Random-Forest-Bagging Broad Learning System With Applications for COVID-19 Pandemic

Choujun Zhan, *Member, IEEE*, Yufan Zheng, Haijun Zhang, *Senior Member, IEEE*, and Quansi Wen, *Member, IEEE*

***Abstract*—The rapid geographic spread of COVID-19, to which various factors may have contributed, has caused a global health crisis. Recently, the analysis and forecast of the COVID-19 pandemic have attracted worldwide attention. In this work, a large COVID-19 data set consisting of COVID-19 pandemic, COVID-19 testing capacity, economic level, demographic information, and geographic location data in 184 countries and 1241 areas from December 18, 2019, to September 30, 2020, were developed from public reports released by national health authorities and bureau of statistics. We proposed a machine learning model for COVID-19 prediction based on the broad learning system (BLS). Here, we leveraged random forest (RF) to screen out the key features. Then, we combine the bagging strategy and BLS to develop a random-forest-bagging BLS (RF-Bagging-BLS) approach to forecast the trend of the COVID-19 pandemic. In addition, we compared the forecasting results with linear regression (LR) model, *K*-nearest neighbors (KNN), decision tree (DT), adaptive boosting (Ada), RF, gradient boosting DT (GBDT), support vector regression (SVR), extra trees (ETs) regressor, CatBoost (CAT), LightGBM (LGB), XGBoost (XGB), and BLS.The RF-Bagging BLS model showed better forecasting performance in terms of relative mean-square error (RMSE), coefficient of determination (R^2), adjusted coefficient of determination (R^2_{adj}), median absolute error (MAD), and mean absolute percentage error (MAPE) than other models. Hence, the proposed model demonstrates superior predictive power over other benchmark models.**

***Index Terms*—Artificial intelligence, broad learning system (BLS), coronavirus disease 2019 (COVID-19) testing capacity, COVID-19, random forest (RF), time-series forecasting.**

Manuscript received October 29, 2020; revised January 25, 2021 and February 23, 2021; accepted March 5, 2021. Date of publication March 17, 2021; date of current version October 22, 2021. This work was supported in part by the Science and Technology Program of Guangzhou, China, under Grant 201904010224; in part by the Natural Science Foundation of Guangdong Province, China, under Grant 2020A1515010761 and Grant 2018A030313351; and in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2020B010166001 and Grant 2019B010137001. *(Corresponding author: Quansi Wen.)*

Choujun Zhan was with the chool of Electronical and Computer Engineering, Nanfang College of Sun Yat-sen University, Guangzhou 510970, China. He is now with the School of Computing, South China Normal University, Guangzhou 510641, China (e-mail: zchoujun2@gmail.com).

Yufan Zheng is with the School of Electrical and Computer Engineering, Nanfang College of Sun Yat-Sen University, Guangzhou 510970, China (e-mail: zhjpre@gmail.com).

Haijun Zhang is with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China (e-mail: hjzhang@hitsz.edu.cn).

Quansi Wen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510641, China, and also with the Jiangmen City Road Traffic Accident Social Relief Fund Management Center, Jiangmeng, China (e-mail: qwen2012@foxmail.com).

Digital Object Identifier 10.1109/JIOT.2021.3066575

I. Introduction

THE NOVEL coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1], has high transmissibility [2]. This new infectious disease spread worldwide in less than half a year in 2020, causing devastation to the human population. The outbreak of COVID-19 has progressed with a tremendous impact on the economic [3], social behavior [4], environment [5], climate [6], etc. Evolutionary virologist found bats may be the origins of SARS-CoV-2 [7], which has a long evolutionary history [8]. However, at present, there is no unified scientific conclusion on the origin of SARS-CoV-2 [9]. Although most countries launched emergency responses early in the outbreak, the COVID-19 still swiftly spread from metropolitan areas to urban areas, from countries to countries. By September 30, 2020, more than 200 countries had been affected, with major outbreaks in the United States, India, Brazil, Russia, Colombia, Peru, and others. A total of 34 488 636 COVID-19 cases and 1 026 176 deaths were reported worldwide [10] and more than 60% of the global population went into coronavirus lockdown [11]. As a result, the World Health Organization (WHO) set COVID-19 to the highest crisis alert level by declaring the COVID-19 outbreak a global pandemic [12].

The tremendous number of COVID-19 cases may be attributable to multiple factors [13], [14]. Presymptomatic and asymptomatic patients play a key role in the spread of COVID-19 [15], [16]. Confirmed COVID-19 cases are mostly quarantined or self-isolated [17], while presymptomatic and asymptomatic patients form a large group of unregistered patients, who can move freely and infect close contacts easily [18]. The population movement of a large group of presymptomatic and asymptomatic patients is a key factor contributing to the spread of COVID-19 [13]. For a country or an area, sufficient COVID-19 tests can greatly decrease the number of unconfirmed infections, reduce the speed of disease transmission [19], and help in an accurate trend analysis to evaluate the pandemic situation [20]. Testing capacity is highly related to the number of confirmed cases and essential to elucidate the progress of the pandemic [19]. However, many low-income countries with comparatively weak health systems have limited resources for conducting massive tests and implementing public health measures to flatten the curve [20], [21]. Hence, the developing level of countries or regions is also related to the spread of COVID-19 [22]. The nonpharmacological interventions

(NPIs) (or public health measures) have been the mainstay for containing the spread of COVID-19 [23]. Additionally, the geographical environment and climate also influence the spread of the COVID-19 [24]. In summarize, the spread of COVID-19 is a nonlinear complex dynamic progress greatly influenced by multiple factors, including COVID-19 testing capacity [19], geographical environment and climate [24], economic level [20], [22], human movements [25], NPIs (or public health measures) [26], air pollution [14], etc.

Machine learning has shown promising results in forecasting nonlinear dynamic progress and has been recognized as a potentially powerful tool for fighting COVID-19 [27]–[29]. However, the new pandemic still brings a number of new challenges, such as predicting the spread of the infection [30], making diagnoses and prognosis [31], [32], searching for treatments and vaccines [33], and social control [34]. Recently, estimates of COVID-19 patient volume are urgently required for local authorities to effectively manage the rising case for restricting the infections [35]. Generally, scholars develop epidemiological models for describing the spread of COVID-19. However, due to the complexity and the high level of uncertainty of the COVID-19, the standard epidemiological models always are a high-dimension nonlinear model with many unknown parameters, which are difficult to determine [25]. Hence, machine learning, which can, in principle, be utilized to build outbreak prediction models, has recently gained attention [36]. Machine learning always requires sufficient pandemic data for training. In contrast, most recent works reporting on using machine learning for a predictive purpose use small samples of only one or several areas, which may be biased and make predictions widely uncertain [37], [38]. Some researchers try to enhance the amount of data set by adding new features, such as social media [36], [39]. However, a large amount of media data inevitably contains a lot of false information or noise, which has to be filtered to create a training set [40]. As a result, it has posed great challenges in developing machine learning models for accurately and reliably forecasting the spread of COVID-19 [41]. Different countries have various attitudes toward COVID-19 and different public health measures. As a result, the daily changes of COVID-19 in different areas are highly volatile and variable, which makes it a challenging task to develop a prediction model, which can be applied to all the countries [37].

To alleviate the problem of lacking data and features, in this work, we developed a data set, including the pandemic data of 184 countries and 1241 areas with a total population of 7 730 029 662, accounting for more than 95% of the global population. Briefly, these countries varied in population size, from less than one million population to more than one billion. Additionally, we collected COVID-19 testing data, economic level data, demographic information, and geographic information to establish a large data set to train machine learning models. A broad learning systems (BLSs) is a new proposed structure neural network without deep architecture [42] and shows good potential in time-series prediction [43]. In this work, we utilize random forest (RF), a popular ensemble learning method, to derive the importance score of each feature. Then, we adopted a set of most important features as a training data set. Combining with Bagging strategy and BLSs, we developed a random-forest-bagging BLSs (RF-Bagging-BLSs) model for predicting the spread of COVID-19 in 184 countries and 1241 areas. For justification, a number of machine learning models, including the linear regression (LR) model, K-nearest neighbors (KNN), decision tree (DT), support vector regression (SVR), adaptive boosting (Ada), RF, gradient boosting DT (GBDT), extra trees (ETs) regressor, CatBoost (CAT), LightGBM (LGB), XGBoost (XGB), and BLS, are adopted to compare with the proposed RF-Bagging-BLS model. Experimental results demonstrate that the RF-Bagging-BLS model outperforms other benchmark models by providing more accurate, stable and robust results.

In this study, our main contributions are as follows.

1) We establish a large data set with comprehensive information on COVID-19 spreading in 184 countries and 1241 areas.
2) RF is adopted for feature importance analysis to improve BLS. A machine learning model, the RF-Bagging-BLS model, is proposed for forecasting the pandemic situation in various countries and areas around the world.
3) We developed prediction models based on traditional machine learning, ensemble learning, and BLS.
4) The new data set is also adopted for multiday-ahead forecasting to evaluate and verify the predictive power of these prediction models in different scenarios.

Our approaches and predictive outcomes can help contain the spread, flatten the curve, and possibly eliminate the current COVID-19 pandemic.

II. Literature Review

Forecasting the spread of COVID-19 in an area has received considerable critical attention. Based on a small data set including pandemic data of 5 countries, a comparative analysis of the predictive performance of machine learning and traditional models, including simple epidemiological and statistical models, is conducted. Two models, the adaptive network-based fuzzy inference system (ANFIS) and multilayered perceptron, showed promising results [28]. A nonauto regressive neural network is trained based on a small data set with 164 samples for global records and 90 scores of nine different countries for predicting the cumulative number of infections and death toll [44]. A modified stacked autoencoder is developed for real-time forecasting the confirmed cases in China from January 11 to Febuary 27, 2020 [45]. However, most of studies focused on prediction of the confirmed cases in just one or few countries, such as American [46], Brazilian [47], Canada [37], China [36], France [48], Hungary [38], Italy [49], India [50], Iran [51], Japan [46], South Korea [51], etc. [52].

