Computers, Materials & Continua
DOI:10.32604/cmc.2021.017435
Article

Tech Science Press

Machine Learning-Based Two-Stage Data Selection Scheme for Long-Term Influenza Forecasting

Jaeuk Moon, Seungwon Jung, Sungwoo Park and Eenjun Hwang[*]

School of Electrical Engineering, Korea University, Seoul, 02841, Korea
[*]Corresponding Author: Eenjun Hwang. Email: ehwang04@korea.ac.kr

Received: 29 January 2021; Accepted: 01 March 2021

Abstract: One popular strategy to reduce the enormous number of illnesses and deaths from a seasonal influenza pandemic is to obtain the influenza vaccine on time. Usually, vaccine production preparation must be done at least six months in advance, and accurate long-term influenza forecasting is essential for this. Although diverse machine learning models have been proposed for influenza forecasting, they focus on short-term forecasting, and their performance is too dependent on input variables. For a country's long-term influenza forecasting, typical surveillance data are known to be more effective than diverse external data on the Internet. We propose a two-stage data selection scheme for worldwide surveillance data to construct a long-term forecasting model for influenza in the target country. In the first stage, using a simple forecasting model based on the country's surveillance data, we measured the change in performance by adding surveillance data from other countries, shifted by up to 52 weeks. In the second stage, for each set of surveillance data sorted by accuracy, we incrementally added data as input if the data have a positive effect on the performance of the forecasting model in the first stage. Using the selected surveillance data, we trained a new long-term forecasting model for influenza and perform influenza forecasting for the target country. We conducted extensive experiments using six machine learning models for the three target countries to verify the effectiveness of the proposed method. We report some of the results.

Keywords: Influenza; data selection; machine learning; forecasting

1 Introduction

Seasonal influenza is one of the most globally prevalent infectious diseases, annually causing tens of millions of respiratory illnesses and hundreds of thousands of deaths worldwide [1]. Most countries have established national health institutes and conducted various activities, including disinfection and quarantine, to reduce the losses. One of the most effective ways to prevent an influenza pandemic is to prepare an influenza vaccine on time [2]. However, due to the time-consuming nature of vaccine production, an elaborate vaccine strategy should be prepared at least six months in advance [3]. Substantial uncertainty exists in such a long-term strategy, which leads to an imbalance in the supply and demand of influenza vaccines during influenza seasons. An insufficient number of vaccines cannot prevent an increase in the number of influenza patients. However, an oversupply of vaccines leads to economic loss because obsolete vaccines must be discarded. Therefore, forecasting how many influenza patients will occur after a long period is required to ensure a smooth vaccine supply [4].

With the recent development of machine learning (ML) technology, various ML-based forecasting models have been proposed to achieve better forecasting accuracy [5–8]. These models take data-driven approaches to identify the influence of various factors represented by the input variables and show superior forecasting performance compared to the compartmental and statistical models [9,10]. However, most ML-based forecasting models for influenza perform short-term forecasting. Diverse data sources exist for short-term influenza forecasting, such as Google searches and microblogs. Their effectiveness has already been proven in many studies [11,12]. In contrast, long-term forecasting faces increasing uncertainty from various sources, such as the accumulation of errors and lack of information [13]. Hence, selecting appropriate input variables is important to guarantee good forecasting performance in ML-based long-term influenza forecasting. Otherwise, the forecasting performance may deteriorate [14]. For long-term influenza forecasting, traditional surveillance data can provide much more comprehensive data with a small time lag, which is easier to maintain and more reliable for a long-term decision-making system than Internet data [9]. Traditional surveillance data generally refers to the number of reported cases over time for a particular disease.

In this paper, we propose an ML-based two-stage data selection scheme for worldwide surveillance data to construct a long-term forecasting model for influenza in a target country after 26 weeks (about six months). In the first stage, based on a simple forecasting model using the country's surveillance data, we measured performance by adding foreign surveillance data shifted by up to 52 weeks. In the second stage, for each set of surveillance data sorted by performance, we incrementally added each set of surveillance data if it positively affects forecasting. Using the resulting surveillance data, we trained another long-term forecasting model of influenza. To evaluate the performance of our proposed method, we conducted extensive experiments using the influenza surveillance data for various countries and several ML models.

The contributions of this paper are as follows:

- We propose a data selection method for foreign surveillance data, which we use as input variables to construct an accurate forecasting model of influenza regardless of the target country or forecasting model.
- We achieved outstanding forecasting accuracy using traditional surveillance data as input variables without external data sources.
- We verified the effectiveness of the proposed method through extensive comparisons with popular ML models.

The remainder of this paper is organized as follows. Section 2 discusses the literature review. Next, Section 3 describes the details of the proposed scheme. Section 4 demonstrates various experiments, and finally, Section 5 presents the conclusions.

2 Related Work

This section presents a brief literature review on various models for influenza forecasting. Work on influenza forecasting can be generally classified into three categories. The first category is based on compartment models, which use differential equations to model infectious disease transmission. For instance, Zhang et al. [15] proposed a spatiotemporal risk assessment model based on the susceptible infected recovered model, evaluating four influencing factors: biological, behavioral, and environmental parameters and infectious sources. They displayed the model output in a set of maps to analyze how these factors affect the spread of infectious diseases in Beijing. Because these models require a relatively small number of parameters, these models are limited because they often have a low forecasting accuracy or cannot find a correlation between various data when dealing with big data [16].

The second category is based on statistical or time-series-based models. For instance, Choi et al. [17] employed a Bayesian maximum entropy method of spatiotemporal statistics to analyze the geographical risk patterns of influenza mortality in California during winter. They found that the high risk of influenza initially occurred during December in the west-central region of the state, and the risk distribution was extended in the south and east-central regions of the state. Choi et al. [18] proposed an autoregressive integrated moving average (ARIMA) model for forecasting influenza activity. They collected the number of deaths related to pneumonia and influenza and used their ratio to measure influenza activity. Although these models are flexible in capturing the trending behavior of the affected populations, they sometimes suffer from poor accuracy because the influence of external factors (e.g., climate and environmental factors) is not captured well [19].

