Original Article

DOI: 10.5582/bst.2017.01257

A comparative study on predicting influenza outbreaks

Jie Zhang*, **Kazumitsu Nawata**

Graduate School of Engineering, University of Tokyo, Tokyo, Japan.

Summary **Worldwide, influenza is estimated to result in approximately 3 to 5 million annual cases of severe illness and approximately 250,000 to 500,000 deaths. We need an accurate time-series model to predict the number of influenza patients. Although time-series models with different time lags as feature spaces could lead to varied accuracy, past studies simply adopted a time lag in their models without comparing or selecting an appropriate number of time lags. We investigated the performance of adopting 6 different time lags in 6 different models: Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting (GB), Artificial Neural Network (ANN), and Long Short Term Memory (LSTM) with hyperparameter adjustment. To the best of our knowledge, this is the first time that LSTM has been used to predict influenza outbreaks. As a result, we found that the time lag of 52 weeks led to the lowest Mean Absolute Percentage Error (MAPE) in the ARIMA, ANN and LSTM, while the machine learning models (SVR, RF, GB) achieved the lowest MAPEs with a time lag of 4 weeks. We also found that the MAPEs of the machine learning models were less than ARIMA, and the MAPEs of the deep learning models (ANN, LSTM) were less than those of the machine learning models. In all the models, the LSTM model of 4 layers reached the lowest MAPE of 5.4%, and the LSTM model of 5 layers with regularization reached the lowest root mean squared error (RMSE) of 0.00210.**

Keywords: Time series, Influenza-Like Illness, time lag, Long Short Term Memory (LSTM)

1. Introduction

Influenza, commonly known as flu, is a contagious respiratory illness caused by influenza viruses (*1,2*). Influenza virus spreads through air from coughs or sneezes as well as by touching surfaces contaminated with influenza virus and then touching mouths or eyes (*3,4*). The strong infectivity and annual outbreak of flu are estimated to result in approximately 3 to 5 million annual cases of severe illness and approximately 250,000 to 500,000 deaths worldwide (*1*). The resultant high levels of worker/school absenteeism and productivity losses lead to direct costs of lost productivity and associated medical treatment and indirect costs of preventative measures. In the U.S., flu is responsible for a total cost of over $10 billion per year, and a future flu pandemic is estimated to cost hundreds of billions of dollars in direct and indirect costs (*5*). Clinics and hospitals are overwhelmed during peak illness periods. The impeding of transmission routes, especially via school closures, and influenza immunizations effectively prevent the propagation of the flu (*6-8*). To help governments, hospitals, clinics, pharmaceutical companies, and others prepare for flu outbreaks efficiently and restrict routes of transmission in a timely manner, we need an accurate time-series model to predict influenza outbreaks.

Time-series models can be categorized into 3 types by using different features. The first type of model is an autoregressive model, which uses the numbers of patients in the past as features ("Xs") and forecasts the number of patients in the future as the response (y). Typical examples include the Auto-Regressive Integrated Moving Average (ARIMA) model and the Vector Auto-Regression model (VAR). The second type of model uses other parameters (such as temperature, humidity, *etc.*) instead of past flu data as features for regression models (*e.g.*, linear regression, random forest, *etc.*). The famous

Released online in J-STAGE as advance publication October 24, 2017.

**Address correspondence to:*
Jie Zhang, Department of Technology Management for Innovations, Graduate School of Engineering, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku Tokyo, 113-8656, Japan.
E-mail: jie-zhang@g.ecc.u-tokyo.ac.jp

example is "Google Flu Trends", which used search engine query data (*9*) as features and a linear regression model. The third type of model is a combination of the first and second types. It uses the numbers of flu patients in the past as features (as in the first type) and regression models (as in the second type) (*10*). In this study, we adopted the third model type and tried 6 different models with hyperparameter adjustments, including: Auto-Regressive Integrated Moving Average (ARIMA), Support Vector Regression (SVR), Random Forest (RF), Gradient Boosting (GB), Artificial Neural Network (ANN), and Long Short Term Memory (LSTM). To the best of our knowledge, this is the first time LSTM has been used to predict influenza outbreaks.

Time-series models with different time lags usually result in different levels of accuracy. The selection of time lags can be essential to improve the accuracy of predications. However, past studies simply adopted a time lag for models without comparing or selecting an appropriate number of time lags, which could make the model misunderstand past outbreak patterns. Therefore, in this study, we investigated the performance of 6 different time lags in each of the models we tried: 2 weeks (approximately 0.5 month), 4 weeks (approximately 1 month), 9 weeks (approximately 2 months), 13 weeks (approximately 3 months), 26 weeks (approximately 6 months), and 52 weeks (approximately 12 months). We hoped we would find some clues from our studies for future studies, which leverage machine learning (ML) and deep learning (DL) models for predicting epidemic outbreaks.

2. Methodology

2.1. *Data*

We collected the U.S. flu season data from the "FluView" Portal of the website for the Centers for Disease Control and Prevention (CDC) (*11*). The data are posted "weekly" with "not available" (N/A) values from the 21st week to the 39th week from 1998, 1999, 2000, 2001, and 2002. Therefore, we only used the U.S. Flu Season Data without any N/As, *i.e.* the U.S. Flu Season Data from the 40th week of 2002 to the 30th week of 2017.

