



OPEN

An interpretable hybrid predictive model of COVID-19 cases using autoregressive model and LSTM

Yangyi Zhang¹, Sui Tang¹✉ & Guo Yu²✉

The Coronavirus Disease 2019 (COVID-19) has had a profound impact on global health and economy, making it crucial to build accurate and interpretable data-driven predictive models for COVID-19 cases to improve public policy making. The extremely large scale of the pandemic and the intrinsically changing transmission characteristics pose a great challenge for effectively predicting COVID-19 cases. To address this challenge, we propose a novel hybrid model in which the interpretability of the Autoregressive model (AR) and the predictive power of the long short-term memory neural networks (LSTM) join forces. The proposed hybrid model is formalized as a neural network with an architecture that connects two composing model blocks, of which the relative contribution is decided data-adaptively in the training procedure. We demonstrate the favorable performance of the hybrid model over its two single composing models as well as other popular predictive models through comprehensive numerical studies on two data sources under multiple evaluation metrics. Specifically, in county-level data of 8 California counties, our hybrid model achieves 4.173% MAPE, outperforming the composing AR (5.629%) and LSTM (4.934%) alone on average. In country-level datasets, our hybrid model outperforms the widely-used predictive models such as AR, LSTM, Support Vector Machines, Gradient Boosting, and Random Forest, in predicting the COVID-19 cases in Japan, Canada, Brazil, Argentina, Singapore, Italy, and the United Kingdom. In addition to the predictive performance, we illustrate the interpretability of our proposed hybrid model using the estimated AR component, which is a key feature that is not shared by most black-box predictive models for COVID-19 cases. Our study provides a new and promising direction for building effective and interpretable data-driven models for COVID-19 cases, which could have significant implications for public health policy making and control of the current COVID-19 and potential future pandemics.

The coronavirus disease 2019 (COVID-19) pandemic has posed a severe threat to global health and economy while producing some of the richest data we have ever seen in terms of infectious disease tracking. The quantity and quality of data placed epidemic modeling and forecasting at the forefront of worldwide public policy making. Compared to previous infectious diseases, COVID-19 shows special transmission characteristics, yielding significant fluctuations and non-stationarity in the new COVID-19 cases. This poses grand challenges in effective prediction, and, on the other hand, draws attention of the global community to epidemic tracking and prediction.

In the last three years, various models and methods have been developed to predict COVID-19 cases (see survey in¹ and references therein). These models can be roughly grouped into two categories: mechanistic models and data-driven models. The mechanistic models aim at directly characterizing the underlying mechanisms of COVID-19 transmission. Typical examples of mechanistic models are based on differential equations, such as the compartmental models SIR and SEIR^{2–5}. The data-driven models formulate the prediction of the COVID-19 cases primarily as a regression problem and exploit fully data-adaptive approaches to understand the functional relationship between COVID-19 cases with a set of observable variables. Data-driven models include classical statistical models such as Autoregressive models (AR)^{6–8}, Support Vector Machine (SVM)^{9–11}, and the deep learning models^{12–18}. In this paper, we will focus on data-driven models.

An Autoregressive model expresses the response variable as a linear function of its previous observations¹⁹. Its simple structure and strong interpretability are found to be powerful in capturing short-term changing trends in time series. AR models have been applied in various application fields, including infectious disease modeling^{20,21}. However, they may fail to capture the highly nonlinear patterns and long-term effects in the data-generating

¹Department of Mathematics, University of California Santa Barbara, Santa Barbara, CA 93106, USA. ²Department of Statistics and Applied Probability, University of California Santa Barbara, Santa Barbara, CA 93106, USA. ✉email: suitang@ucsb.edu; guoyu@ucsb.edu

dynamics. On the other extreme of the predictive model complexity spectrum, deep learning models, particularly LSTM²², have demonstrate impressive power in capturing complex dependence structures in sequential data. LSTM has been used to achieve the best-known results for many problems on sequential data. However, a well-known limitation of the deep learning models is the short of interpretability due to their black-box nature. This lack of interpretability prevents people from drawing useful conclusions from the model outputs, thus hinders effective policy making²³, especially in crucial fields such as public health. This observation motivates us to consider a hybrid model in which the two seemingly distinct types of models join forces while maintaining both good predictive power and certain interpretability.

In this paper, we propose such a hybrid model that *additively combines* the LSTM and the AR model for the task of COVID-19 cases prediction. The proposed hybrid model is formalized as a neural network with an architecture that connects an AR model and a LSTM block, and the relative contribution of these two component models is decided in a totally data-adaptive way in the training procedure. To demonstrate the predictive power of the proposed hybrid model, we consider both county-level and country-level data. Specifically, in 8 counties in the state of California, USA, and 7 other different countries (results available in the Supplementary Material), our method performs favorably compared with either AR or LSTM model alone, as well as other commonly-used predictive models under various evaluation metrics. All codes are accessible through links on the reference page²⁴.

In addition to the predictive accuracy, the importance of predictive models' interpretability has been discussed in plenty of previous works^{23–27}. A higher model interpretability facilitates human's ability to understand its predictions, and thus promotes bias detection and other factors that contribute to policy making. Specifically, we demonstrate how the coefficients from the AR part of the trained hybrid model shed light onto understanding the underlying disease transmission mechanism, and thus could help to predict its prevalence trends, and to inform public health policy makers to improve pandemic planning, resource allocation, and implementation of social distancing measures and other interventions. A long-term mission of this paper is to stretch the application of hybrid models beyond COVID-19 forecasting: toward other fast-moving epidemics and cases that require accurate prediction and interpretability.

Although in this paper we focus on confirmed cases prediction, we note that the proposed framework can be easily extended to tackle other COVID-19 or more general epidemiological tasks (e.g., hot spot prediction). Furthermore, the proposed method has its own research significance from a methodological perspective. For example, it raised the open questions on studying its theoretical guarantees, mathematical quantification of prediction, and interpretability.

Related work. Recently, numerous studies have employed machine learning techniques to investigate various tasks on COVID-19 and achieved impressive results. Examples include using deep learning to detect COVID-19 through CXR images and predicting death status based on food categories to recommend healthy foods during the pandemic^{28–30}. In light of these advances, our research focuses on predicting confirmed cases of COVID-19.

In this section, we provide a more detailed review of data-driven models that formulate the prediction problem as a regression problem. Regression-based models, including simple AR models and more complex models such as Random Forest, Gradient Boosting, and CNN-LSTM, have been widely used for COVID-19 prediction. For example, Mumtaz et al.³¹ used ARIMA to predict the daily confirmed cases in European countries, while Yesilkanat³² used a Random Forest model to predict the number of cases and deaths. Muhammad et al.³³ used a CNN-LSTM model to predict the number of confirmed cases and deaths in Nigeria, South Africa, and Botswana. We summarize a list of recent work from year 2020 to 2022 in Table 1.

