International Journal of Computational Intelligence Systems
Vol. **14(1)**, 2021, *pp.* 821–833
DOI: https://doi.org/10.2991/ijcis.d.210203.004; ISSN: 1875-6891; eISSN: 1875-6883
https://www.atlantis-press.com/journals/ijcis/

Research Article

Forecasting Teleconsultation Demand with an Ensemble Attention-Based Bidirectional Long Short-Term Memory Model

Wenjia Chen[1], Lean Yu[2,*], Jinlin Li[1,*]

[1] *School of Management and Economics, Beijing Institute of Technology, Beijing 100081, China*
[2] *School of Economics and Management, Beijing University of Chemical Technology, Beijing 100029, China*

ARTICLE INFO

Article History

Received 29 Sep 2020
Accepted 31 Jan 2021

Keywords

Teleconsultation
Demand forecast
Holiday effect
Attention mechanism
Deep learning ensemble
Bidirectional long short-term memory (BILSTM)

ABSTRACT

Accurate demand forecast can help improve teleconsultation efficiency. But teleconsultation demand forecast has not been reported in existing literature. For this purpose, the study proposes a novel model based on deep learning algorithm for daily teleconsultation demand forecast to fill in the research gap. Because of the significant effect of holidays on teleconsultation demand, holiday-related variables, and specific prediction technologies were selected to treat it. The technologies attention mechanism and bidirectional long short-term memory (BILSTM) were used to construct a novel forecasting methodology, i.e., ensemble attention-based BILSTM (EA-BILSTM), for the accurate forecasts. Based on actual teleconsultation data, the effectiveness of variable selection is verified by importing different inputs into models, and the superiority of EA-BILSTM is verified by comparison with nine benchmark models. Empirical results show that importing selected variables can lead to better forecasts and EA-BILSTM model can get lowest forecasting errors on two sub-datasets. This indicates that the proposed forecasting model is a high potential approach for teleconsultation demand prediction in the influence of sparse trait, like holiday effects.

1. INTRODUCTION

To omit geographical and functional distance, advanced information and communication technologies (ICTs) are used in telemedicine healthcare services. Telemedicine can often be divided into three modes: teleconsultation, remote monitoring, and remotely supervised treatment or training [1]. Teleconsultations are remote meetings between two or more geographically separated health providers or between health providers and patients for diagnosis or treatment of diseases [2]. At the present stage in National Telemedicine Center of China (NTCC), teleconsultation is a service that doctors at primary hospitals ask for the help of the doctors at class-A tertiary hospitals to diagnose or treat difficult miscellaneous diseases [3]. By this way, teleconsultation can sink high-quality medical resources to underdeveloped areas to alleviate the shortage and uneven distribution of medical resources in China. Because of the function, teleconsultation played an important role in the battle against COVID-19 in the beginning of 2020.

In recent years, more and more hospitals are networked with NTCC, so the demand of teleconsultation has increased gradually. To increase the efficiency of teleconsultation, rational arrangements of triage coordinators and consultation rooms are crucial. To realize this objective, accurate demand forecast can be of help by avoiding unbalanced supply and demand. However, teleconsultation demand forecast hasn't been reported in existing literatures. Previous studies of telemedicine focus on the use introduction [4,5], effect evaluation [6–8], technology development [9], and some operations and management studies. In the operations and management literature, resource allocation [3], appointment and scheduling [10], face-to-face visit prediction [11], blood pressure prediction [12], chronic wound tissue prediction [13], framework development [14], the need survey of healthcare providers [15], and patient preferences [16] are involved.

In previous studies, there is no forecasting model of daily teleconsultation demand. For this purpose, the study tries to develop the forecasting model to fill in the research gap. To build an effective model with high performance, the trait of teleconsultation demand time series data was analyzed firstly. In teleconsultation, the demand is significantly affected by holidays. Specifically, the demand before holidays decreases (pre-effect), and the demand on holiday can decrease to zero. Besides, the demand after holidays is less than the ordinary working days. In all, the holiday effect on teleconsultation demand does not only affect whole holidays but also affect working days near to holidays. In terms of this holiday effect, some related variables were selected as the part of model input, and the attention mechanism and didirectional long short-term memory (BILSTM) were selected to build the forecasting model. For improvement, considering the pre-effect of holidays on

[*] *Corresponding author. Email: yulean@amss.ac.cn; jll_bit@163.com*

teleconsultation demand, a deep learning ensemble framework was proposed to use future information and avoid useless information. Based on this framework, an ensemble attention-based BILSTIM (EA-BILSTM) model was proposed for final prediction purpose.

The main contributions of this paper are described as follows:

(1) This study is the first work to build the forecast model of daily teleconsultation demand for optimizing teleconsultation resources.

(2) This study increases the prediction accuracy of daily teleconsultation demand by elaborating treatment to holiday effect. The holiday effect is tackled by variable selection and model construction.

(3) Deep learning algorithm, attention mechanism, and ensemble method are used to construct a novel forecasting model for teleconsultation demand. This study is also the first work to use those techniques to handle with the holiday effect in demand forecasts.

(4) To illustrate the effectiveness of variable selection and superiority of the proposed model, empirical studies are conducted on actual teleconsultation demand data. Also, the influence of data quantity on the demand forecast is analyzed.

The rest of the paper is organized as follows: Section 2 gives a review of the literature related to demand forecast and holiday effect. Section 3 introduces the ensemble methodology. Section 4 introduces the datasets and experimental designs. Section 5 presents the experimental results and Section 6 gives the discussions. Finally, Section 7 concludes the paper.

2. LITERATURE REVIEW

Demand forecast has been studied in many areas, such as energy [17,18], tourism [19,20], supply chain [21], traffic [22,23], and healthcare [24–30]. In the area of healthcare, most of the demand forecast focused on emergency department (ED). For example, patient visits to ED forecast studies before 2009 were reviewed in [26]. In the identified 9 studies, most of the models used to forecast the number of the patient were either linear regression models including calendar variables or time series models. The study in [27] demonstrated autoregressive integrated moving average (ARIMA) can be used to forecast, at least in the short term, demand for emergency services in a hospital ED. For monthly ED demand forecast, multivariate vector-ARIMA (VARIMA) models provided a more precise and accurate forecast than the ARIMA and Winters' method [28]. For improving forecasts in [29], patients were classified before the ED patients flow were forecasted in the long term and the short term. And for short- and long-term forecast of ED attendances, modified heuristics based on a fuzzy time series model was developed in [30], which was proved to outperform ARIMA and neural network (NN) model.

Demand forecast is more difficult than some time series prediction because of the influence of holidays on the demand change. The holiday effect in healthcare has been identified [31,32]. Teleconsultation demand is also affected by the holidays. Some staffs, like triage nurses and device administrator in NTCC are day off in holidays. Therefore, when holidays are coming, applications for teleconsultation will decrease.

In previous studies, when holiday effect existed, a number of methods were proposed to improve prediction performance. These methods can be divided into two categories, input treatment and model development. To handle holiday effect from input perspective, holidays were seen as an emergency event in [33]. The fluctuation coefficient of emergency event got by trend extrapolation was used to forecast the tourist demand, contributing higher forecasting accuracy. In a multi-stage of the input feature combination in [34], the identified optimal combination of the input features and their appropriate coding can improve the accuracy of passenger demand forecast, not only for the forecasting results on weekdays and weekends, but also for them on national holidays. In the traffic flow data, discrete Fourier transform was used to extract the common trend and support vector regression (SVR) was used to forecast the residual series [23]. Because of the hybrid approach, higher accuracy of traffic demand forecast during the holidays was achieved.

To handle holiday effect from model perspective, high-performance models were proposed by researchers. In [35], a hybrid model, combining SVR model with the adaptive genetic algorithm (GA) and the seasonal index adjustment, was proposed to forecast holiday daily tourist demand. In an attempt to forecast holiday passenger flow more accurately, a modified least squares support vector machine (LSSVM) was proposed in [36]. In this method, an improved particle-swarm optimization (IPSO) algorithm was used to optimize parameters of LSSVM and the pruning algorithm was used to achieve sparseness in the LSSVM solution. The results show that the modified LSSVM model was an effective forecasting approach with higher accuracy than other alternative models. To seize the weekly periodicity and nonlinearity characteristics of short-term ridership in practical application, a combined SVM online model was proposed in [37]. This method combined a support vector machine overall online model (SVMOOL) and a support vector machine partial online model (SVMPOL). The SVMOOL was inserted the weekly periodic characteristics and trained the updated data day by day, and the SVMPOL was inserted the nonlinear characteristics and trains the updated data of the predicted day by time interval. The experimental results demonstrated that the proposed SVM-based online model outperformed the seasonal ARIMA, back-propagation neural network (BPNN), and SVM. In [22], a convolutional neural network (CNN)-based multi-feature forecasting model (MF-CNN) was proposed, which collectively forecasts network-scale traffic flow with multiple spatiotemporal features and external factors.

From the literature review, the holiday effect is not an easy problem to tackle in time series forecasting. On the one hand, considering the holiday effect is lasting for a period of time, how to build and select the holiday-related variables is a state of art. On the other hand, the holidays are sparse distributed, and the date of some Chinese traditional holidays are not fixed. The sparseness and variability require the built model has powerful learning ability to learn the holiday influence. For traditional time series techniques and machine learning methods, the sparseness and variability of holiday effect make them cannot achieve satisfactory forecasting results. Thus, a flexible technique, with long memory on learning results and more attention to specific variables, is more suitable to deal with the sparse distributed and date-changed holiday effect. Among techniques, deep learning method with attention mechanism can be used to construct such a model with this powerful learning capability.

3. METHODOLOGY

To obtain better forecasts for teleconsultation demand in this study, deep learning algorithm, attention mechanism, and ensemble method are combined to deal with the holiday effect. The deep learning algorithm and attention mechanism are introduced in Section 3.1. The ensemble framework is introduced in Section 3.2. And based on the framework, a novel EA-BILSTM model is proposed for teleconsultation demand forecast in Section 3.3.

