Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 1775630, 14 pages
https://doi.org/10.1155/2022/1775630

Research Article

A LASSO-Based Prediction Model for Child Influenza Epidemics: A Case Study of Shanghai, China

Jin Zhu,[1,2] **Yu Xu**,[1] **Guangjun Yu**,[3] **Jie Gao**,[4] **Yuan Liu**,[5] **Dayu Cheng**,[6] **Ci Song**,[2] **Jie Chen**,[2] **and Tao Pei**[2,7,8]

[1]*School of Geography Science and Geomatics Engineering, Suzhou University of Science and Technology, Suzhou, China*
[2]*State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, CAS, Beijing, China*
[3]*Engineering Research Center for Big Data in Pediatric Precision Medicine, Shanghai Children's Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China*
[4]*Department of Infection Control, Shanghai Children's Hospital, Shanghai Jiaotong University, Shanghai, China*
[5]*Shanghai Things-Link Intelligent Technology Co., Ltd, Shanghai, China*
[6]*School of Mining and Geomatics, Hebei University of Engineering, Handan, China*
[7]*University of Chinese Academy of Sciences, Beijing, China*
[8]*Jiangsu Center for Collaborative Innovation in Geographical Information Resource Development and Application, Nanjing, China*

Correspondence should be addressed to Jin Zhu; zhujin@usts.edu.cn and Tao Pei; peit@lreis.ac.cn

Received 21 July 2022; Revised 24 November 2022; Accepted 5 December 2022; Published 12 December 2022

Academic Editor: Zahir Shah

Child influenza is an acute infectious disease that places substantial burden on children and their families. Real-time accurate prediction of child influenza epidemics can aid scientific and timely decision-making that may reduce the harm done to children infected with influenza. Several models have been proposed to predict influenza epidemics. However, most existing studies focus on adult influenza prediction. This study demonstrates the feasibility of using the LASSO (least absolute shrinkage and selection operator) model to predict influenza-like illness (ILI) levels in children between 2017 and 2020 in Shanghai, China. The performance of the LASSO model was compared with that of other statistical influenza-prediction techniques, including autoregressive integrated moving average (ARIMA), random forest (RF), ordinary least squares (OLS), and long short-term memory (LSTM). The LASSO model was observed to exhibit superior performance compared to the other candidate models. Owing to the variable shrinkage and low-variance properties of LASSO, it eliminated unimportant features and avoided overfitting. The experimental results suggest that the LASSO model can provide useful guidance for short-term child influenza prevention and control for schools, hospitals, and governments.

1. Background

Influenza is an infectious respiratory disease which is the leading cause of respiratory illness among children. It can cause serious illness and even death in children. The annual influenza morbidity in children often exceeds 30%, which is higher than that for other age groups [1–4]. This is because children lack prior immunity to this disease. The average hospitalization rate of children attributable to influenza is the highest among all diseases. Further, it is estimated that 9,000–106,000 children younger than 5 years die due to respiratory diseases caused by influenza in 92 countries around the world [5]. During influenza epidemics, the influenza incidence rate in preschool and school-aged children may exceed 40% and 30%, respectively [1, 6, 7]. This results in the highest rates of outpatient medical visits and hospitalizations of young children for acute respiratory diseases [8]. In addition, children are the primary disseminators of

influenza in communities [6]. Influenza also leads to other consequences, including complications, antibiotic treatment, absence from daycare or school, and parental work loss [4]. Consequently, accurate real-time prediction of child influenza is significant and can aid in the adoption of effective measures to reduce harm caused by child influenza.

Over the last few years, several methods have been utilized to predict influenza levels. These methods can be broadly classified into two categories—statistical and mechanistic models. Statistical approaches are based on statistics or machine learning principles, e.g., the autoregressive integrated moving average (ARIMA) model [9, 10], seasonal autoregressive integrated moving average (SARIMA) model [11], which is a variant of ARIMA, elastic net model [12], least absolute shrinkage and selection operator (LASSO) model [13, 14], random forest (RF) model [15], support vector machine (SVM) model [12], long short-term memory (LSTM) model [16], and transformer model [17]. For instance, the ARIMA model has been used to predict the expected morbidity of influenza cases in Ningbo, China [10]. Some studies have reported accurate prediction of influenza-like illness (ILI) in the Netherlands and the USA using the LASSO model, providing early warnings in advance of influenza epidemics [13, 14]. According to research on the forecast of ILI incidence rates at both the national (France) and regional levels (the Brittany region in France), a regression model based on SVM (with a linear kernel) outperformed other models, including the RF and elastic net + residuals fitted by ARIMA (ElasticNet + ARIMA) models. Mechanistic models include compartmental and agent-based models (ABMs). Compartmental models model transitions among subpopulations with different disease states and can represent disease transmission dynamics in a population. Examples of compartmental models include the susceptible-exposed-infectious-recovered (SEIR) model [18] and the susceptible-infectious-recovered-susceptible (SIRS) model [19], both of which are based on differential equations. Compartmental models exhibit some variants and can model the temporal dynamics of other diseases, such as dengue [20, 21] and breast cancer [22], using fractional derivatives. ABMs [23] describe the transmission behavior of each individual in a population. Additionally, by combining multiple models and leveraging the advantages of each, ensemble approaches [24, 25] usually yield better prediction results.

To further improve prediction accuracy, numerous methods often combine various data sources, including influenza surveillance data [14], web search data [26, 27], temperature data [16], air pollutant data [28, 29], Wikipedia access data [30], Twitter posts [31], and electronic medical records [32]. For instance, the ARGO method combines Google search term data and the LASSO model to capture people's dynamic search behavior over time—it has been noted to exhibit excellent influenza prediction performance. Climate [16] and air pollutant data [29] have also been demonstrated to be correlated to the influenza incidence rate to some degree, and the performance of forecast models may be enhanced by accounting for them. Chretien et al. [33], Nsoesie et al. [34], and Reich et al. [35] have provided detailed reviews of influenza prediction approaches in recent years.

Owing to their lack of prior contact with and immunity to influenza viruses, children exhibit the highest influenza infection rate among all age groups. Influenza vaccination is currently the most effective approach to slow down the transmission of the disease, and immunization of children can significantly reduce influenza incidence. In China, influenza vaccination is not routinely recommended, and immunization of children is unusual. Compared to adults, children exhibit different characteristics in terms of influenza immunity and infection.

Although a variety of methods have been proposed for influenza prediction, most existing studies [10–17, 19, 23–28] have focused on influenza prediction in adult or mixed populations rather than in child populations. Among the few studies focusing on the latter, He et al. [9] and Rao et al. [36] employed the ARIMA model to predict the rate of influenza virus infection in children in Wuhan and Suzhou, China. However, the authors did not compare the performance of ARIMA with that of other prediction models. Thus, the current state of research does not reveal the extent to which mainstream influenza prediction methods are applicable to pediatric influenza or their influenza prediction performance on children. To answer these questions and identify an accurate forecasting model for child influenza in China, we performed a case study, thoroughly comparing the child influenza prediction performances of five statistical approaches—ARIMA, OLS (ordinary least squares), RF, LSTM, and LASSO—in Shanghai City.

To this end, we obtained the number of weekly outpatient visits from children with ILI between 2015 and 2020 from Shanghai Children's Hospital. The covariates were taken to be temperature and air pollutant data. The response was taken to be the 1-week-ahead ILI level. The experimental results indicate that LASSO outperformed the other four baseline models. The RMSE (root mean squared error) of LASSO was higher than those of ARIMA, RF, OLS, and LSTM by 9.47%, 22.96%, 26.67%, and 33.70%, respectively. Additionally, owing to the variable shrinkage property of LASSO, the coefficients of the temperature and air pollutant features shrank to zero and had no impact on the predicted child ILI levels.

Thus, estimation using LASSO can be expected to provide adequate guidance on child influenza prevention and control for schools, hospitals, and Centers for Disease Control (CDCs). Hospitals can use the accurate estimates for scientific and timely decision-making regarding public health resource allocation. Schools can also use the reliable estimates to remind students and their households to take precautions against influenza.

2. Methods

The methods used to support the findings of the study are explained in detail in the following sections.

2.1. Data Sources. Three data sources were used to predict child ILI activity: historical data on children's outpatient visits for influenza, temperature, and air pollutant data. The data of historical outpatient visits were used as influenza surveillance

data to represent juvenile ILI activity levels. Previous studies have reported a strong correlation between low temperatures (0–5°C) and high levels of influenza virus infection [37, 38]. Furthermore, much evidence has been reported indicating that air pollution is a risk factor for respiratory diseases, such as influenza, and that air pollutant data ($PM_{2.5}$, PM_{10}, SO_2, CO, NO_2, and O_3) affect the incidence of influenza significantly [39, 40]. Influenza prediction performance can be improved by accounting for these environmental factors [41].

