IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received December 1, 2018, accepted December 13, 2018, date of publication December 19, 2018,
date of current version January 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2018.2888585

A Novel Data-Driven Model for Real-Time Influenza Forecasting

SIVA R. VENNA[1,2], AMIRHOSSEIN TAVANAEI[1,2], RAJU N. GOTTUMUKKALA[2], VIJAY V. RAGHAVAN[1,2], (Life Senior Member, IEEE), ANTHONY S. MAIDA[1], AND STEPHEN NICHOLS[3]

[1]School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, LA 70503, USA
[2]NSF Center for Visual and Decision Informatics, University of Louisiana at Lafayette, Lafayette, LA 70506, USA
[3]Schumacher Clinical Partners, Lafayette, LA 70508, USA

Corresponding author: Raju N. Gottumukkala (raju@louisiana.edu)

This work was supported in part by the Division of Computer and Network Systems under Grant 1650551 and in part by Oak Ridge Associated Universities under Grant 370270.

ABSTRACT We propose a novel data-driven machine learning method using long short-term memory (LSTM)-based multi-stage forecasting for influenza forecasting. The novel aspects of the method include the following: 1) the introduction of LSTM method to capture the temporal dynamics of seasonal flu and 2) a technique to capture the influence of external variables that includes the geographical proximity and climatic variables such as humidity, temperature, precipitation, and sun exposure. The proposed model is compared against two state-of-the-art techniques using two publicly available datasets. Our proposed method performs better than the existing well-known influenza forecasting methods. The results offer a promising direction in terms of both using the data-driven forecasting methods and capturing the influence of spatio-temporal and environmental factors to improve influenza forecasting.

INDEX TERMS Influenza forecasting, LSTM, recurrent neural networks, spatio-temporal data, time series forecasting.

I. INTRODUCTION

Seasonal influenza is a major global health epidemic. According to the Center for Disease Control (CDC) reports [1] in the United States alone, there were 9.2 million to 35.6 million reported illnesses since 2010. Influenza can cause severe illnesses and even deaths for high-risk populations. Prevention and control of influenza spread can be a huge challenge especially without adequate tools to monitor and estimate the intensity of outbreaks in various populations. Predicting influenza is a very difficult task given the stochastic nature of the influenza strain and environmental conditions that affect the severity of the spread. Given the importance of this problem, many researchers have tried different approaches [8]–[18] to model various aspects of influenza outbreaks. Data-driven forecasting models offer a promising direction, especially with availability of real-time data on affected populations, and environmental conditions that contribute to these outbreaks. CDC [2]–[4] and Defense Advanced Research Projects Agency (DARPA) [5], [6] have launched several competitions to solve the problem of real-time forecasting of influenza and other infectious diseases.

Influenza forecasting research may be broadly classified into three categories. The first category includes traditional compartment models such as Susceptible-Infected-Recovered (SIR) [7], [8], Susceptible-Infected-Recovered-Susceptible (SIRS) [9], [10], and Susceptible-Exposed-Infected-Recovered (SEIR) [11], [12]. The compartmental models are intuitive in terms of capturing the different states of infected populations. These models are deterministic and lack flexibility to be recalibrated in terms of capturing the dynamics of influenza spread. The models in the second category employ statistical and time-series based methodologies such as Box-Jenkins applying some variant of Auto-Regression Integrated Moving Average (ARIMA) [13] and Generalized Autoregressive Moving Average (GARMA) [14]. The Box-Jenkins based time-series methods are flexible in terms of capturing the trending behavior of affected populations, but suffer from poor accuracy as the influence of external factors is not well captured in existing forecasting models. The third category models are machine learning methods that have gained prominence in recent years. Some popular machine learning methods include Stacked linear regression [15], Support Vector Regression [16], Binomial Chain [17], Classification and Regression Trees [18]. Machine learning based approaches are data-driven approaches that offer more flexibility in terms of capturing the influence of multiple external variables, but are computationally expensive compared to statistical models. The use of machine learning methods in understanding influenza dynamics are discussed in [19]–[21]. Additionally, a review of existing influenza forecasting methods is provided in [22]–[24].

Recurrent Neural Networks (RNNs), a class of machine learning methods, have the ability to model sequential (temporal) data prediction [25]. However, the conventional RNNs have shown practical difficulties in training the networks faced with long temporal contingencies of input/output sequences [26]. Most recently, a gradient-based method called Long Short Term Memory (LSTM) was introduced to develop a stable recurrent architecture [27]. This new technology supersedes RNNs for time series forecasting. RNNs solve the vanishing/exploding gradient problem and gives much more flexibility to the learning algorithm on when to forget the past or ignore the current input. The deep network architecture of the LSTM cells can provide a powerful model in temporal data processing. Recently, LSTM and deep LSTM have attracted much interest in temporal data prediction such as traffic speed prediction [30] and classification of the diagnoses given intensive care unit time series [31]. *One of the key contributions of the paper is the application of a deep LSTM neural network for the flu prediction problem. The deep architecture can be fulfilled by unrolling the LSTM cells in which the input of the successor cell is provided by the output of the predecessor cell.*

Researchers have attempted to improve the forecasting accuracy of influenza prediction methods by capturing the influence of external environmental variables. Previous studies have clearly identified direct influence of weather variables such as temperature, humidity, precipitation etc. on influenza virus transmission and survival [36]–[38]. As presented in [37], low relative humidity aids in faster evaporation of expelled droplets or particles and longer survival of the airborne virus. Also, geographical regions that are in close proximity to infected regions have high risk of getting infected due to population movements and high-likelihood of social interactions [39]–[41]. The impact of environmental factors has to be integrated effectively into the flu forecasting model to achieve better accuracy with influenza prediction models. Recent work from [14] tried to capture the influence of environmental conditions for flu forecasting using GARMA(3,0) model. Prior experimental studies in [42] and [43], however, demonstrated that temperature and humidity are not linearly correlated with influenza spread. Some of the recent work also includes social media interactions such as Twitter messages [15], [16] Google searches involving flu related words [10], travel patterns [32] to estimate flu risk in a particular region. However, these models, specifically the Google Flu Trends (GFT) [14] were criticized due to lack of reliability that prompted Google to discontinue the model for real-time forecasting. This highlights the gaps in both gathering reliable data and forecasting methods. While both statistical and machine learning methods have been successfully applied for influenza forecasting, one of the known limitations is that they have not been able to capture the influence of external environmental variables to improve influenza forecasting.

We propose a novel LSTM based multi-stage forecasting method that integrates the influence of various external variables into state-of-the-art machine learning models. The first stage of the model employs a time-series forecasting model. During subsequent stages the situational time-lag between the flu occurrence and weather variables, and spatial proximity of different geographical regions are captured to adjust the error introduced by the original forecasting model to further improve the performance of the model. There are two important contributions of the paper. *First, is the use of LSTM model to forecast influenza counts. Second, is the introduction of a novel method to capture the influence of external environmental variables.* The proposed method is compared with existing state-of-the-art models on both GFT and CDC data. *The LSTM model is further improved in terms of its ability to forecast influenza counts at multiple spatial and temporal scales by capturing both the influence of geographical proximity, and the impacts from environmental factors in future stage.* The proposed model performs better than the existing baseline time series based ARIMA model and the EAKF (Ensembled Adjustment Kalman Filter) model. EAKF is a data assimilation method and a recursive filtering technique that combines observations with a temporally-evolving ensemble of model simulations to generate a posterior estimate of the model state [44]. The notations and symbols used in this paper are summarized in Table. 1.