The recurrent neural network (RNN) model is a commonly used model for predicting time series. A number of researchers used long short-term memory (LSTM) to build COVID-19 prediction models [53]–[55]. A modified LSTM model, trained on the 2003 SARS data, is utilized to predict the epidemic in China from January 23 to April 24, 2020 [54]. A data-driven estimation methods based on curve fitting and LSTM is developed for forecasting the number of COVID-19 cases in India [53]. Convolutional neural network (CNN)

TABLE I
INFORMATION ON COVID-19 DATA RELEASED BY 184 COUNTRIES IN SIX CONTINENTS UP TO SEPTEMBER 30, 2020

Continent	Countries (amount)	Areas (amount)	Population (in millions)	Confirmed cases	Recovered cases	Death toll	COVID-19 Tests	Mortality Rate (%)
Asia	45	802	4,465	10,632,594	8,926,323	194,702	134,266,412	1.8312
Europe	46	231	743	5,078,482	2,449,329	224,315	135,205,765	4.4170
North America	24	115	604	9,135,594	5,468,439	319,302	132,616,914	3.4952
South America	11	85	612	7,601,438	6,595,662	240,320	6,351,210	3.1616
Africa	54	0	1,265	1,481,929	1,227,334	35,863	11,189,231	2.4201
Oceania	4	8	39	29,510	27,098	922	8,661,117	3.1244
Total	184	1,241	7,730	33,959,547	24,694,185	1,015,424	428,290,649	2.9901

is also a good candidate for analyzing and predicting the spread of COVID-19 [55]. A model combining mechanistic and machine learning methodologies is developed for alleviating the lack of essential data and real-time forecasting in China [56]. Also, an improved adaptive neurofuzzy inference system (ANFIS) is proposed to estimate and forecast the number of confirmed cases of COVID-19 in the upcoming ten days in China [57]. Ghamizi *et al.* [58] tried to combine the classical susceptible-exposed-infected-recovered (SEIR) model and machine learning to develop an SEIR-HCD model considering the impact of mitigation strategies. A hybrid machine learning method, which is based on a multilayered perceptron-imperialist competitive algorithm (MLP-ICA) and ANFIS, is used to predict the number of confirmed cases and mortality rate in Hungary [38].

III. DATA DESCRIPTION AND FORECASTING PROBLEM

A. Pandemic Data

Based on several public data sets provided by John Hopkins University, local Centres for Disease Control and Prevention (CDC), and other health authorities, we established a large COVID-19 epidemic research data set covering 184 countries and 1241 areas (cities, provinces, states, and other prefectures) spanning from December 8, 2019, to September 30, 2020. The cumulative number of confirmed cases in the 184 countries is shown in Fig. 1(a), while Table I summarizes the information of the COVID-19 data in the six continents. The data set includes the following contents.

1) For each day t, the COVID-19 spreading data set utilized in this study includes the number of confirmed cases $I_C(t)$, fatalities $D(t)$ and recovered cases $R(t)$ for 184 countries and 1241 areas (shown in Table I).
2) The information resultant from COVID-19 tests is a critical factor in ascertaining infection numbers. Additionally, sufficient testing capacity is essential to elucidate the progress of the pandemic [19]: the more tests, the high possibility to identify unconfirmed COVID-19 patients. Hence, test capacity $N_T(t)$, representing the cumulative number of conducted test, is adopted as a feature for forecasting the spread of COVID-19. Note that the world's COVID-19 test capacity increases dramatically in six months from less than 10,000 tests per day on March 1, 2020, to more than 4 million tests per day on September 30, 2020 [shown in Fig. 1(b)].
3) Several studies suggest climate may be one factor that influences the spread of COVID-19 [24]. The environment in an area is closely related to the local location. Hence, the latitude x_{LA} and longitude x_{LO} of each country and region are collected. Fig. 1(a) shows the geographic distribution of confirmed cases in these 184 countries over the world. Additionally, we divide each area into six categories according to the continent where it is located. Then, each region has another feature $x_{CT} = \{1, 2, \ldots, 6\}$, where 1, 2, 3, 4, 5, and 6 represent Asia, Europe, North America, South America, Africa, and Oceania, respectively.
4) Most developed countries have advanced health systems and strong capacity to offset the economic and can apply population-level physical distancing measures to contain the spread of COVID-19, while undeveloped countries may have limited sources to fight with COVID-19. Hence, the economic situation of an area is also an influential factor [20]. According to the World Bank

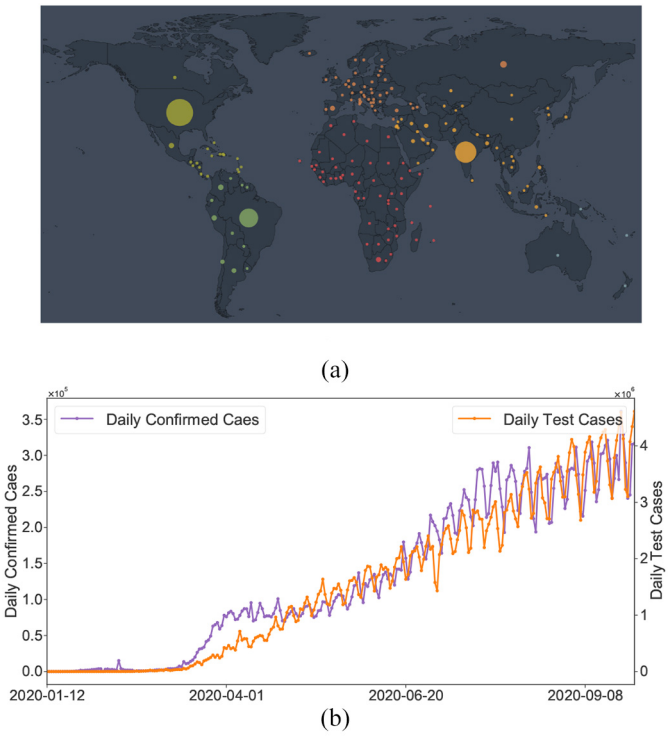


Fig. 1. (a) Cumulative number of confirmed cases in 184 countries up to September, 2020. The size of the solid circles represents the number of confirmed cases. (b) Daily confirmed cases and daily COVID-19 tests over the world up to September 30, 2020.

TABLE II
INPUT FEATURES FOR MACHINE LEARNING MODELS

	Original features
$I_C(t)$	Cumulative confirmed cases at time t
$R(t)$	Total recovered cases at time t
$D(t)$	Death toll at time t
$N_T(t)$	Cumulative COVID-19 tests at time t
x_{LA}	The latitude of the geographical location of an area
x_{LO}	The longitude of the geographical location of an area
x_{CT}	The continent to which an area belongs
x_D	The level of economic development of an area
x_P	The population of an area
	Augmented features
$\Delta I_C(t)$	Daily confirmed cases
$\Delta R(t)$	Daily recovered cases
$\Delta D(t)$	Daily deaths
$I(t)$	Active COVID-19 patients
$\Delta N_T(t)$	Daily COVID-19 tests
$r_I(t)$	Daily growth rate of confirmed cases
$r_R(t)$	Daily growth rate of daily recover cases
$r_D(t)$	Daily growth rate of daily death cases

indicators, we divide countries into developed countries, developing countries, and undeveloped countries. Each region has a feature $x_D = \{1, 2, 3\}$, while 1, 2, and 3 stand for developed, developing, and undeveloped countries.

5) The population x_P of each country or area is also adopted as a feature for forecasting the spread of COVID-19.

Here, public health intervention information is not included in this data set. Each country or even each prefecture in a country would have different control measures. For instance, in the USA, each state implement control measures independently. In [59], the authors summarized country-level public health measures in more than 200 countries from January 1 to October 1, 2020. However, it is difficult to quantify the effort of each public health measures. Hence, we do not consider this factor in this work. Migration is another important feature influencing the spread of the COVID-19. However, few countries or areas provide a daily migration or even weekly migration data [13], [25], [46]. Hence, in our data set, we do not consider migration data either. For developing predictive models, we divided this COVID-19 data set into two parts: 1) the training set (73%), ranging from January 1, 2020, to August 15, 2020 and 2) the test set (27%), from August 16, 2020, to September 30, 2020.

B. Forecasting Problem Formulation

As previously mentioned, we have nine original features, including cumulative confirmed cases, totally recovered cases, death toll, cumulative COVID-19 tests, the latitude and longitude, the continent to which the area belongs, the economic level, and the population of each area. Based on these original features, we can derive the following augmented features.

1) *Daily Confirmed Cases:*

$$\Delta I_C(t) = I_C(t) - I_C(t-1). \tag{1}$$

2) *Daily Recovered Cases:*

$$\Delta R(t) = R(t) - R(t-1). \tag{2}$$

3) *Daily Deaths:*

$$\Delta D(t) = D(t) - D(t-1). \tag{3}$$

4) *Active COVID-19 Cases:*

$$I(t) = I_C(t) - R(t) - D(t) \tag{4}$$

which represents the number of COVID-19 patients, who have not been removed yet.

5) *Daily COVID-19 Tests:*

$$\Delta N_T(t) = N(t) - N(t-1). \tag{5}$$

6) *Daily Growth Rate of Daily Confirmed Cases:*

$$r_I(t) = \frac{\Delta I_c(t) - \Delta I_c(t-1)}{\Delta I_c(t-1)}. \tag{6}$$

7) *Daily Growth Rate of Daily Recover Cases:*

$$r_R(t) = \frac{\Delta R(t) - \Delta R(t-1)}{\Delta R(t-1)}. \tag{7}$$

8) *Daily Growth Rate of Daily Death Cases:*

$$r_D(t) = \frac{\Delta D(t) - \Delta D(t-1)}{\Delta D(t-1)}. \tag{8}$$

All the input features are summarized in Table II. Note that these input features can be classified into two categories

$$\begin{aligned} x_\theta &= \left\{x_{\mathrm{LA}}, x_{\mathrm{LO}}, x_{CT}, x_D, x_p\right\} \\ x(t) &= \{I_C(t), R(t), D(t), N_T(t), \Delta I_C(t), \Delta R(t) \\ &\quad \Delta D(t), \Delta N_T(t), I(t), r_I(t), r_R(t), r_D(t)\} \end{aligned} \tag{9}$$

where $x_\theta \in \mathbb{R}^{5\times 1}$ represents constant features, which is time-independent, while $x(t) \in \mathbb{R}^{12\times 1}$ stands for time-varying features.