3.1. Deep Learning Algorithm and Attention Mechanism

In this section, a typical deep learning method named long short-term memory (LSTM) is introduced, which has advanced gate unites, namely input gate, output gate, and forget gate [38]. The input gate decides what new information can be stored in the cell state; the output gate decides what information can be output based on the cell state; and the forget gate can decide what information will be thrown away from the cell state. These gates units make LSTM suitable for processing and predicting events with very long intervals and delays in time series. Despite of this advantage, LSTM is only able to make use of previous information. In order to overcome this shortcoming, the bidirectional recurrent neural network (BRNN) was introduced by Schuster and Paliwal [39]. This kind of architecture could be trained in both time directions simultaneously, with separate hidden layers. Considering the pre-effect of holiday on teleconsultation demand, the BILSTM may improve the prediction performance. Furthermore, attention mechanism can improve forecasts by readjustment of weights on variables [40]. Attention mechanism has been widely used in natural language processing (NLP), showing outstanding performance [41,42]. And it is gradually applied in time series forecast, showing great potentiality [43].

3.2. The Proposed Attention-Based Deep Learning Ensemble Model

Under the background of significant holiday effect on teleconsultation demand, a deep learning ensemble framework is proposed to use as much useful information as possible for forecast improvement. Normally, forecasting models use history data to predict the future state. However, some information at the forecasted time or after the forecasted time are available, like the holiday information arranged by the government in several months advance. Due to the significant pre-effect of holidays on teleconsultation demand, introducing future holiday data into models can improve forecasts greatly. To use all available and useful information, the attention-based deep learning ensemble model is proposed and shown in Figure 1. In this ensemble model, a typical RNN, like LSTM or BILSTM, is used as the single predictor. In the proposed methodology, there are three main steps, which are elaborated below.

Step 1: Data Divisions

In the beginning, the input data matrix at time t may be in irregular shape because of the different data availability. Data can have different availability at time t because of the different frequencies in data collection or the unknown future information of some variables. According to the different availability, the input data can be divided into different inputting sub-matrices. Variables contained same availability are divided into one group as one input. Variables can be repeatedly used. High available variables can be included in low available variable-groups. Figure 2 gives the example of data division when the data have two availability.

Step 2: Individual Prediction

Each data sub-matrix is input into a RNN with an attention module. Because of the capability of the attention module, the hidden states of RNN are weighted and they are seen as the results of the individual prediction.

Step 3: Ensemble Prediction

Attention-based RNN outputs the weighted hidden states in Step 2. In Step 3, all weighted hidden states are concatenated together as the input of the dense layer. This layer is a regular densely-connected NNs layer. And in this study, the activation of it is the sigmoid. Finally, the dense outputs the final ensemble prediction result of time t.

3.3. The Proposed EA-BILSTM Model

Given the pre-effect of holidays on teleconsultation demand, this study applied attention-based BILSTM to build the forecasting model for teleconsultation demand. Besides, for better use of the future information, an EA-BILSTM model is proposed for the final teleconsultation demand prediction based on the ensemble framework. The proposed EA-BILSTM model in this paper is shown in Figure 3.

There are two available data on the predicted time t, available historical demand data and holiday data, and available future holiday data. Therefore, the input data at time t can be divided into two sub-matrices. Then, these two sub-matrices are respectively introduced into the BILSTM and attention module to obtain individual predictions. Individual prediction results $\boldsymbol{y}_t^{I1}$ and $\boldsymbol{y}_t^{I2}$ of two sub-matrices at time t are given below.

$$\begin{aligned}
\overset{\rightarrow 1}{\boldsymbol{h}}_t &= \overrightarrow{F}_{BL1}\left(\overset{\rightarrow 1}{\boldsymbol{h}}_{t-1}, \overset{\rightarrow 1}{\boldsymbol{c}}_{t-1}, \boldsymbol{x}_t^1\right) \\
\overset{\leftarrow 1}{\boldsymbol{h}}_t &= \overleftarrow{F}_{BL1}\left(\overset{\leftarrow 1}{\boldsymbol{h}}_{t-1}, \overset{\leftarrow 1}{\boldsymbol{c}}_{t-1}, \boldsymbol{x}_t^1\right) \\
\boldsymbol{y}_t^{I1} &= \left[\boldsymbol{W}_{af1} \cdot \overset{\rightarrow 1}{\boldsymbol{h}}_t, \boldsymbol{W}_{ab1} \cdot \overset{\leftarrow 1}{\boldsymbol{h}}_t\right] \\
\overset{\rightarrow 2}{\boldsymbol{h}}_t &= \overrightarrow{F}_{BL2}\left(\overset{\rightarrow 2}{\boldsymbol{h}}_{t-1}, \overset{\rightarrow 2}{\boldsymbol{c}}_{t-1}, \boldsymbol{x}_t^2\right)
\end{aligned} \tag{1}$$

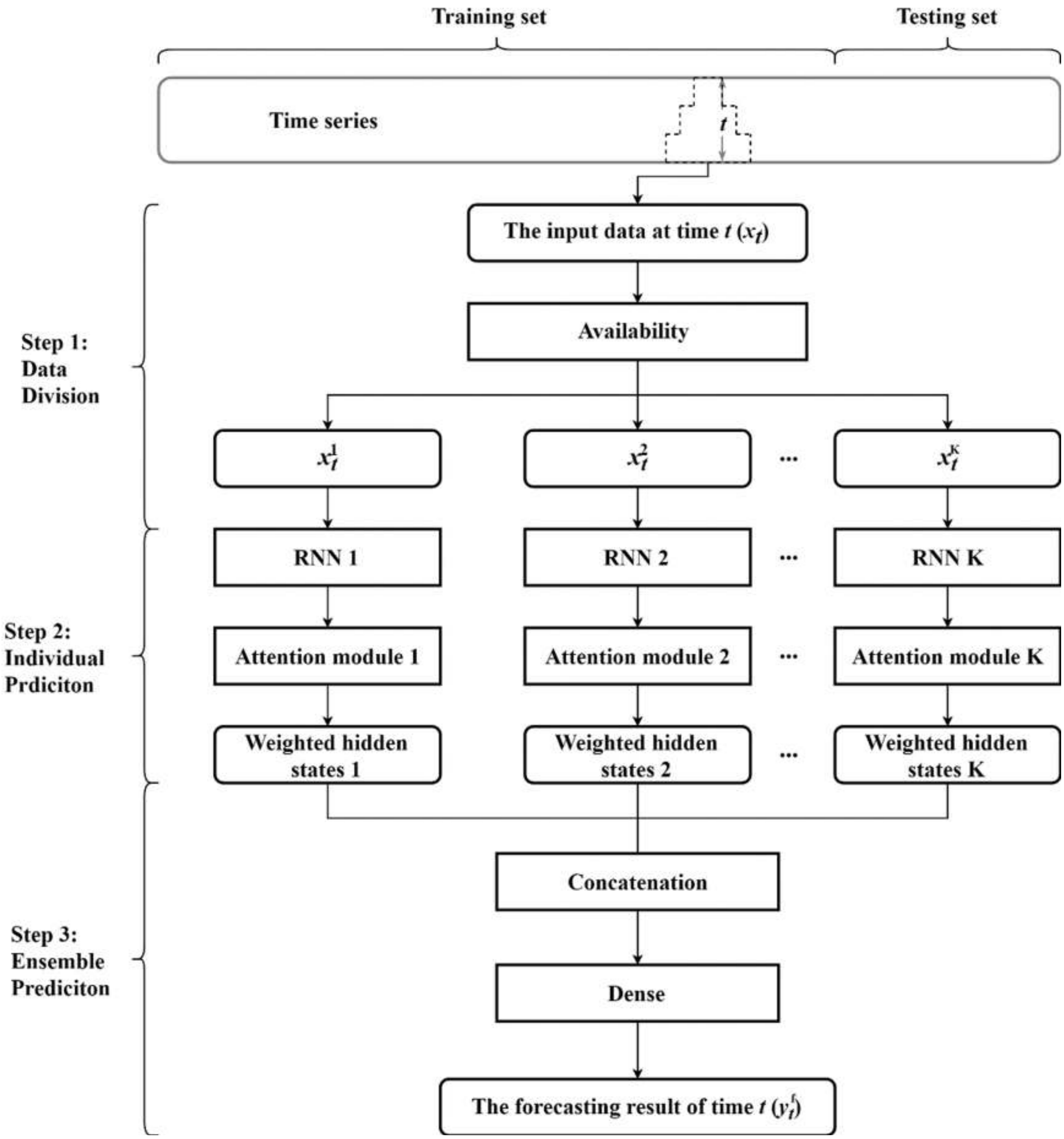

Figure 1 The generic framework of the proposed deep learning ensemble model.

$$\overleftarrow{h}_t^2 = \overleftarrow{F}_{BL2}\left(\overleftarrow{h}_{t-1}^2, \overleftarrow{c}_{t-1}^2, x_t^2\right)$$
$$y_t^{I2} = \left[W_{af2} \cdot \overrightarrow{h}_t^2, W_{ab2} \cdot \overleftarrow{h}_t^2\right] \tag{2}$$

where $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ are the outputs of forward and backward layers, respectively, $\overrightarrow{F}_{BL}$ represents the calculations of forward layers and $\overleftarrow{F}_{BL}$ represents the calculations of backward layers, $\overrightarrow{c}_t$ and $\overleftarrow{c}_t$ denote the cell states of BISLTM, x_t is the input matrix, W_{af} and W_{ab} are the learning weights by attention modules.

To get ensemble predictions, the individual results are aggregated as the input of the dense layer. The final prediction y_t^f at time t is given by

$$y_t^f = \sigma\left(\left[y_t^{I1}, y_t^{I2}\right]\right) \tag{3}$$

4. DATA DESCRIPTION AND EXPERIMENTAL DESIGN

To testify the effectiveness of variable selection and the superiority of the proposed EA-BILSTM model, the actual teleconsultation demand data are used as the sample data. And other forecasting techniques (including traditional models, machine learning models, and single deep learning models) are introduced for comparison. Section 4.1 describes the data and Section 4.2 presents the experimental design.

4.1. Dataset

The dataset in this study is the application records of teleconsultation collected in NTCC [3]. The dataset consists of the number of the application from January 1, 2018, to November 30, 2019, with a total of 668 observations. That is, in these 699 days, the records of 31 days were missing.

