IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 27, 2020, accepted July 29, 2020, date of publication August 13, 2020, date of current version August 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016469

Prediction of Traffic Congestion Based on LSTM Through Correction of Missing Temporal and Spatial Data

DONG-HOON SHIN[1], KYUNGYONG CHUNG[2], AND ROY C. PARK[3]
[1]Department of Computer Science, Kyonggi University, Suwon-si 16227, South Korea
[2]Division of Computer Science and Engineering, Kyonggi University, Suwon-si 16227, South Korea
[3]Department of Information Communication Software Engineering, Sangji University, Wonju-si 26339, South Korea

Corresponding author: Roy C. Park (roypark1984@gmail.com)

This work was supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (Grant 20CTAP-C157011-01).

ABSTRACT With the rapid increase in vehicle use during the fourth Industrial Revolution, road resources have reached their supply limit. Active studies have therefore been conducted on intelligent transportation systems (ITSs) to realize traffic management systems utilizing fewer resources. As part of an ITS, real-time traffic services are provided to improve user convenience. Such services are applied to prevent traffic congestion and disperse existing traffic. Therefore, these services focus on immediacy at the expense of accuracy. As these services typically rely on measured data, the accuracy of the models are contingent on the data collection. Therefore, this study proposes a long short-term memory (LSTM)-based traffic congestion prediction approach based on the correction of missing temporal and spatial values. Before making predictions, the proposed prediction method applies pre-processing that consists of outlier removal using the median absolute deviation of the traffic data and the correction of temporal and spatial values using temporal and spatial trends and pattern data. In previous studies, data with time-series features have not been appropriately learned. To address this problem, the proposed prediction method uses an LSTM model for time-series data learning. To evaluate the performance of the proposed method, the mean absolute percentage error (MAPE) was calculated for comparison with other models. The MAPE of the proposed method was found to be the best of the compared models, at approximately 5%.

INDEX TERMS Long short-term memory, traffic, intelligent transportation system, deep learning, missing data correction, big data-based AI.

I. INTRODUCTION

Based on the core technologies of the fourth industrial revolution, smart vehicles are being produced in diverse forms [1]. The role of the automobile has been extended from a simple means of transportation to a living space and finally, to a type of infotainment system that provides new forms of user convenience [2], [3]. With the increase in the demand for smart automobiles, it is extremely important to collect and process traffic information to enable smooth traffic management. Furthermore, it is necessary to take a qualitative rather than a quantitative approach [4]. To this end, research has been conducted on intelligent transportation systems (ITSs) developed in concert with conventional traffic management systems and information technology [5]–[7]. To improve user convenience, studies on the problems that have arisen with the increased demand for road resources have focused on traffic welfare. Traffic welfare consists of factors including operation service costs, passage of time, accident costs, parking costs, punctuality, and accessibility, with the most important being traffic congestion. As a part of an ITS, traffic surfaces can be put in place to collect traffic information on all roads in real time in order to provide users with information including which regions are congested, traffic volumes, and the locations of traffic accidents. In this way, an ITS can improve the functionality of a road traffic network. An ITS can also provide a real-time traffic-information service. By suggesting an optimal path to each driver, road congestion decreases

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Afzal.

and traffic is dispersed. An ITS thus focuses on immediacy but achieves relatively low accuracy. To solve this problem, active research has been conducted on real-time traffic pattern predictions based on deep learning models and multiple prediction modes, with a particular focus on traffic predictions based on time-series data.

Weilin *et al.* [8] proposed a multi-resolution support vector regression (SVR) traffic flow prediction model based on wavelet decomposition and topological space reconstruction. For their experiment, the researchers utilized data collected from January to December 2011 by performance measurement systems, which collect data in 5-min intervals. The mean absolute percentage error (MAPE) rate for their model was 12.8%.

Filmon *et al.* [9] proposed a nonparametric, data-centric methodology to achieve short-term traffic predictions based on the identification of similar traffic patterns through the improved K-nearest neighbor (K-NN) algorithm. Recently, the weighted Euclidean distance has also been used as a similarity measurement for K-NN. For their experiment, the researchers used 12 datasets from highways in the UK and 24 datasets from highways in the US. A MAPE rate of 22% was achieved.

Zhang [10] proposed a short-term traffic prediction model based on a convolutional neural network (CNN) deep learning framework. In the proposed framework, the optimal input data time delay and amount of spatial data are determined based on the space-time feature selection algorithm. The selected space-time traffic feature is then transformed into a two-dimensional matrix after being extracted from the actual data. The function is learned by the CNN, and a prediction model is constructed. According to a performance analysis, the MAPE rate was approximately 8.3% on average.