One advantage of these models is that they do not require a priori knowledge of the disease dynamics and can capture rich relationships in the data. They have been shown to be effective in predicting COVID-19 cases in various regions around the world. However, COVID-19 data displays rich variability, and therefore a single predictive model may not be sufficient and has its own limitations. For example, one major disadvantage of ARIMA models is that they may not be able to capture non-linear patterns in the data, which can lead to inaccurate predictions. On the other hand, more complex models such as Random Forest and CNN-LSTM may suffer from overfitting, where the model becomes too specialized to the training data and cannot generalize well to new data. These complex models may also lack interpretability, making it difficult to understand the factors driving the predictions and thus provide little to none guidance to actual public health policy making.

Hybrid predictive models that combine different regression models may offer the best of both worlds by capturing both linear and non-linear patterns in the data while maintaining some degrees of interpretability. The idea is to decompose a model into different components that are designed to capture specific characteristics of the data. It has proven to be an effective way of improving empirical predictions in various applications, including those in COVID-19 prediction^{34–38}.

Comparison to previous works on hybrid modeling. The idea of using an additive combination of AR models (or more generally, ARIMA models) with LSTMs has recently appeared in the literature for time series forecasting, with applications in gas and oil well production and sunspot monitoring^{39,40}. However, there is a significant difference between our approach and previous methods: our approach trains the two components in the model jointly, while previous hybrid modeling techniques take a sequential approach to training. Specifically, Zhang⁴¹ proposed a hybrid model of ARIMA and artificial neural networks, aiming to capture more patterns in the data and improve forecasting performance. The preprocessed data is used to fit an ARIMA model first, before the residual term is used as input to train a neural network model. Fan et al.³⁹ followed a similar procedure, using an ARIMA model and an LSTM model. The logistics of these methods is to use an ARIMA model to capture the linear pattern of the data first and rely on the neural networks capture the non-linear pattern in the residuals.

Study	Dataset	Models	Results
Atik ⁴²	Turkey	Support Vector Machine (SVM), Linear Regression, Bagged Tree, Fine Tree	SVM has the highest R value (99%) and the lowest RMSE and MAE values.
Galasso et al. ⁴³	USA counties	Random Forest	MAE below 300 weekly cases/100,000, using easily accessible training features
Ali et al. ⁴⁴	Saudi Arabia	Gradient Tree Boosting, Random Forest, XGBoost, Voting Regressor	The XGBoost and Voting Regressor model outperform the other models
Chumachenko et al. ⁴⁵	Germany, Japan, South Korea, and Ukraine	Random Forest, K-Nearest Neighbors, Gradient Boosting	The Gradient Boosting model has the best performance by related error and MAE
Fang et al. ⁴⁶	USA	ARIMA, XGBoost	The XGBoost model outperforms the ARIMA model on COVID-19 case prediction in the USA
Muhammad et al. ⁴⁷	Mexico	Logistic Regression, Decision Tree, SVM, Naive Bayes, and Artificial Neural Network	Decision tree model has the highest accuracy (94.99%). SVM model has the highest sensitivity (93.34%). Naive Bayes model has the highest specificity (94.30%).
Luo et al. ⁴⁸	America	double-layer LSTM, XGBoost	MAPE of LSTM and XGBoost reaches 2.32% and 7.21%
Vadyala et al. ⁴⁹	USA, State of Louisiana	LSTM, SEIR	RMSE of LSTM reaches 601.20 while the SEIR model has an RMSE of 3615.83 (on 3 Parishes of Louisiana)
Alassafi et al. ¹⁵	Malaysia, Morocco and Saudi Arabia	LSTM, RNN	LSTM models reaches a 98.58% precision accuracy while the RNN models have 93.45%
Tomar and Gupta ⁵⁰	India	LSTM	LSTM's prediction closely matches the official data. By the prediction of the LSTM, social isolation and lockdown can reduce the spread of virus effectively.
Bhandari et al. ⁵¹	India	ARIMA	ARIMA(3,3,1) model has the best performance for cumulative cases and death(predicted RMSE is 668.7 and base RMSE is 5431).

Table 1. A non-exhaustive list of previous works on data-driven models for COVID-19 cases prediction in the past three years. Most commonly used evaluation metrics are RMSE, MAE, and MAPE.

The main goal of these previous works is to explore whether a hybrid model produces better performance than the single models.

In our study, we design a general network architecture that includes both an AR part and an LSTM part additively and trains the entire architecture jointly by minimizing the empirical risk. By doing so, we do not arbitrarily give preference to any of the two additive components. Instead, the relative weights of the interpretable AR part and the predictive LSTM part are determined fully by the data.

In summary, our contributions can be summarized as:

- Development of a novel approach to hybrid modeling for COVID-19 cases prediction: we have designed a general network architecture that combines AR and LSTM models additively and trains the entire architecture jointly, allowing the relative weights of the interpretable AR part and the predictive LSTM part to be fully determined by the data. This approach is a departure from traditional sequential modeling approach and has the potential to contribute to the literature of sequential data prediction.
- Extensive numerical studies on data sets from two sources that displays a rich variety of variability: we have shown that the proposed hybrid model demonstrates better forecasting performance than single models. This finding is important as it shows that the hybrid model is an effective way to combine the strengths of different modeling techniques and can be used as a framework for future research.
- Exploration of interpretability: we have also explored the interpretability of the hybrid model, which is an important contribution as it allows for a better understanding of the model and can lead to improved decision-making based on the model's output. This contribution enhances the practical applicability of our proposed hybrid model.

Methods

In this section, we first overview the two building blocks of our additive hybrid model, namely the AR and the LSTM model, and their relative advantages. Then we present our hybrid model which combines these two building blocks additively, and we intuitively elaborate why it is better than the two individual components.

Autoregressive (AR) models. In time series, we often observe associations between past and present values. For example, by knowing the price of a stock in the past few days, we can often make a rough prediction about its value tomorrow. AR is a simple model that utilized this empirical observation and can yield very accurate prediction in certain applications. It represents the time series values using linear combination of the past values. The number of past values used is called the lag number and often denoted by p . Let ϵ_t denote the Gaussian noise at time t with mean 0 and variance σ^2 . The structure equation of AR(p) model can be represented as

$$Y_t = a_0 + a_1 Y_{t-1} + a_2 Y_{t-2} + \cdots + a_p Y_{t-p} + \epsilon_t \quad (1)$$

where a_0 is the intercept, and $a_1, \dots, a_p$ represent the coefficients. AR model is often effective on stationary data. To ensure stationarity, a common trick is to apply the differencing operation on the time series. A time series value at time t that has been differenced once, $Y^{(1)}$, is defined as follows:

$$Y_t^{(1)} = Y_t - Y_{t-1}, \quad (2)$$

and higher order differencing operation can be defined recursively. However, an AR model is not sufficient to capture the non-linear dependence structure, which is found to be an important feature of the COVID-19 data, indicated by Fig. 1. A purely AR based model is thus often insufficient for the task of COVID-19 cases prediction.

Long short term memory networks (LSTM). RNN (Recurrent Neural Network)⁵² is known to suffer from the long term dependency problem: as the network grows larger through time, the gradient decays quickly during back propagation, making it impossible to train RNN models with long unfolding in time. To solve this problem, Hochreiter and Schmidhuber (1997) introduced a special type of RNN called LSTM with a proper gradient-based learning algorithm²².