The final category is based on ML-based models. Cheng et al. [9] proposed an ensemble approach for short-term influenza forecasting in Taiwan. They used the four ML-based forecasting models random forest (RF), ARIMA, support vector regression, and extreme gradient boosting (XGB) for traditional surveillance data from Taiwan. Then, they integrated their forecasting results using a linear kernel model to produce a more robust model than the individual model. Park et al. [8] used data from other countries to improve the forecasting accuracy of influenza occurrences in a specific country. They obtained the similarity between traditional surveillance data from the target country and other countries with the Euclidean distance. Then, they selected countries with high similarity in influenza patterns and exploited their data as input variables using a light gradient boosting machine (LGBM). Venna et al. [19] proposed a long short-term memory (LSTM)-based multi-stage scheme for influenza forecasting. They employed LSTM to capture the temporal dynamics of seasonal influenza in the first stage. During subsequent stages, the situational time lag between the influenza occurrence and weather variables and the spatial proximity of different geographical regions were captured to adjust the error introduced by the original forecasting model to improve the model performance further. These models take a data-driven approach and grasp the influence of various factors used as input variables, exhibiting superior forecasting results compared to the compartmental and statistical models [9,10].

Most studies on influenza forecasting, including the aforementioned studies, have been for short-term forecasting. In contrast, the studies on long-term influenza forecasting are not yet sufficient. In short-term influenza forecasting, diverse data sources exist, such as Google searches and microblogs, whose effectiveness has already been proven [11,12]. However, long-term forecasting faces growing uncertainty arising from various problems, such as the accumulation of errors and lack of information [13]. Choi et al. [20] presented a long-term influenza forecasting scheme for the United States (US) using influenza activities in other countries. They first collected the data widely used in influenza forecasting, including traditional surveillance data from other countries and search queries from Google Trend (GT). They calculated the cross-correlation between the traditional data from the US and other countries with similar seasonal patterns and influenza outbreaks, and employed highly correlated data as input variables to forecast the next seasonal influenza pattern using ML models. Although they achieved remarkable forecasting accuracy, they manually selected countries with a high correlation to the target country using a statistical method. In this study, we propose a data selection scheme that automatically finds surveillance data from other countries that can contribute to the prediction of the target country using an ML model to improve efficiency and accuracy. We demonstrate that the input configuration based on this selection improves the performance of diverse influenza forecasting models.

3 Method

In this section, we describe the proposed scheme in detail. Fig. 1 illustrates the overall structure of the scheme, which is composed of three parts: (1) data collection and preprocessing, (2) data selection for configuring input variables for the forecasting model, and (3) influenza forecasting for the target country. In the following sections, we first describe the data collection and preprocessing part, and then the data selection and the forecasting part together.

3.1 Data Collection and Preprocessing

In this study, we collected influenza surveillance data from the FluNet database of the World Health Organization (WHO) [21]. FluNet is a global web-based tool for influenza virological surveillance, which was first launched in 1997. Influenza surveillance data have been uploaded to the FluNet database every week. Among the diverse data provided by FluNet, we used the number of influenza patients in 168 countries from the first week in 2010 to the 52nd week in 2018. We collected data from 2010 because influenza showed a peculiar outbreak pattern in 2009 due to the introduction of a new epidemic strain (INF A H1N1 pdm09) [22]. Some countries uploaded surveillance data only when they had high influenza activity. Hence, we replaced the missing data in the dataset with zero. For the collected surveillance data, we performed a min-max normalization for each country, which is necessary to prevent data selection from focusing on a few specific countries with high average occurrences.

In addition to the surveillance data, we also collected time information, such as the year and week in which the surveillance data were collected. As one year consists of 52 or 53 weeks, we represented the week in the ML model using this week number. In particular, to reflect the periodic property of the week, we transformed the week number into two-dimensional data using Eqs. (1) and (2) [23]. *Cycle* represents the period of the week. For instance, when we transform the first week of 2015 into two-dimensional data, *week* and *cycle* are 1 and 53, respectively, and $week_x$ and $week_y$ are $sin((360/53) * 1)$ and $cos((360/53) * 1)$, respectively. Further explanations of this transformation can be found in [24].

Tab. 1 lists the initial input and output variables that we considered in this paper. Initial input variables of the forecasting ML model are necessary to provide basic information such as time and the number of influenza patients in the target country. The initial input variables include the *year*, *week*, $week_x$, $week_y$, and $Occur_{target,t}$. The last variable represents the number of patients in the target country at time t. One output variable, $Occur_{target,t+26}$, represents the number of influenza patients in the target country 26 weeks (6 months) later because we aim for long-term forecasting for the vaccine strategy:

$$week_x = \sin((360/cycle) \times week), \tag{1}$$

$$week_y = \cos((360/cycle) \times week). \tag{2}$$

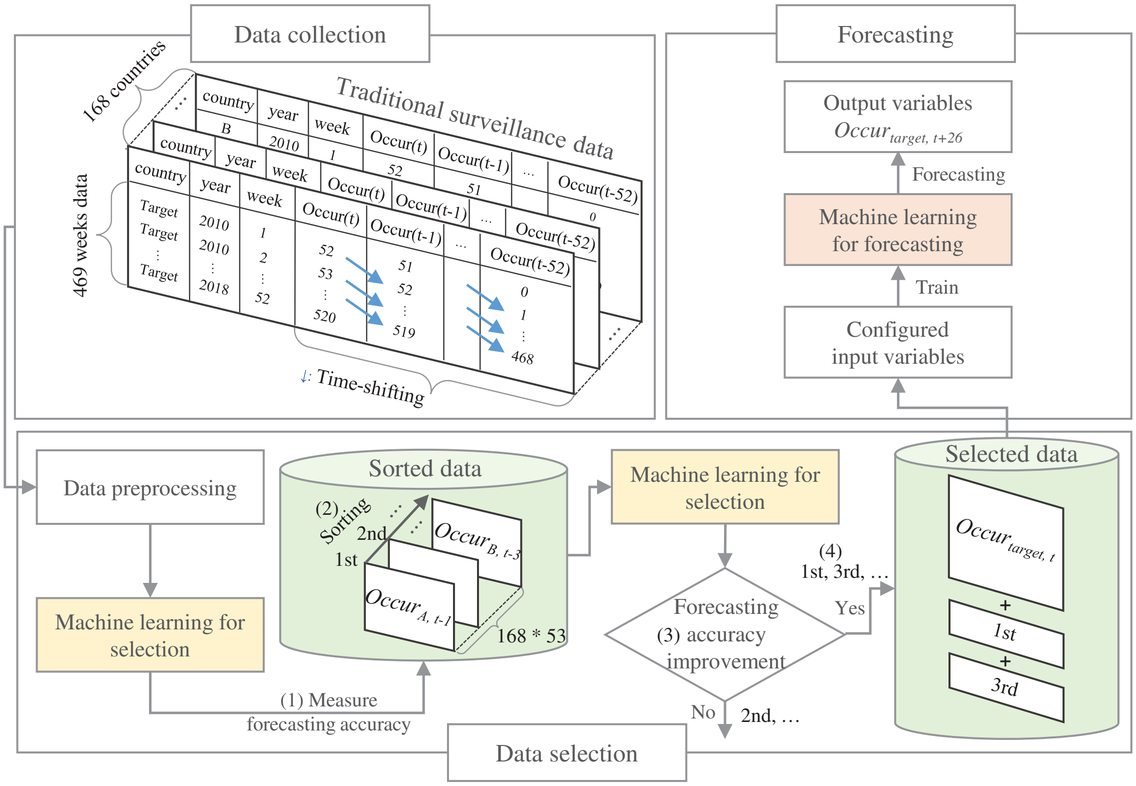


Figure 1: Overall structure of the proposed scheme

3.2 Data Selection and Forecasting

One simple method to construct an influenza forecasting model is to use all surveillance data collected worldwide for training and validation. However, this method takes a long time and can have poor predictive performance. It would be better to consider data that contribute to influenza forecasting from the collected surveillance data to construct a forecasting model more effectively. To do this, we perform data selection in two stages. In the first stage, we construct a simple ML forecasting model using the surveillance data from the target country, which are initial input variables and output variables in Tab. 1. Then, (1) we measure the accuracy of the forecasting model after appending the foreign surveillance data unit to the initial input variables, possibly shifted by up to 52 weeks. In the second stage, (2) we sort all surveillance data units in the order of accuracy. Then, we consider the surveillance data units one by one. (3) If it positively affects forecasting, (4) we select and incrementally add the data to the model as an input variable. Otherwise, we ignore the unit and continue the process. We used this strategy assuming that the data that produced higher forecasting accuracy contain more useful information for forecasting. Finally, the collection of selected surveillance data is used to train the target long-term influenza forecasting model. To evaluate the performance of our proposed method, we conducted extensive experiments using the influenza surveillance data from 168 countries and diverse ML models for data selection and influenza forecasting.

Algorithm 1 describes the overall steps for performing data selection. We considered six ML models for data selection, including the Gradient Boosting Machine (GBM), RF, and linear regression (LR). We also used a validation set to measure the forecasting accuracy. As an evaluation metric, we used the root mean squared error (RMSE).

Table 1: Input and output variables for the data selection model

Input variables	Description	Variable type
$year$	Year value	Continuous
$week$	Week value	Continuous
$week_x$	Sine value for the week	Continuous on [−1, 1]
$week_y$	Cosine value for the week	Continuous on [−1, 1]
$Occur_{target,t}$	Influenza patients in the target country for the week	Continuous
Output variable	Description	Variable type
$Occur_{target,t+26}$	Influenza patients in the target country 26 weeks later	Continuous

```
Algorithm 1: Data selection and input variable configuration
Input: Initial input variables $I_{initial}$
Output: Final input variables $I_{final}$
countries: List of 168 countries
$Occur_{i,t-j}$: Occurrence data of country $i$ that shifted right $j$ times
$SO$: Sorted list of occurrence data in order of forecasting accuracy
$FM(I)$: Forecasting model using input variables $I$
for $i$ in countries do // The first stage of data selection
    for $j = 0$ to 52 do
        measure Accuracy of $FM(I_{initial} \cup Occur_{i,t-j})$ // (1)
        insert $Occur_{i,t-j}$ into $SO \cdots$ (2)
$I_{final} = I_{initial}$; // The second stage of data selection
for $i$ in $SO$ do
        if Accuracy of $FM\ (I_{final}) <$ Accuracy of $FM\ (I_{final} \cup i)$ do // (3)
$I_{final} = I_{final} \cup i$ // (4)
return $I_{final}$
```

4 Experimental Setup

To demonstrate the effectiveness of the proposed scheme, we constructed six different influenza forecasting models based on the ML models used in data selection using the selected surveillance data as input variables. In summary, we have 36 different combinations of data selection and forecasting methods. The six models are composed of two simple models and four ensemble models. The two simple models are the decision tree (DT) and LR, and the four ensemble models are GBM, XGB, LGBM, and RF.