The methods described above tend to achieve higher prediction accuracies than those focusing on immediacy. The prediction modes used in these studies are based on one of three models: SVR [11], [12], CNNs [13], [14], and KNN [15], [16]. Because these models fail to consider the features of time-series data, they may be inappropriate. For prediction, this study therefore utilizes the long short-term memory (LSTM) model, which provides accurate predictions and makes it possible to account for the time-series features of traffic data. The LSTM model solves the problem of the long-term dependence inherent in recurrent neural network (RNN) models [17]–[19]. With the LSTM model, the result of a hidden layer is passed to the same hidden layer as an input. Owing to the recursive construction of hidden layers, it is possible to consider sequential or temporal aspects. For this reason, this model is conducive for learning the time-series features of traffic data. Traffic data include outliers or missing values due to unexpected traffic variables. Outliers and missing values lower model performance and therefore should be corrected when designing an accurate prediction model. The correction can be achieved by removing outliers, correcting missing temporal and spatial values, and applying pattern data. Then a system can be established to provide the predicted traffic information to users. With more accurate data, it is possible to increase the accuracy of predictions and to provide a smooth flow of traffic information to users [20], [21].

This paper is organized into the following sections. Section 2 describes the relevant studies on ITSs and ITS-based traffic predictions. Section 3 details the data collection process, data pre-processing, and model design for traffic congestion prediction. In Section 4, an experiment conducted to evaluate the model performance and its results are described. In addition, a comparison of different methods used to verify the performance and a description of the system implementation are also provided. Finally, Section 5 presents the concluding remarks regarding this study.

II. RELATED WORK

A. RESEARCH OF TRAFFIC CONGESTION PREDICTION

In traffic data, outliers and missing values negatively influence traffic control and traffic congestion prediction in intelligent traffic systems. To address this problem, many missing value correction methods have been proposed. Conventional methods of missing value correction focus on the correction of individual missing values. Although these methods provide a simple and fast estimate for the missing value, they often produce biased results. To resolve this, historical imputation methods (HIMs) that provide multiple estimation values for one missing value have been proposed [22], [23]. In these methods, a missing value is replaced by the mean value of multiple data points collected at the same position and date. Correction methods based on nearest neighbor imputation (NIM) use the mean value from the neighboring roads to estimate a missing value [24], [25]. However, such methods cannot be applied when there is no data from neighboring roads. The missing value correction method proposed in this study makes it possible to correct a missing value and thereby to design complete data, using past data patterns even when there is no information from neighboring roads. In addition, machine learning and deep learning are applied to model more complicated data for traffic prediction. The deep learning model exhibits better performance since it has more functions and more complicated architecture than the conventional model.

Sun *et al.* [26] proposed a traffic prediction method using GPS trajectory data based on an RNN. Their method used the missing values from existing road speed data to estimate the average speeds on stretches of road with GPS trajectory data. However, because an RNN fails to memorize past data features and deletes them with a lapse in time, it has problems dealing with long-term dependency. Accordingly, traffic prediction based on the LSTM model, which resolves problems associated with RNNs, is actively being researched. Mou *et al.* [27] proposed the temporary information improvement (T-LSTM) model to predict the traffic flow on a single stretch of road. In consideration of the similar features exhibited each day by traffic flows at a given time and place,

the model extracted the unique correlation between the traffic flow and time information, thereby improving the prediction accuracy. Yu *et al.* [28] proposed STGCN to solve the problem of previous studies that had ignored spatial and temporal attributes in traffic prediction. They argued that the method was able to obtain a faster training speed with a smaller number of parameters since it formalized the problem in graph and established a model with a complete convolution structure, rather than applying regular convolution and repetition units. Many researchers have tried to increase the accuracy of traffic predictions and reduce the calculation time through their theories and experiments.

B. RESEARCH ON ITS-BASED TRAFFIC PREDICTION

University of Southern California Information Lab has established spatial and temporal data using sensors for road measurements and traffic information (e.g., CCTV and GPS) and uses real-time data and past traffic data to predict on-road traffic [29]. The extent to which a prediction model established using past data depends on the state of real-time traffic is important, and an important task is to evaluate the extent to which models built using past predictions depend on current status data. It is necessary to overcome the limitation of previous data becoming irrelevant in the model over time. To this end, in the USC model, current traffic information is learned in real time and is used as historical data. The framework can predict traffic at an accuracy comparable to that of the most effective prediction-trained model. Fig. 1 shows the transportation prediction system architecture of USC media. The artificial intelligence (AI)-based transportation prediction system offered by Blue Signal in the Republic of Korea provides road map information and predicts traffic flows and accident risk through big data analysis [30]. An AI-based transportation prediction engine was also developed based on transportation theory. Whereas a conventional GPS service provides information such as routes around traffic jams, the shortest travel time, and the shortest path, the prediction engine of Blue Signal predicts the safest and most convenient route. This engine can achieve 98% accuracy for traffic accident prediction on domestic highways.