Here, we provide m-day forecasts for n consecutive days with quantified uncertainty based on machine learning models. The prediction model is

$$\hat{y}(t+n) = f(x(t-m), x(t-m+1), \ldots, x(t-1), x_\theta) \tag{10}$$

where $\hat{y}(t+n)$ represents the predicted value, while $f(\cdot)$ stands for the machine learning model. Then, the prediction problem can be formulated as

$$\begin{aligned} \min_{f(\cdot)} \ & \sum \left\| y(t+n) - \hat{y}(t+n) \right\|_2^2 \\ \text{s.t.} \ & \hat{y}(t+n) = f(x(t-m), x(t-m+1), \ldots, x(t-1), x_\theta) \end{aligned} \tag{11}$$

where $y(t+n)$ is the true value.

IV. METHODS

A. Broad Learning System

Drawing on the idea of the random vector function link neural network (RVFLNN), Chen and Liu [42] proposed a BLS, which is a new flat structure neural network without the need for deep architecture. The BLS simplifies the training procedure for a fast universal approximation and has provided competitive results with deep learning and ensemble learning methods in various fields. In recent studies, BLS has shown impressive performance for specific tasks, including visual-based assessment systems [60], predicting the setting time of cement [61], fatigue detection [62], etc. The BLS has a universal approximation capability and can approximate the loss

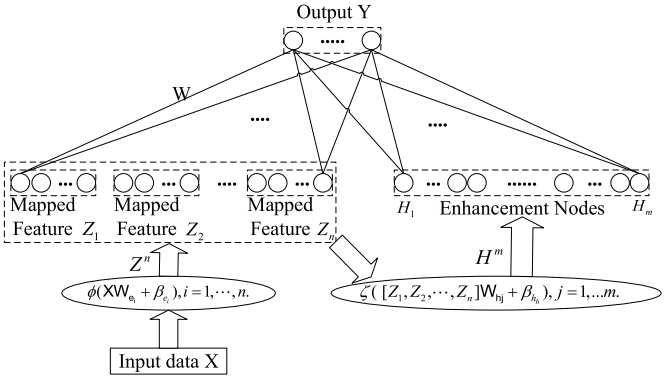


Fig. 2. Simplified structure of a typical BLS.

function globally. Inspired by this work, several structural variations of BLS have been proposed [63]. Additionally, the BLS is a kind of increment learning structure, which can efficiently and effectively update the system using newly added features or data. Fig. 2 illustrates the structure of a typical BLS, which shows the effects of different nodes and information propagation. First, the features are extracted from the training data set by the mapped feature nodes. Then, the adopted features are transformed into enhancement nodes. Finally, the output is a linear combination of all mapping features and the output of enhancement nodes. The connection weights of a typical BLS can be derived from the ridge regression approximation algorithm [64], [65].

Consider the training set $\{(x(i), y(i)|x(i) \in \mathbb{R}^D, y(i) \in \mathbb{R}^C, i = 1, 2, \ldots, N\}$, where D and C are the dimension of each sample and corresponding outputs, respectively. Then, the input pattern is $X \in \mathbb{R}^{s\times D}$, $Y \in \mathbb{R}^{s\times C}$, where $X = [x(1), x(2), \ldots, x(s)]^T$, $Y = [y(1), y(2), \ldots, y(s)]$, and s is the number of input samples. In the feature learning stage, the input matrix X is mapped into n feature nodes by n feature mapping ϕ_i to generates random features. The following feature function generates the mapped feature Z_i:

$$Z_i = \phi(XW_{ei} + \beta_{ei}), \quad i = 1, 2, \ldots, n \tag{12}$$

where weights $W_{ei} \in \mathbb{R}^{D\times k_i}$ and bias term β_{ei} are randomly generated matrices with applicable dimensions from the given proper distribution scope $[-a, a]$. $\phi(\cdot)$ stands for prior activation functions of mapped feature nodes. The outputs of groups of feature nodes can be denoted as

$$Z^n = [Z_1, Z_2, \ldots, Z_n] \tag{13}$$

where $Z_i \in \mathbb{R}^{s\times k_i}$ and $Z^n \in \mathbb{R}^{s\times\sum_{i=1}^n k_i}$ expresse all the mapped features from n feature modes.

Then, Z^n is randomly mapped to enhancement nodes for nonlinear transformation. Assuming that there are m groups of enhancement nodes, the output of the jth group of enhancement node is

$$H_j = \zeta_j\big(Z^n W_{hj} + \beta_{hj}\big), \quad j = 1, 2, \ldots, m \tag{14}$$

where weights W_{hj} and bias term β_{hj} are also generated randomly, and $\zeta_j(\cdot)$ represents the activation function of the jth enhancement node. The overall output of the enhancement layer can be expressed as

$$H^m = [H_1, H_2, \ldots, H_m] \tag{15}$$

where $H^m \in \mathbb{R}^{s\times m}$. Consequently, the output of BLS can be derived as

$$\begin{aligned} Y &= [Z_1, Z_2, \ldots, Z_n|H_m]W_n^m \\ &= \big[Z^n|H_m\big]W_n^m \end{aligned} \tag{16}$$

where W_n^m is the weights connecting the layer of feature and enhancement nodes to the output layer.

Let $A_n^m = [Z^n|H_m] \in \mathbb{R}^{s\times(\sum_{i=1}^n k_i+m)}$, then, the connection weights of a BLS can be rapidly approximated by the ridge regression [66]

$$W_n^m = \Big(\lambda I + \big(A_n^m\big)^T A_n^m\Big)^{-1}\big(A_n^m\big)^T Y \tag{17}$$

where $\lambda \in \mathbb{R}$ is a constant. The BLS has a simple structure, which effectively increases the training procedure and keeps the generalization ability of function approximation.

B. Random Forest Feature Selection

RFs is a popular ensemble learning methods consisting of multiple DTs [67]. Correlation between different DTs can be eliminated via a random adopted strategy. Each DT is developed from a random sample of the original training set. Each tree provides a classification or regression result, and the forest summarizes these results to formulate a more accurate and stable output. Hence, RF shows good performance in solving high-dimensional, nonlinear, and ill-posed classification and regression problems. A highly dimensional problem always has a vast number of input features. In establishing of prediction models, a feature x_j with a high correlation with the objective value y may not be an important feature helping the prediction, while some features with relatively low correlation coefficients could be more important. It is challenging to manually investigate the feature importance and select the most relevant features for prediction. Compared with other feature selection methods, RF is more explanatory and efficient. One key advantage of using RFs is that it can derive the importance score of each feature, which can be utilized to evaluate individual feature importance regarding the prediction results [68]. Hence, RF is adopted to select the important features.

First, we can utilize ordinary RF to derive the importance score of each feature. RF uses the mean-square error (MSE) or mean absolute error (MAE) to develop regression trees and determine regression results in each tree. The MSE at node v, MSE(v), measuring the impurity of v can be derived as

$$\mathrm{MSE}(v) = \sum_{i=1}^{I}\big(\hat{y}_i - y_i\big)^2 \tag{18}$$

where $\hat{y}_i$ is the regression results of sample i recording at node v. I is the number of samples divided for node v.

Again, the MSE of feature x_i for splitting the tree node v is defined as

$$\begin{aligned} \mathrm{Gain}(x_i, v) = \mathrm{MSE}(x_i, v) &- \mathrm{MSE}\big(x_i, v^L\big)W_L \\ &- \mathrm{MSE}\big(x_i, v^R\big)W_R \end{aligned} \tag{19}$$

where $\text{Gain}(x_i, v)$ represents the impurity of node v; v^L and v^R represent the left and right child node of node v, respectively; W_L and W_R stands for a fraction of examples assigned to the left and right child node, respectively. Generally, we adopt the feature maximizing the reduction in impurity as the splitting feature.

Moreover, we can derive the importance score of the tree-j for feature x_i from the $\text{Gain}(x_i, v)$

$$\text{Imp}_i^j = \frac{\sum_{b \in V_{x_i}^j} \text{Gain}(x_i, b)}{\text{Gain}\left(V^j\right)} \tag{20}$$

where V^j is the set of split nodes of the tree-j; $b \in V_{x_i}^j$ is a node set splitting on feature x_i; $\text{Gain}(V^j)$ is the sum of the impurity of all nodes in tree-j.

Normalization of the importance score is defined as

$$\text{NormImp}_i^j = \frac{\text{Imp}_i^j}{\text{Imp}_{\text{sum}}^j} \tag{21}$$

where Imp_i stands for the importance score of x_i and $\text{Imp}_{\text{sum}}^j$ stands for the sum of all features impurity in the tree-j, while the normalized importance score is $0 \le \text{NormImp}_i^j \le 1$.

Finally, in RF, the importance score of x_i is defined as

$$\text{Imp}_i^{RF} = \frac{\sum_{j=1}^{n_{\text{tree}}} \text{NormImp}_i^j}{n_{\text{tree}}} \tag{22}$$

where n_{tree} represents the number of tree.

C. Bagging

Bagging [69] is a kind of ensemble learning strategy that ensemble many weak learners to build a strong learning idea. It is mainly used for model improvement in machine learning and has wide applications in classification and regression tasks in prediction. Boostrap Sampling is a technique in Bagging. Algorithm 2 describes the process of Boostrap Sampling. Suppose one has a data set that contains M samples; then, each sample is randomly selected and put back into the original sample set. Repeat this procedure for N times. A subdata set containing N samples can be obtained. The probability of each sample in the original data set being selected is the same. In this work, we utilized the Boostrap Sampling strategy to establish T subdata sets containing n_s samples. Then, we utilize T subdata sets to establish T weak learners. Finally, a combination strategy is used to combine T weak learners into strong learners.

Assume that the sample data set of COVID-19 information is $D = \{S_1, S_2, \ldots, S_N\}$, where S_i represents a sample of the data set composed of features and predictive value. All the samples can be divided into a training and a test data set, the sizes of which are N_1 and N_2, respectively. A sampling ratio of p is set to determine the number of samples extracted from the original training data set to form a subtraining data set. The bootstrapping technique is adopted to make the selection procedure of the subtraining data set completely random. Bootstrapping technique draws samples after choosing these samples and then puts samples back into the original training data set. The process is shown in Algorithm 2.