To remove any possible variations in populations, we adopted the Influenza-Like Illness (ILI) rates as the response (y) of our models.

$$\text{ILI rate} = \frac{\text{The number of ILI}}{\text{Total number of Illness}}$$

Figure 1(a) illustrates the raw data. The Y-axis represents the weekly ILI rate, and the X-axis represents the time series. The seasonality appears obvious, except in 2009 when swine flu occurred. The swine flu (also called the 2009 flu pandemic) was an influenza pandemic, and the second of two pandemics involving the H1N1 influenza virus (the first was the 1918 flu pandemic), albeit a new variety.

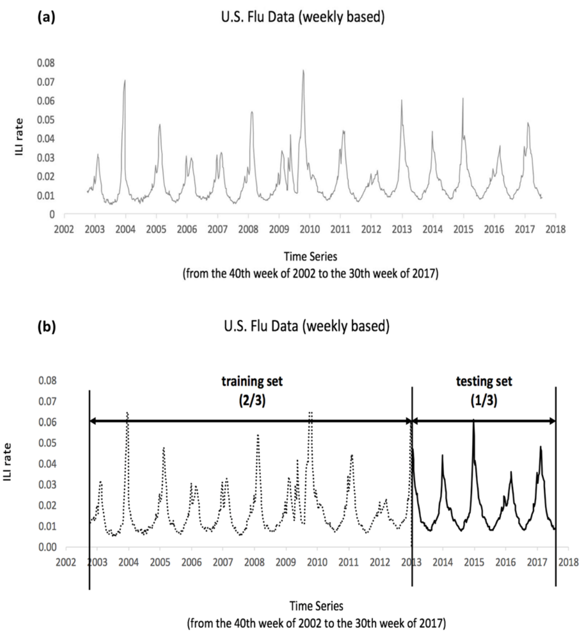

Figure 1. The U.S. flu season data from the 40th week of 2002 to the 30th week of 2017. (a). The Y-axis represents the weekly ILI rate, and the X-axis represents the time series. **(b).** The dashed line is the first 2/3 used for the training set, and the solid line is the last 1/3 used for the testing set. The Y-axis represents the weekly ILI rate, and the X-axis represents the time series.

We split the data into two parts: the first 2/3 was the training set and the last 1/3 was the testing set, as shown in Figure 1(b).

2.2. *Models*

Table 1 illustrates the models, programming languages, libraries, and hyperparameter adjustments we used in this study.

We trained the ARIMA model in R Programming Language (version 3.4.1) and the "forecast" package (version 8.1) (*12*). The function of "auto.arima" in the "forecast" package of R automatically performs a stepwise regression and selects the best hyperparameters based on the Bayesian Inference Criteria (BIC). For SVR, we applied the caret package (Version 6.0-76) in R. For RF and GB, we used Python (Version 3.6.0) and the Scikit-Learn package (Version 0.18.1) with a grid search. For ANN and LSTM, we used Python and the Keras package (Version 2.0.4) based on Tensorflow (Version 1.1.0) and adopted an "early-stopping" algorithm with a "patience" of 100 epochs (for a total of 1000 epochs).

2.3. *Metrics*

We compared different models and different time lags using the Mean Absolute Percentage Error (MAPE) and

Table 1. The models, programming languages, libraries, and hyperparameter adjustments we used in this study

Models	Programming Languages	Programming Libraries	Hyperparameter Adjustment
ARIMA	R (Version 3.4.1)	Forecast (Version 8.1)	● auto.arima
SVR	R (Version 3.4.1)	Caret (Version 6.0-76)	● cross validation (*n* = 3)
RF	Python (Version 3.6.0)	Scikit Learn (Version 0.18.1)	● cross validation (*n* = 3) ● grid search ● n_estimators ● max_features ● max_depth
GB	Python (Version 3.6.0)	Scikit Learn (Version 0.18.1)	● cross validation (*n* = 3) ● grid search ● learning rate ● subsample ● n_estimators ● max_features ● max_depth
ANN	Python (Version 3.6.0)	Keras (Version 2.0.4) Tensorflow (Version 1.1.0)	● different layers (up to 5 layers) ● with/without dropout, ● with/without regularization ● with/without batch normalization
LSTM	Python (Version 3.6.0)	Keras (Version 2.0.4) Tensorflow (Version 1.1.0)	● different layers (up to 10 layers) ● with/without dropout, ● with/without regularization ● with/without batch normalization

Root Mean Squared Error (RMSE) as Key Performance Indicators (KPIs).

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{F_t - A_t}{A_t}\right| * 100\%$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(F_t - A_t)^2}$$

where A_t is the actual value and F_t is the forecasted value.

Figure 2 illustrates the histogram of the weekly ILI rates of the U.S. flu data. In our opinion, comparing models using MAPEs reflects the difference based on the median, and comparing models using RMSE is based on means. In this study, the histogram is right skewed. Furthermore, we performed the Kolmogorov-Smirnov Test to examine the data distribution, and the *p*-value is < 0.001. Therefore, we concluded that the distribution is a non-normal distribution, and we therefore regard the MAPE as the first KPI and the RMSE as an assistant KPI in this study.