The hyperparameters of each model, such as the number of estimators or maximum depth of the ensemble models, were decided using the grid search with the validation set. When the validation set is changed, the hyperparameters of each model are reset and redetermined. We implemented all models using Python 3.7.3 with the scikit-learn library [25].

We used time-series cross-validation to evaluate the accuracy of the forecasting models. This method has the advantage that it is closer to real-world practical applications than the traditional evaluation method, which divides a dataset into training and testing sets to train and test models, respectively [26]. We forecast one year for influenza from 2015 to 2018 and used data from a year before the test year as the validation set. For example, when the test year is 2015, we use data from 2010 to 2013 as the training set and data from 2014 as the validation set. Likewise, if we forecast 2016, data from 2010 to 2014 were used as a training set, and data from 2015 were used as a validation set.

We also calculated the RMSE and mean absolute error (MAE) using Eqs. (3) and (4), respectively, to compare accuracy. Here, N is the number of data samples, and A_n and F_n represent actual influenza occurrence and forecasted influenza occurrence, respectively.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{n}(A_n - F_n)^2}, \tag{3}$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{n}|A_n - F_n|. \tag{4}$$

5 Experiments and Discussion

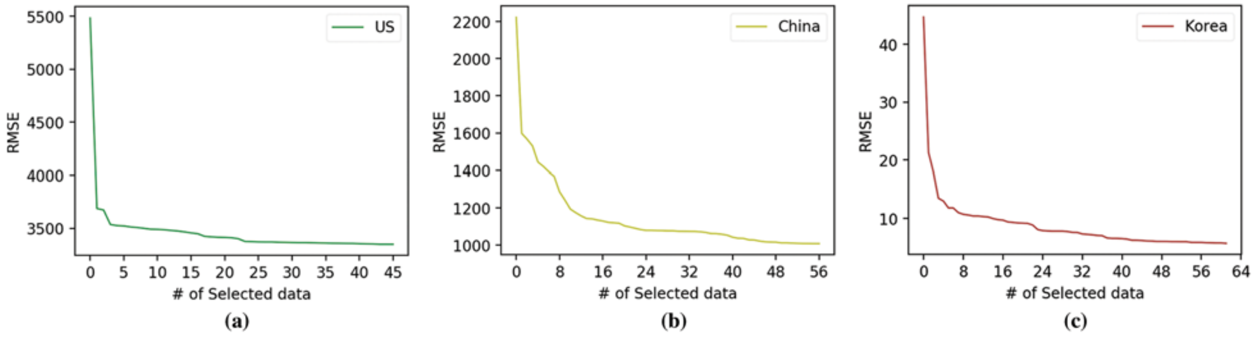
To verify the effectiveness of our data selection scheme, we performed two experiments for three target countries: the US, China, and the Republic of Korea (Korea), and Tab. 2 shows a brief summary of the datasets for the target countries. In the first experiment, we demonstrated why the proposed data selection is valid. In the second experiment, we evaluated the effect of our data selection on the forecasting performance.

Table 2: Dataset summaries of the target countries

	US	China	Korea
Average	2393.47	1297.70	37.33
Maximum	26316	8426	416
Minimum	0	0	0

5.1 Validity of Data Selection

In this experiment, we present two analysis results to justify our data selection scheme: (1) the RMSE changes according to input variables and (2) the comparison between the target country and the first selected country.



(a) (b) (c)

Figure 2: Root mean squared error (RMSE) according to data selection. (a) US. (b) China. (c) Korea.

Fig. 2 depicts the changes in the RMSE of the three target countries according to the input variables when forecasting the influenza occurrence for 2018 using XGB. The x-axis represents the number of selected data, and the y-axis represents the RMSE value. The initial RMSE values in the US, China, and Korea were around 5500, 2200, and 45, respectively, which are forecasting results of ML models constructed using the initial input variables. The number of data points selected by the proposed data selection scheme for the US, China, and Korea is 45, 56, and 61, respectively. All three countries' graphs depict a sharp decline at the beginning and then a very modest decline. This result indicates that using a few well-selected surveillance data points can significantly improve forecasting performance.

Tab. 3 presents detailed information about the selected data, including the country, how many shifts are performed to the initial surveillance data for the three target countries, and the reduction in RMSE by adding the selected data as input. For instance, in the US, the data from Chile 50 weeks ago ($Occur_{Chile,t-50}$) improved forecasting performance the most. Data from Kazakhstan, Belarus, and Congo follow, but these data did not significantly improve the forecasting performance.

Table 3: List of selected data and their root mean squared error (RMSE) with reduced value

Target country											
US				**China**				**Korea**			
Selected data	Country	# of shifts	RMSE (reduced value)	Selected data	Country	# of shifts	RMSE (reduced value)	Selected data	Country	# of shifts	RMSE (reduced value)
0	*X*	*X*	5483.52 (0)	0	*X*	*X*	2221.82 (0)	0	*X*	*X*	44.71 (0)
1	*Chile*	50	3687.12 (−1796.40)	1	*Sri Lanka*	15	1597.67 (−624.15)	1	*Korea*	19	21.29 (−23.42)
2	*Chile*	51	3673.53 (−13.59)	2	*Sri Lanka*	12	1566.54 (−31.13)	2	*Korea*	20	17.89 (−3.40)
3	*Chile*	49	3536.36 (−137.17)	3	*Sri Lanka*	17	1530.26 (−36.27)	3	*Korea*	18	13.48 (−4.40)
4	*Chile*	18	3526.10 (−10.20)	4	*Sri Lanka*	26	1443.61 (−86.64)	4	*Sri Lanka*	3	13.00 (−0.48)
5	*Kazakhstan*	3	3523.06 (−3.05)	5	*Sri Lanka*	46	1420.59 (−23.01)	5	*Ghana*	30	11.81 (−1.19)
6	*Kazakhstan*	2	3513.40 (−9.66)	6	*India*	42	1392.46 (−28.13)	6	*Mali*	38	11.80(−0.01)
7	*Belarus*	41	3508.22 (−5.18)	7	*India*	48	1369.19 (−23.26)	7	*Costa Rica*	38	11.01 (−0.79)
…	…	…	…	…	…	…	…	…	…	…	…
45	*Congo*	18	3349.49 (−0.10)	56	*Angola*	3	1007.81 (−0.10)	61	*Congo*	47	5.69 (−0.06)

