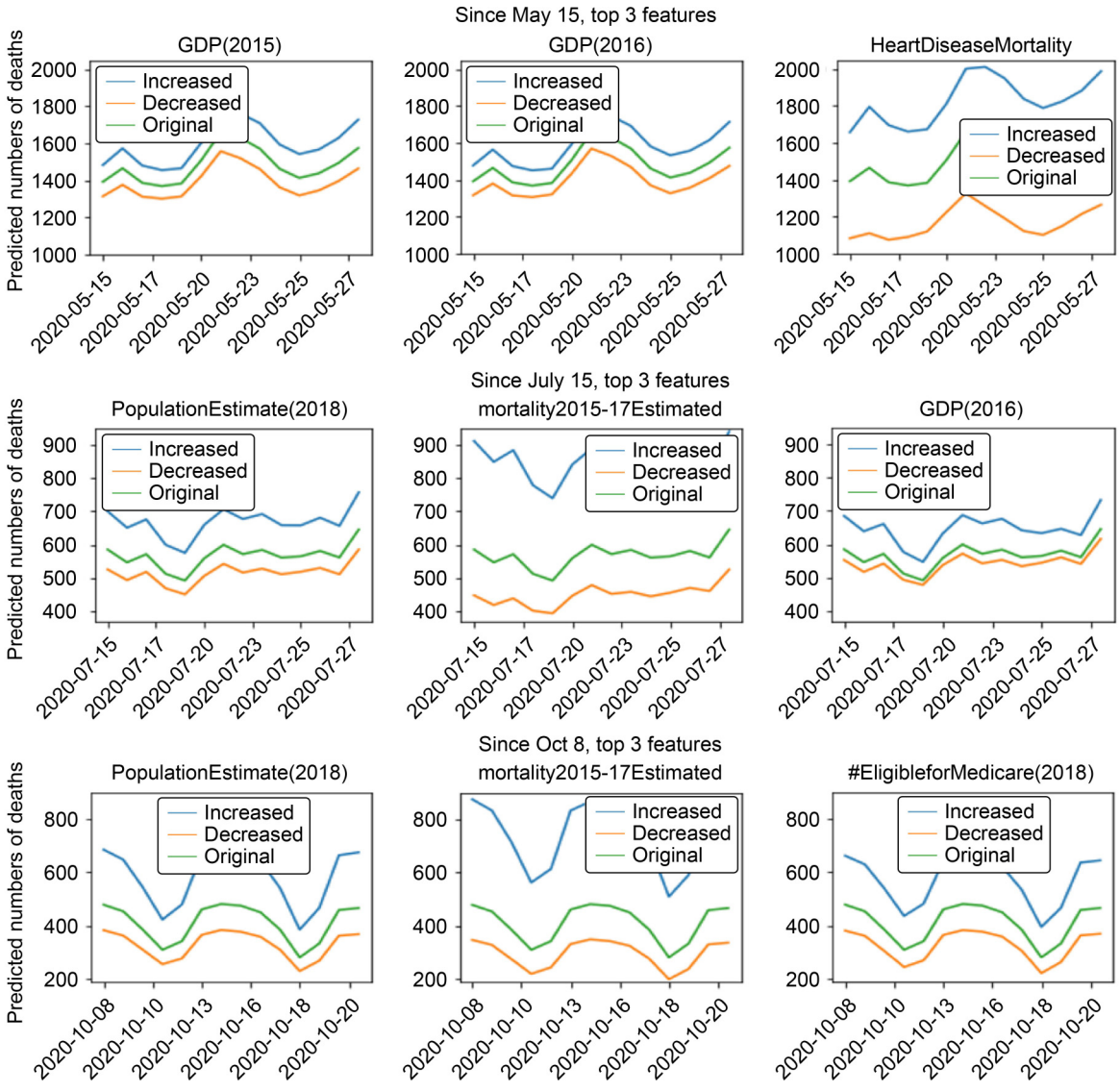

Figure 8. Changes in nationwide prediction after modifying categorical features. For each model making predictions over 2 weeks from May 15 (top), July 15 (middle), and October 8, 2020 (bottom), new predictions after increasing and decreasing a categorical feature are shown in blue and orange, along with the original prediction in green. Each altered feature has been selected based on permutation feature importance from each of the models, and features were either increased or decreased by 3 standard deviations. The values shown in this figure are the predicted numbers of deaths in the United States.

mortality prediction exceeds the actual data significantly at the beginning of 2021. The performance became stable again post mid-February 2021.