We employ a LSTM regression model, which is represented as

$$Y_t = G_\theta(Y_{t-1}, \dots, Y_{t-p}), \quad (3)$$

where we use $Y_{t-1}, \dots, Y_{t-p}$ as the sequential input data; G represents the neural network architecture shown in Supplementary Fig. 1 and θ represents the weight parameters in neural networks.

The core concepts of a LSTM cell are the cell states and the associated gates, as illustrated in Supplementary Fig. 2. The cell state C_{t-1} at time step $t-1$ acts as a transport highway that transfers relative information all the way down the sequence chain, which intuitively characterizes the “memory” of the network. The cell states, in principle, carry relevant information throughout the processing of the sequence. So even information from the earlier time steps can make its way to later time steps, reducing the effects of short-term memory. The Forget Gate decides what information should be kept. The Input Gate decides what information is to be added from the current step and update the cell state C_t at time step t . The Output Gate determines what the next hidden state h_t should be. The four gates comprise a fully connected feed forward neural network.

To achieve optimal prediction results using LSTM model, it is crucial to have a careful hyperparameter tuning, including the choice of units (dimension of the hidden state), the number of cells (i.e. the number of time steps), and layers. This is usually a difficult task in practice. For example, few LSTM cells are unlikely to capture the structure of the sequence, while too many LSTM cells might lead to overfitting. However, just like other neural networks, a well-known limitation of LSTM is its lack of interpretability²³.

The hybrid model. As discussed above, both AR and LSTM have their relative strength and limitations in their prospective domains. We propose to combine the two models additively into one single hybrid model, which is expressed as

$$Y_t = \alpha \text{AR}(p) + (1 - \alpha) G_\theta(Y_{t-1}, \dots, Y_{t-p}), \quad (4)$$

where p is the lag number and α weights the contribution of two components: by tuning the value of α , one can strike a balance between the prediction given by AR and LSTM parts, and thus a prediction of linear and nonlinear signals.

We illustrate the structure of the hybrid model in Fig. 2. The hybrid model is characterized as one neural network architecture where the two composing models are added through the last layer. The AR component captures the linear relationship in time series and the LSTM component would describe the nonlinear patterns. In section “Training” of the Supplementary Material, we show how to train the weights in each of the two components in a fully data-adaptive manner by minimizing the empirical risk. We will compare the contribution of the hybrid model’s AR component and LSTM component in section Results.

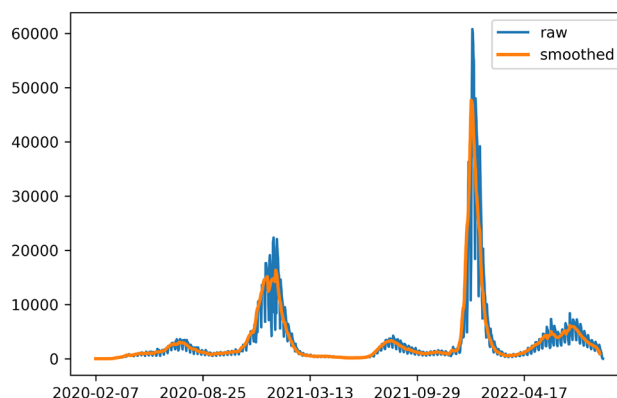


Figure 1. An example of visualizing daily observations, where blue line represents the data before smoothing, orange line represents data after smoothing. The data is collected from the Los Angeles county.

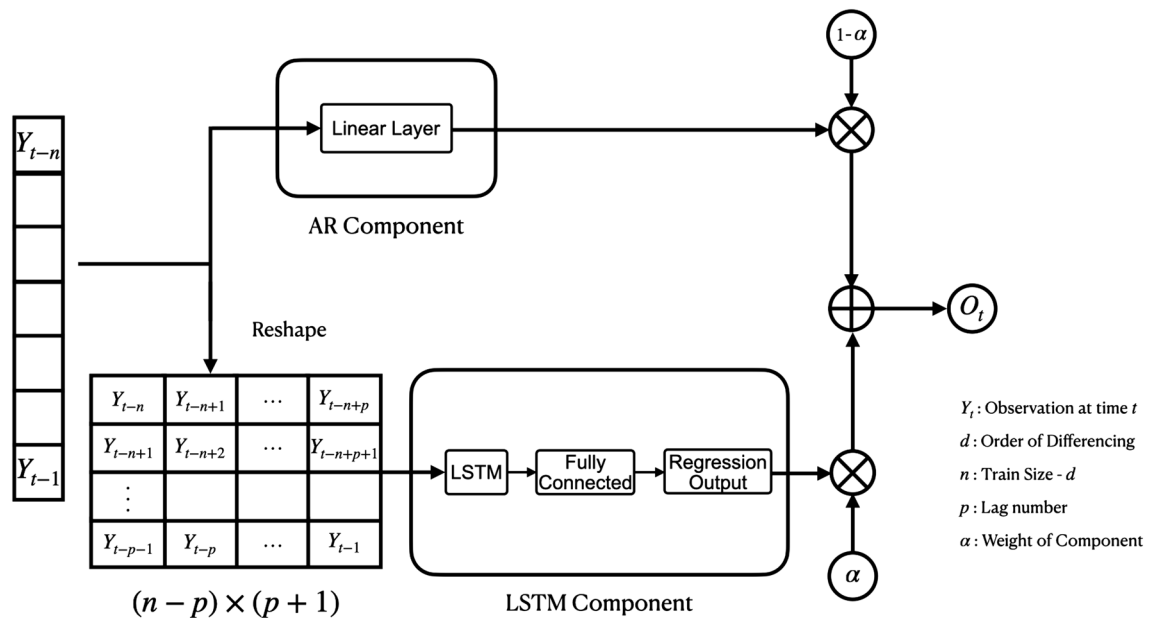


Figure 2. Visualization of the hybrid architecture.

Results

The results include four sections: [Model evaluations](#), [Prediction](#), [Interpretability](#), and [Comparative study on the WHO datasets](#). In [Model evaluations](#), we introduce the metrics we use to evaluate the models and on which we compare the models' performances. In [Prediction](#), we exhibit the visualizations of several interesting trials and compare the numerical predictions and evaluations of the three models. In [Interpretability](#), we compare the AR component of the hybrid model with the AR model. This is to examine how we may interpret the hybrid model. We leave other training details in Supplementary Material. In [Comparative study on the WHO datasets](#), we further examine the performance of the proposed hybrid model by applying it to data of 7 different countries around the world and comparing its performance with that of its component models and 3 additional models.

Data description and statistical analysis. We utilize two primary data sources. The first data source is a dataset specific to California counties, which is available in the CHHS Open Data repository under the title [COVID-19 Time-Series Metrics by County and State](#). This dataset includes information on populations, positive and total tests, number of deaths, and positive cases. We conducted a preliminary statistical analysis to examine correlations between these variables and the number of daily cases. The results of this analysis can be found in Supplementary Fig. 3 in Supplementary Material, and we anticipate that they will provide valuable insights for future research.