Algorithm 1 Derive RF Feature Importance Score

Input: Training dataset: $D_{train} = \{S_1, S_2, \cdots, S_{N_1}\}$; Number of features in D_{train}: n_f; Number of selected features: $n_s (n_s \le n_f)$; Feature set of D_{train}: F_e;
Output: Selected feature set F_e;
Use D_{train} to build a RF model;
Set the number of RF trees: n_{tree};
for $j = 1$ to n_{tree} **do**
  Set $Imp_{\text{sum}}^j = 0$;
  for $i = 1$ to n_f **do**
    Set $Imp_{init} = 0$;
    Set the set of split nodes: V^j;
    Set the set of split nodes on feature x_i: $V_{x_i}^j$;
    for b in $V_{x_i}^j$ **do**
      $Imp_{init} = Imp_{init} + Gain(x_i, b)$;
    end for
    $Imp_i^j = \frac{Imp_{init}}{Gain(V^j)}$;
    $Imp_{\text{sum}}^j = Imp_{\text{sum}}^j + Imp_i^j$;
  end for
  for $i = 1$ to n_f **do**
    $\text{NormImp}_i^j = \frac{Imp_i^j}{Imp_{\text{sum}}^j}$;
  end for
end for
for $i = 1$ to n_f **do**
  Set $\text{NormImp}_i = 0$;
  for $j = 1$ to n_{tree} **do**
    $\text{NormImp}_i = \text{NormImp}_i + \text{NormImp}_i^j$;
  end for
  $Imp_i^{RF} = \frac{\text{NormImp}_i}{n_{\text{tree}}}$;
end for
$Imp^{RF} = \{Imp_1^{RF}, Imp_2^{RF}, Imp_3^{RF}, \cdots, Imp_{n_f}^{RF}\}$;
Sort Imp^{RF} from largest to smallest, get the sorted set $SortedImp^{RF}$;
Base $SortedImp^{RF}$ to extract the feature set F_e of the first n_s scores.
return F_e.

Algorithm 2 Boostrap Sampling

Input: Sample ratio: p;
Number of iterations: T; Training dataset: $D_{train} = \{S_1, S_2, \cdots, S_{N1}\}$;
Output: The sampled sample set D_{bs}
Set the number of subsamples: $N_b = \lfloor N_1 \cdot p \rfloor$;
for $i = 1$ to T **do**
  for $n = 1$ to N_b **do**
    Use random sampling to sample S_n from D_{train};
    S_n goes back to the data sample set D_{train};
    Put S_n into the data set D_{bs_i};
  end for
  $D_{bs_i} = \{S_1, S_2, \cdots, S_{N_b}\}$
  Put D_{bs_i} into the D_{bs};
end for
return $D_{bs} = \{D_{bs_1}, D_{bs_2}, D_{bs_3}, \cdots, D_{bsT}\}$.

D. Random-Forest-Bagging Broad Learning System

Here, the classic RF and ensemble learning-bagging are adopted to enhance the performance of BLS for the prediction of the spread of COVID-19. Here, we leverage the RF feature selection strategy for adopting important features to improve predictive performance. Then, we randomly sample data in these important features to form a number of subtraining data

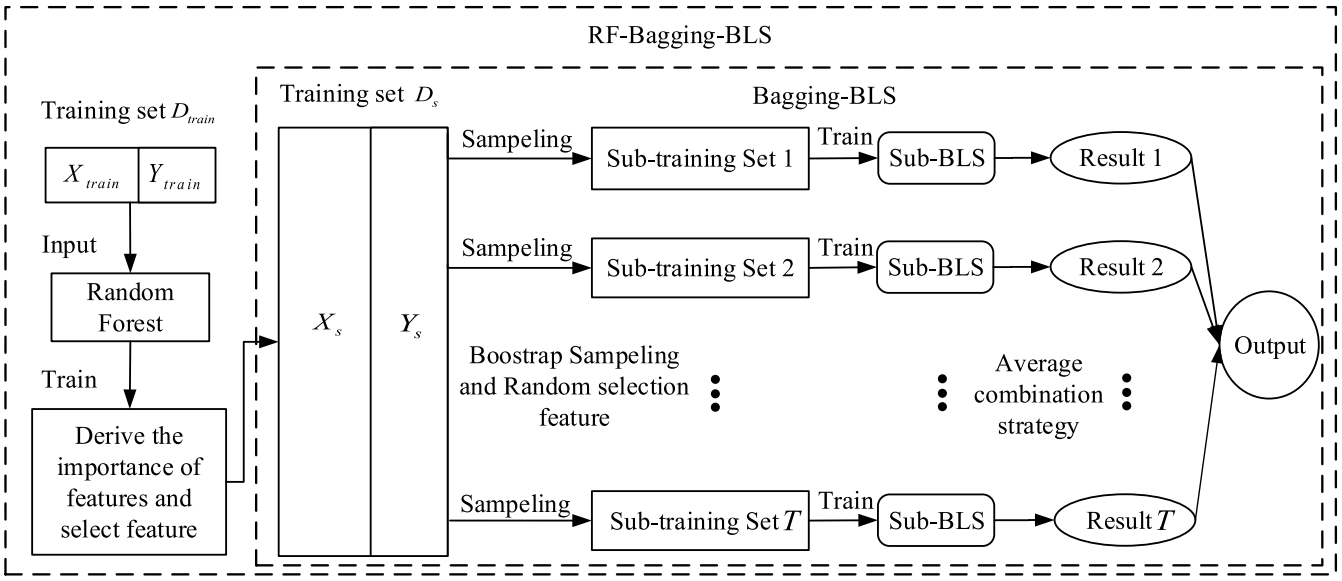


Fig. 3. Structure of bagging-BLS and RF-bagging-BLS.

Algorithm 3 RF-Bagging-BLS

Input: Training dataset :$D_{train} = \{S_1, S_2, \cdots, S_{N_1}\}$; Number of features in D_{train} : n_f; Number of selected features: $n_s, n_s \leq n_f$; Feature set of D_{train} : F; Number of features in D_{train} : n_f; Number of iterations : T; Sample ratio : p; Basic learning model :ξ;

Output: A prediction model H;

1: Enter D_{train}, n_f, n_s and F into Algorithm 1 to get the feature set F_e.
2: Selected data set : $D_s = \{F_e(D)|F_e \subseteq F\}$;
3: Enter p, T and D_s into Algorithm 2 to get the sampled set $D_{bs} = \{D_{bs_1}, D_{bs_2}, D_{bs_3}, \cdots, D_{bs_T}\}$;
4: **for** $i = 1$ to T **do**
5: Randomly generate feature set F_r from F_e;
6: Sub-training set :

$$D^2_{bs_i} = \{F_r(D_{bs_i})|F_r \subseteq F\};$$

7: Using $D^2_{bs_i}$ Training the sub-BLS :

$$h_i = \xi\left(D^2_{bs_i}\right);$$

8: **end for**
9: **return** $H(x) = argmin_{y \in Y} \sum_{i=1}^{T} \|h_i(x) - y(t+n)\|_2^2$.

sets based on bagging strategy and then build multiple independent BLS prediction models based on these subtraining data. Finally, we combine these results to provide a final prediction. The structure of the RF-Bagging-BLS model is shown in Fig. 3, in which X_{train} is the training input, Y_{train} is the expected training output. After the training data pass the feature importance analysis, X_s is the input after the feature selection in data set D_s, Y_s is excepted output in data set D_s.

The establishment of an RF-bagging-BLS includes the following parts.

1) *Feature Selection:* Multiple factors are related to the spread of epidemics. However, part of the features may be less relevant to the spread and redundant, decreasing learning ability. As the multiple features are mutually independent, we adopt an RF feature selection strategy to automatically adopt the most suitable features (shown in Algorithm 1).
2) *Establish Subtraining Data Set:* The whole data set is divided into two data sets: a) a training data set (of size N_1) and b) a test data set (of size N_2). $\lfloor N_1 \cdot p \rfloor$ samples are chosen from the training data set using the Boostrapping technique, where $0 < p < 1$ is the sampling ratio and $\lfloor x \rfloor$ represents the largest integer no more than x. This sampling process is repeated T times to prepare T different subtraining data sets for training submodels.
3) *Build the Sub-BLS Models:* In this model, each sub-BLS model is regarded as a weak learner in the ensemble learning model. Then, we combine multiple weak learners to form strong learners. Finally, the output of the RF-Bagging-BLS model (shown in Algorithm 3) can be computed by

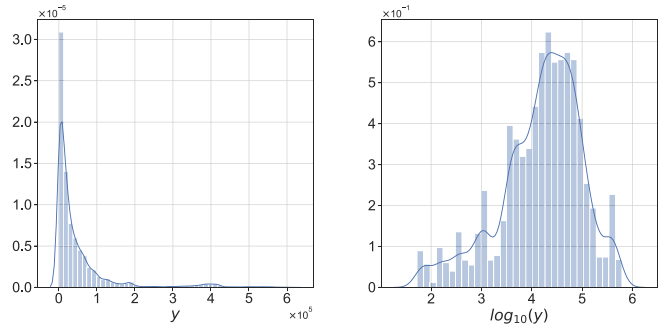
$$y_p = \frac{y_{p_1} + y_{p_2} + \cdots + y_{p_T}}{T} \tag{23}$$

where y_{p_i} is the predicted value of the ith learner, while y_p is final predicted value.

V. Experimental Results

In order to assess the performance of the proposed technique, we adopt several forecasting methods to evaluate the results. We employed LR model, KNN, DT, Ada, RF, GBDT, ETs regressor, SVR, CatBoost (CAT), LightGBM (LGB), XGBoost (XGB), and BLS approaches in the comparisons. We trained each model with data until September 30, 2020, reported by local health authorities in 184 countries and 1241 areas. Meanwhile, multiple evaluation metrics are adopted to evaluate the predictive power of each model.