What is the root cause of the discrepancy? The U.S. started administrating vaccines in December 2021[22]. The large error at the beginning of 2021 could have resulted partially from the lack of this variable in the model. However, given that the U.S. vaccination roll-out during January was still limited (< 4 percent of the total population received at least one dose by the January 22, 2021 [23]), this indicated that vaccination as well as other factors, such as the updated treatment guidelines in hospitalized patients as mentioned above [21]. This provides some insights for future works in applying deep learning methods to tackle disease prediction modeling, and suggests possible ways to prevent a potential reduction in model performance when the input-output relationship has changed over time due to new interventions. First, one might leave out the training data that seems obsolete to the most current predictions. In our case, for example, the input-output mapping learned prior to the first and second wave would inevitably lead to a higher death number after the third wave. Since training the networks with the obsolete data would affect the more present performances, removing the early data might be beneficial, although it would reduce the size of

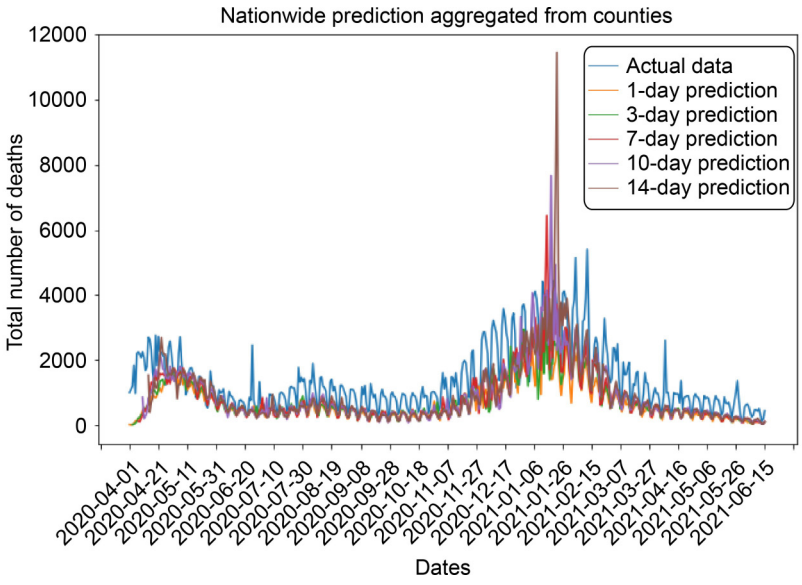

Figure 9. Nationwide prediction on the total number of deaths, extended to June 15, 2021. Different colors labels different prediction intervals as in Fig. 2. The model's prediction overshoots in the beginning of 2021, especially for the long-term predictions. After a few weeks, the model could learn the new input-output mapping and predicted the mortality fairly reasonable.

the training data. Second, one can identify additional critical contributing variables and have the model retrained on all these variables from the pandemic onset. For instance, if a highly effective medication was identified and distributed overnight, it would immediately lower the mortality rate. Models that did not incorporate this new variable and retrained accordingly would inevitably not be able to capture the variance. Using either of these methods would make the model learn the new input-output relationship, either through a fresher data set or with additional features that capture the changes, whether it decreases (effective medications, vaccines, contact tracing/quarantine technology became available) or increases (novel virus strains, hospital overload) the mortality. The aforementioned techniques would be important follow-ups for researchers interested in empowering the nascent deep learning approaches for epidemiology modeling and is beyond the scale of this manuscript for now.

MATERIALS AND METHODS

Data sources and data preprocessing

Our model includes a variety of data that is expected to have correlations with the death toll. The original source of each data is given in Section of Material Availability. The data used in our model fall into one of the following seven categories: (1) COVID-19 mortality and confirmed cases provided by *The New York Times*; (2) demographics and local health resources such as age composition, mortality rate by diseases, and the number of hospitals, provided by the Yu group at University of California, Berkeley; (3) county-wise gross domestic product (GDP) from the Bureau of Economic Analysis; (4) population density and geographical data from 2010 Census; (5) mobility changes in response to COVID-19 provided by Descartes Labs; (6) policy actions in response to COVID-19 such as state of emergency declaration, safe-at-home order, and business closure, from Covidvis team at University of California, Berkeley [24] and the U.S. Department of Health and Human Services; (7) and the U.S. pneumonia and influenza mortality report from the National Center for Health Statistics Mortality Surveillance System in Centers for Disease Control and Prevention (CDC). All the data were county-level, and the latest start date of the time series data was March 1, 2020.