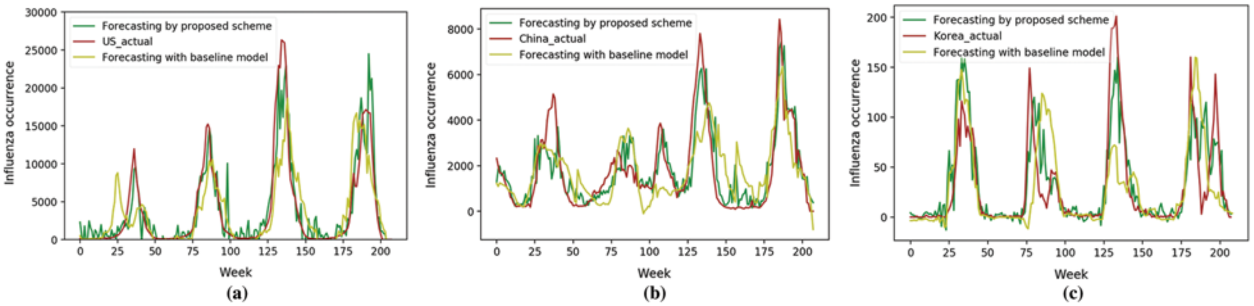


Figure 3: Scaled influenza occurrence of the target country and the first selected country. (a) US (b) China (c) Korea

Fig. 3 illustrates the graphs of the target country after 26 weeks ($Occur_{target,t+26}$) and the first selected country in Tab. 3. The x-axis represents weeks from 2010 to 2018, and the y-axis represents the influenza occurrence scaled from 0 to 1. The first added data ($Occur_{Chile,\ t-50}$ for US, $Occur_{Sri\ Lanka,t-15}$ for China, and $Occur_{Korea,t-19}$ for Korea) exhibit patterns similar to the target data except for the first two years (104 weeks). This result indicates that the proposed scheme can select influenza patterns in foreign countries that are very similar to those of the target country.

5.2 Forecasting Performance

In this experiment, we investigated the effect of the proposed data selection scheme on forecasting performance and then compared the performance of diverse combinations of data selection and forecasting models. Tabs. 4–6 list the RMSE values for the US, China, and Korea, respectively. In the case of the US, the DT exhibits the best performance for data selection in most cases. However, using LR for both data selection and forecasting exhibits the best overall performance. For China, LR exhibits the best performance in data selection, and in Korea, LR has the best performance in data selection in half of all cases.

Table 4: Root mean squared error (RMSE) of data selection and forecasting in the US

		Selection model					
		DT	LR	GBM	LGBM	XGB	RF
Forecasting model	**DT**	**3684.59**	4216.09	4420.13	3982.35	4558.63	4659.43
	LR	3471.61	**2482.14**	3858.34	3466.59	3622.88	3836.29
	RF	**2595.99**	3187.37	3381.01	3003.95	3495.23	3665.54
	GBM	**2819.55**	3128.07	3449.60	3192.83	3398.81	3644.26
	LGBM	**2728.96**	3442.96	3610.31	3100.27	3725.37	3786.16
	XGB	**2862.47**	3369.88	3671.87	3177.51	3789.94	3979.59

Table 5: Root mean squared error (RMSE) of data selection and forecasting in China

		Selection model					
		DT	LR	GBM	LGBM	XGB	RF
Forecasting model	**DT**	1285.77	**1174.32**	1308.35	1351.08	1341.63	1470.95
	LR	1376.33	**1002.86**	1364.88	1312.61	1204.73	1126.13
	RF	961.01	**848.06**	873.61	915.36	913.58	1027.62
	GBM	1073.52	**943.00**	998.66	1037.79	1017.82	953.13
	LGBM	1060.23	**902.97**	955.76	966.64	957.61	1031.86
	XGB	1127.97	**951.46**	1049.73	1011.75	1045.45	1073.39

We also compared the accuracy of diverse forecasting models in terms of input variables in influenza forecasting to verify the effectiveness of the configured input variables. For comparison, we considered four input variable sets: (1) I, initial input variables; (2) $I+GT$, which added GT of the current time to the initial input variables; (3) $I+T$, which added one year of domestic traditional surveillance data of the target country from $Occur_{target,t-1}$ to the target

$Occur_{target,t-52}$ to the initial input variables; and (4) $I+T+GT$, which added both GT and the domestic traditional surveillance data of the target country. The data selection model with the best performance for each forecasting model (bold font in Tabs. 3–5) was used as the performance of our proposed scheme. Tabs. 7–9 represent the forecasting performance (in the RMSE and MAE) of the US, China, and Korea, respectively. Input variables that demonstrated the best performance in each forecasting model are marked in bold font.