The second data source, used for comparative analysis, can be found in the WHO repository at the [WHO Coronavirus \(COVID-19\) Dashboard](#). This resource presents official daily counts of COVID-19 cases, deaths, and vaccine utilization, as reported by countries, territories, and areas. In this study, we use 7 countries: Japan, Canada, Brazil, Argentina, Singapore, Italy, and the United Kingdom.

All datasets generated and analysed during the current study are also available in the [author's Github repository](#)²⁴.

Model evaluations. We use a quantitative measure to evaluate and compare the performance of models: the Mean Absolute Percentage Error (MAPE), defined as:

$$\text{MAPE} = \frac{100}{n} \sum_{t=1}^n \frac{|\hat{Y}_t - Y_{\text{true},t}|}{|Y_{\text{true},t}|} \quad (5)$$

A model with small values of MAPE is preferred.

We examine the performance of the three models (hybrid, AR, and LSTM) on different time periods within the available range. This is essential in our research, since the performance of a model is not constant on different trends; by intuition, a model performs better on smooth curves than it does on steep curves. By repeating our evaluation process on different time periods thus different trends, we wish to understand what trends do the model give the best performance. Such understanding will help us decide to what degrees we may trust the performance of the models. We evaluate the models repeatedly to reduce the influence brought by the instability of model training. Specifically, we leave 7 days between the first date of any two consecutive training data points. Although a larger number of repetitions seems desirable, increasing the repetition number is at the cost of making neighboring training points closer to each other. However, the difference in performance between two



Figure 3. The left panels show the training and testing data. The right panels show the ground truth versus forecasts of the AR, LSTM, and hybrid model, respectively. We display the average prediction (solid line) with 2 times standard error (shaded region). The standard error across 100 runs are reported for LSTM and hybrid. The hybrid model is more stable than the LSTM.

neighboring training points, that are too close to each other, would be attributed more to the instability of model training than to the difference in trend. Such results give us little information about the model performance over

trend. In the end, we let the step number be the same as our lag number. By doing so, we suppose the concept of a week is important in forecasting.

Additional evaluation metrics. In the Supplementary Material, we additionally evaluate and compare above models using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). The evaluation is done on the same dataset across different comparing methods.

Prediction. In this section, we present the numerical results for all three models. We perform a comprehensive comparison of the performance for the three models in multiple counties, showing the advantage of the hybrid model. All predictions are transformed back to the original scale.

Visualization. We compare the three models' performance on COVID-19 case prediction in California 8 counties. For each county, we test the models' performance on several different situations: for example, when the training data has an up trend and the testing data has a down trend. From all trials we practiced, we choose the following trials, presented in Figs. 3 and 4, as representatives of different combinations of training and testing data, since they reflect the general model performances well.

Figure 3a shows models being trained on curved data and being tested on down trend data, as shown on the left and right panel, respectively. Figure 3b shows models being trained on up trend data and being tested on down trend data. Figure 3c shows models being trained on up trend data and being tested on up trend data. Figure 3d shows models being trained on down trend data and being tested on down trend data. Figure 4a,b show models being trained on down trend data and being tested on up trend data, while Fig. 4a has gentle upward testing data and Fig. 4b has sharp upward testing data. Figure 4c show models being trained and tested on jagged data.

To ensure the results above are representative, we run each selected trial 100 times, visualize the mean and standard error of these trials, and present averaged MAPE. While AR outperforms LSTM on some cases, the hybrid model outperforms both in most cases, except that in Fig. 3b and in Fig. 4c. The MAPE, averaged on the 100 trials, shows that LSTM (4.469%) outperforms hybrid (4.993%) slightly in Fig. 3b. However, as shown in the right panel of Fig. 3b, the hybrid model captures the general trend of ground truth better than LSTM does. Similarly, in Fig. 4c, AR (3.675%) outperforms hybrid (3.718%) slightly. Yet, as shown in the right panel of Fig. 4c, the hybrid model captures the general trend of ground truth better than AR does.

Beside, interestingly enough, the hybrid model always seems to capture the ground truth's trend. Actually, the shape of hybrid's forecasts resembles either that of the AR model or that of the LSTM model, or it resembles a combination of both. When AR model captures the trend better than the LSTM does, the hybrid model resembles the AR model in forecast shape: for example, in Fig. 3b, San Francisco 2020-02-17 to 2020-05-14, and in Fig. 4a, Santa Barbara 2022-01-17 to 2022-04-14. When LSTM model captures the trend better than the AR does, the hybrid model resembles the LSTM model in forecast shape: for example, in Fig. 3d, San Francisco 2022-06-10 to 2022-09-05, and in Fig. 4b, Riverside 2022-02-16 to 2022-12-20. On jagged testing data, where AR performs better on some part and LSTM better on the other, the hybrid model presents advantages of both models: for example, in Fig. 4c, the hybrid model resembles AR on the two ends, where AR performs better, and it resembles LSTM in shape between day 5 to day 15, where LSTM seems to capture the trend better.

General performance. We evaluated the model performances numerically, in the 8 California counties across multiple trials. The results are given in Table 2. We observe that the hybrid model outperforms the AR model and the LSTM models almost uniformly: it generally yields the smallest average MAPE. To be specific, the general MAPE of each model (AR, LSTM, LSTM with 2 layers, and hybrid), averaged on the results for all 8 counties, is 5.629%, 4.934%, 6.804%, and 4.173% in order. In general, the hybrid model has the best general performance, and it outperforms the AR model by approximately 1.5%. The LSTM model suffers from overfitting when a second LSTM layer is added. As seen in the Supplementary Material, the proposed hybrid model also yields the lowest RMSE and MAE values.

Interpretability. Interpretability of hybrid models can be defined as the ability to provide insight into the relationships they have learned, as introduced by Murdoch et al.²³. The hybrid model proposed a decomposition approach to decipher the learned model underlying the data-generating mechanism, where the estimated AR model provides the easy-to-understand linear trend. On the other hand, the LSTM is able to capture the long-term and nonlinear trend in the time series data. Our hybrid model aims to strike a balance between interpretability and accuracy, enabling us to gain insights into the underlying data while still achieving high predictive performance.

In this section, we study how AR and LSTM components contribute to the hybrid model when fitting the data. Our purpose is to seek the insights into explaining why the hybrid model enjoys the better performance in general. And more importantly, we seek to use the interpretation from the fitted hybrid model to provide practical guidance to the public health policy making process.

Note that all models are trained on the normalized data as described in section "Training" (Supplementary Material). Consequently all figures below report predictions on the normalized scales.