Preprocessing specifics for each dataset are listed in the following:

• Mobility data from the Descartes Labs was provided for 2,721 counties out of 3,114 counties in total. The missing 393 counties' mobility data were filled in with their corresponding state-level data. Missing dates in mobility data of Descartes Labs were interpolated with the data on the closest existing date of the same day in a week to reflect the weekly pattern inherent in the data. As of November 8, 2020, 14,049 values were missing, out of total 688,413 data points (about 2%). They were filled in using either the aforementioned same-day interpolation or spline interpolation, and the data from the first and the last day of recording was repeated to fill

the missing data before and after the existing data, if necessary. Spline interpolation has an advantage over the same-day interpolation in handling a large number of missing data over a long period of time, whereas the latter can capture the weekly pattern that cannot be maintained in the former method.

• Seasonality feature was extracted from the U.S. pneumonia and influenza data, under the premise that the COVID-19 will follow a seasonality of virus that becomes more dormant during the summer and severe in the winter. More specifically, the multiplicative seasonality was extracted from the state-wide pneumonia and influenza mortality rate during flu seasons from 2013 to 2020 provided by CDC [25].

• Among 64 features of the demographics and local health data, 43 features including the population estimate and the mortality rate of various underlying diseases were selected as the static features.

• For the policy actions features, declaration of the state of emergency, safe-at-home action, and the closure of inessential business were selected as the static features.

All numerical values were standardized before the model training step.

Implementation of condLSTM-Q

The classical LSTM which the condLSTM-Q's backbone was based on is shown in Fig. 10A. LSTM is a type of recurrent neural network (RNN) that can be trained to learn the mapping from the time series features to the time series of interest. Compared to previous architectures of RNN, LSTM has the advantage of remembering long-term dependencies by passing the information through its cell state and hidden states recurrently. However, the classical LSTM architecture cannot take into account non-time series data (hereinafter referred to as categorical data) in a natural way; such features have to be stacked into the same dimension of the time series data during data processing as shown in Fig. 10A, in order to take both types of data into the input stream. This way of data processing undermines the optimal performance of LSTM.

One approach to overcome this limitation is to initialize the initial states of the LSTM units in response to the categorical data. Intuitively, the categorical data should provide a priori information to the prediction, rather than real-time information. Thus, a "conditional" module can be added to get these categorical data and feed in "priors" as the hidden states, as implemented in [9], in contrast to the usual LSTM where the hidden states are initialized to zeros or random noise [26].

This is the way we treat the categorical data in our model, conditional LSTM with Quantile output model (condLSTM-Q; Fig. 10B). In this model, categorical inputs are passed through a fully connected layer into the model as the hidden states to the initial step of the LSTM layer. Through this design, we could input 50 categorical data such as income, age and gender distribution, medicare coverage, and population in their original forms of one scalar per feature into our model. After the initialization step, the cell states and hidden states in LSTM were then updated as in the classical LSTM, based on 8 time series features including mobility of the population, new cases per day, and new deaths per day.

Another feature of condLSTM-Q is that it generates predictions on 10-quantiles ("Q"). To get the sense of the distribution of the predicted values, which help us understand the situation and the quality of the prediction better (a forecast with higher variance results in huge uncertainty in whatever derivatives from the forecast), our model forecasts the death counts in each quantile from 0.1 to 0.9 with 0.1 increments by a variant of the multi-output LSTM. In this model, the outputs from the LSTM layer are fed into 9 parallel dense layers to generate a vector of dimension 9 for each day in the prediction, where each component of the output vector is a forecast of each quantile. A standard measure for the accuracy of such a quantile forecasting is the pinball loss defined as

$$loss = \max(q \cdot \epsilon, (q-1) \cdot \epsilon), \tag{1}$$

where $0 < q < 1$ is the quantile and $\epsilon = y - \hat{y}$ is the difference between the true target value y and the predicted target value $\hat{y}$. As such, our model is trained to minimize the average pinball loss

$$\frac{1}{9}\sum_{i=1}^{9} \max(q_i \cdot \boldsymbol{\epsilon}_i, (q_i - 1) \cdot \boldsymbol{\epsilon}_i), \tag{2}$$

where $q_i = i/10, 1 \leqslant i \leqslant 9$ and $\boldsymbol{\epsilon}_i = \boldsymbol{y} - \hat{\boldsymbol{y}}_i$ is the difference between the true target vector and the predicted target vector for quantile q_i.