Table 6: Root mean squared error (RMSE) of data selection and forecasting in Korea

		Selection model					
		DT	LR	GBM	LGBM	XGB	RF
Forecasting model	**DT**	38.90	42.80	35.21	35.09	**31.96**	42.41
	LR	53.25	**39.57**	49.70	55.99	56.18	53.79
	RF	30.68	**28.90**	31.06	32.06	30.23	31.71
	GBM	34.06	**25.88**	32.69	29.29	30.03	30.67
	LGBM	30.8	32.57	**27.43**	27.47	28.12	28.65
	XGB	29.07	31.31	30.10	**27.77**	28.44	31.96

Table 7: Comparison of forecasting result of baseline models and the proposed scheme in the US

	RMSE					MAE				
	I	$I+GT$	$I+T$	$I+T+GT$	Proposed scheme	I	$I+GT$	$I+T$	$I+T+GT$	Proposed scheme
DT	4820.55	4948.92	5345.37	5522.58	**3684.59**	2557.96	2695.56	3128.96	3114.89	**1852.92**
LR	4700.28	4682.85	5321.12	5361.71	**2482.14**	2902.12	2922.42	3312.52	3466.89	**1581.93**
RF	4625.57	4818.47	4778.30	4839.09	**2595.99**	2415.34	2558.13	2812.05	2683.21	**1293.92**
GBM	4685.97	4921.06	4542.68	4760.07	**2819.55**	2565.70	2681.84	2755.98	2668.51	**1553.18**
LGBM	4561.45	4574.18	5186.42	4618.8	**2728.96**	2678.31	2650.85	3052.57	2706.19	**1388.75**
XGB	4813.54	4582.77	5283.21	4926.31	**2862.47**	2527.90	2525.35	3113.47	2855.15	**1413.85**

Table 8: Comparison of forecasting result of baseline models and the proposed scheme in China

	RMSE					MAE				
	I	$I+GT$	$I+T$	$I+T+GT$	Proposed scheme	I	$I+GT$	$I+T$	$I+T+GT$	Proposed scheme
DT	1672.31	1670.57	2224.67	2187.76	**1174.32**	1258.53	1252.97	1681.95	1566.78	**784.89**
LR	1644.56	1500.89	1611.76	1648.73	**1002.86**	1192.01	1143.45	1235.58	1257.48	**740.64**
RF	1586.33	1599.40	1742.00	1764.19	**848.06**	1191.49	1199.14	1291.99	1310.78	**619.43**
GBM	1585.96	1582.80	1775.28	1805.67	**943.00**	1208.97	1195.35	1366.29	1384.64	**668.38**
LGBM	1586.03	1617.69	1689.74	1678.06	**902.97**	1216.58	1248.67	1308.87	1303.94	**628.05**
XGB	1653.12	1641.77	1814.09	1897.13	**951.46**	1262.76	1221.16	1367.07	1381.80	**687.74**

In the comparison, the proposed scheme exhibited the best performance among all input variable sets we constructed. For China and the US, the forecasting performance deteriorated in several cases when using traditional surveillance data or GT data. Although the GT data have been used in short-term influenza forecasting in many papers [27,28], the data are not effective for long-term forecasting. According to Tab. 3, the top three surveillance data that improved the forecasting performance in Korea were from the same country. As a result, in Korea (Tab. 9), past domestic surveillance data contributed to improving forecasting performance in most cases, which indicates that input variables configured using the proposed scheme are more effective for long-term influenza forecasting than other commonly used input variables.

Table 9: Comparison of forecasting result of baseline models and the proposed scheme in Korea

	RMSE					MAE				
	I	*I* + *GT*	*I* + *T*	*I* + *T* + *GT*	Proposed scheme	*I*	*I* + *GT*	*I* + *T*	*I* + *T* + *GT*	Proposed scheme
DT	53.66	43.41	51.62	52.26	**31.96**	24.61	21.71	28.26	28.85	**16.42**
LR	49.13	39.52	43.62	44.94	**39.57**	39.18	30.04	32.90	34.51	**28.58**
RF	42.56	39.25	40.06	39.45	**28.90**	24.69	22.31	23.31	22.86	**16.18**
GBM	39.68	41.61	42.13	41.82	**25.88**	24.65	23.58	23.89	24.21	**15.50**
LGBM	41.71	43.17	41.43	41.38	**27.43**	23.57	25.06	23.89	24.01	**15.28**
XGB	42.42	42.36	41.72	41.99	**27.77**	23.46	22.74	23.21	23.91	**16.24**

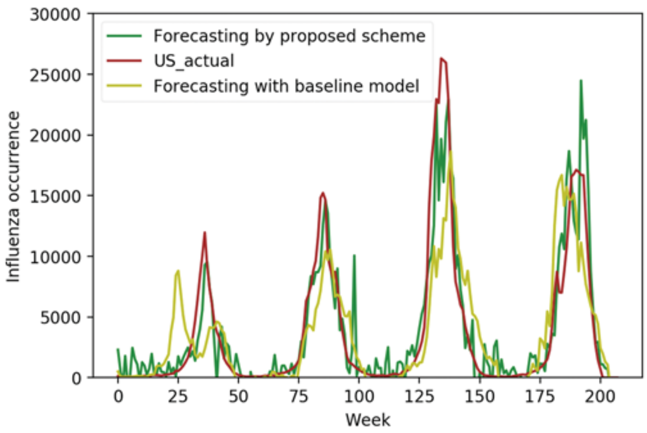


Figure 4: Forecasting in the US using domestic and selected foreign data *vs*. actual data

We compared the actual occurrence data with predicted occurrences using initial input variables and selected surveillance data as input variables, respectively, to investigate the effectiveness of the proposed scheme more closely. Figs. 4–6 display the comparison results for the US, China, and Korea, respectively. For the US, the forecasting results using the initial input variables showed a slight difference in the first peak time prediction, whereas the proposed scheme gave an accurate prediction for the first peak time. It also showed accurate predictions for the remaining peak times. For China, domestic surveillance data were not sufficient to forecast all peaks accurately. However, the proposed scheme showed accurate forecasting performance. In Korea, forecasting using the proposed scheme matched most of the peak times.