As shown in Fig. 1, data are received in real time by the User Interface and Data Interface. Through the adaptive segmentation of the Context Space, the effect of each base

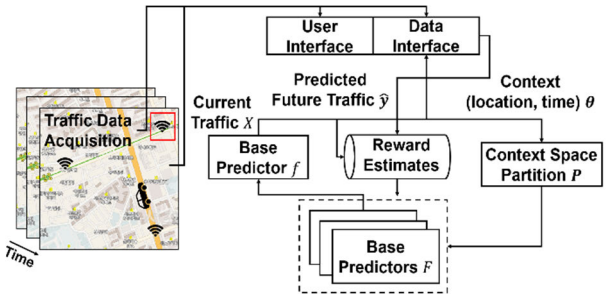

FIGURE 1. Transportation prediction system architecture of USC media.

prediction device is efficiently estimated. In this way, it is possible to predict traffic conditions in diverse situations.

III. PREDICTION OF TRAFFIC CONGESTION BASED ON LSTM THROUGH CORRECTION OF MISSING TEMPORAL AND SPATIAL DATA

The congestion prediction method developed to provide traffic information to users consists of data collection, correction of missing data, and prediction modeling. In this study, the collected data include node/link and traffic speed data provided by an ITS. The node/link data represent a road region or road connection point. The traffic speed data from the ITS are collected by traffic information collectors installed on the roads or along the roadsides. The traffic data include missing values and outliers. An outlier may be generated by an information collection failure, when there are errors in the collectors, or by shaded zones without automobiles travelling in them. The traffic data also include time-series features. For this reason, a missing value makes it difficult to extract the feature values when a deep learning model is used for prediction. Therefore, preprocessing of the outliers and missing values is required [31], [32]. During the data pre-processing, an outlier is processed, and filtering is then applied using the median absolute deviation [33], [34]. Missing data are corrected using spatial trends, temporal trends, and pattern utilization. With the pre-processed data, an LSTM model is used to predict traffic congestion. Fig. 2 shows the entire process of LSTM-based traffic congestion prediction through the correction of temporal and spatial data.

A. OUTLIERS, TYPES OF MISSING VALUES, AND CORRECTION METHODS ACCORDING TO TRAFFIC DATA FEATURES

Traffic speed data include outliers that distort the flow of the average traffic speed and missing values. An outlier represents a value that is either too small or too large in the context of the average traffic flow on each road. Such values are removed to avoid influencing the feature values at the time of prediction. There are two types of missing values. The first type is missing temporal values that occur when not all of the traffic data (which are collected every 5 min) are gathered. The second type is missing spatial values that occur when data are not collected at each road in a given collection interval. Fig. 3 shows the time in a link matrix with examples of an outlier and each type of missing value.

To correct for outliers and missing values in the traffic data, the outlier removal process is first applied. There are a variety of typical outlier removal methods that use, for example, the median absolute deviation, truncated mean, or Winsorized mean. Methods may be combined depending on the features of the roads and traffic data. This study applies the median absolute deviation to identify and remove outliers. That is, the median value of the collected data is used to detect whether a value is abnormally large or small. When a value is identified as an outlier, it is removed. Algorithm 1 shows the outlier removal algorithm. Fig. 4 shows the outlier removal

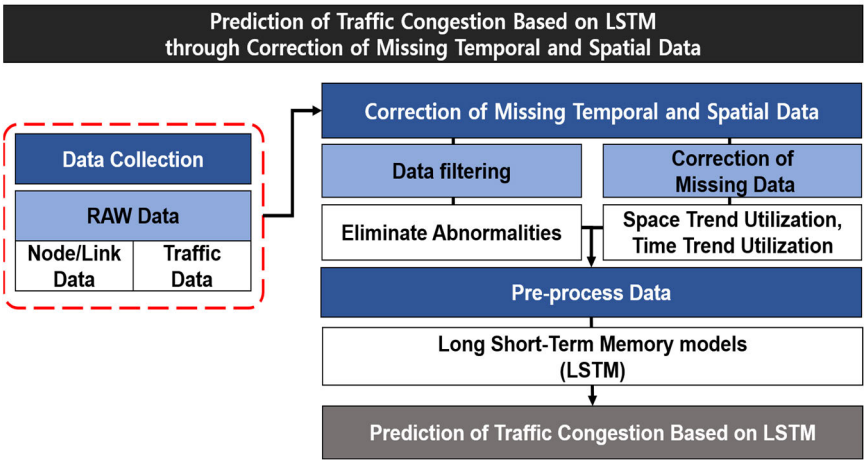

FIGURE 2. Process of LSTM-based traffic congestion prediction through time-space correction.