In Fig. 5, we present three settings with different signal strength ratio (represented by the value of α) of the AR components and LSTM components in the prediction of the hybrid model. Specifically, the larger value of α indicates the AR component dominates the LSTM component in prediction, and the smaller value of α indicates otherwise. We found that the component that has stronger signal characterizes the general trend in the data

	County	AR	LSTM	LSTM (Double)	hybrid
General performance	Los Angeles	5.153 (1.071)	5.120 (1.358)	7.838 (3.414)	3.640 (0.773)
	San Diego	4.446 (0.652)	3.828 (0.669)	5.539(1.549)	3.462 (0.883)
	San Francisco	4.754 (0.421)	4.433 (0.600)	5.350 (1.029)	3.553 (0.317)
	Santa Barbara	7.706 (1.714)	6.157 (1.087)	8.482 (2.677)	5.970 (1.243)
	Fresno	6.549 (1.668)	5.092 (0.933)	6.811 (2.189)	4.188 (0.707)
	Sacramento	5.240 (0.787)	4.707 (0.758)	5.784 (1.182)	3.897 (0.733)
	Ventura	6.525 (0.706)	5.607 (0.913)	7.157 (1.626)	4.687 (0.733)
	Riverside	5.660(0.925)	4.528 (0.651)	7.467 (2.428)	3.985 (0.750)
Latest performance	Los Angeles	4.221	3.607 (0.105)	3.528 (0.161)	3.356 (0.030)
	San Diego	3.749	3.513 (0.055)	3.542 (0.119)	3.343 (0.019)
	San Francisco	5.251	4.115 (0.076)	4.068 (0.068)	4.104 (0.028)
	Santa Barbara	4.731	4.183 (0.034)	4.111 (0.061)	4.160 (0.011)
	Fresno	3.475	3.107 (0.067)	3.525 (0.219)	2.942 (0.009)
	Sacramento	4.685	3.789 (0.048)	3.703 (0.066)	3.880(0.025)
	Ventura	4.143	3.567 (0.029)	3.489 (0.053)	3.410 (0.021)
	Riverside	4.752	3.415 (0.100)	3.283 (0.145)	3.197 (0.044)

Table 2. MAPE (by percentage) for each model on each county. General performance is averaged on all trials. The inconsistent performances of neural networks have been compensated by the small step value, which is 7. The Latest performance is on the latest trial, from 2022-06-10 to 2022-09-05. The results for LSTM, LSTM double and hybrid are each averaged on 100 runs, to compensate the inconsistent performances of neural networks. The value in parenthesis is the standard error. AR has 0 or small standard error for the same trial. The hybrid model is usually the best in performance and has the lowest standard error. Best performances are given in bold.

while the other helps to stabilize the variance. This observation sheds light into why the hybrid model provides better predictive performance in general than a single model.

Moreover, the fitted value of α provides a characterization of the intrinsic nonlinearity of the data, and consequently the difficulty of exploiting interpretation in the linear components of the fitted hybrid model. The smaller the value of α , the higher weight the nonlinear fit using LSTM has in the final prediction. In such a setting, coefficients in the AR components should be given less weight into generating interpretation for policy making. Equivalently, for larger value of α , it is more trustworthy to derive coefficients interpretation from the important AR part. This observation is helpful for public policy maker to distinguish among different virus transmission stages.

Finally, we observe interesting patterns of the coefficients estimates in the AR components of the hybrid model compared with the coefficients in the pure AR model. As shown in Table 3, across the three settings of different values of α , the pure AR model tends to put heavier weight in coefficients of larger lags, say Y_{t-7} . In contrast, the AR component in the hybrid model tends to focus on capturing the short history, i.e., the coefficients associated with smaller lags (e.g., Y_{t-1}) tend to have larger estimates. This indicates that the short history pattern in the data could be well approximated by a simple (say, linear) model, while the longer history in the data possesses more complicated nonlinear structure that requires a LSTM component to fit.

Comparative study on the WHO datasets. In this section, we compare our proposed hybrid model for COVID-19 prediction with its two component models, the ARIMA and LSTM models, as well as three other commonly used models: Support Vector Machines⁵³ (SVM), Random Forest⁵⁴ (RF), and eXtreme Gradient Boosting⁵⁵ (XGBoost). To ensure the effectiveness of our model in different application settings, we use a country-level data for this comparative study, focusing on datasets from seven different countries collected by the World Health Organization.

We provide a brief overview of the three additional comparing methods. Support Vector Machines (SVM)^{42,47} is a machine learning model that identifies the optimal hyperplane in a high-dimensional space that maximally separates data points into different classes. An SVM applies to both classification and regression problems. SVM is known to not perform well on noisy or unbalanced data^{56,57}.

Random Forest^{43–45} is an ensemble learning method that constructs a multitude of decision trees. A Random Forest is very flexible and can handle complex data types. On the other hand, the Random Forests are known for their reduced interpretability, sensitivity to noise, the need for hyperparameter tuning, and potential issues with imbalanced data. These factors may impact their performance in the context of COVID-19 predictions^{58–60}.

Extreme Gradient Boosting (XGBoost)^{44,46,48} has shown exceptional performance in various tasks. XGBoost is an ensemble learning method based on gradient boosting trees. It is known for its efficiency, scalability, and accuracy. However, like other tree-based ensemble methods, it can be more challenging to interpret. This may make it difficult to understand the driving factors behind predictions. In addition, XGBoost can be prone to overfitting, especially with small datasets or when the hyperparameters are not tuned properly^{61,62}.