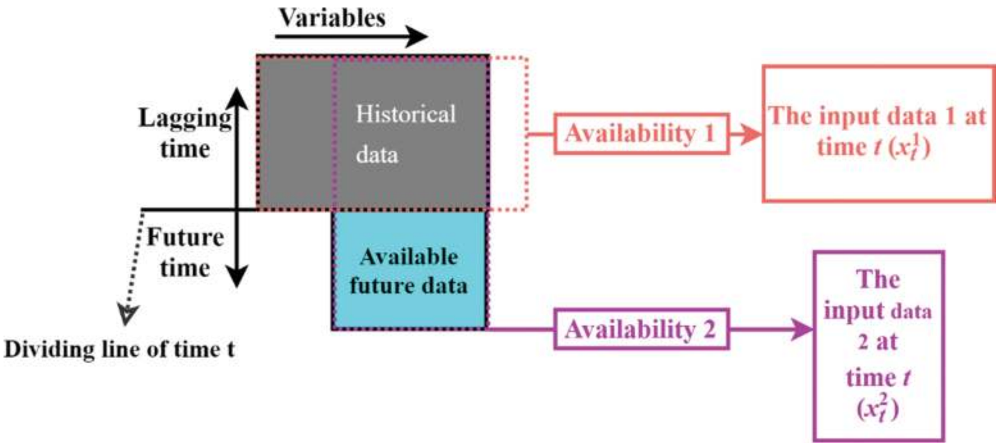

Figure 2 The example of input data division for data with two availability.

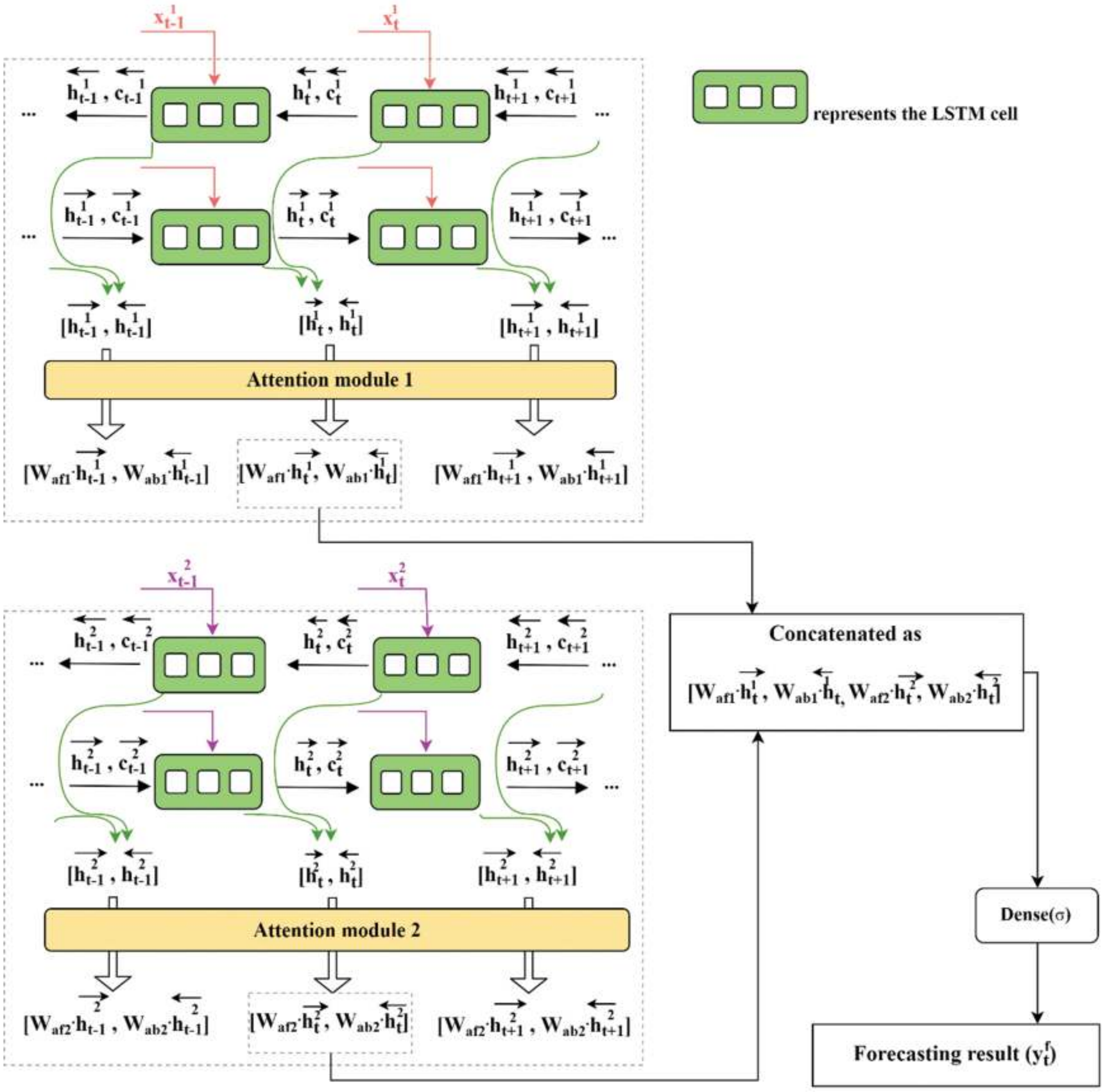
x_{t-1}^1
x_t^1
represents the LSTM cell
$\overleftarrow{h}_{t-1}^1, \overleftarrow{c}_{t-1}^1$
$\overleftarrow{h}_t^1, \overleftarrow{c}_t^1$
$\overleftarrow{h}_{t+1}^1, \overleftarrow{c}_{t+1}^1$
$\overrightarrow{h}_{t-1}^1, \overrightarrow{c}_{t-1}^1$
$\overrightarrow{h}_t^1, \overrightarrow{c}_t^1$
$\overrightarrow{h}_{t+1}^1, \overrightarrow{c}_{t+1}^1$
$[\overrightarrow{h}_{t-1}^1, \overleftarrow{h}_{t-1}^1]$
$[\overrightarrow{h}_t^1, \overleftarrow{h}_t^1]$
$[\overrightarrow{h}_{t+1}^1, \overleftarrow{h}_{t+1}^1]$
Attention module 1
$[W_{af1} \cdot \overrightarrow{h}_{t-1}^1, W_{ab1} \cdot \overleftarrow{h}_{t-1}^1]$
$[W_{af1} \cdot \overrightarrow{h}_t^1, W_{ab1} \cdot \overleftarrow{h}_t^1]$
$[W_{af1} \cdot \overrightarrow{h}_{t+1}^1, W_{ab1} \cdot \overleftarrow{h}_{t+1}^1]$
x_{t-1}^2
x_t^2
$\overleftarrow{h}_{t-1}^2, \overleftarrow{c}_{t-1}^2$
$\overleftarrow{h}_t^2, \overleftarrow{c}_t^2$
$\overleftarrow{h}_{t+1}^2, \overleftarrow{c}_{t+1}^2$
$\overrightarrow{h}_{t-1}^2, \overrightarrow{c}_{t-1}^2$
$\overrightarrow{h}_t^2, \overrightarrow{c}_t^2$
$\overrightarrow{h}_{t+1}^2, \overrightarrow{c}_{t+1}^2$
$[\overrightarrow{h}_{t-1}^2, \overleftarrow{h}_{t-1}^2]$
$[\overrightarrow{h}_t^2, \overleftarrow{h}_t^2]$
$[\overrightarrow{h}_{t+1}^2, \overleftarrow{h}_{t+1}^2]$
Attention module 2
$[W_{af2} \cdot \overrightarrow{h}_{t-1}^2, W_{ab2} \cdot \overleftarrow{h}_{t-1}^2]$
$[W_{af2} \cdot \overrightarrow{h}_t^2, W_{ab2} \cdot \overleftarrow{h}_t^2]$
$[W_{af2} \cdot \overrightarrow{h}_{t+1}^2, W_{ab2} \cdot \overleftarrow{h}_{t+1}^2]$
Concatenated as
$[W_{af1} \cdot \overrightarrow{h}_t^1, W_{ab1} \cdot \overleftarrow{h}_t^1, W_{af2} \cdot \overrightarrow{h}_t^2, W_{ab2} \cdot \overleftarrow{h}_t^2]$
Dense(σ)
Forecasting result (y_t^f)

Figure 3 The proposed ensemble attention-based bidirectional long short-term memory (EA-BILSTM) model for teleconsultation demand forecast.

Besides the missing data, the holiday effect is significant in teleconsultation demand. Taking the data from January 1, 2018, to April 30, 2018, as example, the demand changes with date are shown in Figure 4. The holidays have two main effect on the demand. On the one hand, the demands in the holidays are much less than the ordinary working days. And the demands before and after holidays are also less than ordinary working days. On the other hand, the holidays bring variations on the weekly pattern of teleconsultation demand. In Figure 4, it can be also observed that some missing data are caused by the holidays. Before missing data treatment, the variables relating to the holiday are selected.

Because sample volume affects forecasting results [44], therefore, three sub-datasets with different sample volumes are used in the experiments of this study. They are 365-day dataset (January 1, 2018, to December 31, 2018), 548-day dataset (January 1, 2018, to July 2, 2019), and 699-day dataset (January 1, 2018, to December 30, 2019). In the 365-day dataset and 548-day dataset, 70% observations are used as their training datasets, and remaining 30% observations are used as the testing datasets. In 699-day dataset, 60% observations are used as the training dataset, and remainder 40% observations are used as the testing dataset. The testing set of 699-day dataset accounts for 40% to make the testing dataset include more holidays for the learning effect test of holiday influence.

4.2. Experimental Designs

The experiment designs are presented in this section, in terms of variable selection, missing data treatment, evaluation criteria, benchmark models, and model inputs.