2.1.1. Children's Outpatient Visits for Influenza. We obtained data on children's outpatient visits for influenza from Shanghai Children's Hospital. Shanghai is one of the largest cities in China. It has an area of 6340.5 square kilometers and a population that exceeded 24 million by the end of 2020. The gross domestic product (GDP) per capita was 24,443 US dollars in 2020. The population of children aged 0–14 years is approximately 2.44 million (based on national census data). Shanghai Children's Hospital is located in the central district of Shanghai. It was the first specialized pediatric hospital in Shanghai. It serves almost all children under 14 years of age in the city. In this study, child ILI was defined as the number of outpatient visits from children seeking medical attention for ILI symptoms. Data on weekly outpatient visits were collected from Shanghai Children's Hospital for the period between January 1, 2015 and May 31, 2020.

2.1.2. Temperature Data. Maximum and minimum temperatures for each day in the study period were obtained from a historical weather website [42]. As they exhibit a high correlation, we used them to calculate the average temperature of each day.

2.1.3. Air Pollutant Data. Air pollutant data were also collected from a historical weather website [43]. The data comprised the densities of particular matter $<2.5\,\mu$m ($PM_{2.5}$), particulate matter $<10\,\mu$m (PM_{10}), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), and ozone (O_3) in the air.

Daily temperature and air pollutant data were collected for the period between January 2015 and May 2020. These data contained occasional missing values, which were filled in using linear interpolation. Subsequently, the average weekly temperature and air pollutant data were aggregated based on the daily data. Note that web search data were not used in this study, primarily because children are only a subset of the whole population, and many search terms related to influenza are not directly related to children.

2.2. Predictive Models. In this study, LASSO was utilized as a regularized estimation model to forecast ILI. LASSO is an extension of linear regression and exhibits variable shrinkage and selection. Owing to its sparsity property, the coefficients of some of its features are zero. This property enables LASSO to avoid overfitting and enhances its predictive accuracy.

For comparison, we used four methods: OLS, ARIMA, RF, and LSTM. OLS [44] is perhaps the simplest linear regression method, and LASSO is obtained by adding the L_1 penalty to OLS. Unlike OLS, LASSO exhibits the sparsity property. ARIMA [45] is a classical time-series prediction method that considers temporal autocorrelation and past forecasting errors. RF is an ensemble learning method consisting of multiple small decision trees, RF regression averages their predictive results to obtain accurate predictions [46]. LSTM is a special type of recurrent neural network (RNN) that can learn long-term dependencies in time-series data [16]. It has been commonly utilized in the current deep learning era. The main characteristics of each approach are described below.

2.2.1. LASSO and OLS. Given N observations, (x_i, y_i) and $i = 1, \ldots, N$, where $x_i = (x_{i1}, \ldots, x_{ip})^T$ denotes the p features and y_i denotes the outcome variable, the linear regression model assumes that

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \tag{1}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ denotes the unknown regression parameter and ϵ_i denotes the error term. OLS estimates the β parameter by minimizing the following least-squares objective function:

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}. \tag{2}$$

Equation (2) provides OLS estimates. In general, all OLS estimates are nonzero, i.e., OLS estimates have low bias but high variance. Thus, small changes in inputs affect the estimates significantly.

2.2.2. LASSO Estimates the β Parameter by Solving

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} subject\ to\ \|\beta\|_1 \le t, \tag{3}$$

where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ denotes the ℓ_1 norm of β and $t \ge 0$ denotes a user-specified parameter that controls the degree of variable shrinkage.

The LASSO problem can also be written in the following Lagrangian form:

$$\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda\|\beta\|_1 \right\}, \tag{4}$$

for some $\lambda \ge 0$. By Lagrangian duality, there is a one-to-one relationship between the solutions of equation (3) and that of equation (4).

The most important utility of the ℓ_1 norm is that when t is sufficiently small, LASSO usually shrinks some of the regression coefficients to zero. This enhances the interpretability of LASSO, uncorrelated features are eliminated

from the model, and the model becomes easier to understand. Interpretability is becoming increasingly significant in the era of big data. Currently, we can readily obtain hundreds, thousands, or even millions of features, but we cannot identify in advance the features that are related to the outcome variable. LASSO removes all unrelated features and preserves only important features to avoid this problem.

By excluding unrelated features from the complete OLS model, the prediction variance is decreased, but the bias is increased as a tradeoff [47]. If the reduction in prediction variance exceeds the increase in bias, the accuracy of the model is enhanced. Therefore, LASSO often outperforms OLS in terms of accuracy. By choosing an appropriate λ for LASSO, uncorrelated features are eliminated and accuracy is improved. In this study, the value of λ was optimized using cross-validation (CV) based on the root mean square error (RMSE) metric. In particular, we trained the LASSO model corresponding to different values of the λ parameter, ranging from 0 to 100000 at intervals of 10. The value corresponding to the lowest RMSE score was selected as the optimal LASSO model. It was implemented using Python scikit-learn package version 1.0.9.

2.3. ARIMA. ARIMA is one of the most widely used methods for time series prediction. ARIMA incorporates differencing of lagged values of the prediction variable and lagged errors [45, 46]. Differencing is used to convert a nonstationary time series into a stationary time series. The lagged values of the prediction variable comprise the autoregressive (AR) part of ARIMA and the lagged errors form the moving average (MA) part of ARIMA. ARIMA involves three parameters (p, d, q)–p denotes the order of the autoregressive part, d denotes the degree of first differencing, and q denotes the order of the moving average part. The traditional method of identifying and fitting an ARIMA model is complex, time-consuming, and subjective. Therefore, we used the grid search strategy to identify the optimal ARIMA model based on minimal Akaike's Information Criterion (AIC) via CV. During grid search configuration, the p and q values were taken between 0 and 5, and the d value was selected between 0 and 2. The (p, d, q) parameters with minimal AIC values were chosen to optimize the ARIMA model. The ARIMA model was implemented using Python statsmodels package version 0.13.0.

2.4. Random Forest. RF models nonlinear relationships in data and has been used to predict various types of infectious diseases [46, 48]. It is an ensemble learning method that combines a large number of decorrelated classification and regression trees (CARTs). To this end, it constructs a single tree using the bagging (bootstrap aggregation) technique. Bagging involves sampling with replacement to create decorrelated decision trees during training. RF can be used for classification and regression tasks. As our prediction target is the occurrence of influenza, we used RF regression for prediction, with the average of all individual decision trees yielded as the output. The grid search strategy was applied to optimize hyperparameters by minimizing RMSE using CV. The RF model was applied using Python scikit-learn package version 1.0.9. The n_estimators hyperparameter represents the number of trees in the forest, and its value was varied between 50 and 500 in intervals of 10. The other hyperparameters were set to their default values.

2.5. LSTM. Neural networks (NN) are widely used to model nonlinear relationships in data. As an extension to NN, RNN is designed to deal with sequential data. It combines current inputs and past information to obtain the output. However, RNNs suffer from the gradient vanishing problem. To resolve this, LSTM [16, 49] incorporates an input gate, a forget gate, and an output gate within the RNN cell. These cells transmit past information corresponding to multiple time steps to subsequent time steps. Primarily, LSTM saves old information for later use, thereby avoiding the gradient vanishing problem during dataset training. [40] The LSTM model was applied using the Python Keras package version 2.3.1, which was constructed in TensorFlow package version 2.1.0. The LSTM model comprised two layers. The hiddenNum hyperparameter of each layer was determined using a grid search strategy based on cross-validation. Default values were used for the other hyperparameters. The value of the hiddenNum hyperparameter varied from 8 to 128 in intervals of 8.

2.6. Feature Selection. Although LASSO exhibits variable shrinkage and selection, feature selection was performed to ensure fair comparison between LASSO and other prediction models, and the selected features were transmitted into all models for ILI level prediction. The features were selected using mutual information [50] owing to its suitability for measuring nonlinear relationships between random variables. The mutual information coefficient of each feature was calculated and scaled to the range 0–1. During the calculation of mutual information, the time window of the lag was selected to be 52 weeks (one year) based on previous research [16] and the number of features. Features with mutual information coefficients less than 0.4 were removed. The remaining features were directly transmitted into LASSO, OLS, RF, and LSTM, as these are multivariate models. In contrast, ARIMA is a univariate model that uses only lagged ILI occurrences for prediction.

2.7. Model Assessment. To assess the effectiveness of the aforementioned models, we used a naive method for comparison. The naive method uses the value of the previous week as the predicted value of the current week. It was adopted as the baseline model, and models outperforming the naive method were considered effective.