II. METHOD

The proposed model consists of two stages. In the first stage, a deep learning model based on the LSTM neural network approach is used to estimate an initial forecast. In the second stage the error from the initial forecast is reduced by incorporating two different factors: (1) An impact factor is obtained from the weather variables (humidity, precipitation, temperature, sun exposure) by extracting situational time lags using symbolic time series approach, and (2) a spatio-temporal adjustment factor obtained by capturing the influence of flu spread from neighboring regions within geographical proximity.

The proposed multi-stage forecasting approach includes two following steps. *In the first stage, the LSTM neural network is trained on the flu time series of nodes to forecast the initial flu counts.* A node refers to a geographical region, which could be a HHS region or a GFT city. *In the second stage, the impact of climatic variables and spatio-temporal adjustment factor are added to the flu counts estimated by the LSTM model to reduce the error. The impact component from climatic variables is computed using the time delayed*

TABLE 1. Notations and symbols used in this paper.

Notation/Symbol	Description
$\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_T$	Output sequences from RNN
$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T$	Input sequences for RNN
$\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_T$	Hidden state of RNN
σ	Activation function(Sigmoid)
θ	Activation function(Tanh)
W_*	Weight matrices
b_*	Bias vectors
$F_{n,t}^{\mathrm{LSTM}}$	Forecast value provided my LSTM for node n at time t
$I_{n,t}^{i}$	Impact from climatic variable i at time t on node n
$I_{n,t}^{\mathrm{tot}}$	Total impact on node n from its climatic variables at time t
$\gamma_{n,t}^{\mathrm{i}}$	Impact from spatio-temporal neighbor i at time t on node n.
$\gamma_{n,t}^{\mathrm{tot}}$	Total impact on node n at time t from all its spatio-temporal neighbors
$F_{n,t}^{\mathrm{final}}$	Final value after impacts from climate variables and spatio-temporal neighbors are applied on forecast from LSTM
$BPTT$	Back Propagation Through Time
$LSTM$	Long Short Term Memory
CDC	Center for Disease Contol
$ARIMA$	Auto Regression Integrated Moving Average
GFT	Google Flu Trends
$EAKF$	Ensembled Adjustment Kalman Filter
STL	Situational Time Lag
MSE	Mean Square Error
$MAPE$	Mean Absolute Percentage Error
$RMSE$	Root Mean Squared Error
$RMSPE$	Root Mean Squared Percentage Error

association analysis between each symbolic time series of weather and flu counts. The spatio-temporal adjustment factor is calculated by averaging over the flu variations at nearby data nodes. The proposed model is compared against our baseline LSTM model and two state-of-the-art models namely ARIMA(3,0,3) and Ensembled Adjustment Kalman Filter (EAKF).

A. DEEP LONG SHORT TERM MEMORY NETWORK

1) LSTM CELL

RNN computes an output sequence $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T)$ based on its input sequence $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ and its previous state $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_T)$ as shown in Eq. 1 and Fig. 1.

$$\begin{aligned} \mathbf{h}_t &= \sigma(W_i \cdot \mathbf{x}_t + W_h \cdot \mathbf{h}_{t-1} + \mathbf{b}_h) \\ \mathbf{y}_t &= \theta(W_o \cdot \mathbf{h}_t + \mathbf{b}_y) \end{aligned} \tag{1}$$

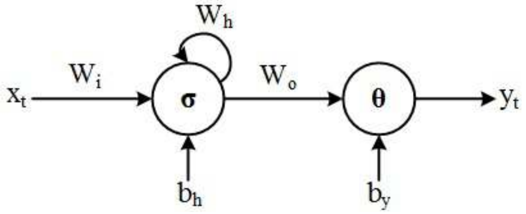


FIGURE 1. Recurrent neural network.

Here σ and θ are the hidden and output activation functions. W and b determine the adaptive weight and bias vectors of the RNN.

LSTM is a variation of RNNs preserving back-propagated error through time and layers. Furthermore, the LSTM learning algorithm is local in both space and time, with computational complexity of $O(1)$ per time step and weight [27], which is faster than the popular RNN learning algorithms (e.g. real-time recurrent learning (RTRL) [47] and back-propagation through time (BPTT) [48]). An LSTM cell performs as a memory to write, read, and erase information according to the decisions specified by the input, output, and forget gates, respectively. The weights associated with the gates are trained (adapted) by a recurrent learning process.

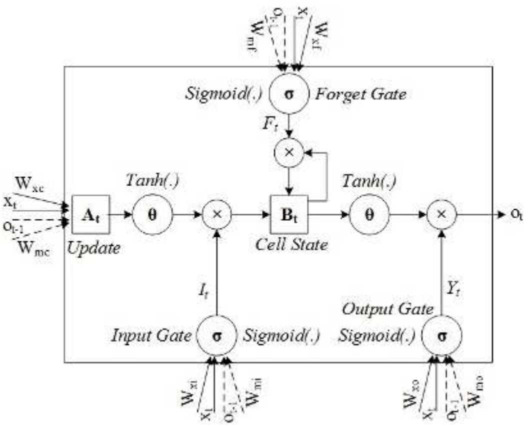


FIGURE 2. An LSTM cell containing the input gate, the forget gate, and the output gate. Each gate receives two vectors as input, $\mathbf{x}_t$, and previous output, $\mathbf{o}_{t-1}$.

The memory cell shown in Fig. 2 is implemented as follows:

$$\mathbf{I}_t = \sigma(W_{xi}\mathbf{x}_t + W_{mi}\mathbf{o}_{t-1} + \mathbf{b_i}) \tag{2}$$

$$\mathbf{F}_t = \sigma(W_{xf}\mathbf{x}_t + W_{mf}\mathbf{o}_{t-1} + \mathbf{b_f}) \tag{3}$$

$$\mathbf{Y}_t = \sigma(W_{xo}\mathbf{x}_t + W_{mo}\mathbf{o}_{t-1} + \mathbf{b_o}) \tag{4}$$

$$\mathbf{A}_t = W_{xc}\mathbf{x}_t + W_{mc}\mathbf{o}_{t-1} + \mathbf{b_c} \tag{5}$$

$$\mathbf{B}_t = \mathbf{F}_t \odot \mathbf{B}_{t-1} + \mathbf{I}_t \odot \theta(\mathbf{A}_t) \tag{6}$$

$$\mathbf{o}_t = \mathbf{Y}_t \odot \theta(\mathbf{B}_t) \tag{7}$$

where W_x and W_m are the adaptive weights, initialized randomly in the range (0,1). $\mathbf{x}_t$ and $\mathbf{o}_{t-1}$ denote the current input and previous output vectors, respectively. $\mathbf{b}$ parameters are bias vectors that are not shown in Fig. 2. The cell state, $\mathbf{B}_t$, is updated by the forget gate, the input gate, and the current input value ($\mathbf{A}_t$). The functions σ and θ determine the *Sigmoid* and *Tanh* activations respectively.