4.2.1. Selection of holiday-related variables

To introduce the holiday information into forecasting models, holiday-related variables are selected. First of all, week and holiday are the commonly used variables, and they are included in the selection. In addition, considering the spillover effect of holidays on teleconsultation demand, other eight variables are adopted, including the length of holiday, the first, second, third day before and after the holiday, and weekends shifted to working days. Variables are selected by significance of coefficients in the constructed linear regression equation, as shown in Table 1. The coefficients of seven variables were significant with the 5% significance level. So those seven variables were selected as the holiday-related part of the input data. Furthermore, the significance of the second and first day before the holiday and the first day after the holiday proves that the holidays have spillover effect on teleconsultation demand.

4.2.2. Treatment of missing data

According to the missing mechanism and missing pattern [45], three kinds of missing data are treated in teleconsultation demand data. And they are discrete missingness at random (DMAR), continuous missingness at random (CMAR), and missing completely at random (MCAR). For these different missing data, feature-driven method and LSTM model are used to treat the missingness, as shown in Table 2. To evaluate the effect of missing data treatments, two evaluation criteria, root mean square error (RMSE), and mean absolute error (MAE) of LSTM are utilized. The treatment with the best forecasting performance of LSTM is selected as the finally applied missingness treatment for later experiments. According to the values of RMSE and MAE, 0 imputation for DMAR and CMAR, and LSTM imputation for MCAR are the best missingness treatment methods. The next models will be built on the dataset treated by these imputation methods.

4.2.3. Evaluation criteria and statistic test

To assess forecasting accuracy, two criteria are applied and they are the root mean squared error (RMSE) and the MAE:

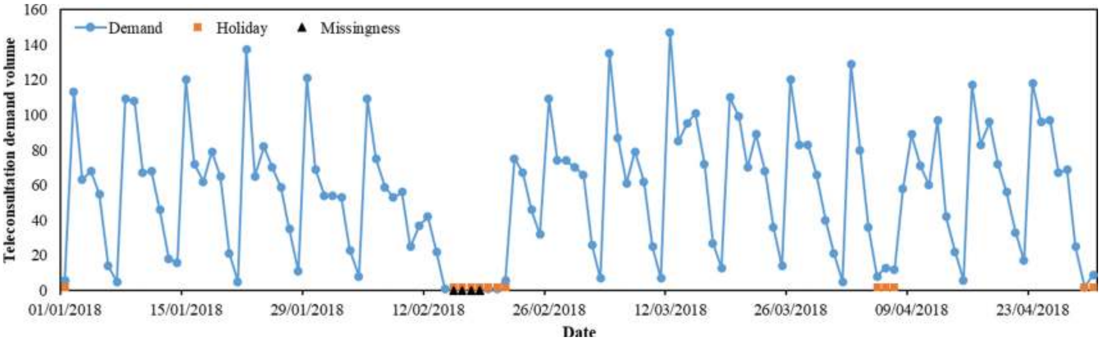

Figure 4 The changes of teleconsultation demand volume with date.

Table 1 Selection of holiday-related variables.

Variables	Week	Holiday	Length of Holiday	The Third	The Second	The First	The First	The Second	The Third	Adjusted as Working Day
				Day Before the Holiday			Day After the Holiday			
Values	1–7	0, 1	3, 4, 7	0, 1	0, 1	0, 1	0, 1	0, 1	0, 1	0, 1
Coefficients	−13.19	−27.88	−3.55	−0.37	−13.39	−33.47	22.11	3.21	3.61	17.21
Significance*	<0.01	<0.01	<0.01	0.943	0.013	<0.01	<0.01	0.524	0.475	<0.01

*is the significance of coefficients in constructed linear regression equation.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(x_t - \hat{x}_t\right)^2} \quad (4)$$

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|x_t - \hat{x}_t\right| \quad (5)$$

where x_t is the actual value, $\hat{x}_t$ is the predicted value, and N is the size of predictions.

To provide statistical evidence of the forecasting ability of the proposed model, the Diebold-Mariano (DM) test is introduced on MAE to identify the significant forecasting differences in models. DM test determines whether there is a significant difference between two prediction models under the assumption that the performance error between them is 0 (the null hypothesis). Under this assumption, DM follows a standard normal distribution. Thus, there is a significant difference between the prediction models if $|DM| > z_{\alpha/2}$, where $z_{\alpha/2}$ is the two-tailed critical value for the standard normal distribution with the α significance (0.1 in this paper). In the test statistics, ±1.65 are the thresholds to identify the 0.1 significance. Negative statistics mean the MAE of benchmark models are inferior to that of test models, and benchmark models outperform test models. Positive statistics mean test models outperform benchmark models.

4.2.4. Benchmark models

To illustrate the superiority of the proposed EA-BILSTM, nine models are applied as the benchmarks. The benchmarks include the traditional econometric model such as ARIMA model, four machine learning models, K nearest neighbor (KNN), SVR, NNs, extreme learning machine (ELM), and four deep learning models, LSTM, BILSTM, attention-based LSTM (A-LSTM), attention-based BILSTM (A-BILSTM). These four deep learning models are respectively used to build their ensemble models, denoted as E-LSTM, E-BILSTM, EA-LSTM, EA-BILSTM.

4.2.5. Inputs of models

To prove the effect of holiday-related variables and future available information, different inputs are constructed for models. Firstly, the historical demand data is the basic input variable. Univariate ARIMA model only uses historical demand information to predict the future demand. For building ARIMA model, characteristic analysis of time series is necessary. First, the nonstationarity of the teleconsultation demand is clearly visible due to the weekly periodic pattern in Figure 4. To further test the nonstationarity of the time series, unit root test is applied. It supposes the time series has a unit root (null hypothesis) and the time series is nonstationary. The test result shows teleconsultation demand data is nonstationary ($p = 0.243$). Because of the nonstationarity, a first difference with 7 lags is applied to the demand series to make it stationary. Second, periodicity can also influence the parameters of ARIMA and the length of the historical data in the inputs of other multivariate models. According to the plots in Figure 3 and the autocorrelation and partial autocorrelation analysis, teleconsultation data had the 7-day cycle. Thus, the maximum time lag in ARIMA is set to 7 and the length of historical data in inputs is set to 7.

For other multivariate models, input dimensions can include different numbers of variables. One group of variables are the historical demand volume and the two commonly used date variables (week and holiday). Another group of variables are the historical demand volume and seven selected variables in Section 4.2.1. Those two groups of variables only use historical information. However, some information at the forecasted time or after the forecasted time are available, like the holiday information. Therefore, the date information of the T day (the forecasted time) and T + 1 day are added into the input data, according to the coefficient significance in Table 1. In this 9-day length input matrix, demands on T day and T + 1 day are replaced by 0 because they are unknown on day T − 1, which introduces useless information into the input. To use future information, ensemble deep learning models have two input matrixes. One is the 7-day length history information and another is the 9-day length holiday-related information. Input data of different models are summarized in Table 3.

5. EMPIRICAL RESULTS

Due to the significant pre-effect of holidays on teleconsultation demand, holiday-related variables are selected and EA-BILSTM is constructed for the accurate demand forecast. To verify the effectiveness of the selected variables and the proposed model, actual teleconsultation demand data collected in NTCC are used as the sample data. In the empirical studies, traditional ARIMA model, machine learning models, single deep learning models are applied as benchmarks. In the two groups of models (machine learning models and single deep learning models), Model-1 represents input-1 is introduced into the corresponding model, Model-2 represents input-2 is introduced into the corresponding model, and Model-3 represents input-3 is introduced into the corresponding model. To compare the performance of models, RMSE and MAE are measured as shown in Table 4.

Table 2 Missing data treatment methods and results.

Missing Data Treatment			LSTM Performance	
DMAR	**CMAR**	**MCAR**	**RMSE**	**MAE**
Deletion	Deletion	LSTM imputation	14.37	11.57
0 imputation	0 imputation	LSTM imputation	<u>14.27</u>	<u>11.33</u>
Deletion	0 imputation	LSTM imputation	15.76	12.55
0 imputation	Deletion	LSTM imputation	14.69	11.45

Note: The underlines are the best performance of each column.
LSTM, long short-term memory; DMAR, discrete missingness at random; CMAR, continuous missingness at random; MCAR, missing completely at random; RMSE, root mean square error; MAE, mean absolute error.

Table 3 Different inputs of forecasting models.

Model	**Input**						
	Size			**Variables**			
		Length (Days)	**Number of Variables**	**Demand**	**Week**	**Holiday**	**Other Five Holiday-Related Variables**[a]
Single model 1	Input 1	7 (T−7 to T−1)[b]	3	√	√	√	
Single model 2	Input 2	7 (T − 7 to T − 1)	8	√	√	√	√
Single model 3	Input 3[c]	9 (T − 7 to T + 1)	8	√	√	√	√
Ensemble model	Input 2	7 (T − 7 to T − 1)	8	√	√	√	√
	Input 4	9 (T − 7 to T + 1)	7		√	√	√

(a) Other five holiday-related variables include length of holiday, the second day before holiday, the day before holiday, the day after holiday, adjusted as working day.
(b) The demand on day T was forecasted.
(c) The demand of T and T + 1 is replaced by 0 in the input-3 matrix.

Table 4 Prediction performance on three datasets.