Hyperparameters and the training

The time series data after preprocessing is an array of shape (#counties, #dates, #features), and the categorical data is an array of shape (#counties, #features). For each date in the time series data, the data is further split into a history window of size 7 and a target window of size 14, so that condLSTM-Q can learn to predict the mortality rate for the next 14 days based on the history of prior 7 days.

For hyperparameter tuning, we held out the last 21 days' worth of data for validation. As of August 4, 2020, we used 136 days' worth of data until July 14, 2020, for

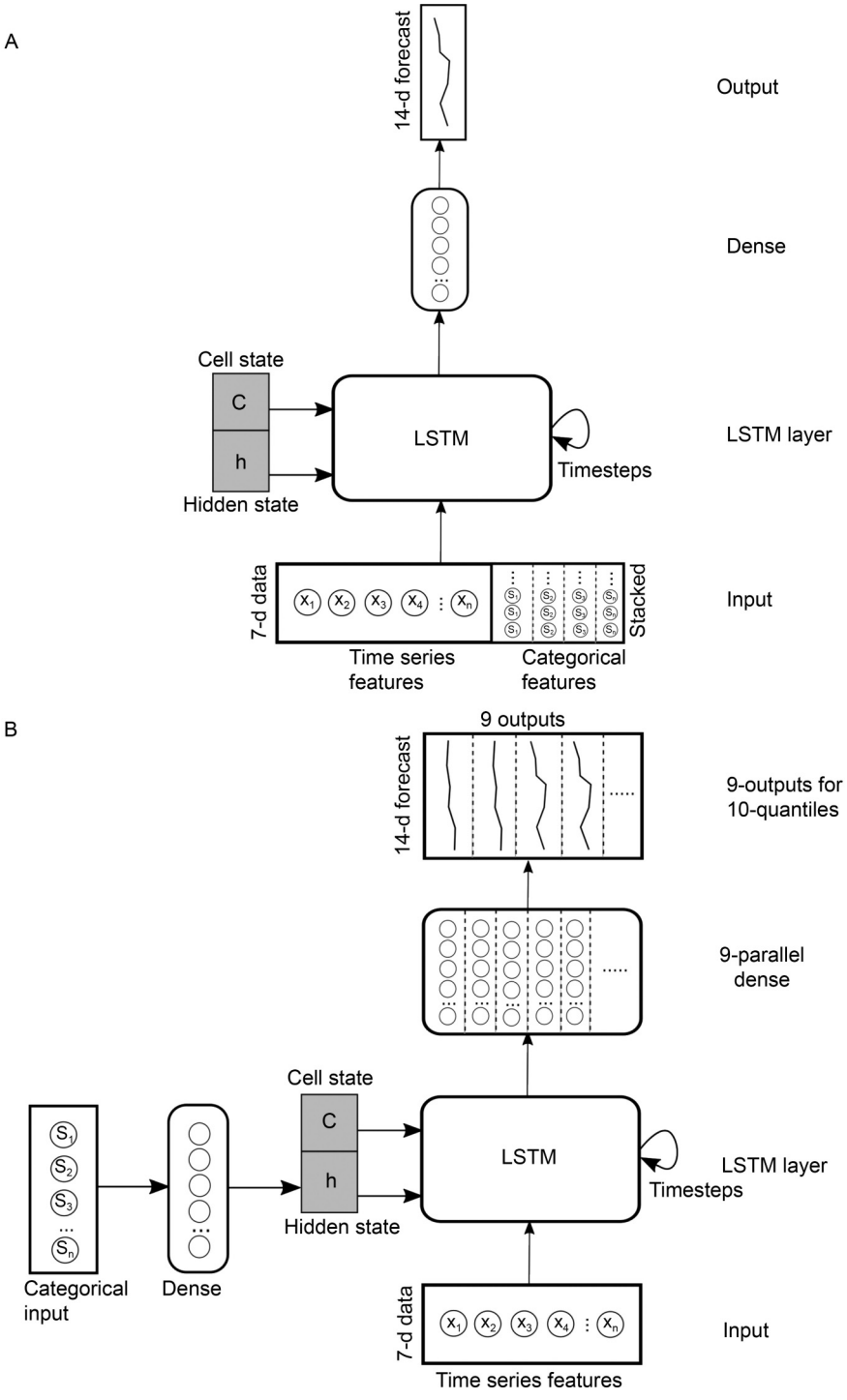

Figure 10. Classical LSTM vs condLSTM-Q Architecture. (A) Classical LSTM architecture. The classical LSTM can predict time series better than the traditional RNN by remembering and passing cell states and hidden states over the timesteps. The hidden states are usually initialized to zero or random values. Categorical features have to be stacked and replicated to the same dimension to match the time series features. The classical structure outputs one time series. (B) The condLSTM-Q architecture. In condLSTM-Q, categorical inputs are passed through a fully connected layer as hidden states to the initial step ("Conditional" based on the static information) of the LSTM layer. A dropout layer is also applied to the dense layer to prevent overfitting to the categorical inputs. The output from the LSTM layer is fed into 9 parallel dense layers to provide 9 outputs for 10-quantiles ("Q"). Each of the 9 parallel outputs is given to the pinball loss function during training, thus the parallel dense layer gets updated to match the corresponding quantiles by minimizing the summed loss.

the training set, and the remaining 21 days' worth of data for the validation set. Since the data was from 3,114 counties, and the input and the target for the model should be 7 and 14 days long, this led to 316,224 samples for training and 3,114 samples for validation.