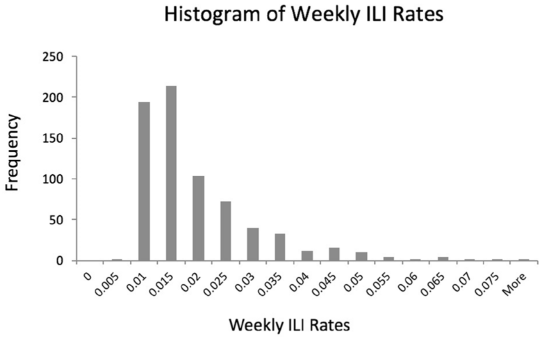

Figure 2. The histogram of weekly ILI Rates from U.S. flu data. The histogram is right skewed. Furthermore, we performed the Kolmogorov-Smirnov Test to examine the data distribution. The *p*-value is < 0.001, and we therefore concluded that the distribution is a non-normal distribution.

2.4. *Feature space*

2.4.1. *Time lags*

Influenza seasonality is an annually recurring time period characterized by the prevalence of outbreaks of influenza. Therefore, in this study, we reviewed a maximum of 52 weeks (approximately 1 year). We tried using time lags of 2 weeks (around half a month), 4 weeks (approximately 1 month), 9 weeks (approximately 2 months), 13 weeks (approximately 3 months), 26 weeks (around half a year), and 52 weeks(approximately 1 year)for model training and compared the results.

2.4.2. *First-order fifferences*

Some previous studies found that first-order differences helped improve the results of the prediction models for influenza data (*13*). We also included the first-order differences as a part of the feature spaces.

In this study, we reviewed a maximum of 52 weeks. In the case of the time lag of 52 weeks, we used (I) the ILI rate of the current week, (II) the ILI rates of the past 52 weeks, and (III) the 52 first-order differences. In total,

Table 2. In the case of the time lag of 52 weeks, in all, we have 105 predictors (I + II + III) for use as feature spaces: (I) the ILI rate of the current week, (II) the ILI rates of the past 52 weeks, and (III) the 52 first-order differences. We dropped the first 52 rows (the first 52 weeks), since we are unable to calculate the first-order differences for the first 52 rows (the first 52 weeks). We also dropped the last row (the last week), since the source data ends at the 30th week of 2017, and we have no data for the 31st week in 2017 and therefore cannot calculate the MAPE or RMSE

			Feature space, *i.e.* Xs										
Time series			Response, *i.e.*, "y"	Current data & historic data					First-order difference				
Year	Week	Monday of week (Year/Month/Date)	ILI rates (1 week previous)	ILI rates (this week)	ILI rates (1 week ago)	ILI rates (2 weeks ago)	…	ILI rates (52 weeks ago)	(This week) - (1 week ago)	(This week) - (2 weeks ago)	…	(This week) - (52 weeks ago)
2002	40	2002/09/30	0.0122	0.0117	N/A	N/A	…	N/A	N/A	N/A	…	N/A
2002	41	2002/10/07	0.0113	0.0122	0.0117	N/A	…	N/A	0.000	N/A	…	N/A
2002	42	2002/10/14	0.0125	0.0113	0.0122	0.0117	…	N/A	- 0.001	0.000	…	N/A
…	…	…	…	…	…	…	…	…	…	…	…	…
2003	39	2003/09/22	0.0096	0.0075	0.0064	0.0064	…	N/A	0.001	0.001	…	N/A
2003	40	2003/09/29	0.0104	0.0096	0.0075	0.0064	…	0.0117	0.002	0.003	…	- 0.0021
2003	41	2003/10/06	0.0105	0.0104	0.0096	0.0075	…	0.0122	0.001	0.003	…	- 0.0017
…	…	…	…	…	…	…	…	…	…	…	…	…
2017	29	2017/07/17	0.0088	0.0085	0.0084	0.0104	…	0.0084	0.000	- 0.002	…	0.0000
2017	30	2017/07/24	N/A	0.0088	0.0085	0.0084	…	0.0083	0.000	0.000	…	0.0004

we have 105 predictors (I + II + III) for use as feature spaces. Since we are unable to calculate the first-order differences for the first 52 rows (the first 52 weeks), we dropped these data. Table 2 illustrates the pretreatment of the source data.

In the case of the time lag of 52 weeks, we had 105 predictors and had to drop the first 52 rows (the first 52 weeks), since we are unable to calculate the first-order differences for the first 52 rows (the first 52 weeks). Similarly, in the case of the time lag of 2, 4, 9, 13, 26 weeks, we had 5, 9, 19, 27, 53 predictors and had to dropped the first 2, 4, 9, 13, 26 rows (the first 2, 4, 9, 13, 26 weeks) since we are unable to calculate the first-order differences for the first 2, 4, 9, 13, 26 rows (the first 2, 4, 9, 13, 26 weeks). As a result, the models with fewer time lags could have more training data, which was considered unfair when we compared the predicting accuracy of the models since models (especially DL models) with more training data usually brought better accuracy (*14*). To fairly compare the predicting accuracy of adopting different time lags, we uniformly removed the first 52 rows (the first 52 weeks) from the training set of all the models.

3. Results

3.1. *ARIMA, SVR, RF, GB, and ANN*

Table 3(a) and Table 3(b) present the MAPE and RMSE of ARIMA, SVR, RF, GB, ANN. When increasing the time lags in the ARIMA models, we found obvious decreases in the MAPE and RMSE. We achieved the lowest MAPE (8.36%) and the lowest RMSE (0.00364) when using the time lag of 52 weeks, and we found a similar phenomenon when performing the ANN models, where we achieved the lowest MAPE (5.79%) and the lowest RMSE (0.002411) when using the time lag of 52 weeks. Regarding the ML models (*i.e.*, SVR, RF, and GB), all of them reached their lowest MAPE (6.75%, 6.75%, 6.58%, respectively) when we used the time lag of 4 weeks. The SVR reached the lowest RMSE (0.002271) when we used the time lag of 52 weeks. The RF reached the lowest RMSE (0.002417) when we used the time lag of 2 weeks. The GB reached the lowest RMSE (0.002351) when we used the time lag of 4 weeks. The cells with the gray background in Table 3(a) are the lowest MAPEs in the ARIMA, SVR, RF, GB, and ANN models, while the cells with the gray background in Table 3(b) are the lowest RMSEs in the ARIMA, SVR, RF, GB, and ANN models.