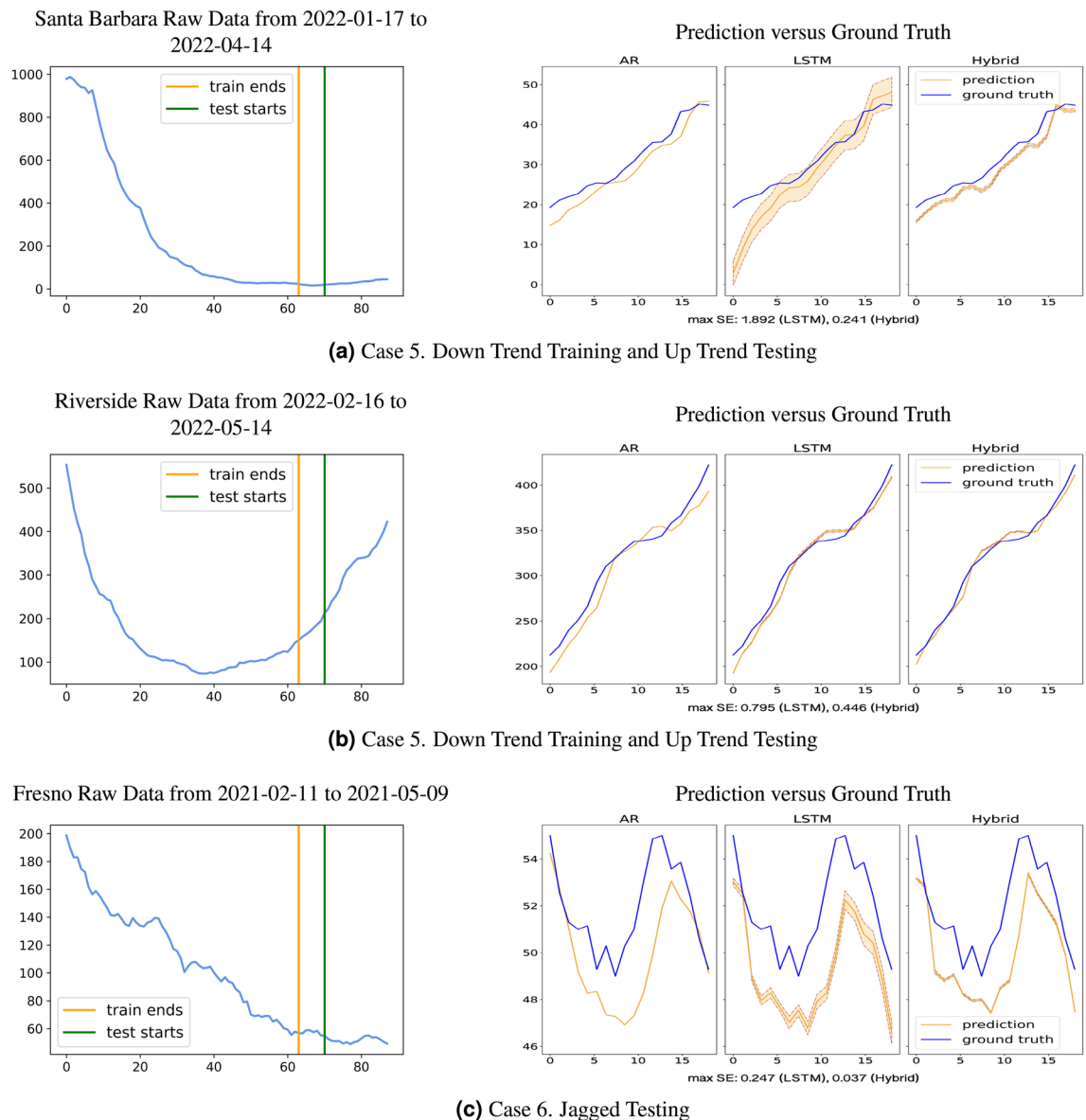


Figure 4. The left panels show the training and testing data. The right panels show the ground truth versus forecasts of the AR, LSTM, and hybrid model, respectively. We display the average prediction (solid line) with 2 times standard error (shaded region). The standard error across 100 runs are reported for LSTM and hybrid. The hybrid model is more stable than the LSTM.

Alpha	Models	Intercept	Y_{t-1}	Y_{t-2}	Y_{t-3}	Y_{t-4}	Y_{t-5}	Y_{t-6}	Y_{t-7}
$\alpha = 0.817$	AR	-0.106	-0.124	0.108	0.063	0.183	0.001	0.039	0.654
	hybrid	-0.008	0.702	0.142	0.044	0.210	0.089	0.114	-0.108
$\alpha = 0.547$	AR	-0.000	-0.019	0.061	-0.077	0.117	-0.041	0.165	0.553
	hybrid	0.022	0.478	-0.298	0.080	-0.003	-0.169	0.049	-0.243
$\alpha = 0.174$	AR	-0.046	0.103	-0.166	0.145	-0.168	0.311	-0.170	0.813
	hybrid	0.025	0.680	0.043	-0.079	-0.138	0.080	-0.156	-0.063

Table 3. Coefficients of AR model v.s. AR coefficients of hybrid model. When AR component dominates the prediction ($\alpha = 0.817$), MAPE for AR and hybrid are 3.088% and 2.593%. When the AR and the LSTM components have similar weight in prediction ($\alpha = 0.547$), MAPE for AR and hybrid are 2.523% and 1.950%. When LSTM component dominates the prediction ($\alpha = 0.174$), MAPE for AR and hybrid are 9.337% and 4.665%. Coefficients with biggest absolute values are given in bold.

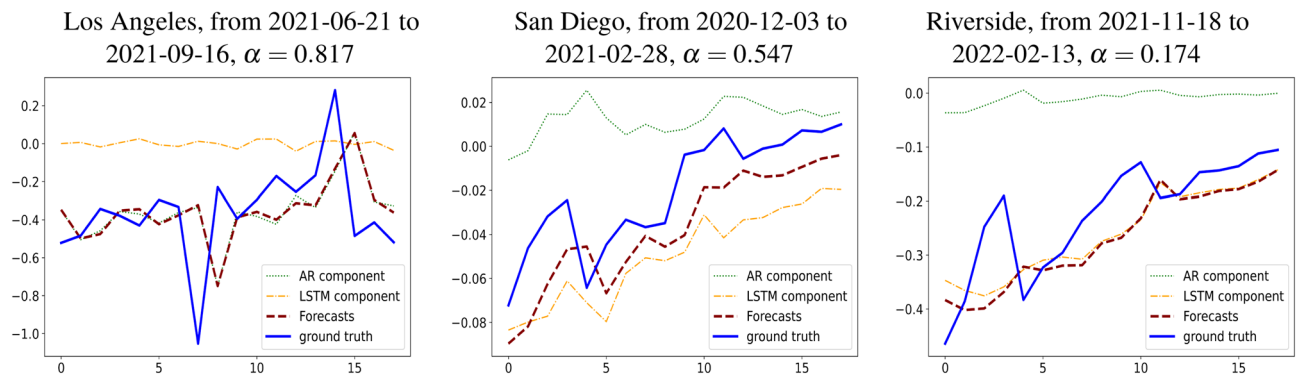


Figure 5. The forecasts of a hybrid model versus the ground truth, and the contribution of its AR and its LSTM component.

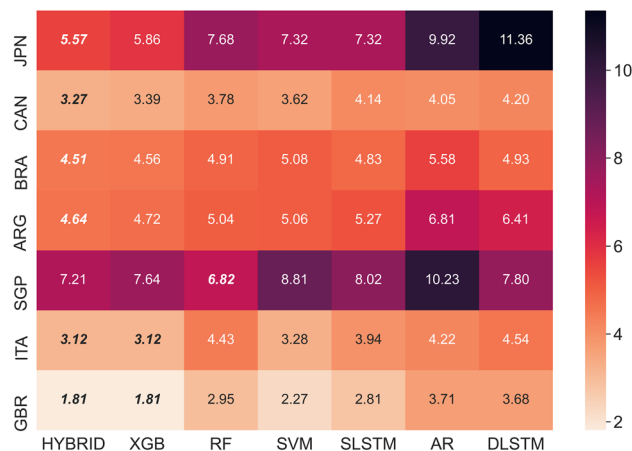


Figure 6. A heatmap exhibiting the performance, measured by MAPE in percentage, of the 7 models from this study and from previous work: AR, Single LSTM(LSTM), Double LSTM(DLSTM), hybrid, SVM, Random Forest(RF), XGBoost(XGB). The assessment has been done on data collected by World Health Organization, from 7 different countries around the world: Japan(JPN), Canada(CAN), Brazil(BRA), Argentina(ARG), Singapore(SGP), Italy(ITA), and The United Kingdom(GBR).

We present the numerical results of the comparative study, which are visualized in Fig. 6. The comparative study is done on data collected by the World Health Organization⁶³ in Japan (JPN), Canada (CAN), Brazil (BRA), Argentina (ARG), Singapore (SGP), Italy (ITA), and the United Kingdom (GBR).

Overall, the proposed hybrid model performs better than the other models in most cases, as evidenced by its lower MAPE. This suggests that our model is effective in various situations and outperforms other commonly used models for COVID-19 prediction.