The optimal hyperparameters selected were 128 units in LSTM, a learning rate of 0.001, and a dropout rate of 0.2. In this setting, overfitting was observed after 30 epochs, so the model was trained with the ADAM optimizer [27] for 20 epochs to avoid overfitting.

Material availability

Code and Model for our study are available in:
https://github.com/cjackal/COVID-SKTW

Demographics and Health Resource dataset from the Yu Group at UC Berkeley:
https://raw.githubusercontent.com/Yu-Group/covid19-severity-prediction/data/county_data_abridged.csv

COVID-19 mortality dataset from *The New York Times*:
https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

Local area GDP dataset from U.S. Bureau of Economic Analysis (BEA):
https://www.bea.gov/news/2019/local-area-gross-domestic-product-2018

Land area and Population density dataset from 2010 Census:
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk

Mobility dataset from DescartesLabs:
https://raw.githubusercontent.com/descarteslabs/DL-COVID-19/master/DL-us-mobility-daterow.csv

COVID-19 related policies dataset from HHS:
https://healthdata.gov/dataset/COVID-19-State-and-County-Policy-Orders/gyqz-9u7n

ACKNOWLEDGEMENTS

The authors thank Prof. Yaser Abu-Mostafa, and the Teaching Assistants of CS156 in Caltech for organizing the COVID19 prediction initiative and for providing the data pipeline for parsing data sources. We thank Isaac Yen-Hao Chu, M.D. for reading the manuscript. Yu-Li Ni was supported by Taipei Veterans General Hospital Yang-Ming University Excellent Physician Scientists Cultivation Program (No.103-Y-A-003).

COMPLIANCE WITH ETHICS GUIDELINES

The authors HyeongChan Jo, Juhyun Kim, Tzu-Chen Huang, and Yu-Li Ni declare that they have no conflict of interest or financial conflicts to disclose. All procedures performed in studies involving animals were in accordance with the ethical standards of the institution or practice at which the studies were conducted, and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