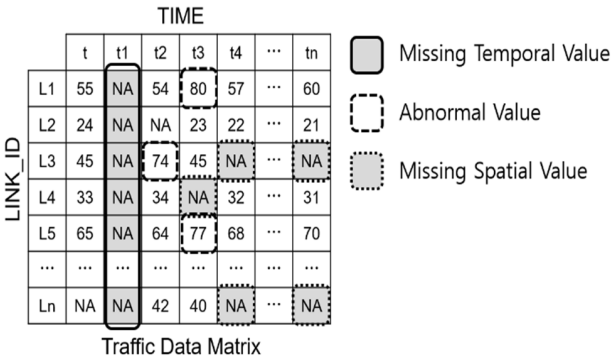

FIGURE 3. Outliers and types of missing values from traffic data.

process using the median absolute deviation. The missing value correction is an algorithm-based filtering process for correcting data that was removed after being identified as outliers. The missing-value correction methods include the application spatial trends from data from regions with a similar traffic pattern, the use of temporal trends to correct the value in question using past data, and the use of pattern data. Each method corrects missing data values from a temporal or spatial perspective.

Algorithm 1 Outlier Removal Algorithm

Input: $[x_1, x_2, \ldots, x_n]$

def *Detection of Outlier*
  $MED \leftarrow Median([x_1, x_2, \ldots, x_n])$
  for x_i *in* $[x_1, x_2, \ldots, x_n]$
    do $x_i' = |x_i - median|$
  $MAD \leftarrow Mean[x_1', x_2', \ldots, x_n']$
  for x_i' *in* $[x_1', x_2', \ldots, x_n']$
  if $\frac{0.6457(x_i - median)}{MAD} >$ ***threshold***
  then $outlier\_set \leftarrow [outlier\_set, x_i]$

Output:$[x_1, x_2, \ldots, x_n]$ - outlier set

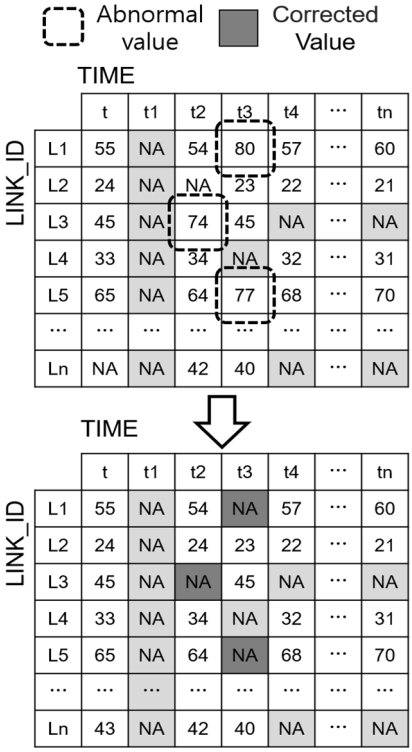

FIGURE 4. Outlier removal process using median absolute deviation.

The spatial trends are used to correct missing values in regions with similar traffic patterns under the assumption that the traffic flow of the upper regions influences that of the lower regions. Algorithm 2 is a missing value correction method based on the use of spatial trends. For instance, if detector x_b has a problem and its data are missing, the mean of the adjoining link data from x_a and x_c is used for the correction. The use of the spatial trend-based missing value correction process is shown in Fig. 5.

If missing data occur at three continuous points, the spatial trend correction is not possible because there are no adjoining

Algorithm 2 Spatial Data Correction Algorithm

Input: X_a (Adjoining Northbound Link) X_b (Target Link)
X_c (Adjoining Southbound Link)

def *Spatial Data Correction*
  if $X_a = Exist$ && $X_b = None$ && $X_c = Exist$
    then $X_b = (X_a + X_c)/2$
  else if $X_a = None$ && $X_b = None$ && $X_c = Exist$
    then $X_b = X_c$
  else if $X_a = Exist$ && $X_b = None$ && $X_c = None$
    then $X_b = X_a$

Output: X_b(Target Link)

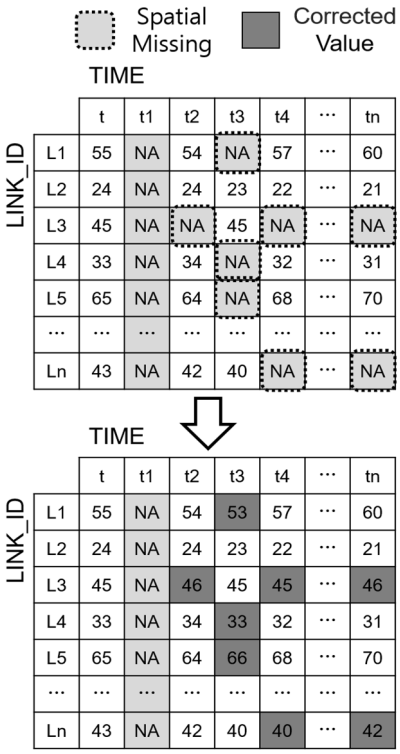

FIGURE 5. Spatial-trend based correction procedure.