The dataset was divided into training and testing sets. Data corresponding to 2015–2016 were used as training data, and data corresponding to 2017–2020 were used as test data. We used the rolling-origin recalibration method for evaluation [51]. The predictions for each week in the test set were obtained by moving the data from the test set to the

training set sequentially. Once the data were updated each week, all prediction models were dynamically retrained to predict the ILI level of the following week. Retrospective estimates of child influenza activity were evaluated corresponding to 2017–2020 using an out-of-sample approach.

Five accuracy evaluation metrics were adopted to compare the performances of the five models—RMSE, mean absolute error (MAE), mean absolute percentage error (MAPE), correlation coefficient, and correlation coefficient of increment.

The following notations are used: y_i denotes the true value of ILI at time t_i, x_i denotes the predicted value of ILI at time t_i, $\overline{y}$ denotes the average value of the time series $\{y_i\}$, and $\overline{x}$ denotes the average value of the time series $\{x_i\}$.

RMSE is a measure of the average difference between the true and predicted values. The RMSE of the two-time series $\{y_i\}$ and $\{x_i\}$ $(i = 1, \cdots, n)$ is defined as follows:

$$\text{RMSE}\left(y_i, x_i\right) = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - x_i\right)^2}{n}}. \tag{5}$$

MAE measures the average absolute difference between the true and predicted values. The MAE of $\{y_i\}$ and $\{x_i\}$ is defined as follows:

$$\text{MAE}\left(y_i, x_i\right) = \frac{\sum_{i=1}^{n}\left|y_i - x_i\right|}{n}. \tag{6}$$

RMSE and MAE are widely used in prediction tasks, and they both measure the average extent of prediction errors. MAE averages the prediction errors directly and can be considered a linear combination of errors, with equal weights corresponding to all errors. However, RMSE squares the prediction errors before computing the average. Therefore, RMSE produces large weights for large errors. As a result, it is particularly suitable in cases where large errors are especially undesirable.

MAPE is a measure of the average percentage difference between the true and predicted values. The MAPE of $\{y_i\}$ and $\{x_i\}$ is defined as follows:

$$\text{MAPE}\left(y_i, x_i\right) = \sum_{i=1}^{n}\frac{\left|y_i - x_i\right|}{y_i}. \tag{7}$$

The correlation coefficient calculates the Pearson correlation coefficient and measures the linear relationship between true and predicted values. The correlation between $\{y_i\}$ and $\{x_i\}$ is defined as follows:

$$\text{Corr}\left(y_i, x_i\right) = \frac{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)\left(x_i - \overline{x}\right)}{\sqrt{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}\sqrt{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}}. \tag{8}$$

Finally, the correlation of increment between $\{y_i\}$ and $\{x_i\}$ is defined as follows:

$$\text{Corr. of increment}\left(y_i, x_i\right) = \text{Corr}\left(y_i - y_{i-1}, x_i - x_{i-1}\right). \tag{9}$$

3. Results

The results obtained in the article is described in the following sections.

3.1. Influenza Prediction Accuracy of LASSO. Retrospective estimates of ILI levels were obtained using LASSO, ARIMA, RF, OLS, LSTM, and the naive model for the period between week 1 of 2017 and week 20 of 2020. We compared the estimates with the ground truth, i.e., the ILI level provided by Shanghai Children's Hospital. The various accuracies of the models are presented in Table 1. The column in Table 1 lists different time periods, including the entire period, the off season, and the four regular flu seasons in 2016–2020. The regular annual flu season lasts from week 40 of each year to week 20 of the following year. The test was conducted for the period 2017–2020. Thus, the start of the 2016-2017 season was taken to be week 1 of 2017. The prediction curves of all models and the ground truth are depicted in Figure 1.

Over the entire period, LASSO outperformed the other models in terms of all metrics except the relative MAPE. Overall, in terms of relative RMSE, MAE, and MAPE, LASSO was the most accurate model in all seasons, including the off-season and the four regular flu seasons. During some regular seasons, the performance of the ARIMA model was slightly better than that of LASSO. For instance, in the 2016–2017 flu season, LASSO (relative RMSE = 1.038 and relative MAE = 1.070) exhibited the second-best performance in terms of relative RMSE and MAE, second to the ARIMA model (relative RMSE = 0.979 and relative MAE = 1.051). The correlation coefficient of LASSO was the highest in all seasons except the 2016–2017 flu season, when it (corr = 82.6%) was slightly inferior to that of ARIMA (corr = 86.0%). Moreover, LASSO exhibited the highest correlation of increments during all periods.

The RMSE values of LASSO were observed to be higher than those of ARIMA, RF, OLS, and LSTM by 9.47%, 22.96%, 26.67%, and 33.70%, respectively. The relative RMSE of LASSO (0.899) and ARIMA (0.993) over the entire period were both less than one, while those of RF (1.167), OLS (1.226), and LSTM (1.356) were not. This suggests that, compared to the naive method, LASSO and ARIMA were effective, but RF, OLS, and LSTM were not.

Although, in terms of the relative RMSE, ARIMA was the second-best model over the whole period, the discrepancy between LASSO and ARIMA during regular flu seasons was slight. However, LASSO (relative RMSE = 0.872) outperformed ARIMA (relative RMSE = 1.201) by a notable margin during the off-season.

Both RF and LSTM are nonlinear models and exhibited poorer performances than the naive method. Figure 1 indicates that the prediction curve of RF does not contain as many undulations as that of LSTM. OLS is similar to LASSO in principle, but its performance was observed to fall between those of RF and LSTM. This suggests that OLS exhibits higher variance than LASSO, which can also be observed in Figure 1. Note that OLS exhibited remarkably large estimation errors on March 23, 2017, December 28, 2017, and February 20, 2020.

The vertical orange dashed line in Figure 1 represents January 23, 2020 during the 2019–2020 flu season. From this day, COVID-19 began to spread rapidly, and the Chinese government closed Wuhan down to prevent the epidemic

Table 1: Influenza estimation accuracies of different models. The best performance corresponding to each accuracy metric in each time period is highlighted in boldface. The reported RMSE, MAE, and MAPE scores are relative to the absolute error of the naive method, i.e., the ratio of the error of a given method to that of the naive method. The numbers in parentheses denote the absolute errors of the naive method.

Metric	Whole period (week 1, 2017–week 20, 2020)	Off-season	Regular flu seasons (week 40 to week 20 of the following year)			
			2016–2017[1]	2017–2018	2018–2019	2019–2020
RMSE						
LASSO	**0.899**	**0.872**	1.038	**0.972**	**0.910**	**0.838**
ARIMA	0.993	1.201	**0.979**	1.066	0.992	0.863
RF	1.167	0.975	1.576	1.281	1.007	1.206
OLS	1.226	1.290	2.604	1.270	0.934	0.875
LSTM	1.356	1.280	1.746	1.642	1.043	1.384
Naive	1 (1822.322)	1 (1232.904)	1 (1496.251)	1 (1561.335)	1 (2217.054)	1 (2546.548)
MAE						
LASSO	**1.037**	**0.928**	1.070	**1.081**	**1.051**	1.088
ARIMA	1.135	1.295	**1.051**	1.082	1.217	**0.976**
RF	1.268	1.056	1.634	1.349	1.189	1.329
OLS	1.461	1.376	2.882	1.440	1.148	1.202
LSTM	1.528	1.481	1.639	1.760	1.362	1.495
Naive	1 (1158.927)	1 (872.169)	1 (1124.75)	1 (1154.000)	1 (1332.273)	1 (1523.909)
MAPE						
LASSO	1.412	**1.001**	**0.999**	1.136	**1.074**	2.076
ARIMA	**1.216**	1.269	1.070	**1.066**	1.218	**1.288**
RF	1.450	1.137	1.443	1.277	1.168	1.860
OLS	1.839	1.434	2.619	1.431	1.149	2.357
LSTM	1.555	1.522	1.436	1.981	1.281	1.566
Naive	1 (7.963)	1 (5.489)	1 (7.582)	1 (6.248)	1 (7.135)	1 (15.159)
Correlation						
LASSO	**0.965**	**0.936**	0.826	**0.956**	**0.844**	**0.985**
ARIMA	0.957	0.893	**0.860**	0.941	0.829	0.981
RF	0.941	0.920	0.758	0.932	0.800	0.963
OLS	0.933	0.876	0.145	0.916	0.838	0.984
LSTM	0.916	0.862	0.402	0.892	0.803	0.950
Naive	0.956	0.916	0.847	0.944	0.818	0.975
Correlation of increment						
LASSO	**0.475**	0.858	**0.322**	0.241	0.282	**0.871**
ARIMA	0.403	0.775	0.302	**0.292**	0.284	0.781
RF	0.368	**0.869**	0.268	0.138	0.117	0.700
OLS	0.358	0.707	−0.198	0.269	0.302	0.788
LSTM	0.357	0.816	−0.202	0.086	**0.478**	0.582
Naive	0.459	0.864	0.353	0.312	0.273	0.790