2) DEEP LSTM ARCHITECTURE

A number of approaches for developing the deep architectures of RNNs and LSTMs have been discussed in [28], [29], [49], and [50]. In this investigation, we construct an LSTM network by unrolling the LSTM cells in time. This model provides a suitable architecture for the time series prediction problems due to its sequential framework. Fig. 3 shows the network architecture consisting of the unrolled

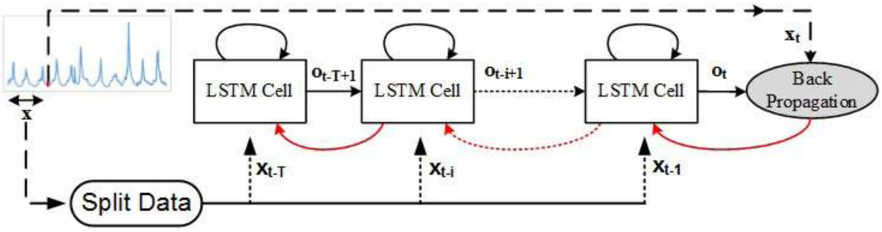


FIGURE 3. LSTM neural network consisting of the unrolled LSTM cells. The red backward arrows show the backpropagation algorithm and are not part of the network architecture.

LSTM cells that are trained by the back-propagation algorithm based on the mean-square-error cost function (training criterion). The corresponding LSTM cell at time $t-i$ receives the flu count calculated by the predecessor cell ($\mathbf{o}_{t-i-1}$) and the input, $\mathbf{x}_{t-i}$, to calculate the flu count at $t-i$, $\mathbf{o}_{t-i}$. This process is repeated for all LSTM cells in the model. The number of LSTM cells denotes the number of time steps, T, before the current time. To calculate the flu count at the current state, t, the data points from T previous time steps are used. After different experimental setups, we selected $T = 20$ time steps.

B. CLIMATIC VARIABLE IMPACT

There is strong evidence from prior literature that the dynamics of flu spread and intensity is influenced by various climatic conditions [33]–[35]. Humidity, sun exposure, precipitation, temperature all have different levels of impact on the flu counts. For example, in Fig. 4, one can observe the strong correlation between the maximum and minimum temperatures with flu counts from CDC data in one of the geographical regions. While the impact of these climatic variables is evident, a linear relationship between a climatic variable and flu count may not be effective. This is because the dynamics of flu spread is not linearly correlated with climatic variables [36]–[38]. One way to capture these nonlinear relationships between the composite climatic variables (i.e. temperature, sun exposure and precipitation) with the flu counts is through a symbolic time series approach. With the symbolic time series approach, the numerical time series is converted to a sequence of symbols [57]–[59]. These symbols can be based on the characteristics of the original time series that include magnitude, change over time, etc. The situational time lag represents the time lag between a climatic variable and flu count.

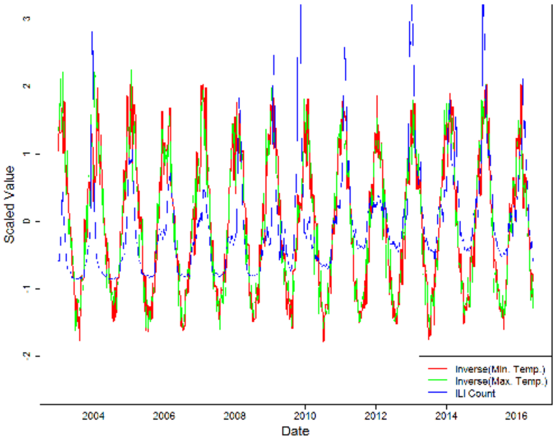


FIGURE 4. A plot showing correlation between minimum and maximum temperatures with flu counts.

Approach to Compute Situational Time Lags (STL):

1) Convert each of numeric time series (i.e. flu counts, temperature, sun exposure, precipitation) into symbolic time series, where the numerical value at each time step is converted to a symbol represented by a tuple (XY), where X ∈ {high, medium, low} and Y ∈ {increasing, decreasing, stable}.
2) Identify frequent symbol associations at different time lags between the climatic variable and the flu counts using the Apriori algorithm [60]. In this context, symbols represent items.
3) From the frequent symbol associations identified in the earlier step, pick the symbol pairs that have high confidence. The confident frequent associated symbol pairs at any time lag represent the situational correlation between the climatic variables and the flu counts.
4) If symbol pairs are confident at multiple time lags, then an average of these time lags is assigned to that particular pair. Also, for symbol pairs missing from the final confident pair list, an overall average time lag is assigned to them by default.
5) Create Situational-Time Lag STL_v table (from step 3&4) for each climatic variable v that includes a symbol pair (XY) and its appropriate situational time-lag.

Once the time lags between flu counts and each weather variable are computed for all the data nodes, total impact, I^{tot}, inflicted at time step t from the weather variables for data node n is estimated using the formula shown in Eq. 8.

$$I^{\text{tot}}_{n,t} = \sum_{i=1}^{D} \left(W_{n,i} \times I^{i}_{n,t} \right) \tag{8}$$

where $I^{\text{tot}}_{n,t}$ is the total impact from all the D climate variables on the node n at time t, $I^{i}_{n,t}$ is the individual impact from climatic variable i on the node n at time t calculated as shown in Eq. 9 and $W_{n,i}$ is the impact weight associated with the node n and climate variable i. The weights, $W_{n,i}$ are trained using Widrow-Hoff learning [51] with mean square error (MSE) criterion as the cost function on the training data with target function as shown in Eq. 8. The target of this Widrow-Hoff learning is to reduce the MSE to obtain the optimum weights ($W_{n,i}$). These weights are exclusive and trained separately for each data node.

$$I^{i}_{n,t} = \frac{\left(V^{i}_{n,t-lag} - V^{i}_{n,t-lag-1}\right)}{\max\left(V^{i}_{n,t-lag}, V^{i}_{n,t-lag-1}\right)} \tag{9}$$

The impact value at node n coming from ith climatic variable at time t is the ratio of change happening before the appropriate situational time-lag (lag) retrieved from the situational time lag table STL_i at time t and symbolic representation of flu count $F_{n,t-1}$ at node n and time $t-1$. V^i_n is the numeric time series (not the symbolic data) of ith weather variable at node n.

C. SPATIO-TEMPORAL ADJUSTMENT FACTOR

Geographical proximity, in general, strongly affects influenza outbreak in a particular region. One can observe similar flu trends between data nodes within spatial proximity as shown in Fig. 5 for both GTF and CDC data. This impact is captured by computing an adjustment factor from the nearby data nodes. Similar to the weather variables, each neighboring data node impacts this data node independently from the other neighboring data nodes. Thus, a weighted summation of individual adjustment factors is used. Here, Widrow-Hoff learning [51] is used to train those weights. Similar to the impact weights, the mean square error (MSE) training criterion is used as the cost function. Adjustment factor coming from each neighboring node is the average of flu variation difference during the previous three time stamps at that node. The adjustment factor, γ, to be applied at data node n on the initial forecast at time step t is the weighted average of changes in the flu counts obtained at other nearby data nodes at time step $t-1$.