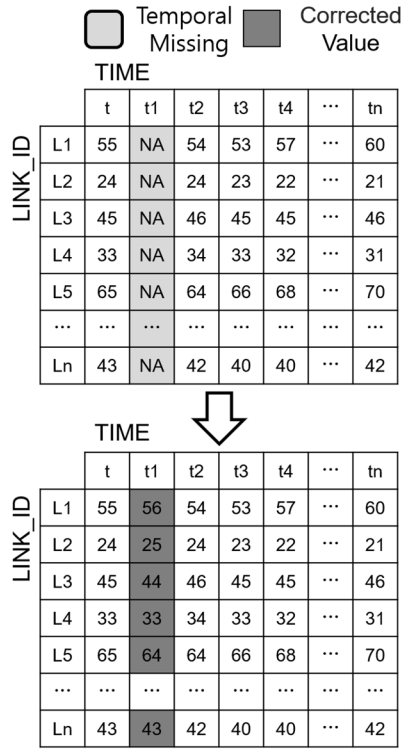

FIGURE 6. Time- trend based correction procedure.

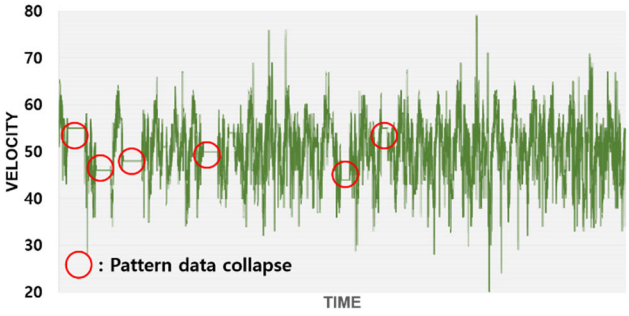

FIGURE 7. Data pattern collapse due to continuous temporal data.

links. In this case, the temporal trend is applied. The temporal method calculates the mean of the n previous observations at missing observations location. Equation 1 shows the correction equation using the temporal trend. In the equation, Ft is the missing value at the current time t and is to be estimated, A_{t-k} is the detected data at time $t - k$, and n is the number of past detected observations. In Fig. 6, the use of the temporal trend-based correction procedure is illustrated. Using the temporal trend, the missing values in the traffic data can be fully corrected. Nevertheless, if there are many sequential missing values, when applying the method, the estimated temporal values are constant, and the data pattern disappears, as shown in Fig. 7.

$$F_t = \frac{A_{t-1} + A_{t-2} + \ldots + A_{t-n}}{n} \quad (1)$$

Therefore, if the temporal trend is not useful, the pattern data are applied. This final method estimates the missing values by applying data collected in the connected parts, such as the data entrance and the entrance access parts. For the pattern data generation procedure, the data from previous days are checked to find the passage features of each day, and the data are saved as one of six types: a special day, Sunday, Saturday, Monday, weekday (Tuesday through Thursday), or Friday. The pattern data of each type are generated every 5 min and are updated by applying a weight to the current collection speed.

B. LSTM-BASED TRAFFIC CONGESTION PREDICTION

For traffic speed prediction, we use time series-based deep learning (LSTM or long-term memory) for modeling [35], [36]. The data used for prediction are pre-processed