[1] As the start of the estimation time duration is the first week of 2017, the regular flu season of 2016–2017 was considered to be from week 1, 2017, to week 20, 2017.

from spreading outside Wuhan. Other cities, including Shanghai, strictly managed the travel of residents—students attended classes at home, and office workers worked from home or went on vacation. Adoption of home isolation measures reduced the exposure of children to influenza, and the number of children infected with influenza decreased sharply. The ILL level was centered around 2000 between the middle of February and May 2020, as depicted in Figure 1. As MAPE calculates the percentage of errors, equation (7) indicates that if the ground truth is a small value, MAPE is very likely to be a relatively large value, even if the prediction result is not very large. For this reason, the relative MAPE of LASSO (2.076) was significantly larger than that of ARIMA (1.288) during this period. Nevertheless, the patterns of previous flu seasons were different from those of the 2019–2020 flu season after the outbreak of COVID-19, and all prediction models performed worse than the naive method in terms of MAPE.

Figure 2 depicts a scatter plot of the ground truth and all prediction results. From LASSO to LSTM, as R^2 decreased, the errors of the prediction models gradually increased, and a tendency towards wider confidence intervals for interval-based prediction was observed.

3.2. Dynamic Regression Coefficients of LASSO. Figure 3 depicts the dynamic regression coefficients of LASSO obtained by applying the rolling-origin recalibration evaluation method to data between week 1 of 2017 and week 20 of 2020. The coefficients of features, num_lag_1 and num_lag_51, were positive and represented the ILI number in the previous week and previous 51 weeks (usually one year contains 52 weeks), respectively. Feature num_lag_1 has been highlighted in dark red to indicate that the ILI number of the previous week exerted the greatest influence on that of the

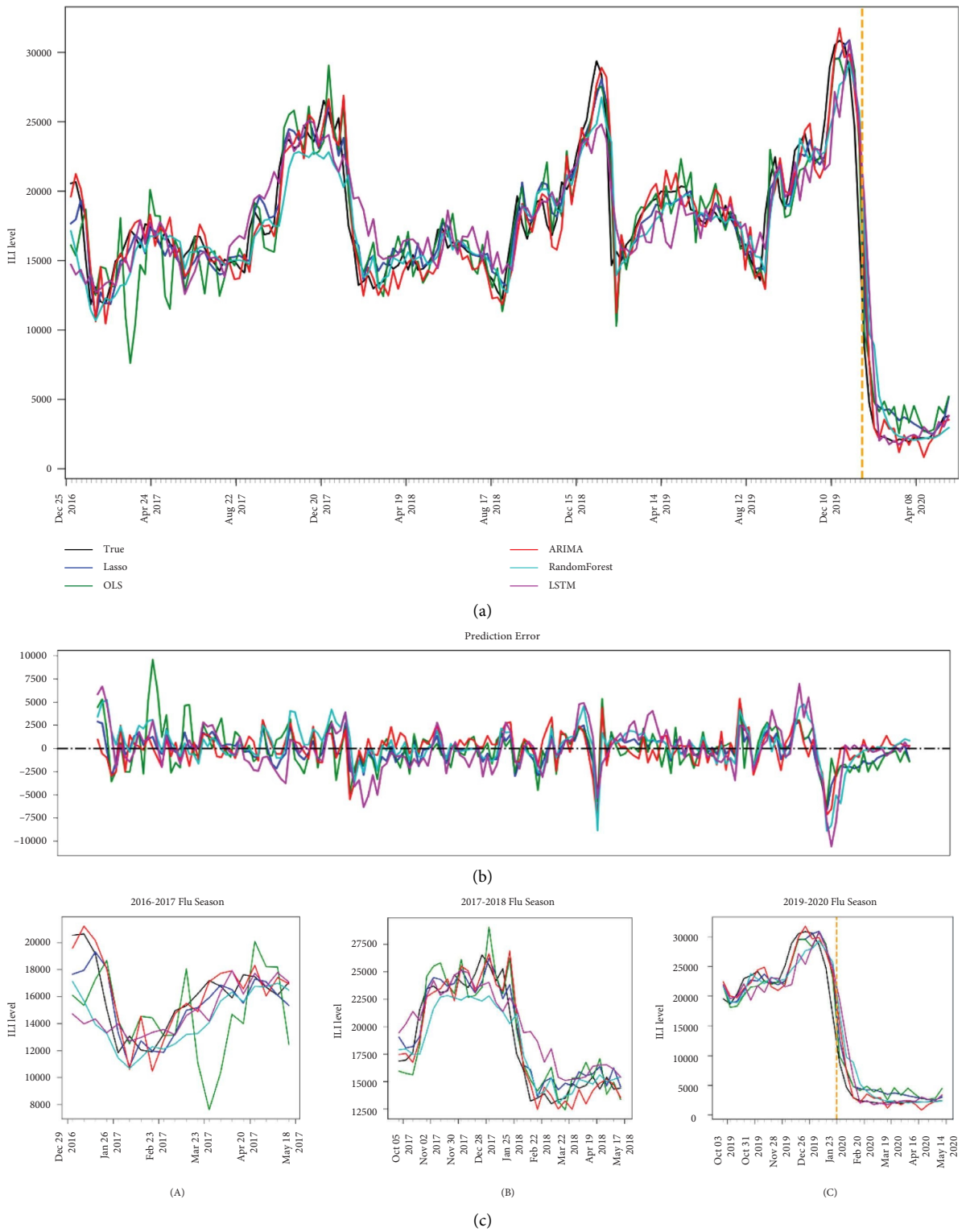

Figure 1: Prediction results. (a) The ILI level predicted by LASSO (blue) and the true ILI level (black), and those predicted by OLS (green), ARIMA (red), RF (cyan), and LSTM (purple). (b) Prediction error is defined to be the difference of the predicted value and the true ILI level. (c) Magnified prediction results for three flu seasons. (A) 2016–2017 flu season, (B) 2017–2018 flu season, and (C) 2019–2020 flu season. The vertical orange dashed line corresponds to Jan 23, 2020, when the Chinese government closed Wuhan down to prevent the COVID-19 epidemic from spreading to other regions.

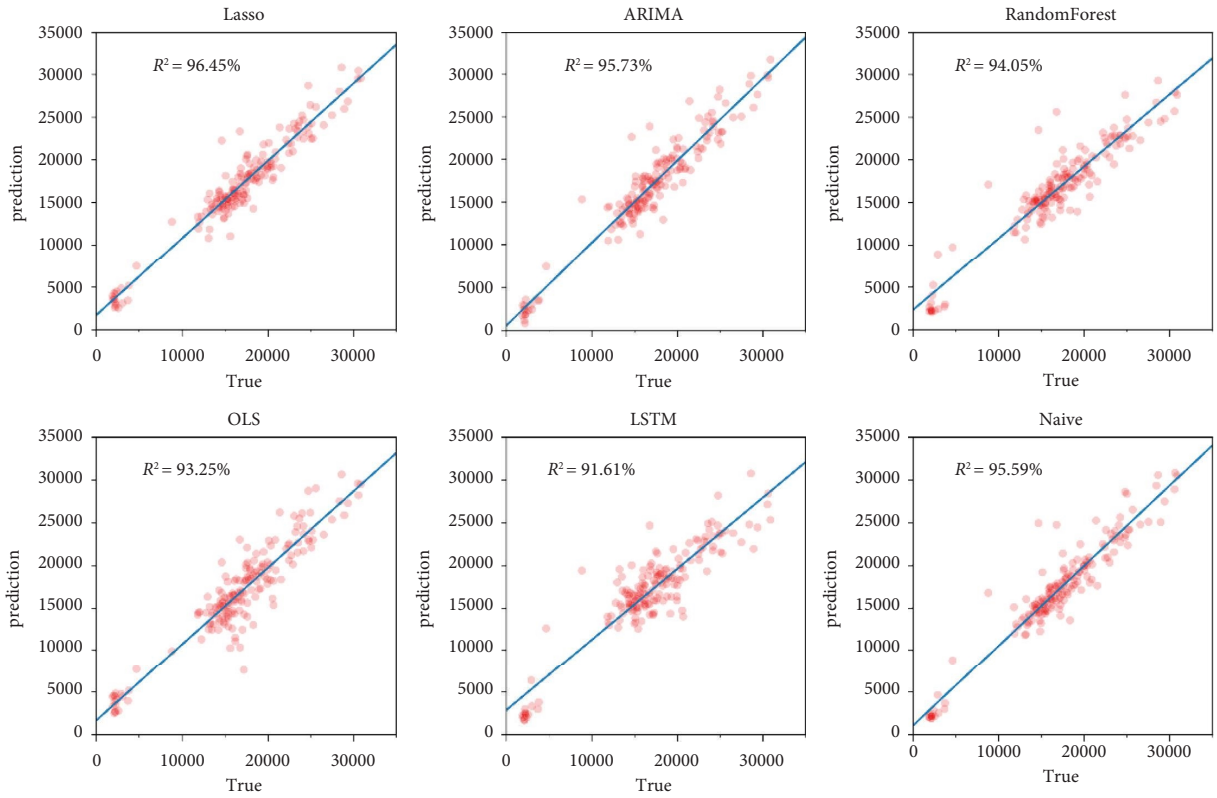

Figure 2: Scatter plot of the ground truth and prediction results obtained using the six prediction methods. The blue line represents the linear regression line. R^2 denotes the Pearson correlation coefficient.