Model	**365-Day Dataset**		**548-Day Dataset**		**699-Day Dataset**	
	Performance		**Performance**		**Performance**	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
ARIMA	19.47	12.07	15.95	11.66	15.72	10.83
KNN-1	20.95	12.89	19.87	13.85	16.14	11.57
SVR-1	18.52	12.77	18.83	13.78	14.82	11.42
NN-1	18.67	12.75	20.41	14.87	20.17	14.25
ELM-1	19.21	12.74	20.66	14.36	15.91	11.19
KNN-2	19.21	13.07	19.78	13.98	16.25	11.90
SVR-2	18.18	13.30	17.44	13.68	15.37	12.26
NN-2	19.58	12.45	19.13	13.74	16.02	11.26
ELM-2	23.80	14.71	21.72	15.01	19.78	13.04
KNN-3	15.50	11.30	16.00	**11.77**	**14.22**	**10.37**
SVR-3	15.77	12.49	**15.99**	12.77	14.42	11.77
NN-3	**14.67**	**10.52**	16.00	11.82	14.32	10.75
ELM-3	18.57	13.09	17.53	13.05	17.63	11.96
LSTM-1	13.62	10.23	19.08	13.16	15.95	11.07
BILSTM-1	13.82	10.27	18.13	13.01	15.70	10.91
A-LSTM-1	14.70	10.20	19.27	12.88	15.83	10.88
A-BILSTM-1	14.26	9.51	16.67	11.28	17.50	10.87
LSTM-2	17.77	12.12	18.02	12.31	14.33	10.24
BILSTM-2	15.01	10.48	16.46	12.02	14.97	10.07
A-LSTM-2	14.40	9.99	17.02	11.53	14.32	10.02
A-BILSTM-2	13.83	9.10	**15.36**	11.12	**13.55**	**9.72**
LSTM-3	14.55	10.34	**15.36**	**11.09**	14.07	10.65
BILSTM-3	14.93	10.36	15.69	11.57	13.66	10.08
A-LSTM-3	15.42	9.93	15.96	11.34	13.97	10.00
A-BILSTM-3	14.34	9.60	15.47	11.12	13.98	9.87
E-LSTM	15.95	10.74	15.98	11.83	13.40	9.82
E-BILSTM	**14.32**	**9.71**	15.51	11.38	13.67	9.93
EA-LSTM	14.94	10.06	16.28	11.46	13.86	9.86
EA-BILSTM	14.85	9.81	15.06	10.60	13.02	9.45

Notes: The bolds are the best performance in each group of models (machine learning models, single deep learning models, and ensemble deep learning models). The underlines are the best performance of each column.
LSTM, long short-term memory; RMSE, root mean square error; MAE, mean absolute error; ARIMA, autoregressive integrated moving average; KNN, K nearest neighbor; SVR, support vector regression; NN, neural network; ELM, extreme learning machine; BILSTM, bidirectional long short-term memory; A-LSTM, attention-based LSTM; A-BILSTM, attention-based BILSTM.

Forecast performance changes with the different models, inputs, and datasets. From model perspective, there are three main important conclusions. First, ARIMA can outperform some single models but can't outperform ensemble models in term of the values of RMSE and MAE. For example, ARIMA has lower MAE than that of SVRs and ELMs, but it has higher MAE than EA-LSTMs and EA-BISLTMs. Second, all single deep learning models outperform the machine learning models. Because the best performance in machine learning models is inferior to that of single deep learning models. In the group of machine learning models, the lowest MAEs in three datasets are 10.52, 11.77, and 10.37. But in the group of single deep learning models, the lowest MAEs in three datasets are 9.10, 11.07, and 9.72. Third, the EA-BISLTM needs enough data quantity to ensure its outperformance. When data quantity increases to 548 and 699, EA-BILSTM obtained the lowest RMSE and MAE.

From input perspective, there are also three main important conclusions. First, introducing more information into the forecasting models can improve prediction performance. In machine learning models, the best performances on three datasets are achieved by introducing Input-3. Similarly, in single deep learning models, the best performances are achieved by introducing Input-2 and Input-3. Second, introducing useless information into deep learning models cannot improve forecasting performance. When Input-3 is introduced, some single deep learning models cannot obtain better performance. Third, introducing useful information and avoiding useless information into prediction models can improve forecasts. In particular, EA-BISLTM can obtain best performance on both 548-day dataset and 699-day dataset. From data perspective, complex structural model is more advantageous when data quantity is large enough for its training. For example, the EA-BISLTM model cannot obtain best performance on 365-day dataset but it can obtain best performance on 548-day dataset and 699-day dataset.

To identify the significance of the performance difference, DM test was applied on MAE. The DM test results are shown in Tables 5–7. In the test statistics, ±1.65 are used for the thresholds to identify the 0.1 significance level. Negative statistics mean the MAE of benchmark models are inferior to that of testing models, and benchmark models outperform testing models. Positive statistics mean test models outperform benchmark models.

DM test results in Table 5 shows that most MAE difference between models are nonsignificant on 365-day dataset. First, ARIMA model has a non-significantly different MAE compared to other models. Second, MAE of most machine learning models aren't significantly different to that of deep learning models when introducing the same input. Besides, ensemble learning models significantly outperform only few machine learning models and single deep learning models. The main significant differences exist between SVR and other models. When using input-3 and introducing more information, KNN and NN can achieve significantly better performance. A-BILSTM-2, achieving the best MAE, is significantly better than LSTM-2, BILSTM-2, ALSTM-2, E-LSTM, and E-ALSTM. The comparison of A-BILSTM-2 and BILSTM-2 demonstrates attention mechanism can significantly improve forecasts.

When data quantity increased to 548-day dataset, as shown in Table 6, the number of MAE significant differences between models increases to 157, which is more than that of 365-day dataset. The significant difference take place in the comparison of ARIMA with machine learning models, deep learning models with machine learning models, and ensemble learning models with machine learning models. For example, the MAE of ARIMA is 11.6, which is significantly better than that of 9 machine learning models and 4 deep learning models. In addition, 7 single deep learning models and 2 ensemble learning models significantly outperform machine learning models when the input of machine learning models is Input-1. When machine learning models use Input-2 as model input, 8 single deep learning models and 4 ensemble models significantly outperform them. Among deep learning models, significant differences of MAE exist in 6 couples of models. The performance improvement is enhanced by introducing attention mechanism in

Table 5 DM test results of MAE for 365-day dataset.

Benchmark	Test models																											
models	K-1	S-1	N-1	E-1	K-2	S-2	N-2	E-2	K-3	S-3	N-3	E-3	L-1	BI-1	AL-1	ABI-1	L-2	BI-2	AL-2	**ABI-2**	L-3	BI-3	AL-3	ABI-3	EL	EBI	EAL	EABI
A	-0.52	-0.80	-0.43	-1.19	-0.56	-0.94	-0.15	-0.92	0.66	-0.38	1.36	-0.83	0.45	1.15	1.49	1.37	-0.02	0.78	0.85	1.25	1.26	1.21	0.99	1.33	0.57	1.28	1.02	1.09
K-1		0.11	0.08	0.14	-0.17	-0.27	0.19	-1.00	1.05	0.20	1.33	-0.10	0.88	1.26	1.61	1.43	0.27	0.85	0.89	1.17	1.29	1.09	1.18	1.06	0.67	1.12	0.97	1.05
S-1			0.02	0.04	-0.28	-0.92	0.17	-0.86	2.06	0.28	2.16	-0.19	2.80	2.35	3.72	2.21	0.29	1.17	1.19	1.61	2.38	1.52	1.62	1.47	0.87	1.59	1.37	1.46
N-1				0.01	-0.29	-0.69	0.23	-1.01	1.45	0.18	1.87	-0.19	1.13	2.12	2.05	1.96	0.39	1.24	1.12	1.65	1.46	1.16	1.42	1.36	0.91	1.52	1.37	1.53
E-1					-0.24	-0.47	0.13	-0.87	1.16	0.17	1.39	-0.19	1.01	1.46	1.56	1.65	0.23	0.91	0.96	1.32	1.32	1.45	1.16	1.27	0.71	1.28	1.10	1.20
K-2						-0.25	0.49	-1.18	1.66	0.33	1.96	-0.01	1.26	1.79	2.62	1.84	0.53	1.32	1.17	1.59	1.60	1.32	1.45	1.35	0.97	1.50	1.36	1.45
S-2							0.75	-0.64	3.25	0.87	4.08	0.16	12.34	3.98	4.15	3.01	0.81	2.28	1.80	2.39	3.80	2.13	2.17	2.39	1.55	2.44	2.00	2.10
N-2								-0.99	0.80	-0.02	1.30	-0.31	1.19	1.38	1.32	1.55	0.29	1.30	1.13	1.65	1.05	0.90	1.17	1.13	0.90	1.29	1.19	1.22
E-2									1.43	0.73	1.79	0.66	1.49	1.60	1.78	1.63	0.91	1.29	1.20	1.46	1.49	1.34	1.41	1.33	1.08	1.40	1.33	1.28
K-3										-1.42	1.03	-1.56	1.32	1.08	1.33	1.48	-0.57	0.68	0.75	1.42	1.19	0.83	1.22	1.09	0.37	1.35	0.81	0.83
S-3											2.32	-0.48	6.52	2.42	1.81	2.60	0.22	2.21	2.01	2.90	3.12	2.84	2.95	2.98	1.53	4.02	2.00	2.19
N-3												-2.32	0.27	0.23	0.22	0.69	-1.24	0.03	0.31	0.91	0.21	0.12	0.44	0.62	-0.15	0.64	0.16	0.25
E-3													1.76	1.47	1.19	1.63	0.60	1.78	1.54	2.05	2.14	2.08	1.94	2.06	1.24	2.05	1.74	1.61
L-1														-0.04	0.07	0.75	-1.09	-0.20	0.18	0.89	-0.28	-0.12	0.24	0.59	-0.37	0.53	0.16	0.38
BI-1															1.69	1.05	-1.11	-0.16	0.18	0.84	-0.10	-0.06	0.24	0.45	-0.34	0.47	0.17	0.43
AL-1																1.11	-1.03	-0.18	0.13	0.69	-0.58	-0.12	0.15	0.41	-0.31	0.34	0.54	0.95
ABI-1																	-1.38	-0.63	-0.33	0.34	-0.78	-0.53	-0.30	-0.07	-0.85	-0.17	-0.54	-0.54
L-2																		1.82	1.35	2.01	0.98	0.95	1.99	1.32	1.29	1.65	2.24	1.93
BI-2																			0.62	1.92	0.11	0.10	1.20	0.81	-0.87	0.98	0.50	1.04
AL-2																				2.37	-0.24	-0.27	0.07	0.38	-1.54	0.34	-0.35	0.21
ABI-2																					-0.85	-0.95	-1.45	-0.57	-6.89	-1.03	-2.09	-1.64
L-3																						-0.03	0.30	0.94	-0.26	0.58	0.24	0.39
BI-3																							0.34	1.16	-0.27	0.67	0.27	0.42
AL-3																								0.34	-2.14	0.43	-0.37	0.61
ABI-3																									-1.05	-0.18	-0.66	-0.21
EL																										1.47	0.80	1.76
EBI																											-0.85	-0.25
EAL																												0.74