REFERENCES

1. IHME COVID-19 health service utilization forecasting team, Murray, C. J. M. (2020) Forecasting COVID-19 impact on hospital bed-days, ICU-days, ventilator-days and deaths by US state in the next 4 months. medRxiv, doi: https://doi.org/10.1101/2020.03.27.20043752
2. Abu-Mostafa, Y. S. (2020) Caltech CS156 COVID-19 Model. https://cs156.caltech.edu/. Accessed: Accessed: February 1, 2021
3. Khan, N., Shahid, S., Juneng, L., Ahmed, K., Ismail, T. and Nawaz, N. (2019) Prediction of heat waves in Pakistan using quantile regression forests. Atmos. Res., 221, 1−11
4. Brennan, A., Cross, P. C. and Creel, S. (2015) Managing more than the mean: using quantile regression to identify factors related to large elk groups. J. Appl. Ecol., 52, 1656–1664
5. IHME. (2020) Institute for health metrics and evaluation COVID-19 estimate. http://www.healthdata.org/covid/data-downloads. Accessed: February 1, 2021
6. Hochreiter, S. and Schmidhuber, J. (1997) Long short-term memory. Neural Comput., 9, 1735–1780
7. Pal, R., Sekh, A. A., Kar, S. and Prasad, D. K. (2020) Neural network based country wise risk prediction of COVID-19. Appl. Sci. (Basel), 10, 6448
8. Melin, P., Monica, J. C., Sanchez, D. and Castillo, O. (2020) Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: The case of Mexico. Healthcare (Basel), 8, 181
9. Rémy, P. (2020) Conditional RNNs made easy with Tensorflow and Keras. https://github.com/philipperemy/cond_rnn. Accessed: February 1, 2021
10. Karpathy, A. and Li, F. (2015) Deep visual-semantic alignments for generating image descriptions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3128−3137
11. Breiman, L. (2001) Random forests. Mach. Learn., 45, 5−32
12. Fisher, A., Rudin, C., and Dominici, F. (2019) All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn Res., 20, 177
13. Analytics, COVID (2020) DELPHI Model. https://www.covida-

nalytics.io/projections/. Accessed: February 1, 2021
14. Los Alamos National Laboratory. (2020) LANL Model. https://covid-19.bsvgateway.org/. Accessed: February 1, 2021
15. Tan, W. and Aboulhosn, J. (2020) The cardiovascular burden of coronavirus disease 2019 (COVID-19) with a focus on congenital heart disease. Int. J. Cardiol., 309, 70–77
16. Chen, R., Liang, W., Jiang, M., Guan, W., Zhan, C., Wang, T., Tang, C., Sang, L., Liu, J., Ni, Z., *et al.*, (2020) Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. Chest, 158, 97–105
17. Chen, Y., Gong, X., Wang, L. and Guo, J. (2020) Effects of hypertension, diabetes and coronary heart disease on COVID-19 diseases severity: a systematic review and meta-analysis. medRxiv, doi: 10.1101/2020.03.25.20043133
18. Chen, C., Chen, C., Yan, J. T., Zhou, N., Zhao, J. P. and Wang, D. W. (2020) Analysis of myocardial injury in patients with COVID-19 and association between concomitant cardiovascular diseases and severity of COVID-19. Zhonghua Xin Xue Guan Bing Za Zhi (in Chinese), 48, 567–571
19. Zhang, Y., Tian, H., Zhang, Y., and Chen, Y. (2020) Is the epidemic spread related to GDP? Visualizing the distribution of COVID-19 in Chinese mainland. arXiv, 2004.04387
20. U.S. Bureau of Economic Analysis (BEA). (2020) Local area gross domestic product, 2018. https://www.bea.gov/system/files/2019-12/lagdp1219.pdf. Accessed: February 1, 2021
21. National Institutes of Health. (2019) COVID-19 treatment guidelines panel. Coronavirus disease 2019 (COVID-19) treatment guidelines. National Institutes of Health. https://www.covid19treatmentguidelines.nih.gov/. Accessed: February 1, 2021
22. U.S. Food and Drug Administration Pfizer-BioNTech COVID-19 Vaccine. https://www.fda.gov/emergencypreparedness-and-response/coronavirus-disease-2019-covid-19/pfizer-biontech-covid-19-vaccine. Accessed: February 1, 2021
23. Centers for Disease Control and Prevention. COVID-19 vaccinations in the United States. https://covid.cdc.gov/covid-data-tracker/vaccinations. Accessed: February 1, 2021
24. The Covidvis team. (2020) Covidvis. https://covidvis.berkeley.edu/. Accessed: February 1, 2021
25. Centers for Disease Control and Prevention. (2020) Pneumonia and influenza mortality surveillance from the national center for health statistics mortality surveillance system. https://gis.cdc.gov/grasp/fluview/mortality.html. Accessed: February 1, 2021
26. Zimmermann, H.-G., Tietz, C. and Grothmann, R. (2012) Forecasting with recurrent neural networks: 12 tricks. In Neural Networks: Tricks of the Trade. pp. 687–707. Springer
27. Kingma, D. P. and Ba, J. (2017) Adam: A Method for Stochastic Optimization. arXiv, 1412.6980