current week, and feature num_lag_51 exhibited a much weaker effect. The coefficients of features, num_lag_2 and num_lag_52, were negative, and their absolute values were generally lower than those of features, num_lag_1 and num_lag_51, respectively. The former pair seemed to slightly offset the latter pair. The coefficients of features concerning the temperature, PM2.5, PM10, SO_2, NO_2, and O_3 were constantly zero. Compared to the autoregressive features of ILI number, they exhibited almost no impact on the ILI number of the current week; thus, their coefficients were reduced to zero by LASSO.

The green vertical dashed line represents the outbreak of COVID-19. Subsequently, the coefficients of num_lag_2 and num_lag_52 became remarkably smaller. These coefficients are directly proportional to the predicted ILI number. Owing to home isolation measures, the number of children infected with influenza decreased sharply.

3.3. Density of Regression Coefficients for LASSO and OLS. Figure 4 depicts a violin plot of the two groups of regression coefficients (excluding the intercept term) of LASSO and OLS. Without loss of generality, the two groups were derived based on the training process for Week 1 of 2019. The *Y*-axis represents the values of the regression coefficients. The dots in the interior of the violin represent the coefficients, and the outline of the violin describes the density of the coefficients. The dot at the top of the LASSO violin plot corresponds to the num_lag_1 feature. Owing to the variable shrinkage property of LASSO, most of the LASSO coefficients were exactly zero, while the others were near zero. In contrast, the OLS coefficients were rarely zero and were distributed roughly over a wide range, ranging from −200 to 200. This increased the variance of the OLS model.

3.4. The Distribution of Coefficients of the Most Important Features. Figure 5 displays box plots of the coefficients of the ten features with the largest mean absolute values of the coefficients. The features were ordered from left to right in terms of the mean absolute values of the coefficients. As all of these features are autoregressive, the coefficient values of these features represent their importance. The four most important features were observed to be num_lag_1, num_lag51, num_lag_2, and num_lag_52. The negative coefficients of the features, num_lag_2 and num_lag_52, slightly offset the positive coefficients of num_lag_1 and num_lag_51. This relationship is illustrated in Figure 3. The mean absolute values of the coefficients of the other six features did not exceed 0.02; thus, these features were deemed trivial compared to the four most important features.

3.5. The Variable Shrinkage Property of LASSO. In equation (3), the parameter t represents the possible size range of $\sum_{j=1}^{p} |\beta_j|$. When t is large, the constraint on the ℓ_1 norm of β is relaxed and the estimated coefficients can be large. In particular, when t is greater than $\sum_{j=1}^{p} |\widehat{\beta}_j^o|$ ($\widehat{\beta}_j^o$ denote the OLS coefficients), the estimated solution is the OLS solution.

Figure 3: Heat map of dynamic regression coefficients of LASSO between week 1 of 2017 and week 20 of 2020. These regression coefficients were computed using the rolling-origin recalibration evaluation method. The X-axis represents the date and the Y-axis corresponds to the regression coefficient of features. Positive coefficients are indicated in red, negative coefficients are indicated in blue, and zero is indicated in white. The feature, num_lag_x, denoted the ILI number with lag (x) and temp_lag_x denoted the average temperature with lag (x). The green vertical dashed line represents January 23, 2020, when the Chinese government closed Wuhan down to prevent the COVID-19 epidemic from spreading outside Wuhan.

In contrast, if t is small, the constraint on $\sum_{j=1}^{p} |\beta_j|$ is strict, and the estimated coefficients are also small. When t is sufficiently small, some of the estimated coefficients become zero: this is the variable shrinkage property of LASSO. This property can be visualized using the definition of the shrinkage ratio, s:

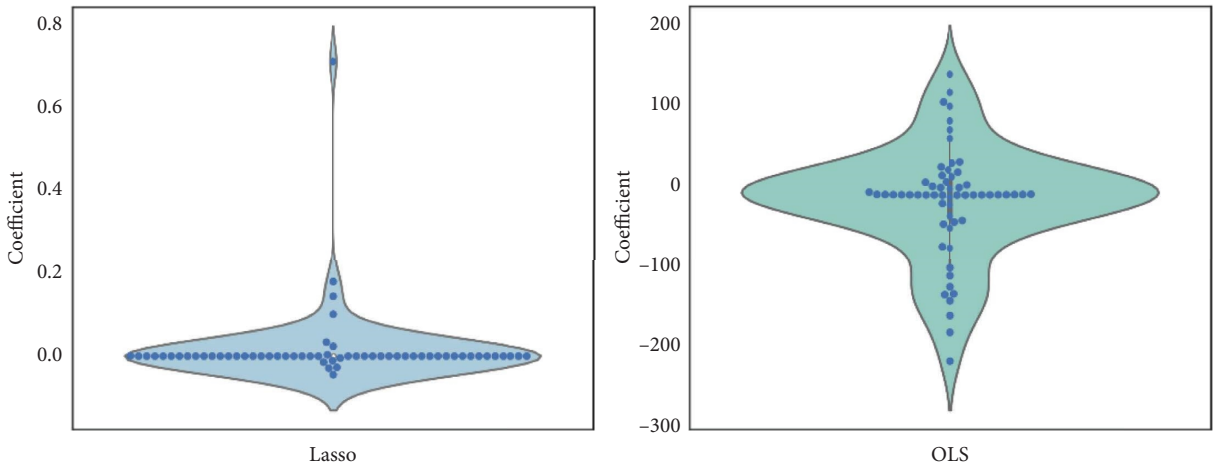

Figure 4: Violin plot of regression coefficients for LASSO and OLS.

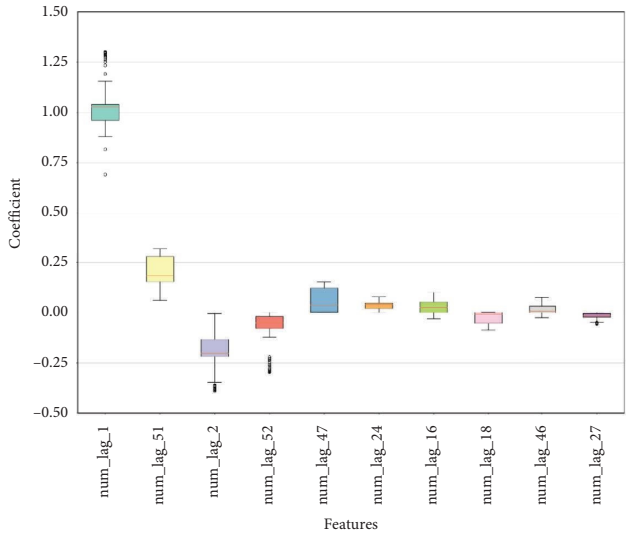

Figure 5: Box plots of the coefficients of ten features with the highest mean absolute values of coefficients. The features were ordered from left to right.