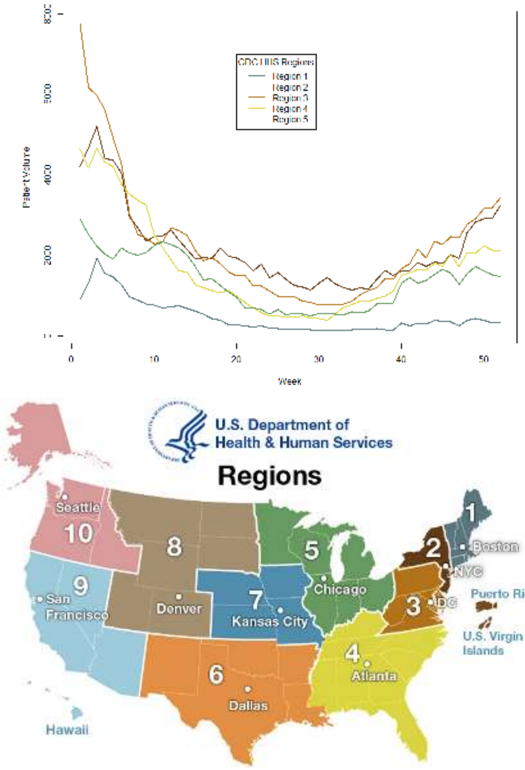


FIGURE 5. A plot showing similar trends in flu counts in 2015 for different CDC regions (top). A map showing the CDC-HHS regions (bottom).

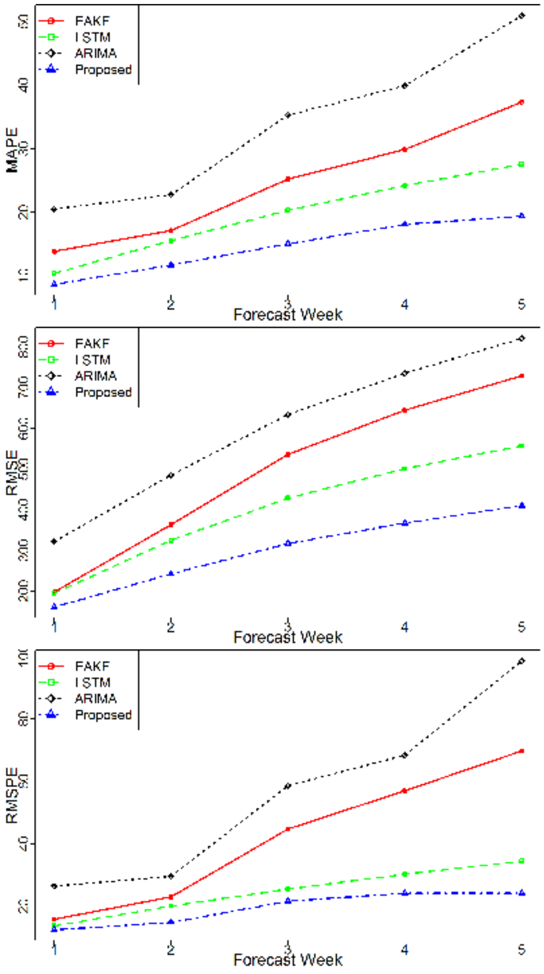


FIGURE 6. Comparison of MAPE, RMSE and RMSPE of the flu prediction models for 1 to 5 weeks ahead forecasts with the CDC-ILI dataset.

$$\gamma^{\mathrm{tot}}_{n,t} = \sum_{i\in N}\left(W_{n,i} \times \gamma^{\mathrm{i}}_{n,t}\right) \tag{10}$$

$$\gamma^{\mathrm{i}}_{n,t} = \frac{1}{y}\sum_{j=1}^{y}(F_{i,t-j} - F_{i,t-j-1}) \tag{11}$$

Total adjustment $\gamma^{tot}_{n,t}$ at data node n and time t is the average weighted summation of the individual adjustments $\gamma^{i}_{n,t}$ coming from all its neighbors that are in geographical proximity of n. Similar to the impact weights, adjustment weights($W_{n,i}$) are also trained using the Widrow-Hoff algorithm on the historical data from this node as well as its neighbors. Here $F_{i,t-j}$ is the actual flu count at neighbor i to n at time $t-j$. In our experiments we selected y to be 3 as it gave us optimum results.