Discussion

In this paper, we introduce a novel hybrid model that borrows strength from a highly structured Autoregressive model and a LSTM model for the task of COVID-19 cases prediction. Through intensive numerical experiments, we conclude that the hybrid model yields more desirable predictive performance than considering the AR or the LSTM counterpart alone. In principle, the hybrid model enjoys the advantages of each of its two building blocks: the expressive power of LSTM in representing nonlinear patterns in the data and the interpretability from the simple structures in AR. Consequently, the proposed hybrid model is useful in simultaneously providing accurate prediction and shedding light into understanding the transition of the virus transmission phases, and thus providing guidance to the public health policy making process.

It is also noteworthy that the predictive performance of the proposed hybrid model can be further improved by properly choosing the hyperparameters. Furthermore, while we considered LSTM as the nonlinear component in the hybrid model, it can be substituted by any other deep learning models.

Received: 11 November 2022; Accepted: 17 April 2023

Published online: 25 April 2023

References

1. Rahimi, I., Gandomi, A. H. & Chen, F. A review on covid-19 forecasting models. *Neural Comput. Appl.* <https://doi.org/10.1007/s00521-020-05626-8> (2021).
2. He, S., Peng, Y. & Sun, K. Seir modeling of the covid-19 and its dynamics. *Nonlinear Dyn.* **101**, 1667–1680 (2020).
3. Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B. & Sledge, D. The challenges of modeling and forecasting the spread of covid-19. *Proc. Natl. Acad. Sci. USA* **117**, 16732–16738 (2020).
4. Ndairou, F., Area, I., Nieto, J. J. & Torres, D. F. Mathematical modeling of covid-19 transmission dynamics with a case study of Wuhan. *Chaos Solitons Fractals* **135**, 109846 (2020).
5. O’Dea, E. B. & Drake, J. M. A semi-parametric, state-space compartmental model with time-dependent parameters for forecasting covid-19 cases, hospitalizations and deaths. *J. R. Soc. Interface* **19**, 20210702 (2022).
6. McDonald, D. J. *et al.* Can auxiliary indicators improve covid-19 forecasting and hotspot prediction?. *Proc. Natl. Acad. Sci. USA* **118**, e2111453118 (2021).
7. Maleki, M., Mahmoudi, M. R., Wraith, D. & Pho, K.-H. Time series modelling to forecast the confirmed and recovered cases of covid-19. *Travel Med. Infect. Dis.* **37**, 101742. <https://doi.org/10.1016/j.tmaid.2020.101742> (2020).
8. Sioofy Khoojine, A., Shadabfar, M., Hosseini, V. R. & Kordestani, H. Network autoregressive model for the prediction of covid-19 considering the disease interaction in neighboring countries. *Entropy* **23**, 1267 (2021).
9. Guhathakurata, S., Kundu, S., Chakraborty, A. & Banerjee, J. S. 18: A novel approach to predict covid-19 using support vector machine. In *Data Science for COVID-19* (eds Kose, U. *et al.*) 351–364 (Academic Press, 2021).
10. Yadav, M., Perumal, M. & Srinivas, M. Analysis on novel coronavirus (covid-19) using machine learning methods. *Chaos Solitons Fractals* **139**, 110050. <https://doi.org/10.1016/j.chaos.2020.110050> (2020).
11. Tiwari, D., Bhati, B. S., Al-Turjman, F. & Nagpal, B. Pandemic coronavirus disease (covid-19): World effects analysis and prediction using machine-learning techniques. *Expert Syst.* **39**, e12714 (2022).
12. Shahid, F., Zameer, A. & Muneeb, M. Predictions for covid-19 with deep learning models of lstm, gru and bi-lstm. *Chaos Solitons Fractals* **140**, 110212. <https://doi.org/10.1016/j.chaos.2020.110212> (2020).
13. Zeroual, A., Harrou, F., Dairi, A. & Sun, Y. Deep learning methods for forecasting covid-19 time-series data: A comparative study. *Chaos Solitons Fractals* **140**, 110121 (2020).
14. Davahli, M. R., Fiok, K., Karwowski, W., Aljuaid, A. M. & Taiar, R. Predicting the dynamics of the covid-19 pandemic in the United States using graph theory-based neural networks. *Int. J. Environ. Res. Public Health* **18**, 3834 (2021).
15. Alassafi, M. O., Jarrah, M. & Alotaibi, R. Time series predicting of covid-19 based on deep learning. *Neurocomputing* **468**, 335–344. <https://doi.org/10.1016/j.neucom.2021.10.035> (2022).
16. Chimmula, V. K. R. & Zhang, L. Time series forecasting of covid-19 transmission in Canada using lstm networks. *Chaos Solitons Fractals* **135**, 109864 (2020).
17. Hawas, M. Generated time-series prediction data of covid-19’s daily infections in Brazil by using recurrent neural networks. *Data Brief* **32**, 106175 (2020).
18. Long, J., Khaliq, A. & Furati, K. M. Identification and prediction of time-varying parameters of covid-19 model: A data-driven deep learning approach. *Int. J. Comput. Math.* **98**, 1617–1632 (2021).
19. Box, G. E., Jenkins, G. M., Reinsel, G. C & Ljung, G. M. *Time Series Analysis: Forecasting and Control* (Wiley, 2015).
20. Allard, R. Use of time-series analysis in infectious disease surveillance. *Bull. World Health Organ.* **76**(4), 327–333 (1998).
21. Michael A., Johansson, N. G. R., Aditi Hota, J. S. B., Santillana, M. (2016) Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico. *Sci. Rep.* <https://doi.org/10.1038/srep33707> (2016).
22. Sepp Hochreiter, J. S. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
23. Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **116**, 22071–22080. <https://doi.org/10.1073/pnas.1900654116> (2019).
24. Zhang, Y. <https://github.com/yangyi-zhang/covid-forecasting> (2022).
25. Xuhong Li, H. X. *et al.* Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* **64**, 3197–3234 (2022).
26. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl.* **32**, 18069–18083 (2020).
27. Forough Poursabzi-Sangdeh, D. G. G., Hofman, J. M., Vaughan, J. W. & Wallach, H. M. Manipulating and measuring model interpretability. *CoRR abs/1802.07810*. [arXiv:1802.07810](https://arxiv.org/abs/1802.07810) (2018).
28. ElAraby, M. E., Elzeki, O. M., Shams, M. Y., Mahmoud, A. & Salem, H. A novel gray-scale spatial exploitation learning net for covid-19 by crawling internet resources. *Biomed. Signal Process. Control* **73**, 103441 (2022).
29. Shams, M. Y. *et al.* Hana: A healthy artificial nutrition analysis model during covid-19 pandemic. *Comput. Biol. Med.* **135**, 104606. <https://doi.org/10.1016/j.compbiomed.2021.104606> (2021).
30. Harleen Kaur, B. A., Ahsaan, S. U. & Chang, V. A proposed sentiment analysis deep learning algorithm for analyzing covid-19 tweets. *Inf. Syst. Front.* **23**, 1417–1429 (2021).
31. Awan, T. M. & Aslam, F. Prediction of daily covid-19 cases in European countries using automatic arima model. *J. Public Health Res.* **9**, 1765 (2020).
32. Yeşilkanat, C. M. Spatio-temporal estimation of the daily cases of covid-19 in worldwide using random forest machine learning algorithm. *Chaos Solitons Fractals* **140**, 110210 (2020).
33. Muhammad, L., Haruna, A. A., Sharif, U. S. & Mohammed, M. B. Cnn-lstm deep learning based forecasting model for covid-19 infection cases in Nigeria, South Africa and Botswana. *Health Technol.* **23**, 1259–1276 (2022).
34. Azeez, A., Obaromi, D., Odeyemi, A., Ndege, J. & Muntabayi, R. Seasonality and trend forecasting of tuberculosis prevalence data in Eastern Cape, South Africa, using a hybrid model. *Int. J. Environ. Res. Public Health* **13**, 757. <https://doi.org/10.3390/ijerph13080757> (2016).
35. Yan, W., Xu, Y., Yang, X. & Zhou, Y. A hybrid model for short-term bacillary dysentery prediction in Yichang City, China. *Jpn. J. Infect. Dis.* **63**, 264–270. <https://doi.org/10.7883/yoken.63.264> (2010).
36. Zhao, W., Chen, Y., Li, Y. & Guan, W. Prediction of covid-19 data using hybrid modelling approaches. *Front. Public Health* **10**, 923978 (2022).
37. Ala’raj, M., Majdalawieh, M. & Nizamuddin, N. Modeling and forecasting of covid-19 using a hybrid dynamic model based on seird with arima corrections. *Infect. Dis. Model.* **6**, 98–111 (2021).
38. Zheng, N. *et al.* Predicting covid-19 in China using hybrid AI model. *IEEE Trans. Cybern.* **50**, 2891–2904 (2020).
39. Fan, D. *et al.* Well production forecasting based on arima-lstm model considering manual operations. *Energy* **220**, 119708 (2021).
40. Fathi, O. Time series forecasting using a hybrid arima and lstm model. *Velvet Consult.* **2019**, 1–7 (2019).
41. Zhang, G. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* **50**, 159–175. [https://doi.org/10.1016/S0925-2312\(01\)00702-0](https://doi.org/10.1016/S0925-2312(01)00702-0) (2003).
42. Atik, I. Performance comparison of regression learning methods: Covid-19 case prediction for turkey. *Int. J. Mech. Eng.* **7**, 6297–6308 (2022).