Due to the large amount of human and financial resources required to achieve a comprehensive picture of the spread of COVID-19 in an area, the data released by many local



(a) (b)

Fig. 4. (a) Probability distribution of $y(t+n)$. (b) Probability distribution of $\log 10(y(t+n))$.

authorities will occasionally have some errors. For instance, the number of daily confirmed cases or the daily increase in the number of recovered people is negative. Hence, we have to process the data set by removing the abnormal data or filling in missing data. In this study, the predictive value $y(t+n)$ is the cumulative number of confirmed cases in an area, which increases monotonically. Here, we adopt $m = 7$ and $n = 7$; namely, we provide 7-days forecasts for 7-consecutive days in an area (cities, provinces, states, and other prefectures). According (10), we have more than 100 different candidate features. Studies indicate that, in some scenarios, data-driven methods perform well when the output data have a distribution close to a uniform or normal distribution [70]. We hope that the predicted value distribution is closer to the normal distribution to improve the model's generalization ability after training. However, the distribution of $y(t+n)$ is far from a normal distribution [shown in Fig. 4(a) with $n = 7$], while the distribution of $\log_{10}(y(t+n))$ is similar to a normal distribution [shown in Fig. 4(b)]. Here, we consider two scenarios: under scenario-I, first, machine learning models are developed to predict $\log_{10}(y(t+n))$ to achieve prediction value $\log_{10}(\hat{y}(t+n))$, and then perform the reverse operation to achieve the predictive value of $\hat{y}_{\log_{10}}(t+n) = 10^{\log_{10}(\hat{y}(t+n))}$; under scenario-II, we establish machine learning models predict $y(t+n)$ directly.

A. Correlation Analysis

In our case, there exists a large number of features that may influence the spread of COVID-19. In the proposed methods, we use RF for feature selection, while we adopted correlation analysis for feature selection for other classical models for a fair comparison. In order to adopt features highly correlated with the predictive value as the input of machine learning models, the violin chart is utilized. First, we derived all the correlation coefficients between input feature x_j and the predictive value y. Then, the violin chart (Fig. 5) shows the frequency (or probability distribution) of the absolute value of correlation coefficients. The correlation coefficients range from 0.0004 to 0.9953, while upper and lower quartiles are 0.0436 and 0.8052, respectively. The median of the correlation coefficient is about 0.2. Note that the violin chart is clearly separated into two parts (the dashed line in Fig. 5): one class of the features with the absolute value of the correlation coefficient is greater than 0.5, while the other is less than 0.5. Then, according to this observation, $\gamma_{\text{thr}} = 0.5$ was taken as the threshold. The feature with a correlation coefficient greater than 0.5 after taking the absolute value is taken as the input of machine learning models. Finally, 50 features are selected as the input of classical models except for the proposed RF-BLS and RF-Bagging-BLS. The 50 features are divided into two categories, including original and augmented features (shown in Table III).

TABLE III
50 IMPORTANT FEATURES FOR CLASSICAL MODELS

	Selected features	Description
Original features	$I_C(t-m)$	Cumulative confirmed cases at day $t-m$ $(m=1,\cdots,7)$
	$R(t-m)$	Total recovered cases at day $t-m$ $(m=1,\cdots,7)$
	$D(t-k)$	Death toll at day $t-m$ $(m=1,\cdots,7)$
	$N_T(t-m)$	Cumulative COVID-19 tests at day $t-m$ $(m=1,\cdots,7)$
	x_P	The population of an area
Augmented features	$I(t-m)$	Activate COVID-19 patients at day $t-m$ $(m=1,\cdots,7)$
	$\Delta I_C(t-m)$	Daily confirmed cases at day $t-m$ $(m=1,\cdots,7)$
	$\Delta N_T(t-m)$	Daily COVID-19 tests at day $t-m$ $(m=1,\cdots,7)$

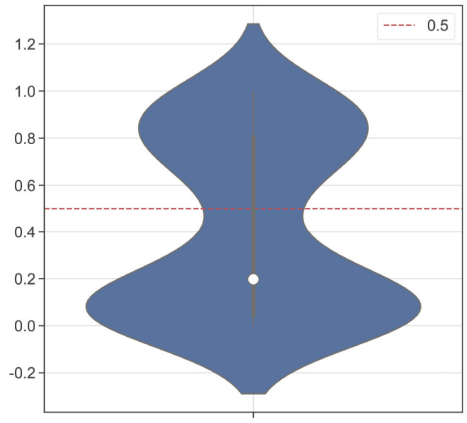


Fig. 5. Absolute value of the correlation coefficient between features and the predictive value: the white point is the median value of data; the upper and lower bounds of the middle black box represent the upper and lower quartiles of the correlation coefficients, respectively; the upper and lower bounds of the middle black line represent the maximum and minimum values of the correlation coefficients; the shape of the violin indicates the frequency (or estimated probability distribution) of correlation coefficients.

B. Experimental Results by Classical Methods

After selecting the relevant features, we used all the models mentioned above to build the predictive models. Since the scale of different features varies in a large range, we first adopted the Z-score standardized method to normalize the feature data on the same scale

$$x^* = \frac{x - \bar{x}}{\sigma} \tag{24}$$

where $\bar{x}$ and σ represent the mean of feature and the standard deviation of feature, respectively.

TABLE IV
SCENARIO-I: THE EVALUATION VALUE OF DIFFERENT MODELS WITH THE PREDICTIVE VALUE $\hat{y}_{\log 10}(t+n)$

Method	RMSE	MAE	R^2	MAD	R^2_{adj}	MAPE
LR	57387.6106e+9	76255.7453e+8	-1.5816e+17	21649.1423	-1.6344e+17	2.5650+9
KNN	32589.3837	13272.0041	0.9498	3875.6634	0.9472	13.7311
DT	20441.1788	9758.9613	0.9799	3664.0001	0.9792	31.8415
SVR	119751.7181	43284.9216	0.3113	6676.6545	0.2883	28.2169
Ada	27126.1988	9364.6234	0.9646	1033.5000	0.9634	5.9561
RF	23413.5558	8136.2741	0.9736	1090.4159	0.9727	5.2950
GBDT	25739.2592	9643.7920	0.9681	1791.4312	0.9671	8.3891
ET	21978.5323	**7612.5017**	0.9768	1292.7426	0.9760	6.1060
CAT	31788.7347	11952.3085	0.9515	1533.8260	0.9498	8.4482
LGB	24249.7400	8033.8196	0.9717	**939.8207**	0.9708	**5.4479**
XGB	29203.4126	9389.2712	0.9590	1107.6260	0.9576	5.5478
BLS	35079.3674	16463.1951	0.9409	9145.3791	0.9389	27.6222
RF-BLS	39131.9767	23896.6042	0.9265	9328.3896	0.9255	40.3023
Bagging-BLS	**13539.3023**	9882.3092	**0.9911**	7690.6102	**0.9909**	23.0960
RF-Bagging-BLS	35793.3709	19923.3754	0.9384	8895.5277	0.9376	39.5285

TABLE V
SCENARIO-II: THE EVALUATION VALUE OF DIFFERENT MODEL WITH PREDICTIVE VALUE $y(t+n)$

Method	RMSE	MAE	R^2	MAD	R^2_{adj}	MAPE
LR	3394.6494	1601.8012	0.9994	642.4868	0.9994	3.2467
KNN	32267.9550	16772.8781	0.9483	6280.7837	0.9483	22.3130
DT	24733.7204	10304.4542	0.9706	1423.0000	0.9696	10.8748
SVR	53319.4198	23591.0204	0.8634	8708.3930	0.8589	63.6835
Ada	17861.0283	5933.6822	0.9846	1196.0000	0.9842	5.5883
RF	23770.2149	8554.5544	0.9728	1218.9372	0.9719	6.7004
GBDT	21103.7782	7325.3585	0.9786	977.8010	0.9778	7.2672
ET	19635.1575	6997.1890	0.9814	1469.4092	0.9808	6.7076
CAT	25744.2812	8687.9153	0.9681	1351.8562	0.9671	7.3150
LGB	25470.3411	9120.9554	0.9688	973.4710	0.9678	7.9324
XGB	24354.7123	8455.0691	0.9715	1082.3067	0.9705	6.8996
BLS	2269.8906	1362.8407	0.9997	766.4345	0.9997	6.7834
RF-BLS	2042.4723	1018.2183	0.9997	565.5706	0.9997	6.4979
Bagging-BLS	2872.1029	1475.1688	0.9996	697.3510	0.9995	3.9903
RF-Bagging-BLS	**1989.1970**	**952.5739**	**0.9998**	**432.3244**	**0.9998**	**3.0090**

Each machine learning model has a large volume of hyperparameters to be adjusted to achieve a satisfying performance. Here, a grid search method is adopted for searching the optimal hyperparameters [71]. The regression task evaluation metrics, including MAE, relative MSE (RMSE), coefficient of determination (R^2), adjusted coefficient of determination (R^2_{adj}), median absolute error (MAD), and mean absolute percentage error (MAPE), were adopted to evaluate the performance of each model. Suppose that the test set has n samples, the characteristics of the model input are k, $\hat{y}$ is the predicted value of sample i, y_i is the actual value of the sample, $\text{median}_n(.)$ represents the median value of n samples

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2} \tag{25}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - y_i\right| \tag{26}$$

$$\text{MAD} = \text{median}_n\left(\left|\hat{y}_i - \hat{y}_i\right|\right) \tag{27}$$

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{28}$$

$$\text{SS}_{\text{res}} = \sum\left(\hat{y}_i - \hat{y}_i\right)^2 \tag{29}$$

$$\text{SS}_{\text{tot}} = \sum\left(\hat{y}_i - \bar{y}\right)^2 \tag{30}$$

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}} = 1 - \frac{\sum\left(y_i - \hat{y}_i\right)^2}{\sum\left(y_i - \bar{y}\right)^2} \tag{31}$$

$$\text{R}_{\text{adj}}{}^2 = 1 - \left[\frac{\left(1 - R^2\right)(n-1)}{n-k-1}\right] \tag{32}$$

where SS_{res} is residual sum of squares and SS_{tot} represents total sum of squares.