Notes: (1) Model abbreviations: A: ARIMA, K-1: KNN-1, S-1: SVR-1, N-1: NN-1, E-1: ELM-1, K-2: KNN-2, S-2: SVR-2, N-2: NN-2, E-2: ELM-2, K-3: KNN-3, S-3: SVR-3, N-3: NN-3, E-3: ELM-3, L-1: LSTM-1, BI-1: BILSTM-1, AL-1: A-LSTM-1, ABI-1: A-BILSTM-1, L-2: LSTM-2, BI-2: BILSTM-2, AL-2: A-LSTM-2, ABI-2: A-BILSTM-2, L-3: LSTM-3, BI-3: BILSTM-3, AL-3: A-LSTM-3, ABI-3: A-BILSTM-3, EL: ELSTM, EBI: E-BILSTM, EA-L: EALSTM, EABI: EA-BILSTM. (2) The underlines are out of range (-1.65, 1.65), meaning the MAE difference are significant. (3) The bold model gets the lowest MAE. (4) Table 6 and Table 7 have the same notes with Table 5.

the comparison of BILSTM-1 and A-BILSTM-1, BILSTM-2, and A-BILSTM-2. In the comparison of different inputs, the effect of Input-1 and Input-2 on deep learning models are only significantly different on A-BILSTM model. Input-3 doesn't have significant better effect on deep learning models compared to that of input-2 and input-1. In all models, EA-BILSTM achieved the best performance, and it significantly outperforms 21 models.

When data quantity further increased to 699 observations, the number of significant MAE differences between models is 234 in Table 7, which is more than that in Table 6. In detail, the significant differences between ARIMA and other models decrease. ARIMA significantly outperforms three machine learning models, while three models, A-BILSTM-2, EA-LSTM, and EA-BILSTM, significantly outperform ARIMA. Furthermore, the significant

Table 6 DM test results of MAE for 548-day dataset.

Benchmark	Test models																											
models	K-1	S-1	N-1	E-1	K-2	S-2	N-2	E-2	K-3	S-3	N-3	E-3	L-1	BI-1	AL-1	ABI-1	L-2	BI-2	AL-2	ABI-2	L-3	BI-3	AL-3	ABI-3	EL	EBI	EAL	**EABI**
A	-2.90	-2.83	-2.71	-2.56	-2.52	-3.00	-1.91	-2.08	-1.10	-1.75	-1.04	-1.63	-2.49	-2.33	-1.70	0.93	-1.72	-1.23	0.79	0.66	0.43	0.93	0.73	0.81	-1.21	0.98	0.89	1.13
K-1		0.14	-0.87	-0.73	-0.15	0.18	0.09	-0.65	1.52	0.71	1.29	0.72	1.12	1.29	0.95	2.62	1.63	1.37	1.79	1.89	1.74	1.52	1.67	1.78	1.50	1.55	1.73	2.00
S-1			-0.97	-0.84	-0.22	0.15	0.04	-0.66	1.64	0.75	1.45	0.72	1.43	1.25	1.28	3.03	2.36	1.50	1.99	2.10	1.92	1.83	2.03	2.08	1.65	1.72	1.98	2.30
N-1				0.48	0.85	1.11	1.32	-0.10	2.29	1.43	2.31	1.28	1.54	1.74	1.75	3.22	2.38	2.84	3.96	3.33	2.71	2.72	2.60	2.85	2.60	2.47	2.70	3.29
E-1					0.49	0.63	0.59	-0.35	1.85	0.98	1.62	1.09	1.53	1.84	1.64	2.63	2.20	1.71	2.31	2.18	2.16	1.80	1.94	1.98	1.74	1.86	2.01	2.25
K-2						0.34	0.26	-0.60	2.36	0.93	1.77	0.89	1.07	1.02	1.19	2.63	2.57	2.01	2.91	2.75	2.42	2.01	2.24	2.18	2.02	2.20	2.61	2.48
S-2							-0.07	-0.78	2.59	1.29	2.64	0.82	0.60	0.69	1.01	3.78	3.52	2.47	3.53	3.52	2.79	4.16	4.23	3.66	3.32	3.08	4.18	4.03
N-2								-1.07	1.65	0.75	1.55	0.79	0.45	0.56	0.76	2.05	1.45	1.95	3.08	2.36	2.19	2.06	2.09	1.95	2.01	2.20	2.09	2.47
E-2									1.64	1.12	1.77	1.04	0.95	1.04	1.02	2.08	1.43	1.99	2.32	2.13	1.93	1.99	2.00	1.99	1.98	1.95	1.95	2.47
K-3										-1.80	-0.08	-1.30	-1.15	-0.99	-1.08	0.48	-0.96	-0.35	0.39	1.41	1.50	0.29	0.74	0.76	-0.10	0.75	1.43	1.74
S-3											10.39	-0.27	-0.26	-0.16	-0.09	1.39	0.52	0.89	1.60	2.60	2.61	2.77	2.97	2.63	2.09	2.83	3.54	3.33
N-3												-1.20	-0.86	-0.80	-0.87	0.48	-0.49	-0.23	0.35	0.92	1.20	0.35	0.61	0.73	-0.01	0.64	0.58	1.86
E-3													-0.09	0.03	0.26	1.35	1.56	0.98	1.67	1.96	1.84	1.51	1.76	1.85	1.32	1.66	1.71	2.11
L-1														0.44	0.35	2.14	1.09	0.94	1.35	1.50	1.41	1.11	1.32	1.41	0.98	1.18	1.37	1.73
BI-1															0.14	2.47	0.83	0.81	1.22	1.40	1.35	1.12	1.32	1.44	0.91	1.03	1.23	1.71
AL-1																1.47	1.22	0.74	1.29	1.78	1.73	1.08	1.27	1.36	0.89	1.26	1.39	1.69
ABI-1																	-1.25	-0.77	-0.27	0.02	0.16	-0.35	-0.07	0.08	-0.56	-0.08	-0.19	0.78
L-2																		0.45	1.07	1.65	1.45	0.83	1.23	1.21	0.66	1.17	1.45	1.62
BI-2																			1.15	1.71	1.09	0.80	1.25	1.09	0.42	0.90	1.30	2.01
AL-2																				0.89	0.61	-0.06	0.27	0.30	-0.54	0.26	0.14	1.72
ABI-2																					0.32	-0.47	-0.13	-0.09	-1.16	-0.23	-0.70	0.93
L-3																						-0.68	-0.41	-0.28	-1.05	-0.46	-0.62	0.55
BI-3																							0.66	0.69	-0.51	0.28	0.19	1.96
AL-3																								0.14	-0.98	-0.07	-0.24	1.37
ABI-3																									-0.93	-0.13	-0.36	2.41
EL																										0.85	0.96	1.83
EBI																											-0.17	1.08
EAL																												2.61

Notes: (1) Model abbreviations: A: ARIMA, K-1: KNN-1, S-1: SVR-1, N-1: NN-1, E-1: ELM-1, K-2: KNN-2, S-2: SVR-2, N-2: NN-2, E-2: ELM-2, K-3: KNN-3, S-3: SVR-3, N-3: NN-3, E-3: ELM-3, L-1: LSTM-1, BI-1: BILSTM-1, AL-1: A-LSTM-1, ABI-1: A-BILSTM-1, L-2: LSTM-2, BI-2: BILSTM-2, AL-2: A-LSTM-2, ABI-2: A-BILSTM-2, L-3: LSTM-3, BI-3: BILSTM-3, AL-3: A-LSTM-3, ABI-3: A-BILSTM-3, EL: ELSTM, EBI: E-BILSTM, EA-L: EALSTM, EABI: EA-BILSTM. (2) The underlines are out of range (-1.65, 1.65), meaning the MAE difference are significant. (3) The bold model gets the lowest MAE.

Table 7 DM test results of MAE for 699-day dataset.