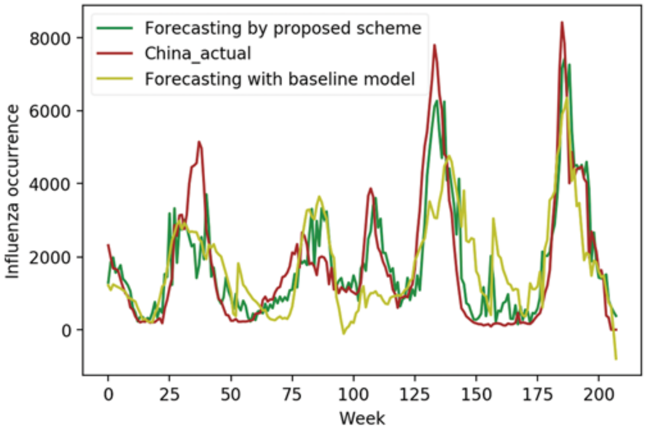


Figure 5: Forecasting in China using domestic and selected foreign data *vs*. actual data

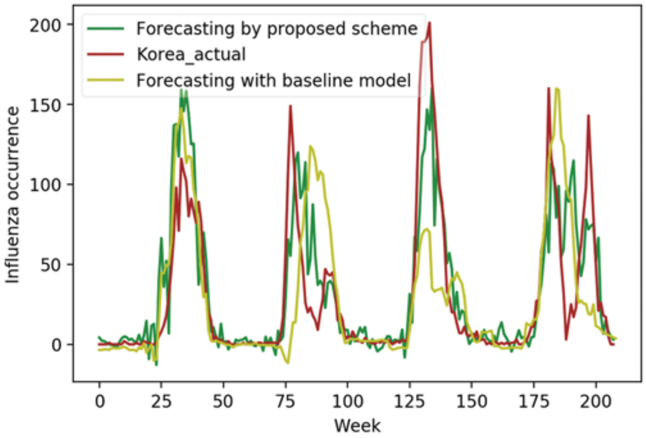


Figure 6: Forecasting in Korea using domestic and selected foreign data *vs*. actual data

6 Conclusion

In this paper, we proposed a two-stage data selection scheme for foreign surveillance data to improve the performance of long-item influenza forecasting. We evaluated each foreign surveillance data using a ML-based model constructed for the target country based on domestic surveillance data in the first stage. In the second stage, we evaluated each surveillance data unit in the order of accuracy and incrementally added it into the current model if it improved forecasting performance. We constructed diverse ML-based forecasting models for three countries using the selected data as input variables to evaluate the effect of the proposed data selection scheme. We performed extensive experiments to determine the effects of different combinations of data selection and forecasting models. The experimental results demonstrated that our data selection scheme was remarkably effective in constructing an influenza forecasting model for the target country. Furthermore, the input variable set configured by the proposed scheme stably enhanced the forecasting accuracy compared to the input variable sets using traditional surveillance data of the target country or GT data popularly used in influenza forecasting.

However, the limitation of our data selection scheme is that it evaluates the suitability of input variables through the validation set. If the validation set has many different patterns from the testing set, the input variables configured by the proposed scheme may not improve the forecasting performance. In future work, we will design an influenza forecasting scheme that does not need the validation set for input variable configuration and is more applicable for the real world than this scheme.

Funding Statement: This research was supported by a government-wide R&D fund project for infectious disease research (GFID), Republic of Korea (Grant Number: HG19C0682).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] World Health Organization, “Influenza fact sheet,” 2018. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal).

[2] K. L. Nichol, J. D. Nordin, D. B. Nelson, J. P. Mullooly and E. Hak, “Effectiveness of influenza vaccine in the community-dwelling elderly,” *New England Journal of Medicine*, vol. 357, no. 14, pp. 1373–1381, 2007.

[3] J. K. Agor and O. Y. Özaltın, “Models for predicting the evolution of influenza to inform vaccine strain selection,” *Human Vaccines & Immunotherapeutics*, vol. 14, no. 3, pp. 678–683, 2018.

[4] M. Biggerstaff, M. Johansson, D. Alper, L. C. Brooks, P. Chakraborty *et al.,* “Results from the second year of a collaborative effort to forecast influenza seasons in the United States,” *Epidemics*, vol. 24, no. 21, pp. 26–33, 2018.

[5] A. Signorini, A. M. Segre and P. M. Polgreen, “The use of Twitter to track levels of disease activity and public concern in the U.S. during the Influenza A H1N1 pandemic,” *PLoS One*, vol. 6, no. 5, pp. 1–10, 2011.

[6] S. C. Lemon, J. Roy, M. A. Clark, P. D. Friedmann and W. Rakowski, “Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression,” *Annals of Behavioral Medicine*, vol. 26, no. 3, pp. 172–181, 2003.

[7] J. Moon, S. Jung, H. Kim and E. Hwang, “Daily occurrence prediction of regional infectious diseases using random forest,” in *The Korean Institute of Information Scientists and Engineers*, Pyeongchang, Korea, pp. 335–337, 2019.

[8] S. Park, J. Moon, S. Jung and E. Hwang, “Explainable influenza forecasting scheme using feature selection and SHAP,” in *The 6th Int. Conf. on Next Generation Computing*, Busan, Korea, pp. 289–292, 2020.

[9] H. Y. Cheng, Y. C. Wu, M. H. Lin, Y. L. Liu, Y. Y. Tsai *et al.,* “Applying machine learning models with an ensemble approach for accurate real-time influenza forecasting in Taiwan: Development and validation study,” *Journal of Medical Internet Research*, vol. 22, no. 8, pp. e15394, 2020.

[10] 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, pp. 890–909, 2020.

[11] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie *et al.,* “Combining search, social media, and traditional data sources to improve influenza surveillance,” *PLoS Computational Biology*, vol. 11, no. 10, pp. e1004513, 2015.