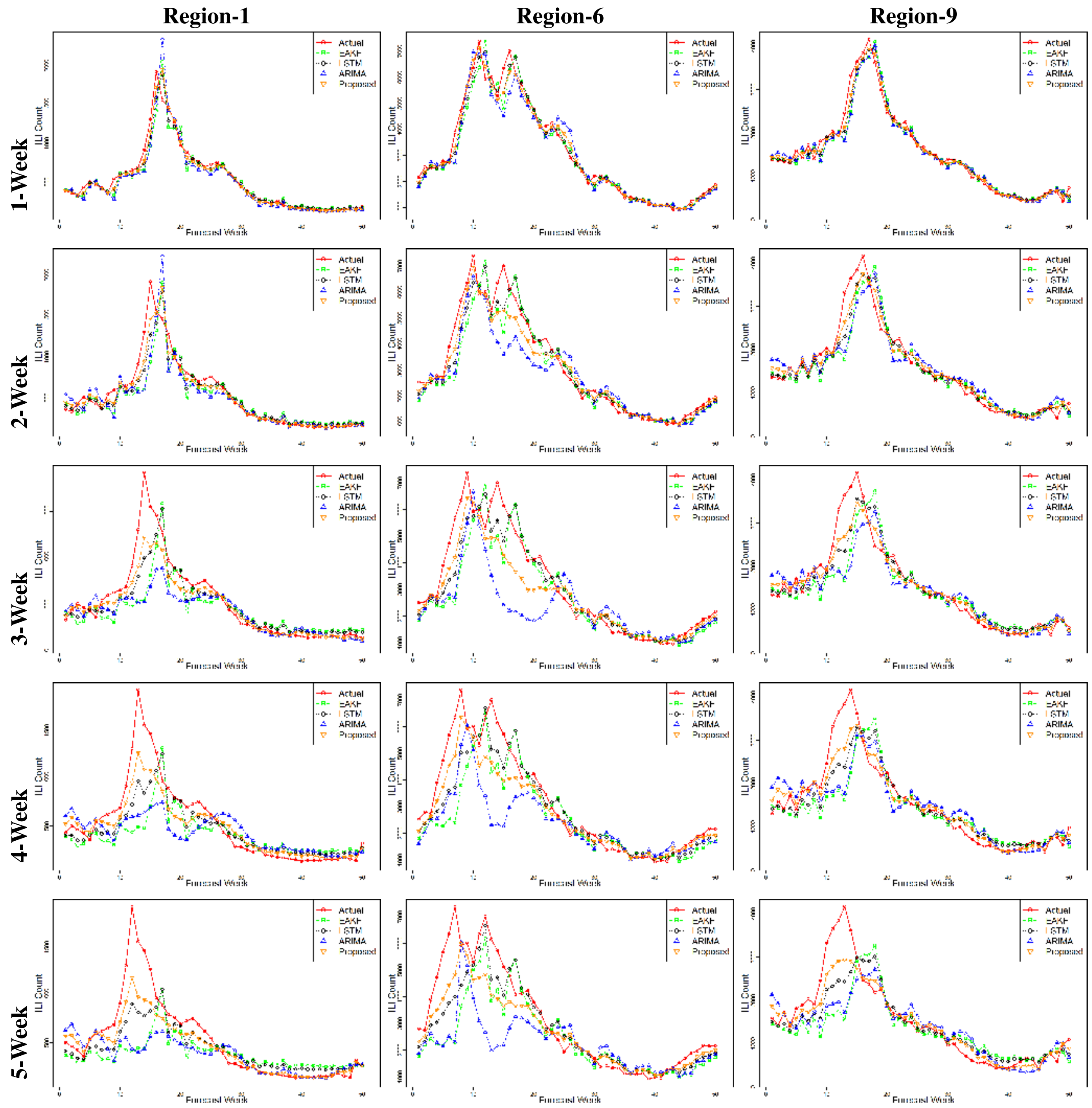


FIGURE 7. Comparison of Actual and predicted ILI counts for CDC-HHS Regions 1, 6, and 9 while forecasting 1 to 5 weeks ahead for an entire flu-season.

D. FORECAST VALUE ESTIMATION

Final forecast after applying impact factor $I_{n,t}^{\text{tot}}$ from weather variables and adjustment factor $\gamma_{n,t}$ from spatio-temporal neighbors as computed in Eq. 8 and 10, $F_{n,t}^{\text{final}}$, of data node n at time t is computed as shown in the Eq. 12.

$$F_{n,t}^{\text{final}} = \left(1+\gamma_{n,t}\right) \times \left(1+I_{n,t}^{\text{tot}}\right) \times F_{n,t}^{\text{LSTM}} \qquad (12)$$

III. EXPERIMENTS AND RESULTS

The baseline LSTM model and the new proposed model are compared against two state of the art models ARIMA and EAKF on two different publicly available data sets related to influenza counts, namely the CDC and GFT data sets. Both data sets represent a broad sample in terms of spatio-temporal granularity. The model was evaluated on three widely accepted evaluation metrics, namely Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE) and Root Mean Square Percentage Error (RMSPE) used in [52] and [53]. Each of the models were implemented in R [54]. The LSTM model was implemented using the Tensorflow library [55]. Computational complexity of BPTT and LSTM are both $O(W)$ where W is the number of adaptive weights. However, LSTM, unlike the BPTT, is local in time and space and does not need to store unlimited activation values [27]. The computational complexity of both the Widrow-Hoff models trained for Eq. 8 and 10 is dependent on the size of the weight vector W and the number of iterations required for their convergence. The equations in Eq. 9, 11 and 12 are computed in linear time $O(1)$ as they are simple additions and subtractions. A personal computer with Intel I7-6700k processor, 16 gigabytes of RAM and an NVIDIA 1070 GTX GPU was used for the experiments. The LSTM model for each training dataset takes 18-20 minutes to converge, and the overall model takes approximately 24 to 25 minutes to train. The prediction takes less than 2 seconds on the same hardware.

A. DATA DESCRIPTION

For influenza activity, two different real-world data sets were chosen. The CDC-reported Influenza Like Illness (ILI) data from CDC for all ten HHS regions between 2002-2016 [1] is the only national level dataset available for the United States. Google Flu Trends (GFT) [45] data (available from 2003 to 2014) is a weekly estimate of influenza activity derived from aggregated search query data. A subset of the GFT dataset including the flu count trends reported for 6 cities from Texas and Louisiana (Austin, Dallas, Houston, San Antonio, Baton Rouge and New Orleans) is selected as a sample for our experiments. The weather data is downloaded from Climate Data Online (CDO) [46] that provides free access to National Climatic Data Center (NCDC) archive of historical weather and climate data. The weather variables that were used include precipitation, maximum temperature, minimum temperature, and sun exposure. For each city from the GFT dataset, all available stations from the CDO within that city's geographical limits are downloaded. For the CDC dataset, all the stations within each HHS region boundaries are downloaded from the CDO. The data collected from the CDO for the both datasets are then aggregated for each city or region by averaging into single weekly summarized time-series with respect to each climatic variable. This aggregated data is then cleaned to treat any further missing values using simple moving average based smoothing. At this time all collected datasets ILI, GFT and respective weather variables are weekly summarized time series. For each experiment a combination of training (80%) and testing set (20%) is used, where training and testing sets are in sequence and mutually exclusive. For LSTM the dataset is divided into training (60%), validation (20%) and testing (20%) sets. During each of the training exercises approximately 560 samples are used for training and/or validation and the last 140 samples are used for testing with respect to CDC dataset. At the same time for GFT dataset the training and/or validation, testing sample sizes are approximately 480 and 120 respectively.

B. EVALUATION CRITERIA

The prediction performance of the proposed system is evaluated using the following three metrics:

Mean absolute percentage error (MAPE) measures the average percent of absolute deviation between actual and forecasted values.

$$MAPE = \frac{1}{N}\sum \frac{|A-F|}{|A|} \times 100 \tag{13}$$

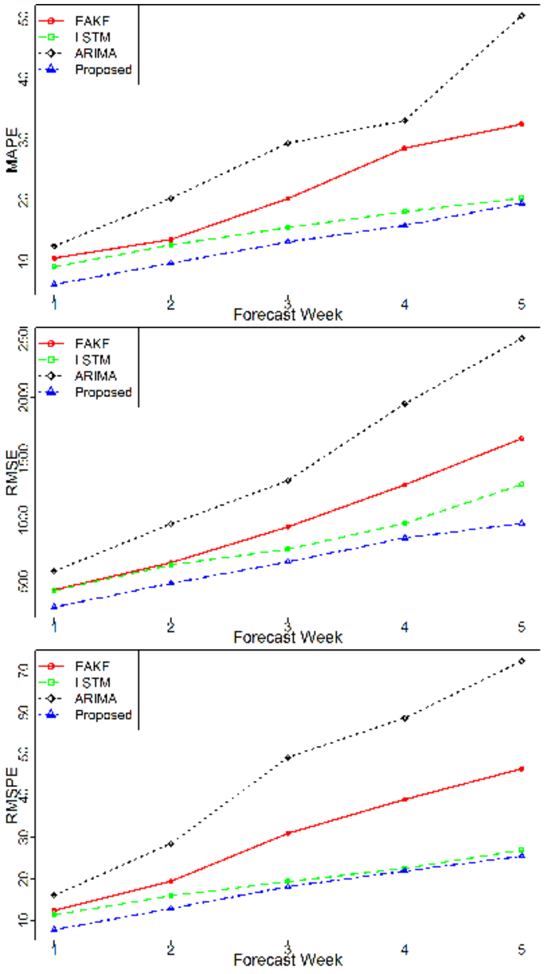


FIGURE 8. **Comparison of MAPE, RMSE and RMSPE of the flu prediction models for 1 to 5 weeks ahead forecasts with the GFT dataset.**

Root mean squared error (RMSE) captures the square root of average of squares of the difference between actual and forecasted values.

$$RMSE = \sqrt{\frac{1}{N}\sum (A-F)^2} \tag{14}$$

Root mean squared percentage error (RMSPE) captures percentage of square root of average of squares of the deviation between actual and forecated values.

$$RMSPE = \sqrt{\frac{1}{N}\sum (\frac{A-F}{A})^2} \times 100 \tag{15}$$

where, N is the number of test samples, A is the actual flu count, and F is its respective forecasted value.