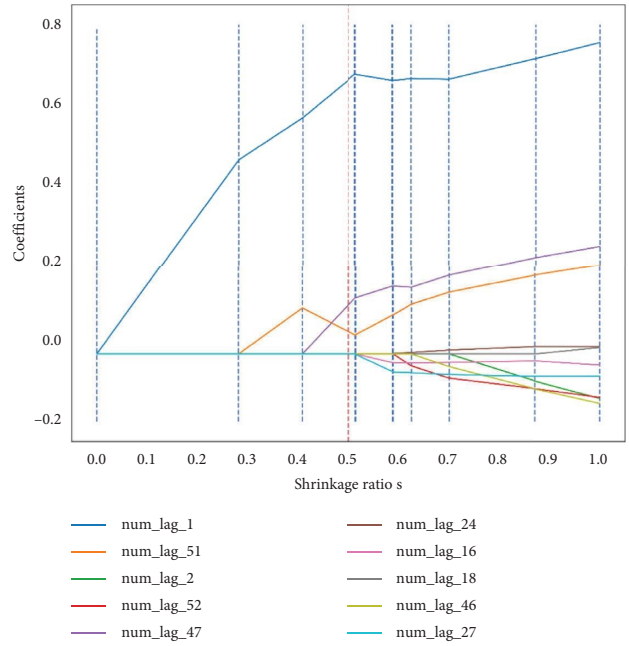

Figure 6: The solution path of LASSO for week 1 of 2017. The X-axis represents the shrinkage ratio and the Y-axis represents the estimated coefficients. The ten most important features are displayed in the figure. The red dashed line represents $s = 0.5$.

$$s = \frac{t}{\sum_{j=1}^{p} \left| \widehat{\beta}_j^o \right|}. \tag{10}$$

When s lies between 0 and 1, the corresponding coefficient of the solution can be obtained. The collection of coefficients for all s $(0 < s \leq 1)$ was generated for LASSO. The generated coefficients of LASSO for Week 1 of 2017 are depicted in Figure 6. As the overall number of features was large, only the ten most important features are depicted. When the shrinkage ratio s was gradually increased, the most important feature, num_lag_1, first entered the LASSO model. Subsequently, the features, num_lag_51, and num_lag_47 were entered into the model. These features exerted a smaller influence on the predictive target than num_lag_1. In the figure, s is represented on the X axis, and the outcome of variable shrinkage can be observed. For instance, corresponding to $s = 0.5$, the three features, num_lag_1, num_lag_47, and num_lag_51 were used to predict the ILI number, and their coefficients were determined by the intersections of the coefficient curve and the red dashed vertical line. The coefficients of the other features were zero.

3.6. The Effect of the Parameter, λ, on LASSO. LASSO involves only one parameter λ in equation (4) or t in equation (3), and the effects of λ and t are identical. To assess the effect of parameter λ on LASSO, the RMSEs for various nonnegative values of λ are depicted in Figure 7. The value of λ was varied from 0 to 100000 in intervals of 10. When $\lambda = 0$, LASSO degenerated to OLS and exhibited high RMSE. When λ was increased, the RMSE decreased quickly. When $\lambda = 48700$, the lowest RMSE (1637.43) was observed. When λ was between

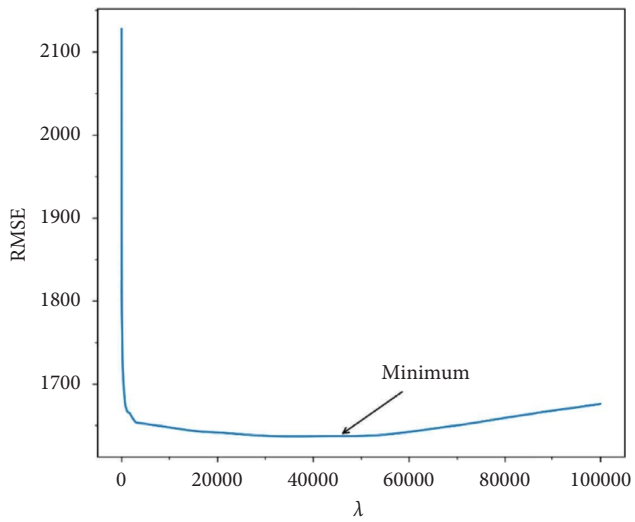

Figure 7: RMSE for various values of λ.

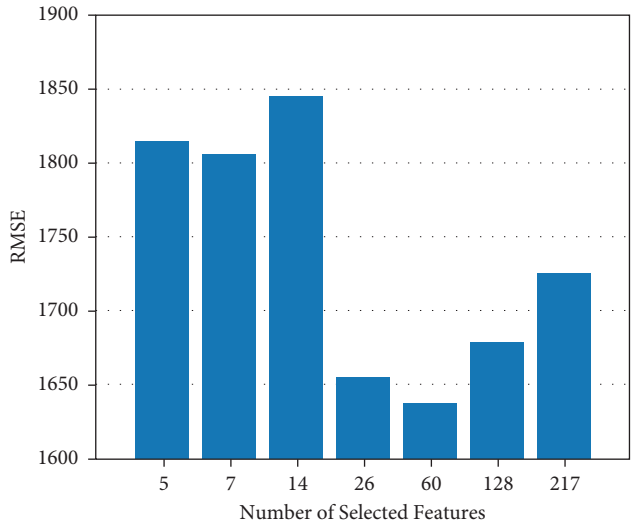

Figure 8: The RMSE values of LASSO corresponding to different numbers of features selected using the mutual information.

20000 and 50000, the RMSE did not change significantly. When λ was increased further, the RMSE gradually increased. This behavior can be explained by the fact that, when λ became too large, the amount of variable shrinkage was also large. This shrunk the coefficients of some important features down, even to zero in some cases. Consequently, the performance of LASSO was degraded. Therefore, the selection of λ is critical in LASSO. As the computational efficiency of LASSO is very high and λ is the only parameter, parameter selection for LASSO can be performed efficiently.

3.7. The Effect of Feature Selection on Prediction Accuracy. The effects of selecting different numbers of features on LASSO were assessed in this case study. Feature selection was performed using mutual information, and features with mutual information values exceeding a certain threshold were selected and input into LASSO. The thresholds were set to 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, and 0.2, and the corresponding numbers of selected features were 5, 7, 14, 26, 60, 128, and 217, respectively. Different LASSO models were trained using these groups of features. For each group of selected features, the LASSO model was trained with CV performed using λ values ranging from 1 to 100000 in intervals of 10.

The minimal RMSE for each group of selected features is depicted in Figure 8. The effects of selected features on LASSO were not negligible. The RMSE values ranged from 1637 (60 selected features) to 1844 (14 selected features). This result is extremely interesting, considering that LASSO exhibits variable shrinkage and selection. In particular, the RMSE did not increase or decrease monotonically as the number of selected features increased. The RMSE was minimal, corresponding to 60 selected features. Therefore, despite the variable shrinkage property of LASSO, thorough feature selection is still beneficial.

4. Discussion

In this study, a prediction method based on LASSO was proposed to track incidence of influenza in children in Shanghai, China. To the best of our knowledge, this is the first attempt to evaluate state-of-the-art prediction models thoroughly in the case of influenza incidence in children and identify the superior model. LASSO was used to discard unnecessary features and select the most significant features based on lagged ILI levels, temperature, and air pollutant data. Owing to its variable shrinkage property, LASSO eliminated unimportant features and outperformed other influenza tracking models, including ARIMA, RF, OLS, and LSTM, in terms of prediction accuracy. Compared to adults, children exhibit different characteristics of influenza immunity and infection. Previous studies have demonstrated that LASSO is effective in predicting influenza in adults. However, the predictive accuracy of LASSO for pediatric influenza remains unclear. This study revealed that LASSO also achieves accurate prediction of influenza trends in children.

The interpretability of LASSO is enhanced by its variable shrinkage properties. The ILI level from the previous week was observed to have a significant impact on the ILI level of the current week, and ILI levels with lags 51, 2, and 52 provided information of decreasing importance. This reflects a strong temporal autocorrelation and yearly cycle pattern in child influenza data that is corroborated by the ARGO (Autoregression with Google search data) model [14]. In ARGO, the current week's ILI level is significantly influenced by the ILI levels of the previous week and those from half a year ago and one year ago.

The proposed approach is not directly compared to ARGO as they use different datasets and features. However, both this study and ARGO utilized the relative RMSE metric with respect to the naive method. The relative RMSE of the proposed approach (0.899) was observed to be much larger than that of ARGO (0.608).The difference could be attributed to the differences in the adopted features. In addition to the lagged ILI features, the proposed approach considered temperature and air pollutant features, whereas ARGO uses Google-search-term features. In our approach, lagged temperature and air pollutant features were not selected by LASSO, and their coefficients were decreased to zero. However, the coefficients

of many Google search features are not zero, and these features play an important role in ARGO. It is likely that if we accounted for child influenza-related search data, the accuracy of our approach could be further improved.