Figure 3(a), 3(b), 3(c), 3(d), and 3(e) compares the actual and the predicted outcomes when we used the time lag of 52 weeks in ARIMA, the time lag of 4 weeks in SVR, the time lag of 4 weeks in RF, the time lag of 4 weeks in GB, and the time lag of 52 weeks in ANN. All the time lags we adopted in Figure 3 achieved the lowest MAPE in the respective model.

Table 3(a). The MAPEs of the testing set for ARIMA, SVR, RF, GB, and ANN. When performing the ARIMA models, we achieved the lowest MAPE (8.36%) when using the time lag of 52 weeks. We achieved the lowest MAPE (6.75%, 6.75%, and 6.58%) of all the ML models when we used the time lag of 4 weeks. When performing the ANN models, we achieved the lowest MAPE (5.79%) when using the time lag of 52 weeks. The cells with the gray background are the lowest MAPEs in the ARIMA, SVR, RF, GB, and ANN models

	Time lags (Weeks)	2	4	9	13	26	52
Models	ARIMA MAPE (%)	13.46	11.90	9.14	8.72	8.58	8.36
	SVR MAPE (%)	6.76	6.75	6.99	6.90	6.85	6.86
	RF MAPE (%)	7.36	6.75	6.95	7.82	7.07	6.92
	GB MAPE (%)	6.96	6.58	7.24	6.92	7.67	7.02
	ANN MAPE (%)	6.65	6.50	6.32	6.34	6.16	5.79

Table 3(b). The RMSEs of the testing set for ARIMA, SVR, RF, GB, ANN. The ARIMA reached the lowest RMSE (0.003285) when we used the time lag of 13 weeks. The SVR reached the lowest RMSE (0.002271) when we used the time lag of 52 weeks. The RF reached the lowest RMSE (0.002417) when we used the time lag of 2 weeks. The GB reached the lowest RMSE (0.002351) when we used the time lag of 4 weeks. The ANN reached the lowest RMSE (0.002411) when we used the time lag of 4 weeks. The cells with the gray background are the lowest RMSEs in the ARIMA, SVR, RF, GB, and ANN models

	Time lags (Weeks)	2	4	9	13	26	52
Models	ARIMA RMSE	0.004437	0.004099	0.003665	0.003285	0.003428	0.003642
	SVR RMSE	0.002558	0.002554	0.002536	0.002526	0.002514	0.002271
	RF RMSE	0.002417	0.002522	0.002582	0.002692	0.002554	0.002591
	GB RMSE	0.002378	0.002351	0.002652	0.002586	0.002732	0.002511
	ANN RMSE	0.002590	0.002552	0.002574	0.002549	0.002516	0.002411