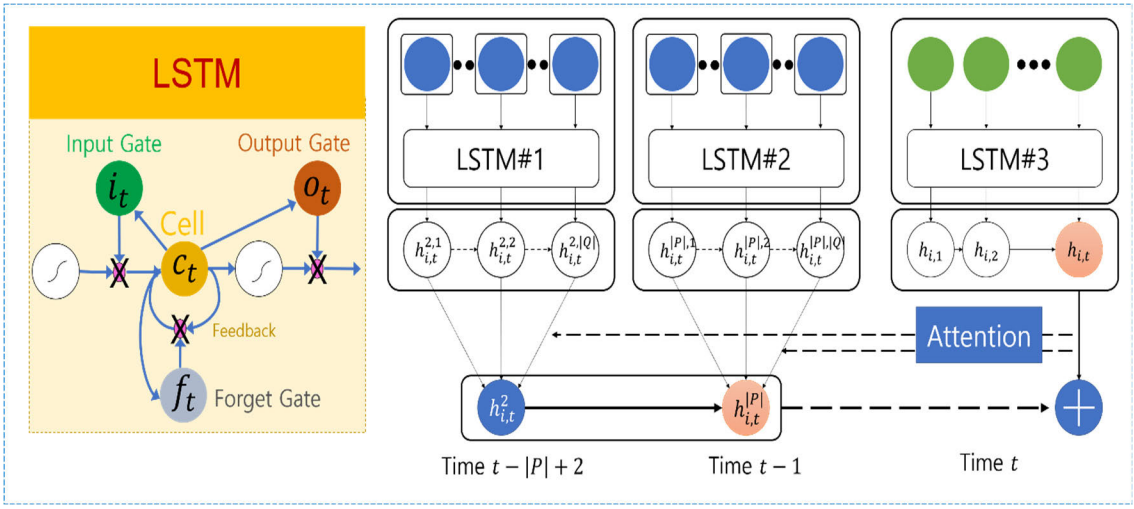

FIGURE 8. LSTM-based traffic prediction process.

using the method described in the previous section. The input data used for modeling are the mean speeds from 10 min and 5 min earlier, the current speed, and the speed of adjoining upper region. The output data is the predicted speed 5 min after the current time. Fig. 8 shows the LSTM-based traffic prediction process proposed in this study. An LSTM cell consists of a memory cell and gates. Input information is saved in the memory cell, and a gate controls the saved information. The parameters of the proposed LSTM model are shown in Table 1. The learning rate is a Hyper parameter in an optimization algorithm that determines the step size at each iteration while moving toward a minimum of a loss function. [37]. Dropout is used to prevent overfitting, which can occur during the learning process [38]. In other words, dropout is used when a model lacks flexibility due to overfitting (which means that the error is small when testing with the learning data and large when testing with the test data) and can therefore not be generalized. The batch size represents the data input to the model concurrently with the training data. The optimization function is an algorithm for updating the weights. The number of hidden neurons and layers, the number of epochs, and the loss function, all of which affect performance, are frequently changed to induce improved performance [39], [40].

TABLE 1. Hyper parameter values.

Hyper parameter	Value
Learning rate	0.001
Dropout	0.005
Batch size	100
Optimization	RMSprop
Epoch	500

IV. EXPERIMENT AND RESULTS

The LSTM-based traffic congestion prediction method proposed in this study was implemented using the following hardware and operating system: Windows 10 Pro, an AMD Ryzen 5 1600 6-Core processor, an NVIDIA GeForce GTX 1070, and 16 GB of RAM. In terms of software, a TensorFlow back-end engine and the deep learning library Keras were used in the design. The traffic speed data used in this study was collected in Gangnam-gu, Seoul during one month in November, 2018. There are a total of 1,630 links in Gangnam-gu, and data was collected at each link [41], [42]. There were a total of 8,640 observations collected at each link according to the collection cycle (5 min $*$ 30 days) and the collection period. Some data were missing; data may have failed to be collected due to a sensor or software error in the process of data collection. The average missing rate of Gangnam-gu traffic speed data is approximately 33%.

A. IMPLEMENTATION OF TRAFFIC CONGESTION PREDICTION SYSTEM

In this study, a system for pre-processing traffic data and a traffic congestion prediction model were established. Fig. 9 shows the pre-processing system for the traffic data. The table in Fig. 9 shows an example of the speed data for all regions in Gangnam-gu. By entering a LINK_ID in the Setting field at the bottom right, selecting a pre-processing method (outlier removal, correction of missing spatial or temporal values, or the use of pattern data), and clicking the Start button, data pre-processing is applied. The pre-processed region data appear in the bottom left of Fig. 9. It is possible to save the pre-processed data by clicking the Save button. region is selected in the region selection window in the top-left of the prediction system. In the LINK overview window, the description of the selected region (LINK_ID, LINK_NAME, Velocity) is displayed. In the data collection