Previous studies have reported the relationships between temperature/air quality data and ILI levels and employed these data to promote the performance of predictive models. In this study, before transmitting the features into LASSO, they were selected via feature selection using the mutual information criterion. Many of the temperature and air pollutant features were selected by feature selection, but their final coefficients were reduced to zero by LASSO as depicted in Figure 3. This demonstrates the primary advantage of LASSO; even among the selected and potentially useful features, relatively unimportant features were eliminated by LASSO while retaining significant features. In this study, the temperature and air pollutant features were observed to be relatively unimportant compared to the autoregressive features of the ILI number. Thus, our results indicate that air pollution does not have a significant impact on child influenza. However, we only used a dataset corresponding to Shanghai, and, thus, the result may be a consequence of the limited data sample. Therefore, further data collection and research are required.

Although LASSO exhibits variable shrinkage and selection, the impact of the features input into LASSO on its estimation accuracy was not negligible. This was attributed primarily to the variable shrinkage property of LASSO. When several unrelated features were input into LASSO, their coefficients were often close to zero but not equal to 0. This increased the variance and decreased the estimation accuracy of LASSO. Therefore, cautious feature selection remains an important factor influencing LASSO's accuracy.

LASSO involves a single hyperparameter, λ, which is another merit of LASSO. Furthermore, its training process is highly efficient. Therefore, hyperparameter search for LASSO is a trivial task. ARIMA, RF, and LSTM all involve more than three hyperparameters, and the training processes for these models are less efficient than those of LASSO. Therefore, hyperparameter search is tedious in these cases.

This study has several limitations. Owing to availability issues, we only used a child influenza dataset pertaining to Shanghai, China. The vaccination rates of influenza vaccines in children vary over regions and countries. Moreover, children from different countries exhibit different levels of immunity against influenza. Thus, the effectiveness of LASSO as a predictor of the child influenza incidence in other cities or on larger scales, such as states and countries, requires further research. Additionally, the child influenza dataset used in this study corresponded to a relatively short period, from January 1, 2015 to May 31, 2020. Finally, the predictive target was taken to be the 1-week-ahead ILI level. We only focused on short-term forecasts in this study, and other long-term forecasts were not considered.

5. Conclusions

In this study, the feasibility of using the LASSO model to predict child ILI activity level based on data corresponding to the period 2017–2020 in Shanghai, China, was demonstrated. The proposed model leverages data from multiple input data sources, including lagged ILI number, lagged temperatures, and lagged air pollutant data. Owing to the variable shrinkage property of LASSO, the coefficients of the unimportant features (lagged temperature and air pollutant features) are decreased to zero. On the contrary, autoregressive ILI number features are preserved as important features. The proposed LASSO model outperforms the other candidate models assessed in the study. Although there are some distinctions between child and adult influenza, this study demonstrates that LASSO is effective and accurate for child influenza prediction, making it a powerful tool for providing guidance on child influenza prevention and control for schools, hospitals, and the CDC. Although LASSO exhibits variable shrinkage, feature selection continues to have a significant impact on its performance. Thus, cautious feature selection can further improve its prediction accuracy. In future works, we intend to study the deep relationship between feature selection and LASSO, and investigate long-term forecasts, such as the 1-month-ahead ILI level. In addition, we wish to evaluate the feasibility of the LASSO model using child influenza datasets in other cities or regions.

Abbreviations:

LASSO: Least absolute shrinkage and selection operator
ARIMA: Autoregressive integrated moving average
RF: Random forest
OLS: Ordinary least squares
LSTM: Long short-term memory
ARGO: Autoregression with Google search data
ILI: Influenza-like illnesses
RMSE: Root-mean-squared error
MAE: Mean absolute error
MAPE: Mean absolute percentage error
CV: Cross-validation.

Data Availability

The datasets used during the study can be obtained from the corresponding author upon reasonable request.

Ethical Approval

This study was approved by the Ethics Committee at Shanghai Children's Hospital, China.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Jin Zhu conceptualized the study; Yu Xu developed the methodology; Dayu Cheng helped with software; Ci Song validated the study; Tao Pei carried out formal analysis; Yuan Liu and Jie Chen investigated the study; Guangjun Yu and Jie Gao collected the resources; Jin Zhu wrote and prepared the original draft; Ci Song, Jie Chen, and Tao Pei wrote,

reviewed, and edited the manuscript; Yu Xu and Dayu Cheng visualized the study; Tao Pei supervised the study; Jin Zhu and Tao Pei administrated the project; Yuan Liu and Tao Pei funded acquisition. All authors have read and agreed to the final manuscript. Jin Zhu, Yu Xu, and Guangjun Yu contributed equally to this work.

Acknowledgments

The authors gratefully acknowledge Shanghai Children's Hospital for providing children influenza outpatient visit data. This work was funded by the National Natural Science Foundation of China (nos. 42071436, 42071435, and 71874110), Grant of State Key Laboratory of Resources and Environmental Information System (no. 201816), and Research Foundation for Talent Introduction of Suzhou University of Science and Technology (no. 331511203).

References

[1] A. S. Monto, K. M. S. Sullivan, and K. M. Sullivan, "Acute respiratory illness in the community. Frequency of illness and the agents involved," *Epidemiology and Infection*, vol. 110, no. 1, pp. 145–160, 1993.

[2] W P. Glezen, L. H. Taber, A L. Frank et al., "Influenza virus infections in infants," *The Pediatric Infectious Disease Journal*, vol. 16, no. 11, pp. 1065–1068, 1997.

[3] E. Hurwitz, M. Haber, M. Ginsberg et al., "Studies of the 1996–1997 Inactivated Influenza Vaccine among Children Attending Day Care: Immunologic Response, Protection against Infection, and Clinical Effectivenessnactivated influenza vaccine among children attending day care: immunologic response, protection against infection, and clinical effectiveness," *The Journal of Infectious Diseases*, vol. 182, no. 4, pp. 1218–1221, 2000.

[4] T. Heikkinen, P. Toikka, R. Vainionpaa et al., "Burden of influenza in children in the community," *The Journal of Infectious Diseases*, vol. 190, no. 8, pp. 1369–1373, 2004.

[5] A D. Iuliano, B. J. Cowling, H H. Chang et al., "Estimates of global seasonal influenza-associated respiratory mortality: a modelling study," *The Lancet*, vol. 391, no. 10127, pp. 1285–1300, 2018.

[6] I. M. Longini, J. S. Koopman, A. S. Monto et al., "Estimating household and community transmission parameters for influenza," *American Journal of Epidemiology*, vol. 115, no. 5, pp. 736–751, 1982.

[7] W P. Glezen, W. A. Keitel, L. H. Taber et al., "Age Distribution of Patients with Medically-Attended Illnesses Caused by Sequential Variants of Influenza A/H1N1: Comparison to Age-Specific Infection Rates, 1978–1989istribution of patients with medically-attended illnesses caused by sequential variants of influenza A/H1N1: comparison to age-specific infection rates, 1978–1989," *American Journal of Epidemiology*, vol. 133, no. 3, pp. 296–304, 1991.

[8] F M. M. Munoz, "The impact of influenza in children," *Seminars in Pediatric Infectious Diseases*, vol. 13, no. 2, pp. 72–78, 2002.

[9] Z. He, H. Tao, and H. Tao, "Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study nine-year retrospective study," *International Journal of Infectious Diseases*, vol. 74, pp. 61–70, 2018.

[10] C. Wang, G. Xu, H. Ni et al., "Epidemiological Epidemiological Features and Forecast Model Analysis for the Morbidity of Influenza in Ningbo, China, 2006–," *IJERPH*, vol. 14, no. 6, p. 559, 2017.

[11] R. Cong and W. Xie, "Predicting Seasonal Influenza Based on SARIMA Model, in Mainland China from 2005 to 2018," *IJERPH*, vol. 16, no. 23, p. 4760, 2019.

[12] C. Poirier, A. Lavenu, E. Chazard et al., "Real Real Time Influenza Monitoring Using Hospital Big Data in Combination with Machine Learning Methods: Comparison Studyime influenza monitoring using hospital big data in combination with machine learning methods: comparison study," *JMIR Public Health Surveill*, vol. 4, Article ID e11361, 2018.

[13] P P. Schneider, C. J. van Gool, P. Spreeuwenberg et al., "Using web search queries to monitor influenza-like illness: an exploratory retrospective analysis, Netherlands, 2017/18 influenza season," *Euro Surveillance*, vol. 25, no. 21, Article ID 1900221, 2020.

[14] S. Yang, M. Santillana, S. C. Kou, and S. C. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14473–14478, 2015.

[15] M. J. Kane, N. Price, M. Scotch et al., "Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks," *BMC Bioinformatics*, vol. 15, no. 1, p. 276, 2014.