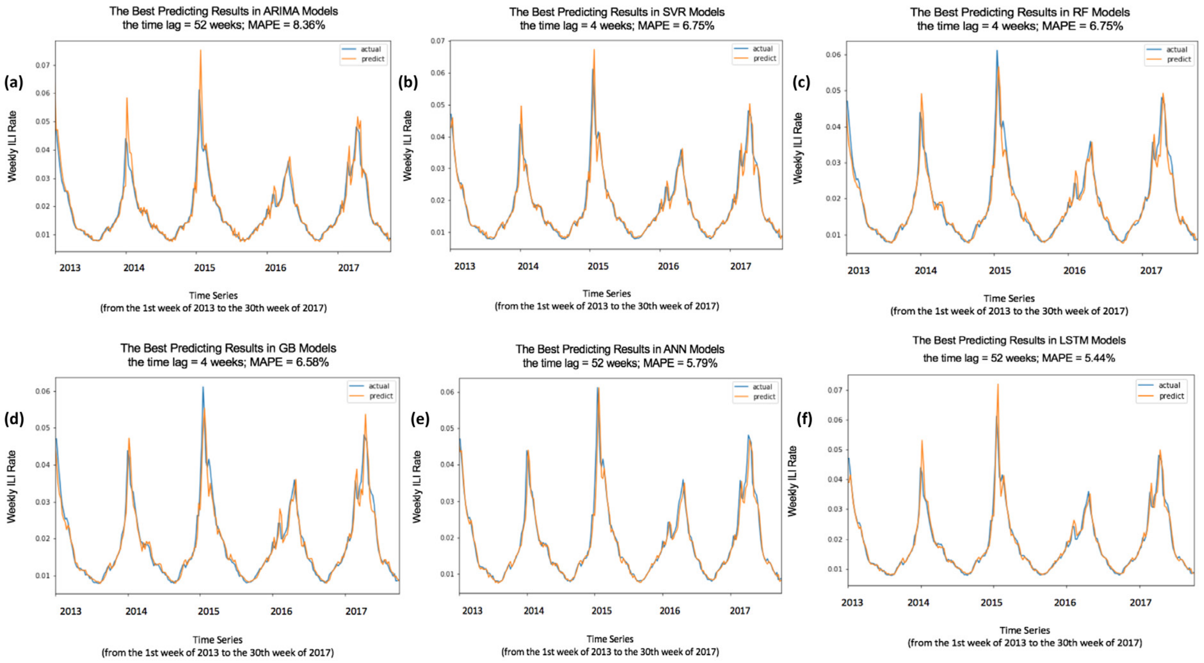

Figure 3. The actual and the predicted outcomes when we use the time lag of 52 weeks in ARIMA, the time lag of 4 weeks in SVR, the time lag of 4 weeks in RF, the time lag of 4 weeks in GB, the time lag of 52 weeks in ANN, and the time lag of 52 weeks in the LSTM model (4 layers). All the time lags we adopted achieved the lowest MAPE in ARIMA, SVR, RF, GB, and ANN, respectively. The X-axis represents the time series (from the 1st week of 2013 to the 30th week of 2017) of the testing set. The Y-axis represents the weekly ILI rates.

3.2. *LSTM results*

We performed a variety of LSTM models: 3 layers, 4 layers, 4 layers with dropout, 4 layers with regularization, 5 layers, 5 layers with regularization, 6 layers with regularization, and 10 layers with regularization. Table 4(a) and Table 4(b) present the MAPEs and RMSEs of all the LSTM models with different hyperparameters, as previously mentioned. All the LSTM models achieved the lowest MAPEs (6.71%, 5.44%, 6.27%, 5.45%, 6.28%, 5.53%, 5.46%, and 5.72%) when we adopted a time lag of 52 weeks. The cells with the gray background are

Table 4(a). The MAPEs of all the LSTM models: 3 layers, 4 layers, 4 layers with dropout, 4 layers with regularization, 5 layers, 5 layers with regularization, 6 layers with regularization, and 10 layers with regularization. All the LSTM models achieved the lowest MAPEs (6.71%, 5.44%, 6.27%, 5.45%, 6.28%, 5.53%, 5.46%, and 5.72%) when we adopted a time lag of 52 weeks. The cells with the gray background are the lowest MAPEs in the different LSTM models

	Time lags	2	4	9	13	26	52	Mean MAPE of Different LSTM Structures
LSTM Structures	3 layers MAPE (%)	6.80	7.00	7.00	6.87	6.93	6.71	6.89
	4 layers MAPE (%)	6.69	6.42	6.28	6.17	6.06	5.44	6.18
	4 layers with dropout MAPE (%)	7.62	7.17	7.26	7.18	6.56	6.27	7.01
	4 layers with regularization MAPE (%)	6.74	6.32	6.22	6.09	6.07	5.45	6.15
	5 layers MAPE (%)	6.85	6.61	7.20	6.64	6.53	6.28	6.69
	5 layers with regularization MAPE (%)	6.56	6.38	6.11	6.01	5.91	5.53	6.08
	6 layers with regularization MAPE (%)	6.61	6.52	6.20	6.12	5.91	5.46	6.14
	10 layers with regularization MAPE (%)	6.46	6.42	5.98	5.90	5.75	5.72	6.04
	Mean MAPE of Different Time Lags MAPE (%)	6.79	6.61	6.53	6.37	6.22	5.86	

Table 4(b). The RMSEs of all the LSTM models: 3 layers, 4 layers, 4 layers with dropout, 4 layers with regularization, 5 layers, 5 layers with regularization, 6 layers with regularization, and 10 layers with regularization. All the LSTM models achieved the lowest MAPEs when we adopted a time lag of 52 weeks, except for the 4 layers with regularization and 5 layers with regularization, which reached their lowest MAPEs when we used a time lag of 13 weeks. The cells with the gray background are the lowest RMSEs (0.002102, 0.002431, 0.002352, 0.002499, 0.002099, 0.002439, 0.002438, 0.002274) in the different LSTM models

	Time lags	2	4	9	13	26	52	Mean MAPE of Different LSTM Structures
LSTM Structures	3 layers RMSE	0.002534	0.002535	0.002497	0.002490	0.002411	0.002102	0.002428
	4 layers RMSE	0.002611	0.002581	0.002570	0.002572	0.002516	0.002431	0.002547
	4 layers with dropout RMSE	0.002528	0.002517	0.002518	0.002499	0.002462	0.002352	0.002479
	4 layers with regularization RMSE	0.002621	0.002563	0.002505	0.002499	0.002632	0.002559	0.002563
	5 layers RMSE	0.002504	0.002486	0.002458	0.002434	0.002408	0.002099	0.002398
	5 layers with regularization RMSE	0.002663	0.002549	0.002556	0.002439	0.002590	0.002566	0.002560
	6 layers with regularization RMSE	0.002593	0.002561	0.002503	0.002479	0.002521	0.002438	0.002516
	10 layers with regularization RMSE	0.002460	0.002390	0.002312	0.002325	0.002296	0.002274	0.002343
	Mean MAPE of Different Time Lags	0.002564	0.002523	0.002490	0.002467	0.002480	0.002353	

the lowest MAPEs in the different LSTM models. All the LSTM models achieved the lowest MAPEs when we adopted a time lag of 52 weeks, except the 4 layers with regularization and the 5 layers with regularization, which reached their lowest MAPEs when we used a time lag of 13 weeks. The cells with the gray background are the lowest RMSEs (0.002102, 0.002431, 0.002352, 0.002499, 0.002099, 0.002439, 0.002438, and 0.002274) in the different LSTM models.