43. Galasso, J., Cao, D. M. & Hochberg, R. A random forest model for forecasting regional covid-19 cases utilizing reproduction number estimates and demographic data. *Chaos Solitons Fractals* **156**, 111779. <https://doi.org/10.1016/j.chaos.2021.111779> (2022).
44. Ali, A. A., Usmani, A. & Sher, R. Covid-19 cases prediction in Saudi Arabia using tree-based ensemble models. *Intell. Autom. Soft Comput.* **32**, 298–400 (2022).
45. Chumachenko, D., Meniaillov, I., Bazilevych, K., Chumachenko, T. & Yakovlev, S. Investigation of statistical machine learning models for covid-19 epidemic process simulation: Random forest, k-nearest neighbors, gradient boosting. *Computation* **10**, 86. <https://doi.org/10.3390/computation10060086> (2022).
46. Fang, Z.-G., Yang, S.-Q., Lv, C.-X., An, S.-Y. & Wu, W. Application of a data-driven xgboost model for the prediction of covid-19 in the USA: A time-series study. *BMJ Open* **12**, e056685. <https://doi.org/10.1136/bmjopen-2021-056685> (2022).
47. Muhammad, L. J., Algehhyne, E. A. & Usman, S. S. Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. *SN Comput. Sci.* <https://doi.org/10.1007/s42979-020-00394-7> (2020).
48. Luo, J., Zhang, Z., Fu, Y. & Rao, F. Time series prediction of covid-19 transmission in America using lstm and xgboost algorithms. *Results Phys.* **27**, 104462. <https://doi.org/10.1016/j.rinp.2021.104462> (2021).
49. Vadyala, S. R., Betgeri, S. N., Sherer, E. A. & Amritphale, A. Prediction of the number of covid-19 confirmed cases based on k-means-lstm. *Array* **11**, 100085. <https://doi.org/10.1016/j.array.2021.100085> (2021).
50. Tomar, A. & Gupta, N. Prediction for the spread of covid-19 in India and effectiveness of preventive measures. *Sci. Total Environ.* **728**, 138762. <https://doi.org/10.1016/j.scitotenv.2020.138762> (2020).
51. Bhandari, S. *et al.* Evolving trajectories of covid-19 curves in India: Prediction using autoregressive integrated moving average modeling. *Mol. Biol.* **2020**, 1–15. <https://doi.org/10.21203/rs.3.rs-40385/v1> (2020).
52. Lipton, Z. C. A critical review of recurrent neural networks for sequence learning. CoRR. [arXiv:1506.00019](https://arxiv.org/abs/1506.00019) (2015).
53. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
54. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
55. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* <https://doi.org/10.1145/2939672.2939785> (ACM, 2016).
56. Akbani, R., Kwek, S. & Japkowicz, N. Applying support vector machines to imbalanced datasets. In *Machine Learning: ECML 2004: 15th European Conference on Machine Learning. Proceedings 15* 39–50 (Springer, 2004).
57. Fung, G. & Mangasarian, O. L. Proximal support vector machine classifiers. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 77–86 (2001).
58. Antoniadis, A., Lambert-Lacroix, S. & Poggi, J.-M. Random forests for global sensitivity analysis: A selective review. *Reliab. Eng. Syst. Saf.* **206**, 107312 (2021).
59. Aria, M., Cuccurullo, C. & Gnasso, A. A comparison among interpretative proposals for random forests. *Mach. Learn. Appl.* **6**, 100094 (2021).
60. Biau, G. & Scornet, E. A random forest guided tour. *Test* **25**, 197–227 (2016).
61. Pesantez-Narvaez, J., Guillen, M. & Alcañiz, M. Predicting motor insurance claims using telematics data-xgboost versus logistic regression. *Risks* **7**, 70. <https://doi.org/10.3390/risks7020070> (2019).
62. Li, W., Yin, Y., Quan, X. & Zhang, H. Gene expression value prediction based on xgboost algorithm. *Front. Genet.* **10**, 1077 (2019).
63. World Health Organization. *WHO-COVID-19-Global-Data* (2023).

Author contributions

S.T. and Y.G. provided the initial idea and research plan, Y.Z. collected data, implement algorithms and performed simulations. All authors participated in the analysis and discussion of the results, and participated in the writing of the manuscript.

Funding

Y. Zhang was partially supported by Raymond L Wilder Award sponsored by University of California Santa Barbara and Hellman Family Faculty Fellowship. S.T. was partially supported by Regents Junior Faculty fellowship, Faculty Early Career Acceleration grant sponsored by University of California Santa Barbara, Hellman Family Faculty Fellowship and the NSF DMS-2111303. G.Y. was partially supported by Regents Junior Faculty fellowship, Faculty Early Career Acceleration grant sponsored by University of California Santa Barbara.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-33685-z>.

Correspondence and requests for materials should be addressed to S.T. or G.Y.

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) 2023