[16] S. Chae, S. Kwon, D. Lee, and D. Lee, "Predicting Predicting Infectious Disease Using Deep Learning and Big Datanfectious disease using deep learning and big data," *IJERPH*, vol. 15, no. 8, p. 1596, 2018.

[17] L. Li, Y. Jiang, B. Huang, and B. Huang, "Long-term prediction for temporal propagation of seasonal influenza using Transformer-based model," *Journal of Biomedical Informatics*, vol. 122, Article ID 103894, 2021.

[18] S. Mwalili, M. Kimathi, R. Mbogo et al., "SEIR model for COVID-19 dynamics incorporating the environment and social distancing," *BMC Research Notes*, vol. 13, no. 1, p. 352, 2020.

[19] J. Shaman, A. Karspeck, M. Lipsitch et al., "Real-time influenza forecasts during the 2012–2013 season," *Nature Communications*, vol. 4, no. 1, p. 2837, 2013.

[20] R. Jan, S. Boulaaras, and S. Boulaaras, "Analysis of fractional-order dynamics of dengue infection with non-linear incidence functions," *Transactions of the Institute of Measurement and Control*, vol. 44, no. 13, pp. 2630–2641, 2022.

[21] S. Boulaaras, R. Jan, A. Khan, M. Ahsan, A. Khan, and M. Ahsan, "Dynamical analysis of the transmission of dengue fever via Caputo-Fabrizio fractional derivative," *Chaos, Solitons & Fractals: X X.*vol. 8, Article ID 100072, 2022.

[22] T Q. Tang, Z. Shah, E. Bonyah et al., "Modeling and Modeling and Analysis of Breast Cancer with Adverse Reactions of Chemotherapy Treatment through Fractional Derivativenalysis of breast cancer with adverse reactions of chemotherapy treatment through fractional derivative," *Computational and Mathematical Methods in Medicine*, vol. 2022, pp. 1–19, 2022.

[23] E. Nsoesie, M. Mararthe, J. Brownstein, and J. Brownstein, "Forecasting Forecasting peaks of seasonal influenza epidemics.eaks of seasonal influenza epidemics," *PLoS Curr*, vol. 5, 2013.

[24] E. L. Ray, N. G. L. Reich, and N. G. Reich, "Prediction of infectious disease epidemics via weighted density ensembles,"

PLoS Computational Biology, vol. 14, no. 2, Article ID e1005910, 2018.

[25] N G. Reich, G. C. Gibson, T K. Yamana et al., "Accuracy of real-time multi-model ensemble forecasts for seasonal influenza in the US," *PLoS Computational Biology*, vol. 15, no. 11, Article ID e1007486, 2019.

[26] S. Cook, C. Conrad, A L. Fowlkes, M. H. Mohebbi, A. L. Fowlkes, and M. H. Mohebbi, "Assessing Google Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemiclu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. Cowling bj, editor," *PLoS One*, vol. 6, no. 8, Article ID e23610, 2011.

[27] Q. Yuan, E. O. Nsoesie, B. Lv et al., "Monitoring Monitoring Influenza Epidemics in China with Search Query from Baidunfluenza epidemics in China with search query from baidu. Cowling BJ," *PLoS One*, vol. 8, no. 5, Article ID e64323, 2013.

[28] C T. Yang, Y. A. Chen, Y W. Chan et al., "Influenza-like illness prediction using a long short-term memory deep learning model with multiple open data sources," *The Journal of Supercomputing*, vol. 76, no. 12, pp. 9303–9329, 2020.

[29] G J. Yu, J. L. Gu, W B. Cui et al., "Identifying and characterizing the effects of calendar and environmental conditions on pediatric admissions in Shanghai," *Journal of Big Data*, vol. 6, no. 1, p. 14, 2019.

[30] G. De Toni, C. Consonni, A. Montresor, and A. Montresor, "A general method for estimating the prevalence of influenza-like-symptoms with Wikipedia data," *PLoS One*, vol. 16, no. 8, Article ID e0256858, 2021.

[31] H. Woo, H. Sung Cho, K. Lee et al., "Identification of Identification of Keywords From Twitter and Web Blog Posts to Detect Influenza Epidemics in Koreaeywords from twitter and web blog posts to detect influenza epidemics in korea," *Disaster Medicine and Public Health Preparedness*, vol. 12, no. 3, pp. 352–359, 2018.

[32] M. Santillana, A. T. Nguyen, J S. Brownstein et al., "Cloud-based Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillancelectronic health records for real-time, region-specific influenza surveillance," *Scientific Reports*, vol. 6, no. 1, Article ID 25732, 2016.

[33] J. P. Chretien, D. George, J. Shaman, R. A. Chitale, and F. E. McKenzie, "Influenza forecasting in human populations: a scoping review," *PLoS One*, N. G. Reich, Ed., vol. 9, Article ID e94130, 2014.

[34] E O. Nsoesie, J. S. Brownstein, N. Ramakrishnan et al., "A systematic review of studies on forecasting the dynamics of influenza outbreaks," *Influenza Other Respi Viruses*, vol. 8, no. 3, pp. 309–316, 2014.

[35] N G. Reich, T. K. Yamana, S J. Fox et al., "A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States," *Proceedings of the National Academy of Sciences*, vol. 116, no. 8, pp. 3146–3154, 2019.

[36] X. Rao, Z. Chen, Y. Yan et al., "Epidemiology of influenza in hospitalized children with respiratory tract infection in Suzhou area from 2016 to 2019," *Journal of Medical Virology*, vol. 92, no. 12, pp. 3038–3046, 2020.

[37] J. Park, W. Son, O. Kwon et al., "Effects of temperature, humidity, and diurnal temperature range on influenza incidence in a temperate region," *Influenza Other Respi Viruses*, vol. 14, no. 1, pp. 11–18, 2020.

[38] E. Lofgren, N. H. Fefferman, Y N. Naumov et al., "Influenza Seasonality: Underlying Causes and Modeling Theorieseasonality: underlying causes and modeling theories," *Journal of Virology*, vol. 81, no. 11, pp. 5429–5436, 2007.

[39] X X. Liu, X. Zheng, K. Zhao et al., "Effects of air pollutants on occurrences of influenza-like illness and laboratory-confirmed influenza in Hefei, China," *International Journal of Biometeorology*, vol. 63, no. 1, pp. 51–60, 2019.

[40] Y. Zheng, K. Wang, L. Zhang, and L. Wang, "Study on the relationship between the incidence of influenza and climate indicators and the prediction of influenza incidence," *Environmental Science and Pollution Research*, vol. 28, no. 1, pp. 473–481, 2021.

[41] L. Liu, M. Han, Y. Zhou, and Y. Wang, "LSTM recurrent neural networks for influenza trends prediction," in *Bioinformatics Research and Applications*, F. Zhang, Z. Cai, P. Skums, and S. Zhang, Eds., Springer International Publishing, Berlin, Germany, pp. 259–264, 2018.

[42] tianqihoubao, "Shanghai historical weather database," 2022, http://www.tianqihoubao.com/lishi/shanghai.html.

[43] tianqihoubao, "Shanghai historical AQI database," 2022, http://www.tianqihoubao.com/aqi/shanghai.html.

[44] G. James, D Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer, Berlin,Germany, 2013.

[45] D. Benvenuto, M. Giovanetti, M. Ciccozzi et al., "Application of the ARIMA model on the COVID-2019 epidemic dataset," *Data in Brief*, vol. 29, Article ID 105340, 2020.

[46] T. Petukhova, D. Ojkic, B. McEwen, R. Deardon, and Z. Poljak, "Assessment of autoregressive integrated moving average (ARIMA), generalized linear autoregressive moving average (GLARMA), and random forest (RF) time series regression models for predicting influenza A virus frequency in swine in Ontario, Canada," *PLoS One*, J. Shaman, Ed., vol. 13, Article ID e0198313, 2018.

[47] V. Isham, N. Keiding, T. Louis, N. Reid, R. Tibshirani, and H. Tong, *Subset selection in regression*, Chapman and Hall/CRC, London, UK, 2002.

[48] X. Fang, W. Liu, Y. Wu et al., "Forecasting incidence of infectious diarrhea using random forest in Jiangsu Province, China," *BMC Infectious Diseases*, vol. 20, no. 1, p. 222, 2020.

[49] J. Gu, Y. Zhang, N. He et al., "A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China," *Scientific Reports*, vol. 9, no. 1, Article ID 17928, 2019.

[50] A. Kraskov, H. Stögbauer, P. Grassberger, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, Article ID 066138, 2004.

[51] C. Bergmeir and J. M. Benítez, "On the use of cross-validation for time series predictor evaluation," *Information Sciences*, vol. 191, pp. 192–213, 2012.