[12] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski *et al.,* “Detecting influenza epidemics using search engine query data,” *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[13] J. Tikka and J. Hollmén, “Sequential input selection algorithm for long-term prediction of time series,” *Neurocomputing*, vol. 71, no. 13–15, pp. 2604–2615, 2008.

[14] F. Jiménez, J. Palma, G. Sánchez, D. Marín, M. D. F. Palacios *et al.,* "Feature selection based multivariate time series forecasting: An application to antibiotic resistance outbreaks prediction," *Artificial Intelligence in Medicine*, vol. 104, no. 3, pp. e101818, 2020.

[15] N. Zhang, H. Huang, M. Duarte and J. J. Zhang, "Dynamic population flow based risk analysis of infectious disease propagation in a metropolis," *Environment International*, vol. 94, no. 39–42, pp. 369–379, 2016.

[16] Y. Wu, Y. Yang, H. Nishiura and M. Saitoh, "Deep learning for epidemiological predictions," in *The 41st Int. ACM SIGIR Conf. on Research & Development in Information Retrieval*, MI, USA, pp. 1085–1088, 2018.

[17] K. M. Choi, H. L. Yu and M. L. Wilson, "Spatiotemporal statistical analysis of influenza mortality risk in the State of California during the period 1997–2001," *Stochastic Environmental Research and Risk Assessment*, vol. 22, no. 1, pp. 15–25, 2008.

[18] K. Choi and S. B. Thacker, "An evaluation of influenza mortality surveillance, 1962–1979. I. Time series forecasts of expected pneumonia and influenza deaths," *American Journal of Epidemiology*, vol. 113, no. 3, pp. 215–226, 1981.

[19] S. R. Venna, A. Tavanaei, R. N. Gottumukkala, V. V. Raghavan, A. S. Maida *et al.,* "A novel data-driven model for real-time influenza forecasting," *IEEE Access*, vol. 7, pp. 7691–7701, 2018.

[20] S. B. Choi, J. Kim and I. Ahn, "Forecasting type-specific seasonal influenza after 26 weeks in the United States using influenza activities in other countries," *PLoS One*, vol. 14, no. 11, pp. e0220423, 2019.

[21] World Health Organization, "Flumart outputs," 2021. [Online]. Available: http://apps.who.int/flumart/Default?ReportNo=12.

[22] S. Caini, W. J. Alonso, C. E. Séblain, F. Schellevis and J. Paget, "The spatiotemporal characteristics of influenza A and B in the WHO European Region: Can one define influenza transmission zones in Europe?," *Eurosurveillance*, vol. 22, no. 35, pp. 1–11, 2017.

[23] J. Moon, S. Park, S. Rho and E. Hwang, "A comparative analysis of artificial neural network architectures for building energy consumption forecasting," *International Journal of Distributed Sensor Networks*, vol. 15, no. 9, pp. 1–19, 2019.

[24] J. Moon, J. Kim, P. Kang and E. Hwang, "Solving the cold-start problem in short-tern forecasting using tree-based methods," *Energies*, vol. 13, no. 4, pp. 886–922, 2020.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion *et al.,* "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] J. Moon, J. Park, S. Han and E. Hwang, "Power consumption forecasting scheme for educational institutions based on analysis of similar time series data," *Journal of KIISE*, vol. 44, no. 9, pp. 954–965, 2017.

[27] S. Volkova, E. Ayton, K. Porterfield and C. D. Corley, "Forecasting influenza-like illness dynamics for military populations using neural networks and social media," *PLoS One*, vol. 12, no. 12, pp. e0188941, 2017.

[28] C. Comito, A. Forestiero and C. Pizzuti, "Improving influenza forecasting with web-based social data," in *2018 IEEE/ACM Int. Conf. on Advances in Social Networks Analysis and Mining*, Barcelona, Spain, pp. 963–970, 2018.

Appendix A. Comparison of Forecasting Results Using PCC Values

We also measured pearson correlation coefficient (PCC) in the same way as the comparison experiment in Section 5.2. Tabs. A1–A3 list the obtained PCC values for the US, China, and Korea, respectively. In the tables, the input variables that demonstrated the best performance in each forecasting model are marked in bold font. The proposed scheme achieved the best in all cases.

Table A1: Comparison of PCC value of baseline models and the proposed scheme in the US

	I	$I+GT$	$I+T$	$I+T+GT$	Proposed scheme
DT	0.64	0.63	0.52	0.57	**0.79**
LR	0.69	0.70	0.47	0.47	**0.91**
RF	0.65	0.66	0.60	0.63	**0.89**
GBM	0.65	0.64	0.50	0.65	**0.92**
LGBM	0.65	0.66	0.50	0.63	**0.88**
XGB	0.64	0.67	0.53	0.63	**0.90**

Table A2: Comparison of PCC value of baseline models and the proposed scheme in the China

	I	$I+GT$	$I+T$	$I+T+GT$	Proposed scheme
DT	0.50	0.50	0.23	0.20	**0.77**
LR	0.59	0.59	0.45	0.41	**0.83**
RF	0.53	0.52	0.33	0.31	**0.85**
GBM	0.56	0.52	0.31	0.28	**0.89**
LGBM	0.52	0.50	0.39	0.40	**0.86**
XGB	0.50	0.51	0.31	0.27	**0.87**

Table A3: Comparison of forecasting result of baseline models and the proposed scheme in the US

	I	$I+GT$	$I+T$	$I+T+GT$	Proposed scheme
DT	0.63	0.59	0.60	0.58	**0.76**
LR	0.63	0.63	0.60	0.61	**0.71**
RF	0.59	0.62	0.59	0.61	**0.81**
GBM	0.62	0.59	0.60	0.58	**0.81**
LGBM	0.57	0.52	0.62	0.61	**0.80**
XGB	0.58	0.59	0.63	0.63	**0.85**