Benchmark	Test models																											
models	K-1	S-1	N-1	E-1	K-2	S-2	N-2	E-2	K-3	S-3	N-3	E-3	L-1	BI-1	AL-1	ABI-1	L-2	BI-2	AL-2	ABI-2	L-3	BI-3	AL-3	ABI-3	EL	EBI	EAL	**EABI**
A	-1.06	-0.79	-4.00	-0.70	-1.40	-2.01	-0.77	-2.09	0.49	-0.95	0.24	-1.01	-0.34	-0.12	-0.10	-0.04	0.56	1.40	1.22	1.70	1.02	0.97	1.42	1.59	1.21	1.20	1.69	1.95
K-1		0.26	-3.31	0.67	-0.62	-1.06	0.65	-1.48	1.63	-0.23	1.37	-0.39	0.92	1.34	0.78	1.06	1.58	2.31	2.36	2.29	1.96	1.97	2.43	2.84	2.11	2.41	2.51	2.75
S-1			-3.43	0.53	-0.82	-2.71	0.30	-1.54	3.05	-0.80	1.41	-0.63	1.36	5.58	0.73	1.13	2.16	2.84	5.96	6.80	3.34	4.05	6.04	13.63	3.93	4.68	6.30	8.52
N-1				5.56	3.50	3.02	3.53	1.92	3.82	2.50	6.08	5.24	5.34	5.12	6.09	6.25	4.21	4.52	5.18	4.83	3.97	5.65	5.15	4.70	5.00	4.56	5.26	5.05
E-1					-1.31	-2.58	-0.13	-2.35	1.13	-0.77	0.92	-0.98	0.38	2.41	1.22	0.90	2.17	2.07	2.35	2.41	1.86	1.90	2.51	2.46	2.02	2.08	2.41	2.91
K-2						-0.74	1.25	-1.28	2.44	0.17	2.29	-0.07	2.29	3.28	1.98	2.18	3.08	3.49	4.46	3.32	1.63	1.85	4.08	3.27	3.14	2.90	3.48	3.36
S-2							1.81	-0.73	5.77	1.26	4.28	0.45	7.61	4.94	3.36	4.74	7.25	5.29	6.15	13.98	5.92	4.06	9.49	9.68	10.47	6.21	9.03	12.32
N-2								-2.07	1.21	-0.63	0.85	-0.74	0.33	0.62	0.23	0.51	2.59	4.48	2.56	2.58	1.77	1.85	3.37	2.23	2.05	3.17	2.57	3.09
E-2									2.03	0.91	2.19	1.17	2.30	3.91	3.04	3.17	2.40	2.73	2.83	2.82	2.40	2.38	2.61	2.46	2.32	2.52	2.56	2.67
K-3										-9.97	-0.47	-1.91	-1.16	-1.13	-0.74	-0.87	0.24	0.43	1.00	2.47	-0.65	0.91	0.99	1.26	1.47	1.08	1.61	2.87
S-3											2.24	-0.21	1.23	1.91	1.40	1.89	2.37	3.10	5.96	11.44	6.40	8.88	6.18	5.79	5.45	5.98	8.94	10.09
N-3												-2.56	-2.38	-1.01	-0.45	-1.21	0.98	1.03	1.78	1.80	1.79	1.45	1.28	1.89	1.65	1.47	2.10	2.01
E-3													1.66	1.94	1.92	3.03	2.15	1.89	2.86	2.76	1.71	2.67	2.36	2.10	2.74	2.66	2.84	2.90
L-1														0.82	0.69	0.56	1.73	1.54	2.24	2.56	0.66	2.02	2.06	2.12	2.11	2.14	2.23	2.89
BI-1															0.09	0.15	1.71	1.39	2.33	3.12	0.50	2.45	2.41	2.23	2.44	2.25	3.03	3.42
AL-1																0.02	1.42	1.45	1.70	2.06	0.32	1.35	1.65	1.76	1.47	1.48	1.67	2.35
ABI-1																	0.91	0.89	1.74	2.09	0.42	1.68	1.34	1.44	1.83	1.65	1.91	2.30
L-2																		0.63	1.07	1.33	-0.60	0.37	1.11	1.13	0.77	0.82	0.97	2.34
BI-2																			0.10	0.59	-0.65	-0.02	0.23	0.58	1.26	0.42	1.67	1.70
AL-2																				1.57	-1.35	-0.24	0.18	0.55	0.66	0.66	0.77	3.40
ABI-2																					-2.04	-1.42	-2.09	-0.35	-1.26	-1.36	-1.06	1.56
L-3																						1.81	1.27	1.32	2.13	1.60	1.95	2.76
BI-3																							0.29	0.61	0.79	0.60	1.09	2.54
AL-3																								0.54	0.27	0.26	0.46	2.68
ABI-3																									0.18	-0.22	1.19	1.87
EL																										-0.12	-0.25	2.18
EBI																											0.26	2.62
EAL																												1.73

Notes: (1) Model abbreviations: A: ARIMA, K-1: KNN-1, S-1: SVR-1, N-1: NN-1, E-1: ELM-1, K-2: KNN-2, S-2: SVR-2, N-2: NN-2, E-2: ELM-2, K-3: KNN-3, S-3: SVR-3, N-3: NN-3, E-3: ELM-3, L-1: LSTM-1, BI-1: BILSTM-1, AL-1: A-LSTM-1, ABI-1: A-BILSTM-1, L-2: LSTM-2, BI-2: BILSTM-2, AL-2: A-LSTM-2, ABI-2: A-BILSTM-2, L-3: LSTM-3, BI-3: BILSTM-3, AL-3: A-LSTM-3, ABI-3: A-BILSTM-3, EL: ELSTM, EBI: E-BILSTM, EA-L: EALSTM, EABI: EA-BILSTM. (2) The underlines are out of range (-1.65, 1.65), meaning the MAE difference are significant. (3) The bold model gets the lowest MAE.

Table 8 EA-BILSTM performance under different cross-validation ratios on 365-day dataset.

Cross-Validation Ratio	0.0	0.1	0.2	0.3	0.4
RMSE	14.85	13.98	13.35	14.34	14.42
MAE	9.81	9.54	9.49	9.70	9.94

Note: The underlines represent the best performance.

differences of machine learning models are got from comparison of different forecasting technologies. In addition, KNN-3 and NN-3 can obtain significantly better performance than other machine learning models. Single deep learning models can achieve significantly better MAE than machine learning models. Among deep learning models, A-BILSTM-2 get lowest MAE (i.e., 9.72), which is totally significantly better than 19 models. Among ensemble deep learning models, EA-BILSTM performs the best, which is significantly better than all other models except A-BILSTM-2. Totally, EA-BILSTM significantly outperforms 27 models, more than that of A-BILSTM-2. This infers that EA-BILSTM outperforms A-BILSTM-2.

6. DISCUSSIONS

From above empirical results, the forecasting performance is affected by forecasting technology, data quantity, and input information. The impact of forecasting technology on prediction performance, caused by the algorithm principle, is inherent. One technology has its own advantages and disadvantages in forecasting tasks. For example, deep learning models can achieve better forecasting performance on stock data [44], but they get inferior forecasting performance of them on air pollution data [46]. As mentioned in Section 3.1, BILSTM is suitable for teleconsultation demand forecast for two reasons. On the one hand, the holidays are sparse distributed and one-year cycled, which requires the model to keep learning results for a long time. On the other hand, the holidays have pre-effect on the demand and the learning is bidirectional in BILSTM. Thus, the impact of forecasting technology on prediction performance isn't discussed here. The impact of data quantity and input information on overall better-performed deep learning models, including single models and ensemble models, are analyzed.

The impact of sample quantity on forecasting results can be explained by the learning level. Usually, sample volume strongly affects predictions, and deep learning performs well when applied to large data [44,47]. While deep learning models often suffers from overfitting problems when training data is insufficient. In this paper, the training set quantity of 365-day dataset is 251, similarly the training set quantity of 548- and 699-day dataset are 379 and 417. The training set quantity of the two latter datasets are more than 365. Because the holiday cycle is 365-day, the advantages of single and ensemble deep learning models are stronger than the training set with more than 365 samples. On 699-day dataset, many models are used to test the dataset with more holidays, which needs powerful holiday effect learning ability. In these models, EA-BLSTM shows the powerful holiday effect learning ability to achieve the best performance.

To improve the forecasting performance of EA-BLSTM on 365-day dataset, different cross-validation ratios are set in the model. The forecasting results are presented in Table 8. Proper cross-validation ratio can make EA-BLSTM better on the 365-day dataset. When cross-validation ratio increases from 0 to 0.2, RMSE and MAE of EA-BLSTM become smaller, reaching the lowest RMSE of 13.35 and MAE of 9.49. When the cross-validation ratio further increases to 0.4, the performance of EA-BLSTM become worse because less data are used for training but more data are used for validation. Therefore, setting proper cross-validation ratio is important for deep learning models in forecast tasks when data quantity is relatively insufficient.

As for the different inputs, introducing more variables means that more parameters are needed to be tuned and adjusted. Usually, large learning task needs enough data volume for training to avoid overfitting problems. In addition, attention-based models need to learn the weights of hidden states, leading to larger learning tasks, but it increases the weights assigned to holiday-related hidden states. This weight adjustment can lead to better forecasting performance near holidays and on holidays. Compared to single models using Input 3, the ensemble models can avoid introducing useless information and exclusively learn the changes of holidays. Compared to single models using Input 2, the ensemble models introduced more available holiday-related information. Therefore, EA-BILSTM achieved the best performance on 548-day dataset and 699-day dataset. Compared to Input 1, effect of Input 2 and Input 3 are better on 548-day dataset and 699-day dataset. It is can be implied that increasing variables under enough data quantity can improve forecasts without overfitting problems.

Although EA-BLSTM could achieve significant better forecasts, its performance was affected by the data quantity without enough robustness. To build more robust forecasting model, the influence of holidays on time series forecast can be investigated from data level or model level. For example, time series decomposition can be applied to redisplay the influence of sparse factors, or hybrid model can be built to take the advantages of different techniques.

From above results and discussions, we can draw the following four main conclusions. (1) Increasing data quantity can make deep learning models significantly better than machine learning models and make ensemble deep learning models significantly outperform deep learning models. (2) Introducing variables selected based on the expanded holiday effect can lead to better forecasts of teleconsultation demand. But introducing future information and useless information together may not lead to better forecasts. (3) In model construction, the involvement of attention mechanism can significantly improve forecasts. (4) By using attention mechanism and only introducing useful future information, EA-BILSTM can get best performance and it can significantly outperform benchmark models on dataset with enough volume, indicating the superiority of the proposed deep learning models.

7. CONCLUSIONS

To improve the efficiency of limited resources, the paper studies the daily teleconsultation demand forecast. In teleconsultation, the demand is significantly affected by holidays. Considering this influence, related variables are selected and an EA-BLSTM model is proposed for accurate forecast. In this advanced method, the ensemble deep learning framework can make full use of all available information and avoid any useless information. And the attention mechanism can increase the weights of holiday-related hidden states in BILSTM. Furthermore, the BILSTM can keep learning results of the holiday pre-effect for a long-time span. Empirical results demonstrate the effectiveness of variable selection, and the superiority of the proposed EA-BLSTM method over benchmarks. It is worth noting that sufficient training data samples are necessary to guarantee the superiority of EA-BLSTM. Despite of this limitation, the ensemble attention-based deep learning model shows high prediction potentiality to deal with the influence of sparseness, like holiday effect, in time series forecasting.

CONFLICTS OF INTEREST

The authors have no conflicts of interest to declare.

AUTHORS' CONTRIBUTIONS

W.C., L.Y. and L.J. developed the idea for the study. W.C. did the experiments. And all authors analysed the results and were involved in writing the manuscript.

ACKNOWLEDGMENTS

This research is supported by General Programs of National Natural Science Foundation of China (71972012) and Key Program of National Natural Science Foundation of China (71432002). Authors are grateful to the staffs in National Telemedicine Center of China who provided help for the study. Also, authors would like to express their sincere appreciation to the editor and the two independent referees in making valuable comments and suggestions to the paper. Their comments have improved the quality of the paper immensely.

REFERENCES

[1] B. Klaassen, B.J.F. van Beijnum, H.J. Hermens, Usability in telemedicine systems—a literature survey, Int. J. Med. Informat. 93 (2016), 57–69.