We compared our results with two state-of-the-art models namely ARIMA and EAKF. The four models compared in the results section are as follows:

- EAKF (Flu count estimated using the state-of-art Ensembled Adjustment Kalman Filter)
- LSTM (The value predicted by LSTM (F^{LSTM}) alone, that is without the variable impact and adjustment factor applied to it)

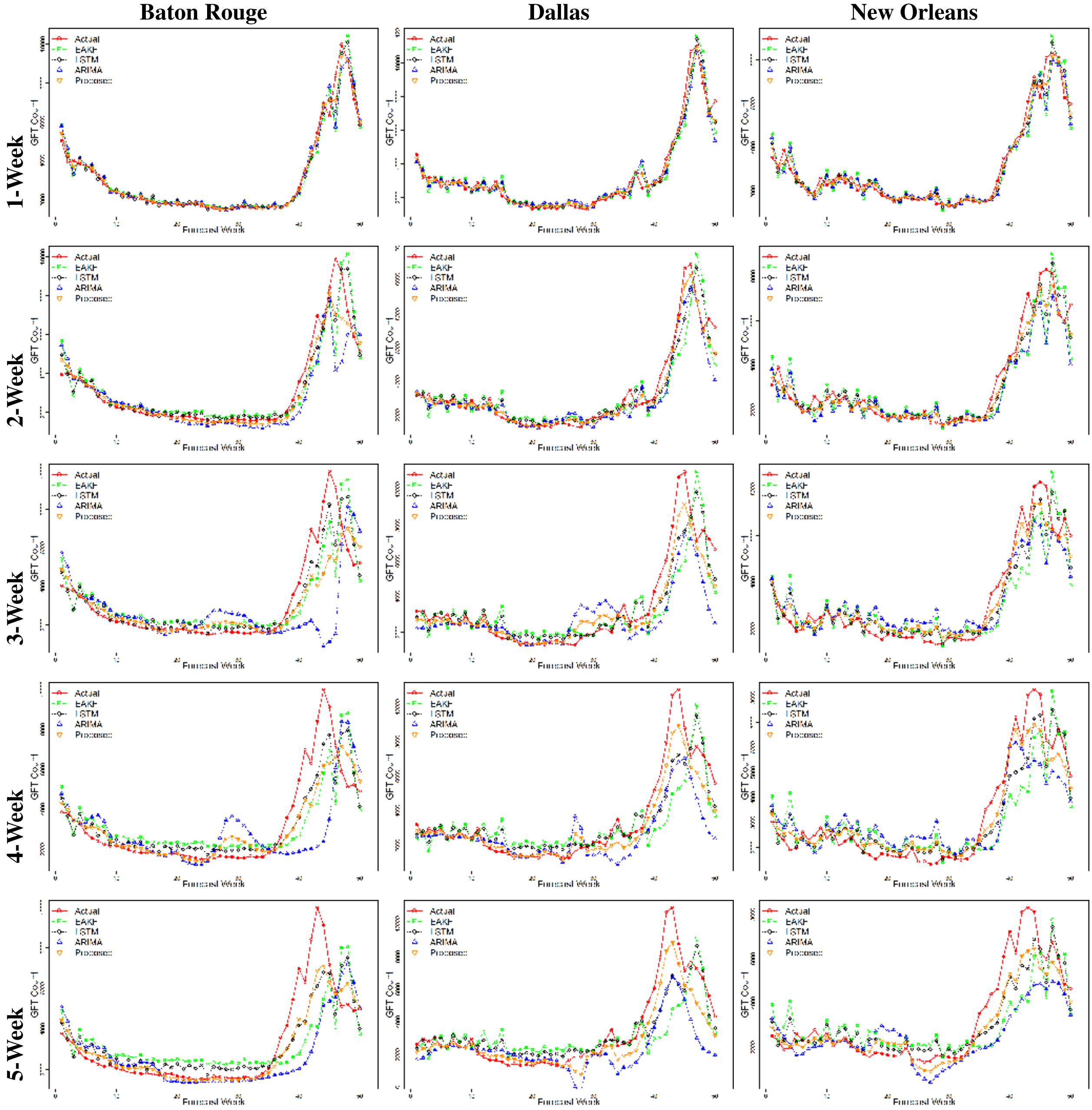


FIGURE 9. Comparison of Actual and predicted GFT counts for cities of Baton Rouge, Dallas and New Orleans while forecasting 1 to 5 weeks ahead for an entire flu-season.

- ARIMA (Flu count estimated using the state-of-art ARIMA)
- Proposed (This is the final forecast value ($F_{n,t}^{\text{final}}$) after both climatic variable impact factor and spatio-temporal adjustment factor are added to LSTM as computed in Eq. 12. This is the proposed approach)

C. RESULTS

Plots from Fig. 6 and Fig. 8 show the errors for various models for both CDC and GFT for weekly forecasting ranging from 1 week to 5-weeks. The prediction error presented in the figures is the average forecast across all the geographical regions. From both tables, one can observe that for both CDC and GFT data, the error increases with increase in forecast length (i.e. 1 week to 5 weeks in advance) for all four models (i.e. EAKF, LSTM, ARIMA model and the proposed model). The charts in Fig. 7 and Fig. 9 show the flu distribution and forecasts over a one-year time period for three regions from CDC data and three cities from GFT data respectively. One can observe that the error in the forecast is typically high when there is a sudden increase or decrease in flu observations. The prediction errors for both LSTM and EAKF models are less compared to the baseline ARIMA model for both datasets. ARIMA has been extensively used in the past because the data sample was not too large. We now have a sizable dataset of 14 years of flu data from CDC for 10 HHS regions, and 11 years of data from GFT. Given the large sample size, we observe that both LSTM and EAKF models outperform the baseline ARIMA model. We were also able to reliably quantify the impact of weather on influenza spread, the impact of neighboring regions at a regional and city scale. This enabled us to further improve the baseline LSTM model by adjusting the error of the baseline forecast by incorporating the impact of climatic influence and spatio-temporal flu patterns. This error adjustment leads to a better forecast compared to the baseline LSTM model and the EAKF model. The primary limitation of the model is the requirement of sufficient training data both for capturing the influence of external variables and training the baseline forecast model. So, this model may not be effective when sample sizes are small.

The plots in Fig. 7 compares the predicted values from the four models with actual data for 1 to 5 week-ahead forecasts and HHS-CDC regions 1, 6 and 9 separately. For regions 1 and 9, all four models are successful in predicting the peak flu season for 1-week and 2-week ahead forecasts; however ARIMA fails to identify peak for 4-week and 5-week ahead forecasts. The other 3 models could identify peaks up to 5-week ahead forecasts. The proposed model's prediction is closest to the actual observed peak. In Region-6, there are two different peaks during the flu season. ARIMA failed to identify the second peak after 1-week ahead forecast, whereas the other 3 models identified both peaks up to 3 week-ahead forecasts, with the proposed approach being the most accurate.