Figure 3(f) compares the actual and the predicted outcomes when we used the time lag of 52 weeks in the LSTM model (4 layers). The X-axis represents the time series (from the 1st week of 2013 to the 30th week of 2017) of the testing set. The Y-axis represents the weekly ILI rates

4. Discussion

4.1. *Time lag selection in ARIMA, SVR, RF, GB, and ANN*

The MAPEs of the ARIMA model decreased significantly (from 13.46% to 8.36%) when we increased the time lags from 2 weeks to 52 weeks. (Figure 4) The probable explanation for this phenomenon is that ARIMA is an autoregressive model focusing on seasonality. The closer the feature spaces to a complete seasonality, the lower the MAPE will be. In other words, when training ARIMAs for time-series prediction, at the least, we need a complete duration. Similar to those of the ARIMA models, the MAPEs of the ANN models also decreased (from 13.46% to 8.36%) when we increased the time lag from 2 weeks to 52 weeks. (Figure 4)

Regarding the ML models (SVR, RF, and GB), the MAPEs were always approximately 7%, with almost no changes as we increased the time lags (Figure 4), likely because the ML models usually cannot learn the seasonality but can learn the trend of the data by inputting the first-order differences into the training and testing.

4.2. *With and without regularization*

We calculated the standard deviations of the MAPEs of the LSTM models of 3, 4, and 5 layers without regularization and of 4, 5, 6, and 10 layers with regularization when using the time lags of 2, 4, 9,13, 26, and 52 weeks. (Table 5) We found the standard deviations of the MAPEs of the LSTM models with regularization were less than those of the LSTM models without regularization when we used almost all the time lags except the time lag of 2 weeks. (Figure 5a) The

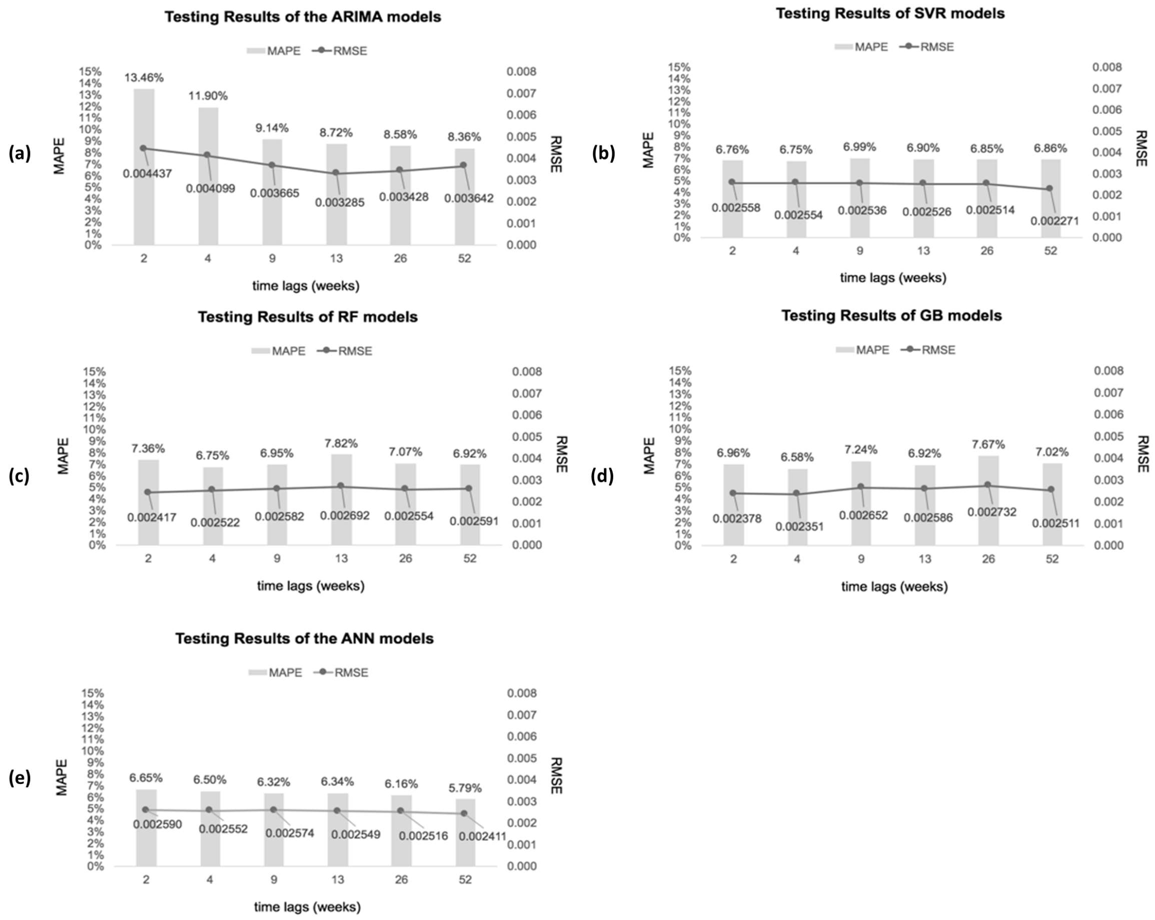

Figure 4. The MAPEs and RMSEs of the ARIMA, SVR, RF, GB, and ANN models with the different time lags as the feature spaces.