Tables IV and V show the regression task evaluation metrics for each model under two scenarios. Without considering the proposed RF-BLS and RF-Bagging-BLS methods, LR takes a good effect in predicting $y(t+n)$, with 3394.6494 for RMSE, 1601.8012 for MAE, and 0.9994 for R^2. DT makes a better effect to predict $\hat{y}_{\log 10}(t+n)$. The RMSE and MAE of DT in case two are 20441.1788 and 9758.9613, respectively. However, DT shows better robustness in predicting $y(t+n)$, with 1423 for MAD. For ensemble learning models, Ada has the best predictive effect compared to other models. In case one for predicting $y(t+n)$, RMSE and MAE are 17861.0283 and 5933.6822, respectively. LGB has the best robustness among the classical ensemble learning models with MAD equals 939.8207. Among all the classical methods, the predictive performance of BLS is better than all the other methods. In scenario-II, The RMSE, MAE, and R^2 of BLS are 2269.8906, 1362.8407, and 0.9997, respectively. It means BLS has a better fitting effect on the test set. By comparing Tables IV and V, we find that DT and LGB have better results in predicting $\hat{y}_{\log_{10}}(t+n)$, which means that the data has a normal distribution and a certain lifting effect for some models.

C. Experimental Results by RF-Bagging-BLS

1) RF-BLS: The RF-BLS is established by BLS using the features adopted through RF. Compared with classical BLS, the predicted RF-BLS results are slightly improved, indicating that the feature selection strategy base on RF helps improve BLS performance. Table V shows the predicted results of RF-BLS and BLS. The RMSE, MSE, and MAE of RF-BLS predicting $y(t+n)$ are 4171692.9897, 2042.4723, and 1018.2183, respectively. It can be observed that the RMSE, MSE and MAE of the RF-BLS are lower than that of BLS.

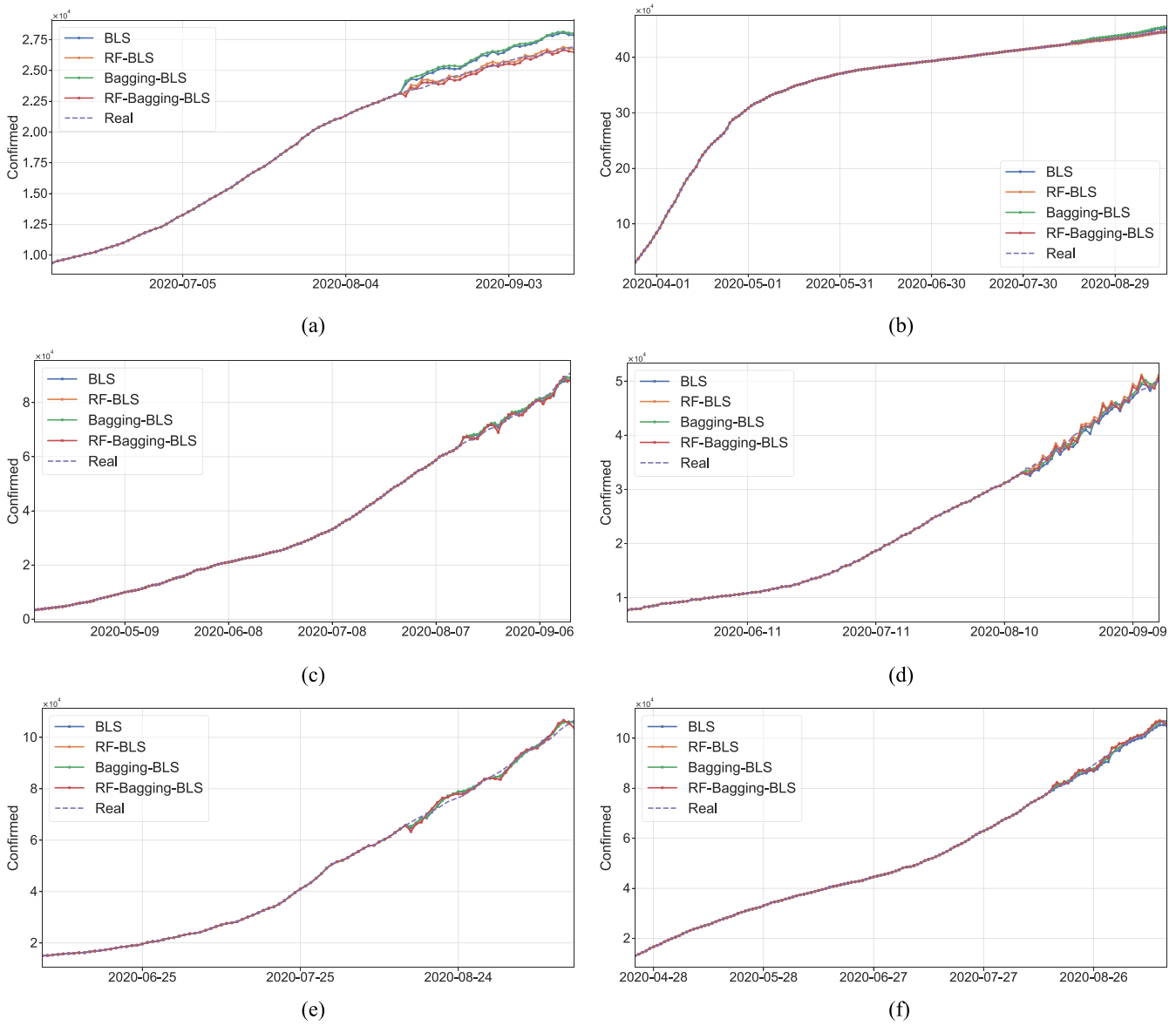


Fig. 6. Forecast results generated by BLS, RF-BLS, Bagging-BLS, and RF-Bagging-BLS; (a) New Mexico. (b) New York. (c) Wiscosin. (d) Kansas. (e) Missouri. (f) Indiana.

Meanwhile, compared with BLS, RF-BLS is also more robust and the MAD is 565.5706.

2) Bagging-BLS: The Bagging-BLS is a combination model of Bagging and BLS [72]. Fig. 3 shows the structure of Bagging-BLS. It is worth noting that the feature selection used by Bagging-BLS is still a correlation analysis. Through RMSE and MAE in Table V, Bagging-BLS also has better robustness than classical BLS. The MAD of Bagging-BLS predicts $y(t+n)$ is 565.5706. However, by comparing RMSE and MAE, the Bagging-BLS model only increase the predictive accuracy slightly.

3) RF-Bagging-BLS: Experimental results show that the proposed RF-Bagging-BLS achieve the best value among all the comparison algorithms. All the evaluation metric of BLS are all the best (shown in Table V). In case one, the RMSE, MAE R^2 of RF-Bagging-BLS predicts $y(t+n)$ are 1989.1970, 952.5739 and 0.9998, respectively. Additionally, the metric MAD shows that RF-Bagging-BLS has great robustness. Fig. 6 shows the prediction results of New Mexico, New York, Wisconsin, Kansas, Missouri, and Indiana by BLS, RF-BLS, Bagging-BLS, and RF-Bagging-BLS, respectively. Several researchers utilize CNN, LSTM, and GRU models to forecast the spread of COVID-19 based on a data set of one or a few countries. In this work, we also test CNN, LSTM, and GRU based on our data set. However, these methods are easily overfitting. For instance, LSTM (with three layers, 64 nodes), GRU (four layers, 128 nodes) would converge in 1000 epochs, and 1-D CNN (one layer, three kernel size, one stride) would converge in 100 epochs. GRU achieves 25021.441 RMSE in training data, while 147259.83 RMSE in testing data; LSTM achieves 18975.209 RMSE in training data, while 169766.4 RMSE in testing data; 1-D CNN achieves 11937.837 RMSE in training data, but performed poorly in the testing data. Experimental results indicate that these neural network models

are easily overfitting based on a small data set and provide poor performance in this task.

Previous work points outs that in some cases that data-driven methods perform well when the output data have a distribution close to a uniform or normal distribution. Hence, in this work, we tested two scenarios: under scenario-I, first, we develop machine learning models to predict $\log_{10}(y(t+n))$ to achieve the predictive value $\log_{10}(\hat{y}(t+n))$, then derive $\hat{y}(t+n)$, while under scenario-II, machine learning models predict the cumulative confirmed cases $y(t+n)$ directly. However, experimental results indicate that, under scenario one, all the models achieve a worse performance in forecasting $y(t+n)$. Hence, the predictive output $\log_{10}(y(t+n))$ with a normal distribution cannot help improve the performance. Additionally, according to the experimental results presented in Tables IV and V, most machine learning models can achieve relatively better performance under scenario-II. The proposed RF-Bagging BLS show the best performance in predicting $y(t+n)$, which is also the best result under two scenarios. Experiment results indicate that the distribution of predictive value affects the performance of data-driven models [73]. However, in our case, the predictive value with a uniform or normal distribution may not help in improving the predictive performance. BLS algorithm is efficient for training. In our work, the mean training time of RF-BLS, Bagging-BLS, and RF-Bagging-BLS, is 15.346, 0.500, and 17.175 s, respectively. After training, the time required to predict a result is 0.273, 0.287, and 0.467 s for RF-BLS, Bagging-BLS, and RF-Bagging-BLS, respectively.

VI. Conclusion

COVID-19 has become a global public health threat and spread to more than 200 countries by September 30, 2020. How to contain the spread of COVID-19 becomes a challenging task for policymakers to assess health care requirements to estimate the present trends, determine public health measures, and flatten the COVID-19 curve shortly. Until now, most researchers just developed classical models based on epidemic spreading data covering one or several countries. However, the behavior of the COVID-19 outbreak varies from region-to-region. The number of confirmed cases released by the local authorities could be influenced by multiple factors, such as testing capacity and other related factors. Additionally, classical epidemiological models usually do not consider extra details, such as testing capacity, population, geographic information, etc. With limited testing capacity and human resources, significant delays in identifying, isolating, and reporting cases due to the magnitude of the epidemic are unavoidable, which has a negative impact on the predictive performance. Hence, whether these models can be applied to other countries or not is a question.