The plots in Fig. 9 compare the predicted values from the four models with actual data for 1 to 5 week-ahead forecasts and GFT cities Baton Rouge, Dallas and New Orleans separately. For Baton Rouge ARIMA fails to identify peaks after 2-weeks ahead forecasts, while the other 3 could predict the peaks up to 5-weeks ahead forecasts with the proposed approach being the most accurate.

IV. CONCLUSION

In this paper, we proposed a new data-driven approach for influenza forecasting. The first key contribution is the applicability of the LSTM based deep-learning method which is shown to perform well compared to existing time series forecasting methods. We further reduced the error of the deep learning based forecasting method by introducing an approach to integrate the impact from climatic variables and spatio-temporal factors. We evaluated the proposed approach on publicly available CDC-HHS ILI and GFT datasets. The proposed method offers a promising direction to improve the performance of real-time influenza forecasting models.

In this paper, we have implemented separate learning components for the climatic variables and for the geospatially proximal variables. Our future study seeks to develop an end-to-end learning model incorporating all the modules. This could be done by using a convolutional LSTM [56] to learn spatio-temporal patterns.

REFERENCES

[1] (Oct. 19, 2018). *Influenza (Flu), Centers for Disease Control and Prevention*. Accessed: Nov. 20, 2018. [Online]. Available: http://www.cdc.gov/flu/weekly/overview.htm

[2] M. Biggerstaff *et al.*, "Results from the centers for disease control and prevention's predict the 2013–2014 influenza season challenge," *BMC Infectious Diseases*, vol. 16, no. 1, p. 357, 2016.

[3] Centers for Disease Control and Prevention. *CDC Competition Encourages Use of Social Media to Predict Flu*. Accessed: Nov. 20, 2018. [Online]. Available: https://www.cdc.gov/flu/news/predict-flu-challenge.htm

[4] Centers for Disease Control and Prevention. *Flu Activity Forecasting Website Launched*. Accessed: Nov. 20, 2018. [Online]. Available:https://www.cdc.gov/flu/news/flu-forecast-website-launched.htm

[5] InnoCentive. *DARPA Forecasting Chikungunya Challenge | InnoCentive Challenge*. Accessed: Nov. 20, 2018. [Online]. Available: https://www.innocentive.com/ar/challenge/9933617?

[6] Defense Advanced Research Projects Agency. *CHIKV Challenge Asks Teams to Forecast the Spread of Infectious Disease*. Accessed: Nov. 20, 2018. [Online]. Available: https://www.darpa.mil/news-events/2014-08-15

[7] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Rev.*, vol. 42, no. 4, pp. 599–653, 2000.

[8] M. J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton, NJ, USA: Princeton Univ. Press, 2008.

[9] M. B. Hooten, J. Anderson, and L. A. Waller, "Assessing North American influenza dynamics with a statistical sirs model," *Spatial Spatio-Temporal Epidemiol.*, vol. 1, no. 2, pp. 177–185, 2010.

[10] J. Shaman, A. Karspeck, W. Yang, J. Tamerius, and M. Lipsitch, "Real-time influenza forecasts during the 2012–2013 season," *Nature Commun.*, vol. 4, Dec. 2013, Art. no. 2837.

[11] G. Chowell, M. A. Miller, and C. Viboud, "Seasonal influenza in the united states, France, and Australia: Transmission and prospects for control," *Epidemiol. Infection*, vol. 136, no. 6, pp. 852–864, 2008.

[12] G. Chowell, H. Nishiura, and L. M. Bettencourt, "Comparative estimation of the reproduction number for pandemic influenza from daily case notification data," *J. Roy. Soc. Interface*, vol. 4, no. 12, pp. 155–166, 2007.

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

[14] A. F. Dugas *et al.*, "Influenza forecasting with Google flu trends," *PLoS ONE*, vol. 8, no. 2, p. e56176, 2013.

[15] J. C. Santos and S. Matos, "Analysing Twitter and Web queries for flu trend prediction," *Theor. Biol. Med. Model.*, vol. 11, no. 1, p. S6, 2014.

[16] 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, p. e19467, 2011.

[17] H. Nishiura, "Real-time forecasting of an epidemic using a discrete time stochastic model: A case study of pandemic influenza (H1N1-2009)," *Biomed. Eng. Online*, vol. 10, no. 1, p. 15, 2011.

[18] 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," *Ann. Behav. Med.*, vol. 26, no. 3, pp. 172–181, 2003.

[19] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[20] W. Xu, Z.-W. Han, and J. Ma, "A neural netwok based approach to detect influenza epidemics using search engine query data," in *Proc. IEEE Int. Conf. Mach. Learn. (ICMLC)*, vol. 3, Jul. 2010, pp. 1408–1412.

[21] S. I. Hay, D. B. George, C. L. Moyes, and J. S. Brownstein, "Big data opportunities for global infectious disease surveillance," *PLoS Med.*, vol. 10, no. 4, p. e1001413, 2013.

[22] E. O. Nsoesie, J. S. Brownstein, N. Ramakrishnan, and M. V. Marathe, "A systematic review of studies on forecasting the dynamics of influenza outbreaks," *Influenza Other Respiratory Viruses*, vol. 8, no. 3, pp. 309–316, 2014.

[23] E. Christaki, "New technologies in predicting, preventing and controlling emerging infectious diseases," *Virulence*, vol. 6, no. 6, pp. 558–565, 2015.

[24] N. Perra and B. Gonçalves, "Modeling and predicting human infectious diseases," in *Social Phenomena*. Cham, Switzerland: Springer, 2015, pp. 59–83.

[25] H. T. Siegelmann and E. D. Sontag, "Turing computability with neural nets," *Appl. Math. Lett.*, vol. 4, no. 6, pp. 77–80, 1991.

[26] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[28] X. Li and X. Wu, "Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2015, pp. 4520–4524.

[29] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. INTERSPEECH*, 2014, pp. 338–342.

[30] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote microwave sensor data," *Transp. Res. C, Emerg. Technol.*, vol. 54, pp. 187–197, May 2015.

[31] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell. (2015). "Learning to diagnose with LSTM recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1511.03677

[32] C. Viboud, M. A. Miller, B. T. Grenfell, O. N. Bjørnstad, and L. Simonsen, "Air travel and the spread of influenza: Important caveats," *PLoS Med.*, vol. 3, no. 11, p. e503, 2006.

[33] J. Shaman and M. Kohn, "Absolute humidity modulates influenza survival, transmission, and seasonality," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 9, pp. 3243–3248, 2009.

[34] R. Tellier, "Review of aerosol transmission of influenza a virus," *Emerg. Infectious Diseases*, vol. 12, no. 11, pp. 1657–1662, 2006.

[35] C. Fuhrmann, "The effects of weather and climate on the seasonality of influenza: What we know and what we need to know," *Geogr. Compass*, vol. 4, no. 7, pp. 718–730, 2010.