Table 5. The standard deviations of the MAPEs of the LSTM models of 3, 4, and 5 layers without regularization and of 4, 5, 6, and 10 layers with regularization

	Time lags	2	4	9	13	26	52	Mean MAPE of Different LSTM Structures
LSTM Structures	3 layers MAPE (%)	6.80	7.00	7.00	6.87	6.93	6.71	6.89
	4 layers MAPE (%)	6.69	6.42	6.28	6.17	6.06	5.44	6.18
	5 layers MAPE (%)	6.85	6.61	7.20	6.64	6.53	6.28	6.69
	Standard Deviation of MAPEs of LSTM without Regularization of 3, 4, 5 Layers (%)	0.08	0.30	0.49	0.36	0.44	0.64	0.37
LSTM Structures	4 layers with regularization MAPE (%)	6.74	6.32	6.22	6.09	6.07	5.45	6.15
	5 layers with regularization MAPE (%)	6.56	6.38	6.11	6.01	5.91	5.53	6.08
	6 layers with regularization MAPE (%)	6.61	6.52	6.20	6.12	5.91	5.46	6.14
	10 layers with regularization MAPE (%)	6.46	6.42	5.98	5.90	5.75	5.72	6.04
	Standard Deviation of MAPEs of LSTM with Regularization of 4, 5, 6, 10 Layers MAPE (%)	0.12	0.08	0.11	0.10	0.13	0.12	0.05

probable explanation for this finding is that regularization made the models more robust, and the robust models made the results (*i.e.*, the MAPEs in this study) relatively stable.

Although we achieved the lowest MAPE (5.44%) when we used the 4-layer LSTM model without regularization, the gap between the MAPEs of the 4-layer LSTM model without and with regularization is very limited (5.45% - 5.44% = 0.01%). Considering that unstable models may lead to poor accuracy if we changed the testing data, we recommend the use of the model with regularization for U.S. flu prediction.

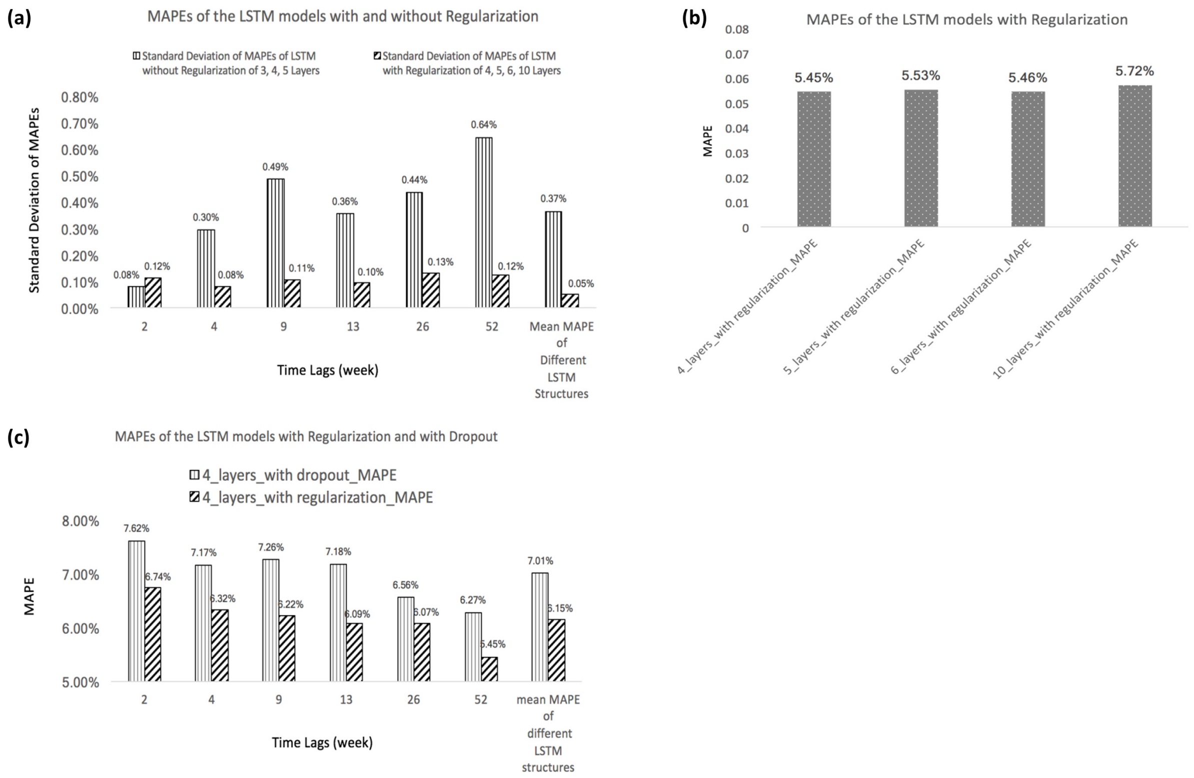

Figure 5. Comparison of LSTM Predicting Accuracy. (a). MAPEs of the LSTM models with and without regularization. The standard deviations of the MAPEs of the LSTM models with regularization were less than those of the LSTM models without regularization when we used almost all the time lags except the time lag of 2 weeks. **(b).** MAPEs of the LSTM models with different number of layers with regularization. **(c).** MAPEs of the LSTM models with Regularization and with Dropout.

4.3. *LSTM structure (Layers)*

After comparing the MAPEs of the LSTM models with and without regularization, we compared the different layers for the LSTM with regularization. (Figure 5b) We found extra layers (more than 4 layers) contributed little to improve the predicting accuracy. In other words, the LSTM models of 4 to 5 layers are considered sufficient for U.S. flu prediction.