Due to the complex nature of forecasting the COVID-19 trend, we suggest machine learning as an effective technique to model the outbreak. Recently, most of the local authorities provide COVID-19 information, which scatters on dozens of public databases. In this work, we collect and unified these data sets into one comprehensive data set, including the epidemic spread data, geographic information, economic information, population, COVID-19 testing information of 184 countries and 1241 areas (cities, provinces, states, and other areas). Then, a hybrid machine learning model of RF-Bagging-BLS is developed for predicting the COVID-19 trend in more than 180 countries and 1200 areas. The proposed models showed promising performance in timely short-term forecasts without the requirement of epidemiological models. RF-Bagging-BLS model outperformed other models by delivering accurate results on validation samples. Experimental results show that the proposed method presents the best result in all evaluation criteria, indicating the RF-Bagging-BLS model is suitable with this training data set. An accurate prediction of the pandemic situation can help the authority to evaluate the hospital capacity needs and provide effective help for the government to adopt epidemic prevention policies [74], [75]. However, inaccurate predictions of cases may lead to a loosening of containment policies, leading to the emergence of another wave of infection and a rapid increase in the number of infected cases [76]. The effective implementation of public health interventions, such as social distancing, lockdown, and personal protection, will be critical to bringing the epidemic under control. Short-term forecasting of COVID-19 can help policymakers, including health managers, public health officials, etc., to prepare medical resources, organize health care to confront the epidemic, plan nonpharmaceutical interventions required to mitigate an outbreak, finally contain the epidemic outbreak or even eliminate the pandemic. This work can help local authorities to make suitable decisions in the future.

The world today is connected with smart devices [77]–[79]. Data is recorded and shared between the regions in an unprecedented way than ever before [80]. The availability of timely and high-quality pandemic data can help scholars develop data-driven methods to analyze the pandemic situation. Weather and air quality are other important factors. However, the proposed data set did not include the detailed weather and air quality data, such as daily temperature, wind speed, etc. We replace the weather data with geographic location data, which may be too rough in this work. Experimental results show that the proposed method present the best result in all the evaluated criterion, indicating the proposed method is suitable with this training data set. In the future work, we would quantify the effort of public health measures, and consider migration data and weather data. Machine learning has been shown as a powerful tool in healthcare. We would explore the true capability of the proposed hybrid model.

References

[1] R. Wölfel *et al.*, "Virological assessment of hospitalized patients with COVID-2019," *Nature*, vol. 581, no. 7809, pp. 465–469, 2020.
[2] X. He *et al.*, "Temporal dynamics in viral shedding and transmissibility of COVID-19," *Nat. Med.*, vol. 26, no. 5, pp. 672–675, 2020.
[3] S. Ou *et al.*, "Machine learning model to project the impact of COVID-19 on U.S. motor gasoline demand," *Nat. Energy*, vol. 5, pp. 666–673, Jul. 2020.
[4] H. S. Badr, H. Du, M. Marshall, E. Dong, M. M. Squire, and L. M. Gardner, "Association between mobility patterns and COVID-19 transmission in the USA: A mathematical modelling study," *Lancet Infectious Diseases*, vol. 20, no. 11, pp. 1247–1257, 2020.

[5] Y. Wang, Y. Yuan, Q. Wang, C. Liu, Q. Zhi, and J. Cao, "Changes in air quality related to the control of coronavirus in China: Implications for traffic and industrial emissions," *Sci. Total Environ.*, vol. 731, Aug. 2020, Art. no. 139133.

[6] P. M. Forster *et al.*, "Current and future global climate impacts resulting from COVID-19," *Nat. Climate Change*, vol. 10, pp. 913–919, Oct. 2020.

[7] S. K. Lau *et al.*, "Possible BAT origin of severe acute respiratory syndrome coronavirus 2," *Emerg. Infectious Diseases*, vol. 26, no. 7, p. 1542, 2020.

[8] M. F. Boni *et al.*, "Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic," *Nat. Microbiol.*, vol. 4, pp. 1408–1417, Jul. 2020, doi: 10.1038/s41564-020-0771-4.

[9] T. T.-Y. Lam *et al.*, "Identifying SARS-CoV-2-related coronaviruses in malayan pangolins," *Nature*, vol. 583, pp. 282–285, Jul. 2020.

[10] Worldometer. (2020). *COVID-19 Coronavirus Pandemic*. [Online]. Available: https://www.worldometers.info/coronavirus/

[11] Agence France Presse. (2020). *Coronavirus: 4.5 Billion People Confined*. [Online]. Available: https://www.barrons.com/news/coronavirus-4-5-billion-people-confined-01587139808

[12] D. Cucinotta and M. Vanelli, "Who declares COVID-19 a pandemic," *Acta Bio-Medica Atenei Parmensis*, vol. 91, no. 1, pp. 157–160, 2020.

[13] C. Zhan, C. Tse, Y. Fu, Z. Lai, and H. Zhang, "Modelling and prediction of the 2019 coronavirus disease spreading in china incorporating human migration data," *PLoS ONE*, vol. 15, no. 10, 2020, Art. no. e0241171, doi: 10.1371/journal.pone.0241171.

[14] D. Fattorini and F. Regoli, "Role of the chronic air pollution levels in the COVID-19 outbreak risk in Italy," *Environ. Pollution*, vol. 264, Sep. 2020, Art. no. 114732.

[15] Y. Bai *et al.*, "Presumed asymptomatic carrier transmission of COVID-19," *J. Amer. Med. Assoc.*, vol. 323, no. 14, pp. 1406–1407, 2020.

[16] C. Rothe *et al.*, "Transmission of 2019-NCOV infection from an asymptomatic contact in Germany," *New England J. Med.*, vol. 382, no. 10, pp. 970–971, 2020.

[17] J. Hellewell *et al.*, "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," *Lancet Global Health*, vol. 8, no. 4, pp. 488–496, 2020.

[18] S. Mallapaty, "Antibody tests suggest that coronavirus infections vastly exceed official counts," *Nature*, to be publilshed.

[19] J. Peto, "COVID-19 mass testing facilities could end the epidemic rapidly," *BMJ*, vol. 368, Mar. 2020, Art. no. m1163.

[20] World Health Organization. (2020). *COVID-19 Strategy Update 2020*. [Online]. Available: https://www.who.int/docs/default-source/coronaviruse/covid-strategy-update-14april2020.pdf

[21] E. Callaway, "The unequal scramble for coronavirus vaccines-by the numbers," *Nature*, vol. 584, no. 7822, pp. 506–507, 2020.

[22] P. G. Walker *et al.*, "The impact of COVID-19 and strategies for mitigation and suppression in low-and middle-income countries," *Science*, vol. 369, no. 6502, pp. 413–422, 2020.

[23] S. T. Ali *et al.*, "Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions," *Science*, vol. 369, no. 6507, pp. 1106–1109, 2020.

[24] R. E. Baker, W. Yang, G. A. Vecchi, C. J. E. Metcalf, and B. T. Grenfell, "Susceptible supply limits the role of climate in the early SARS-CoV-2 pandemic," *Science*, vol. 369, no. 6501, pp. 315–319, 2020.

[25] C. Zhan, Y. Zheng, Z. Lai, T. Hao, and B. Li, "Identifying epidemic spreading dynamics of COVID-19 by pseudocoevolutionary simulated annealing optimizers," *Neural Comput. Appl.*, to be published.

[26] K. Prem *et al.*, "The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: A modelling study," *Lancet Public Health*, vol. 5, no. 5, pp. 261–270, 2020.

[27] C. J. Carlson, E. Dougherty, M. Boots, W. Getz, and S. J. Ryan, "Consensus and conflict among ecological forecasts of zika virus outbreaks in the united states," *Sci. Rep.*, vol. 8, no. 1, pp. 1–15, 2018.

[28] S. F. Ardabili *et al.*, "COVID-19 outbreak prediction with machine learning," in *Proc. SSRN*, 2020, Art. no. 3580188.

[29] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight COVID-19," *Phys. Genom.*, vol. 52, no. 4, pp. 200–202, 2020.

[30] M. H. D. M. Ribeiro, R. G. da Silva, V. C. Mariani, and L. dos Santos Coelho, "Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for brazil," *Chaos Solitons Fractals*, vol. 135, Jun. 2020, Art. no. 109853.

[31] L. Yan *et al.*, "An interpretable mortality prediction model for COVID-19 patients," *Nat. Mach. Intell.*, vol. 2, pp. 283–288, May 2020.

[32] C. Iwendi *et al.*, "COVID-19 patient health prediction using boosted random forest algorithm," *Front. Public Health*, vol. 8, p. 357, Jul. 2020.

[33] B. R. Beck, B. Shin, Y. Choi, S. Park, and K. Kang, "Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 784–790, Mar. 2020.

[34] A. Rivas, "Drones and artificial intelligence to enforce social isolation during COVID-19 outbreak," *Medium Towards Data Sci.*, vol. 26, Mar. 2020.

[35] A. Vespignani *et al.*, "Modelling COVID-19," *Nat. Rev. Phys.*, vol. 2, no. 6, Jun. 2020, pp. 279–281.

[36] N. Zheng *et al.*, "Predicting COVID-19 in China using hybrid Ai model," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 2891–2904, Jul. 2020.

[37] V. K. R. Chimmula and L. Zhang, "Time series forecasting of COVID-19 transmission in Canada using LSTM networks," *Chaos Solitons Fractals*, vol. 135, Jun. 2020, Art. no. 109864.

[38] G. Pinter, I. Felde, A. Mosavi, P. Ghamisi, and R. Gloaguen, "COVID-19 pandemic prediction for hungary; a hybrid machine learning approach," *Mathematics*, vol. 8, no. 6, p. 890, 2020.

[39] K. Jahanbin and V. Rahmanian, "Using Twitter and Web news mining to predict COVID-19 outbreak," *Asian Pac. J. Tropical Med.*, vol. 13, no. 8, pp. 378–380, 2020.

[40] D. Lazer, R. Kennedy, G. King, and A. Vespignani, "The parable of Google flu: Traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[41] W. Naudé, "Artificial intelligence vs COVID-19: Limitations, constraints and pitfalls," *Ai Soc.*, vol. 35, no. 3, pp. 761–765, 2020.

[42] C. P. Chen and Z. Liu, "Broad learning system: An effective and efficient incremental learning system without the need for deep architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 1, pp. 10–24, Jan. 2017.