[2] K. Deldar, K. Bahaadinbeigy, S.M. Tara, Teleconsultation and clinical decision making: a systematic review, Acta Informat. Med. 24 (2016), 286–292.

[3] Y. Qiao, L. Ran, J. Li, Optimization of teleconsultation using discrete-event simulation from a data-driven perspective, Telemed. E-Health. 26 (2020), 114–125.

[4] R.W. Hall, P.J. Dehnel, J.J. Alexander, D.M. Bell, M. Bunik, B.L. Burke, J.A. Kahn, J.R. Kile, Telemedicine: pediatric applications, Pediatrics. 136 (2015), 293–308.

[5] J.E. Given, B.P. Bunting, M.J. O'Kane, F. Dunne, V.E. Coates, Tele-Mum: a feasibility study for a randomized controlled trial exploring the potential for telemedicine in the diabetes care of those with gestational diabetes, Diabetes Technol. Ther. 17 (2015), 880–888.

[6] S.M. Ihorn, P. Arora, Teleconsultation to support the education of students with visual impairments: a program evaluation, J. Educ. Psychol. Consul. 28 (2018), 319–341.

[7] S.E.R. Oest, M.B. Swanson, A. Ahmed, N.M. Mohr, Perceptions and perceived utility of rural emergency department telemedicine services: a needs assessment, Telemed. E-Health. 26 (2019), 855–864.

[8] F. Lopez Segui, J. Vidal-Alaball, M. Sagarra Castro, A. Garcia-Altes, F. Garcia Cuyas, General practitioners' perceptions of whether teleconsultations reduce the number of face-to-face visits in the catalan public primary care system: retrospective cross-sectional study, J. Med. Internet Res. 22 (2020), 1–8.

[9] J.G. Zhang, J.N. Stahl, H.K. Huang, X.Q. Zhou, S.L. Lou, K.S. Song, Real-time teleconsultation with high-resolution and large-volume medical images for collaborative healthcare, IEEE Trans. Inf. Technol. Biomed. 4 (2000), 178–185.

[10] S.A. Erdogan, T.L. Krupski, J.M. Lobo, Optimization of telemedicine appointments in rural areas, Serv. Sci. 10 (2018), 261–276.

[11] F. Lopez Segui, R.A. Egg Aguilar, G. de Maeztu, A. Garcia-Altes, F. Garcia Cuyas, S. Walsh, M. Sagarra Castro, J. Vidal-Alaball, Teleconsultations between patients and healthcare professionals in primary care in catalonia: the evaluation of text classification algorithms using supervised machine learning, Int. J. Environ. Res. Pub. Health. 17 (2020), 1–9.

[12] E.W.-Y. Kwong, H. Wu, G.K.-H. Pang, A prediction model of blood pressure for telemedicine, Health Informat. J. 24 (2018), 227–244.

[13] C. Chakraborty, B. Gupta, S.K. Ghosh, D.K. Das, C. Chakraborty, Telemedicine supported chronic wound tissue prediction using classification approaches, J. Med. Syst. 40 (2016), 1–12.

[14] S. AlDossary, M.G. Martin-Khan, N.K. Bradford, N.R. Armfield, A.C. Smith, The development of a telemedicine planning framework based on needs assessment, J. Med. Syst. 41 (2017), 1–9.

[15] N. Maarop, K.T. Win, Understanding the need of health care providers for teleconsultation and technological attributes in relation to the acceptance of teleconsultation in Malaysia: a mixed methods study, J. Med. Syst. 36 (2012), 2881–2892.

[16] H. Park, Y. Chon, J. Lee, I.-J. Choi, K.-H. Yoon, Service design attributes affecting diabetic patient preferences of telemedicine in South Korea, Telemed. E-Health. 17 (2011), 442–451.

[17] S. Wang, L. Yu, L. Tang, S. Wang, A novel seasonal decomposition based least squares support vector regression ensemble learning approach for hydropower consumption forecasting in China, Energy. 36 (2011), 6542–6554.

[18] M. De Felice, A. Alessandri, P.M. Ruti, Electricity demand forecasting over Italy: potential benefits using numerical weather prediction models, Electr. Power Syst. Res. 104 (2013), 71–79.

[19] Y.-L. Huang, C.-T. Lin, Developing an interval forecasting method to predict undulated demand, Qual. Quant. 45 (2011), 513–524.

[20] H. Liu, Y. Liu, Y. Wang, C. Pan, Hot topics and emerging trends in tourism forecasting research: a scientometric review, Tour. Econ. 25 (2019), 448–468.

[21] P. Egri, Information elicitation for aggregate demand prediction with costly forecasting, Auton. Agent Multi Agent Syst. 30 (2016), 681–696.

[22] D. Yang, S. Li, Z. Peng, P. Wang, J. Wang, H. Yang, MF-CNN: traffic flow prediction using convolutional neural network and multi-features fusion, IEICE Trans. Inf. Syst. E102.D (2019), 1526–1536.

[23] X.L. Luo, D.Y. Li, S.R. Zhang, Traffic flow prediction during the holidays based on DFT and SVR, J. Sensors. 2019 (2019), 1–10.

[24] L. Yu, G. Hang, L. Tang, Y. Zhao, K.K. Lai, Forecasting patient visits to hospitals using a WD&ANN-based decomposition and ensemble model, Eurasia J. Math. Sci. Technol. Educ. 13 (2017), 7615–7627.

[25] B. Klute, A. Homb, W. Chen, A. Stelpflug, Predicting outpatient appointment demand using machine learning and traditional methods, J. Med. Syst. 43 (2019), 1–10.

[26] M. Wargon, B. Guidet, T.D. Hoang, G. Hejblum, A systematic review of models for forecasting the number of emergency department visits, Emerg. Med. J. 26 (2009), 395–399.

[27] F. Kadri, F. Harrou, S. Chaabane, C. Tahon, Time series modelling and forecasting of emergency department overcrowding, J. Med. Syst. 38 (2014), 1–20.

[28] P. Aboagye-Sarfo, Q. Mai, F.M. Sanfilippo, D.B. Preen, L.M. Stewart, D.M. Fatovich, A comparison of multivariate and univariate time series approaches to modelling and forecasting emergency department demand in Western Australia, J. Biomed. Informat. 57 (2015), 62–73.

[29] M. Afilal, F. Yalaoui, F. Dugardin, L. Amodeo, D. Laplanche, P. Blua, Forecasting the emergency department patients flow, J. Med. Syst. 40 (2016), 1–18.

[30] T. Jilani, G. Housley, G. Figueredo, P.S. Tang, J. Hatton, D. Shaw, Short and long term predictions of hospital emergency department attendances, Int. J. Med. Informat. 129 (2019), 167–174.

[31] E. Buckingham-Jeffery, R. Morbey, T. House, A.J. Elliot, S. Harcourt, G.E. Smith, Correcting for day of the week and public holiday effects: improving a national daily syndromic surveillance service for detecting public health threats, BMC Pub. Health. 17 (2017), 1–9.

[32] M.G. Jahromi, R. Goudarzi, V. Yazdi-Feyzabadi, S. Amini, J. Nazari, M. Amiresmaili, Effect of new year holidays on hospital mortality: a time series study, Int. J. Emerg. Med. 12 (2019), 20–20.

[33] F. Qian, C. Han, H.Y. Meng, Intelligent model system for tourism flow prediction: a study of Xi'an Museum, in Proceedings of the 2016 International Conference on Intelligent Information Processing, Wuhan, China, 2016.

[34] L.J. Liu, R.C. Chen, Q.F. Zhao, S.Z. Zhu, Applying a multistage of input feature combination to random forest for improving MRT passenger flow prediction, J. Ambien Intell. Humaniz. Comput. 10 (2019), 4515–4532.

[35] R. Chen, C.-Y. Liang, W.-C. Hong, D.-X. Gu, Forecasting holiday daily tourist flow based on seasonal support vector regression with adaptive genetic algorithm, Appl. Soft Comput. 26 (2015), 435–443.

[36] S.S. Liu, E.J. Yao, Holiday passenger flow forecasting based on the modified least-square support vector machine for the metro system, J. Trans. Eng. Part A-Syst. 143 (2017), 1–8.

[37] X.M. Wang, N. Zhang, Y.L. Zhang, Z.B. Shi, Forecasting of short-term metro ridership with support vector machine online model, J. Adv. Transport. 2018 (2018), 1–13.

[38] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (1997), 1735–1780.

[39] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Trans. Signal Process. 45 (1997), 2673–2681.

[40] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv: Computation and Language, 2014. https://arXiv:1409.0473

[41] J. Xie, B. Chen, X. Gu, F. Liang, X. Xu, Self-attention-based BiLSTM model for short text fine-grained sentiment classification, IEEE Access. 7 (2019), 180558–180570.

[42] G. Wu, G. Tang, Z. Wang, Z. Zhang, Z. Wang, An attention-based BiLSTM-CRF model for chinese clinic named entity recognition, IEEE Access. 7 (2019), 113942–113949.

[43] Y.G. Cinar, H. Mirisaee, P. Goswami, E. Gaussier, A. Aït-Bachir, Period-aware content attention RNNs for time series forecasting with missing values, Neurocomput. 312 (2018), 177–186.

[44] L. Chen, Z. Qiao, M. Wang, C. Wang, R. Du, H.E. Stanley, Which artificial intelligence algorithm better predicts the Chinese stock market?, IEEE Access. 6 (2018), 48625–48633.

[45] Y. Yuan, Multiple Imputation for Missing Data: Concepts and New Development, SAS Institute Inc., Rockville, MD, USA, 2010. http://support.sas.com/rnd/app/stat/papers/multipleimputation.pdf

[46] Y.-S. Chang, S. Abimannan, H.-T. Chiao, C.-Y. Lin, Y.-P. Huang, An ensemble learning based hybrid model and framework for air pollution forecasting, Environ. Sci. Pollut. Res. 27 (2020), 38155–38168.

[47] J. Baek, K. Sohn, Deep-learning architectures to forecast bus ridership at the stop and stop-to-stop levels for dense and crowded bus networks, Appl. Artif. Intell. 30 (2016), 861–885.