4.4. *Regularization and dropout*

In addition to regularization, dropout can also usually help prevent overfitting and make the model more robust. We found that the MAPE of the LSTM models with regularization is obviously lower than those with dropout. "Dropout" randomly drops neurons, while "Regularization" selectively drops neurons. Although both suppress the number of neurons, in this study, the selective dropping performed much better than the random dropping.

4.5. *Feature spaces*

Comparing results, we found that the MAPE of ARIMA > MAPEs of SVR, RF, and GBM > MAPEs of ANN and LSTM. Although the different models have totally different algorithms, the increasing feature space and the increasing model parameters are considered other factors that impact the models' accuracy. In ARIMA, we only have very limited feature spaces. The number of features is equal to the lag times, *i.e.*, 3, 5, 10, 14, 27, or 53. In the ML models (SVR, RF, and GB), we added the first-order differences, including more information in the models, and we clearly found that the ML models with 105 features resulted in lower MAPEs. In DL models, we used 255 neurons in every LSTM layer, which included more parameters in the models. To achieve better epidemic predictions, more neurons can perform more linear and non-linear combinations of past data, create more "artificial" feature spaces, and provide better results.

4.6. *More time lags*

Why not adopt more time lags such as 104 weeks (around 2 years) or more? For one things, the models with the time lag of 52 weeks (around 1 year) have brought an accuracy of about 95% (*i.e.* 1 - MAPE). For another thing, we have to drop more training data (the first 104 rows) if we adopt a time lag of 104 week or more. Although a longer time lag might help improve the accuracy, the less training data would also setback the accuracy. We suppose whether the accuracy would be better or worse depends on different data. However, a time lag including a complete periodicity is recommended for ARIMA, ANN, and LSTM.

5. Conclusion

In this study, we performed ARIMA, SVM, RF, GB, ANN, and LSTM models with different time lags (2, 4, 9, 13, 26, 52 weeks) to forecast the weekly ILI rate of U.S. flu data. We found the ARIMA, ANN and LSTM models with a lag time of 52 weeks (*i.e.*, the periodicity of the flu season) resulted in the best MAPEs, while SVR, RF, and GB performed with almost no changes when we used the time lags. We also found the MAPEs of the ML models (SVR, RF, and GB) with the first differences were lower than those of ARIMA, and the MAPEs of the deep learning models (ANN and LSTM) with multiple layers were lower than those of the ML models (SVR, RF, and GB). To the best of our knowledge, this is the first time LSTM has been used to predict influenza outbreaks. In all the models (with different model types, different hyperparameters, and different time lags), the LSTM model of 4 layers reached the lowest MAPE of 5.4%, and the LSTM model of 5 layers with regularization reached the lowest RMSE of 0.00210. Additionally, the LSTM models with 4 ~ 6 layers with regularization resulted in very low MAPEs of approximately 5.4 ~ 5.5%, and more than 6 layers contributed little to improving the predictive accuracy.

References

1. World Health Organization. Influenza (Seasonal) fact sheet. *http://www.who.int/mediacentre/factsheets/fs211/en/* (accessed August 23, 2017).
2. Longo D. Harrison's principles of internal medicine. (18th ed., Fauci A, Kasper D, Hauser S, Jameson J, Loscalzo J, eds.). McGraw-Hill, New York, 2012; 187:1442.
3. Centers for disease control and prevention. Key facts about influenza (Flu). *https://www.cdc.gov/flu/keyfacts.htm* (accessed Aug 23, 2017).
4. Brankston B, Gitterman L, Hirji Z, Lemieux C, Gardam M. Transmission of influenza A in human beings. Lancet Infect Dis. 2007; 4:257-265.
5. Statement from president george W. Bush on Influenza". *http://www.presidency.ucsb.edu/ws/index.php?pid=65259* (accessed August 23, 2017).
6. Earn DJ, He D, Loeb MB, Fonseca K, Lee BE, Dushoff J. Effects of school closure on incidence of pandemic influenza in Alberta, Canada. Ann Intern Med. 2012; 156:173-181
7. Cauchemez S, Valleron AJ, Boëlle PY, Flahault A, Ferguson NM. Estimating the impact of school closure on influenza transmission from Sentinel data. Nature. 2008; 452:750-754.
8. Heymann A, Hoch I, Valinsky L, Kokia E, Steinberg DM. School closure may be effective in reducing transmission of respiratory viruses in the community. Epidemiol Infect. 2009; 137:1369-1376.
9. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457:1012-1014
10. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and random forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics. 2014; 15:276.
11. "FluView" portal of centers for disease control and prevention (CDC). *https://www.cdc.gov/flu/weekly/fluviewinteractive.htm* (accessed Jun 7, 2017).
12. Hyndman R. Package 'forecast' Ver 8.1. CRAN. *https://cran.r-project.org/web/packages/forecast/index.html* (accessed Oct 2, 2017).
13. Wu H, Cai Y, Wu Y, Zhong R, Li Q, Zheng J, Lin D, Li Y. Time series analysis of weekly influenza-like illness rate using a one-year period of factors in Random forest regression. Biosci Trends. 2017. 11:292-296
14. Andrew NG, What Data Scientists Should Know about Deep Learning. *https://www.slideshare.net/ExtractConf/andrew-ng-chief-scientist-at-baidu* (accessed Sep 21, 2017).

(Received September 25, 2017; Revised October 9, 2017; Accepted October 20, 2017)