[43] M. Xu, M. Han, C. L. P. Chen, and T. Qiu, "Recurrent broad learning systems for time series prediction," *IEEE Trans. Cybern.*, vol. 50, no. 4, pp. 1405–1417, Apr. 2020.

[44] N. M. Ghazaly, M. A. Abdel-Fattah, and A. Abd El-Aziz, "Novel coronavirus forecasting model using nonlinear autoregressive artificial neural network," *J. Adv. Sci.*, vol. 29, no. 5s, pp. 1831–1849, 2020.

[45] Z. Hu, Q. Ge, S. Li, E. Boerwinkle, L. Jin, and M. Xiong, "Forecasting and evaluating multiple interventions for COVID-19 worldwide," *Front. Artif. Intell.*, vol. 3, p. 41, May 2020.

[46] C. Zhan, K. T. Chi, Z. Lai, X. Chen, and M. Mo, "General model for COVID-19 spreading with consideration of intercity migration, insufficient testing and active intervention: Application to study of pandemic progression in Japan and USA," *JMIR Public Health Surveillance*, vol. 6, no. 3, 2020, Art. no. e18880.

[47] R. G. da Silva, M. H. D. M. Ribeiro, V. C. Mariani, and L. dos Santos Coelho, "Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables," *Chaos Solitons Fractals*, vol. 139, Oct. 2020, Art. no. 110027.

[48] D. Fanelli and F. Piazza, "Analysis and forecast of COVID-19 spreading in China, Italy and France," *Chaos Solitons Fractals*, vol. 134, May 2020, Art. no. 109761.

[49] N. Chintalapudi, G. Battineni, and F. Amenta, "COVID-19 disease outbreak forecasting of registered and recovered cases after sixty day lockdown in Italy: A data driven model approach," *J. Microbiol. Immunol. Infection*, vol. 53, no. 3, pp. 396–403, Jun. 2020.

[50] S. Tiwari, S. Kumar, and K. Guleria, "Outbreak trends of coronavirus (COVID-19) in India: A prediction," *Disaster Med. Public Health Preparedness*, vol. 14, no. 5, pp. e33–e38, 2020.

[51] C. Zhan, K. T. Chi, Z. Lai, T. Hao, and J. Su, "Prediction of COVID-19 spreading profiles in South Korea, Italy, and Iran by data-driven coding," *PLoS ONE*, vol. 15, no. 7, 2020, Art. no. e0234763, doi: 10.1371/journal.pone.0234763.

[52] R. Chowdhury *et al.*, "Dynamic interventions to control COVID-19 pandemic: A multivariate prediction modelling study comparing 16 worldwide countries," *Eur. J. Epidemiol.*, vol. 35, no. 5, pp. 389–399, 2020.

[53] A. Tomar and N. Gupta, "Prediction for the spread of COVID-19 in India and effectiveness of preventive measures," *Sci. Total Environ.*, vol. 728, Aug. 2020, Art. no. 138762.

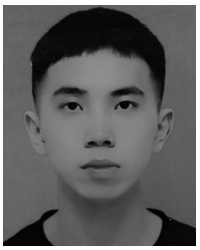
[54] Z. Yang *et al.*, "Modified SEIR and ai prediction of the epidemics trend of COVID-19 in China under public health interventions," *J. Thoracic Disease*, vol. 12, no. 3, p. 165, 2020.

[55] C.-J. Huang, Y. Shen, P.-H. Kuo, and Y.-H. Chen, "Novel spatiotemporal feature extraction parallel deep neural network for forecasting confirmed cases of coronavirus disease 2019," *Socio-Econ. Plan. Sci.*, Nov. 2020, Art. no. 100976.

[56] C. Poirier *et al.*, "Real-time forecasting of the COVID-19 outbreak in Chinese provinces: Machine learning approach using novel digital data and estimates from mechanistic models," *J. Med. Internet Res.*, vol. 22, no. 8, 2020, Art. no. e20285.
[57] M. A. Al-qaness, A. A. Ewees, H. Fan, and M. Abd El Aziz, "Optimization method for forecasting confirmed cases of COVID-19 in China," *J. Clin. Med.*, vol. 9, no. 3, p. 674, 2020.
[58] S. Ghamizi, R. Rwemalika, M. Cordy, Y. Le Traon, and M. Papadakis, "Pandemic simulation and forecasting of exit strategies: Convergence of machine learning and epidemiological models," Interdiscipl. Centre Security Rel. Trust, Univ. Luxembourg, Luxembourg City, Luxembourg, Rep., 2020.
[59] S. Porcher, "Response2COVID19, a dataset of governments responses to COVID-19 all around the world," *Sci. Data*, vol. 7, no. 1, pp. 1–9, 2020.
[60] B. Sheng, P. Li, Y. Zhang, L. Mao, and C. P. Chen, "GreenSea: Visual soccer analysis using broad learning system," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1463–1477, Mar. 2021.
[61] J. Guo, L. Wang, K. Fan, and B. Yang, "An efficient model for predicting setting time of cement based on broad learning system," *Appl. Soft Comput.*, vol. 96, Nov. 2020, Art. no. 106698.
[62] Y. Yang, Z. Gao, Y. Li, Q. Cai, N. Marwan, and J. Kurths, "A complex network-based broad learning system for detecting driver fatigue from eeg signals," *IEEE Trans. Syst., Man, Cybern., Syst.*, early access, Dec. 6, 2019, doi: 10.1109/TSMC.2019.2956022.
[63] S. Feng and C. L. P. Chen, "Fuzzy broad learning system: A novel neuro-fuzzy model for regression and classification," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 414–424, Feb. 2020.
[64] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[65] C. L. P. Chen, "A rapid supervised learning neural network for function interpolation and approximation," *IEEE Trans. Neural Netw.*, vol. 7, no. 5, pp. 1220–1230, Sep. 1996.
[66] F. Stulp and O. Sigaud, "Many regression algorithms, one unified model: A review," *Neural Netw.*, vol. 69, pp. 60–79, Sep. 2015.
[67] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[68] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Comput. Stat. Data Anal.*, vol. 52, no. 4, pp. 2249–2260, 2008.
[69] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
[70] D. Pyle, *Data Preparation for Data Mining*. London, U.K.: Morgan Kaufmann, 1999.
[71] I. Syarif, A. Prugel-Bennett, and G. Wills, "SVM parameter optimization using grid search and genetic algorithm to improve classification performance," *Telkomnika*, vol. 14, no. 4, p. 1502, 2016.
[72] M. Wang, Q. Ge, H. Jiang, and G. Yao, "Wear fault diagnosis of aeroengines based on broad learning system and ensemble learning," *Energies*, vol. 12, no. 24, p. 4750, 2019.
[73] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[74] G. E. Weissman *et al.*, "Locally informed simulation to predict hospital capacity needs during the COVID-19 pandemic," *Ann. Internal Med.*, vol. 173, no. 1, pp. 21–28, 2020.
[75] N. Pearce, J. P. Vandenbroucke, T. J. VanderWeele, and S. Greenland, "Accurate statistics on COVID-19 are essential for policy guidance and decisions," *Amer. J. Public Health*, vol. 110, no. 7, pp. 949–951, 2020.
[76] Y. Hazem, S. Natarajan, and E. Berikaa. (2020). *Hasty Reduction of COVID-19 Lockdown Measures Leads to the Second Wave of Infection*. [Online]. Available: https://doi.org/10.1109/10.1101/2020.05.23.20111526
[77] A. Rahman, M. S. Hossain, N. A. Alrajeh, and F. Alsolami, "Adversarial examples—Security threats to COVID-19 deep learning systems in medical IoT devices," *IEEE Internet Things J.*, early access, Aug. 3, 2020, doi: 10.1109/JIOT.2020.3013710.
[78] J. Zhang, *et al.*, "Joint resource allocation for latency-sensitive services over mobile edge computing networks with caching," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4283–4294, Jun. 2019.
[79] X. Wang, Z. Ning, S. Guo, and L. Wang, "Imitation learning enabled task scheduling for online vehicular edge computing," *IEEE Trans. Mobile Comput.*, early access, Jul. 28, 2020, doi: 10.1109/TMC.2020.3012509.
[80] Z. Ning *et al.*, "Mobile edge computing enabled 5G health monitoring for Internet of medical things: A decentralized game theoretic approach," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 463–478, Feb. 2021.

Choujun Zhan (Member, IEEE) received the B.S. degree in automatic control engineering from Sun Yat-sen University, Guangzhou, China, in 2007, and the Ph.D. degree in electronic engineering from the City University of Hong Kong, Hong Kong, in 2012.

After graduation, he worked as a Postdoctoral Fellow with the Hong Kong Polytechnic University, Hong Kong. Since Fall 2016, he has been an Associate Professor with the Department of Electronic Communication and Software Engineering, Nanfang College of Sun Yat-sen University, Guangzhou. He is currently a Professor with the School of Computer, South China Normal University, Guangzhou. His research interests include complex networks, time-series modeling and prediction, epidemic spreading, information diffusion, and machine learning.



Yufan Zheng is currently pursuing the B.S. degree with the Nanfang College of Sun Yat-sen University, Guangdong, China.

His recent research interests include data mining of time series, machine learning, deep learning, and epidemic spreading based on complex networks and artificial intelligence.

Haijun Zhang (Senior Member, IEEE) received the B.Eng. and master's degrees from Northeastern University, Shenyang, China, in 2004 and 2007, respectively, and the Ph.D. degree from the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, in 2010.

He was a Postdoctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada, from 2010 to 2011. Since 2012, he has been with the Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen, China, where he is currently a Professor of Computer Science. His current research interests include neural networks, data mining, machine learning, computational advertising, and service computing.

Prof. Zhang is currently an Associate Editor of *Neurocomputing*, *Neural Computing and Applications*, and *Pattern Analysis and Applications*.

Quansi Wen (Member, IEEE) received the bachelor's and master's degrees from RMIT University, Melbourne, VIC, Australia, in 2011 and 2013, respectively, and the Ph.D. degree from the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China, in 2020.

Since 2021, she has been with the School of Computer Science and Engineering, South China University, China, and also with the Jiangmen City Road Traffic Accident Social Relief Fund Management Center, Jiangmen, China. Her current research interests includes information diffusion, access control, and network security.