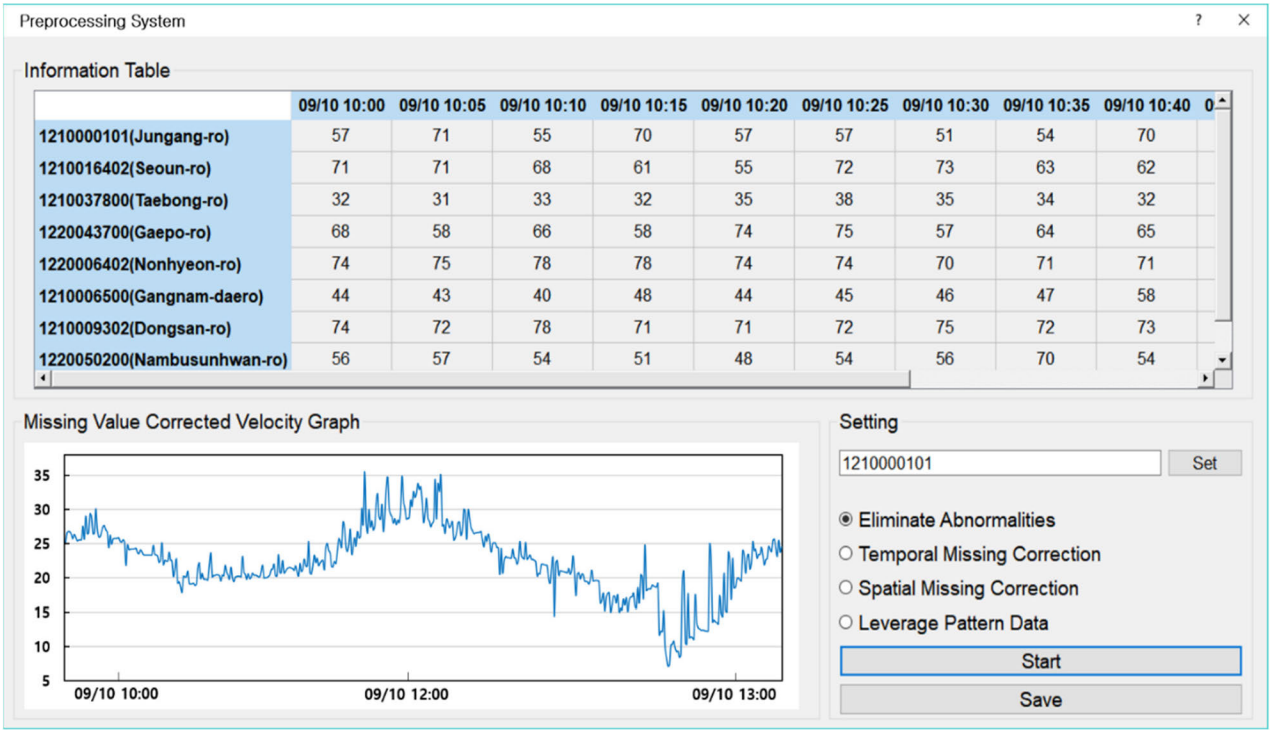

FIGURE 9. Pre-processing system for LSTM-based traffic congestion prediction.