[36] R. P. Soebiyanto, F. Adimi, and R. K. Kiang, "Modeling and predicting seasonal influenza transmission in warm regions using climatological parameters," *PLoS ONE*, vol. 5, no. 3, p. e9450, 2010.

[37] A. C. Lowen, S. Mubareka, J. Steel, and P. Palese, "Influenza virus transmission is dependent on relative humidity and temperature," *PLoS Pathog*, vol. 3, no. 10, p. e151, 2007.

[38] A. C. Lowen and J. Steel, "Roles of humidity and temperature in shaping influenza seasonality," *J. Virol.*, vol. 88, no. 14, pp. 7692–7695, 2014.

[39] M. E. Wilson, "The traveller and emerging infections: Sentinel, courier, transmitter," *J. Appl. Microbiol.*, vol. 94, no. s1, pp. 1–11, 2003.

[40] A. J. Tatem, D. J. Rogers, and S. Hay, "Global transport networks and infectious disease spread," *Adv. Parasitol.*, vol. 62, pp. 293–343, Jan. 2006.

[41] J. S. Brownstein, C. J. Wolfe, and K. D. Mandl, "Empirical evidence for the effect of airline travel on inter-regional influenza spread in the united states," *PLoS Med.*, vol. 3, no. 10, p. e401, 2006.

[42] G. J. Harper, "Airborne micro-organisms: Survival tests with four viruses," *J. Hygiene*, vol. 59, no. 4, pp. 479–486, 1961.

[43] F. L. Schaffer, M. E. Soergel, and D. C. Straube, "Survival of airborne influenza virus: Effects of propagating host, relative humidity, and composition of spray fluids," *Arch. Virol.*, vol. 51, no. 4, pp. 263–273, 1976.

[44] J. Shaman and A. Karspeck, "Forecasting seasonal outbreaks of influenza," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 50, pp. 20425–20430, 2012.

[45] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[46] *Climate Data Online: Dataset Discovery*. Accessed: Nov. 20, 2018. [Online]. Available: https://www.ncdc.noaa.gov/cdo-web/datasets

[47] A. Robinson and F. Fallside, "The utility driven dynamic error propagation network," Dept. Eng., Univ. Cambridge, Cambridge, U.K., Tech. Rep. CUED/F-INFENG/TR.1, 1987.

[48] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," *Neural Netw.*, vol. 1, no. 4, pp. 339–356, Oct. 1988.

[49] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. (Dec. 2013). "How to construct deep recurrent neural networks." [Online]. Available: https://arxiv.org/abs/1312.6026

[50] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech signal Process.*, May 2013, pp. 6645–6649.

[51] B. Widrow and M. E. Hoff, "Adaptive switching circuits," *IRE WESCON Conv. Rec.*, vol. 4, no. 1, pp. 96–104, 1960.

[52] N. G. Reich, J. Lessler, K. Sakrejda, S. A. Lauer, S. Iamsirithaworn, and D. A. T. Cummings, "Case study in evaluating time series prediction models using the relative mean absolute error," *Amer. Statist.*, vol. 70, no. 3, pp. 285–292, 2016.

[53] R. Fildes, "The evaluation of extrapolative forecasting methods," *Int. J. Forecasting*, vol. 8, no. 1, pp. 81–98, 1992.

[54] *R: A Language Environment for Statistical Computing*, R Found. Statist. Comput., Vienna, Austria, 2013.

[55] M. Abadi *et al.* (Mar. 2015). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467

[56] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[57] J. G. Brida and L. F. Punzo, "Symbolic time series analysis and dynamic regimes," *Structural Change Econ. Dyn.*, vol. 14, no. 2, pp. 159–183, 2003.

[58] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, vol. 7. Cambridge, U.K.: Cambridge Univ. Press, 2004.

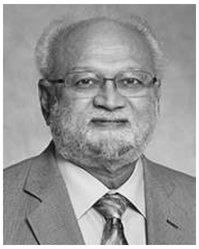
[59] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," *Proc. 8th ACM SIGMOD Workshop Res. Issues Data Mining Knowl. Discovery (DMKD)*, 2003, pp. 2–11.

[60] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in *Proc. Int. Conf. Very Large Data Bases*, vol. 1215, 1994, pp. 487–499.

SIVA R. VENNA received the B.Tech. degree in computer science and engineering from GITAM University, India, in 2008, and the M.Tech. degree in computer science from NIT, Warangal, India, in 2010. He is currently pursuing the Ph.D. degree with the Center for Advanced Computer Studies, University of Louisiana at Lafayette. His research interests include machine learning, data mining, time-series analysis and graph mining, and network time series.

AMIRHOSSEIN TAVANAEI received the M.S. degree in computer science-AI from the Sharif University of Technology, Tehran, Iran, in 2012, and the M.S. and Ph.D. degrees in computer science from the Center for Advanced Computer Studies, University of Louisiana at Lafayette, LA, USA, in 2016 and 2018, respectively. His fields of interest include artificial intelligence, machine learning, deep learning, data mining, and bio-inspired spiking neural networks.

RAJU N. GOTTUMUKKALA received the Ph.D. degree in computer science from Louisiana Tech University, LA, USA. He is currently the Director of Research of the Informatics Research Institute, University of Louisiana at Lafayette. He is also the Site Director of NSF CVDI—an NSF Industry/University Cooperative Research Center in the area of big data. He has led various efforts in the area of big data platforms, system resilience, modeling and verification of distributed systems, software-defined networks, visual analytics, and evolutionary networks. His research interests include the broader area of cyber-physical systems—specifically addressing real-world informatics and integrated systems modeling issues.



VIJAY V. RAGHAVAN is currently the Alfred and Helen Lamson Endowed Professor of computer science with the Center for Advanced Computer Studies, UL Lafayette; the Director of the Center for Visual and Decision Informatics, a multi-institutional, NSF-sponsored, Industry/University Research Center; and the Co-Director of the Laboratory for Internet Computing. His research interests include information retrieval and extraction, data and web mining, multimedia retrieval, data integration, and link discovery.

ANTHONY S. MAIDA received the B.A. degree in mathematics, the Ph.D. degree in psychology, and the M.S. degree in computer science from the University of Buffalo, in 1973, 1980, and 1981, respectively, the Ph.D. degree from Brown University, and the Ph.D. degree from the University of California, Berkeley. He was a member with the Computer Science Faculty, Penn State University, from 1984 to 1991. He has been a member with the Center for Advanced Computer Studies, University of Louisiana at Lafayette, since 1991. His current research interests include neural implementations of machine-learning algorithms, including artificial neural networks, bio-inspired artificial intelligence, deep learning, and brain simulation.

STEPHEN NICHOLS received the M.D. degree. He has been a Staff Physician with the Emergency Department, Brownwood, TX, USA, for 20 years, where he was the Medical Director for the past decade. He has been the Chief of Clinical Operations with Hospital Physician Partners, Inc., since 2015. He is currently the Chief of Clinical Operations Performance–Emergency Medicine and Hospital Medicine with The Schumacher Group. He has over 25 years of experience in emergency medicine.