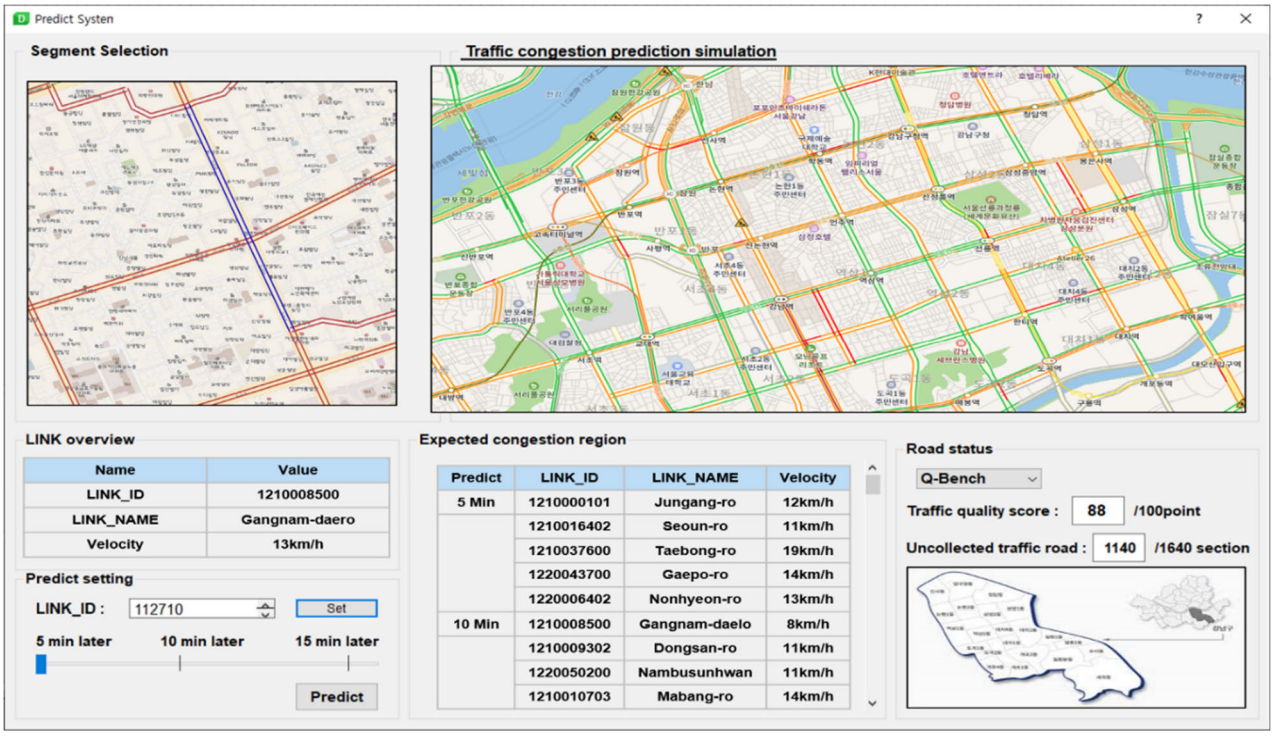

FIGURE 10. LSTM-based traffic congestion prediction system.

window, the speed data collected in the selected region are shown for the date provided. When '15 min later' is selected, and the Predict button is clicked in the Status window, the overall congestion results for the Gangnam-gu region of Seoul are displayed. The traffic congestion criteria differ depending on the road type. For general roads, 'smooth'

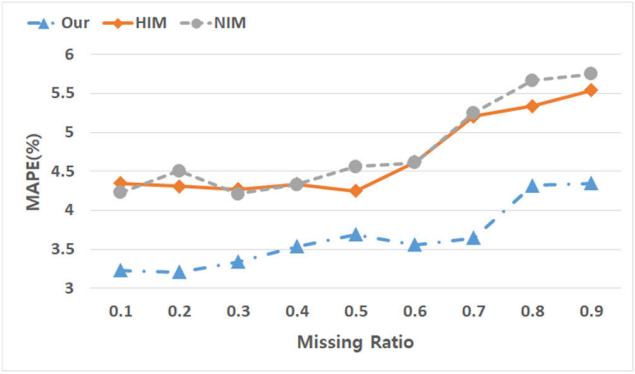

FIGURE 11. Results of performance evaluation according to missing value correction method.

refers to speeds of 30 km/h or higher, 'congested' to 15 km/h ∼ 30 km/h, and 'very congested' to less than 15 km/h. For highways, 'smooth' refers to speeds of 70 km/h or higher, 'congested' to 40 km/h ∼ 70 km/h, and 'very congested' to less than 40 km/h. These criteria are suggested by the Ministry of Land, Infrastructure and Transport [41]. Numerical information for the expected congestion region is provided in the table below the simulation map. Fig. 10 shows the LSTM-based traffic congestion prediction system [43], [44].

B. COMPARATIVE EVALUATION OF PERFORMANCE ACCORDING TO THE MISSING VALUE CORRECTION METHOD

If a model learns on data that includes missing values, the predication ability can be diminished. For this reason, it is necessary to correct missing values, and the model accuracy may change according to the correction method.

We therefore evaluated the performance of our correction methods through repeated experiments varying the missing rate. In the experiments, historical imputation methods (HIM) and nearest neighbor imputation (NIM) are used as conventional missing value correction methods for comparison with the missing value correction method proposed in this study. The performance comparison was conducted through the data missing rate based MAPE. The data missing rate ranged from 10% to 90% in increments of 10%. Fig. 11 shows the results of the performance evaluation for each of the missing value correction methods.

As shown in Fig. 11, the proposed method performed better in terms of MAPE than the conventional missing value correction methods. HIM corrects temporal missing values but fails to correct spatial missing values. In addition, its performance deteriorates when a large proportion of the data is missing. Unlike the HIM, the NIM cannot correct the temporal missing value, but it is possible to correct the data when the data in the neighboring space is not recorded. In contrast, the data correction method proposed in this study is able to correct both spatial and temporal data and exhibits excellent performance in terms of MAPE.

C. PERFORMANCE EVALUATION OF PREDICTION MODEL

For the performance evaluation and loss function of the model used in this study, the MAPE was used [45], [46]. The MAPE can be applied to overcome the effect of size-dependent error and represents the mean of the absolute error between the actual and predicted values. It was used for the loss function because it is sensitive to small values in low-speed regions such as congested areas. It was also used for the performance evaluation of the proposed method. The MAPE can be calculated by Equation (2), where A_i is an actual value and F_i is the predicted value. The MAPE is expressed as a percentage by subtracting the actual value from the predicted value and dividing the result by the actual value; this quantity is summed for all of the observations, and the sum is dividing by n. The lower a MAPE value is, the higher the model accuracy is.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \tag{2}$$

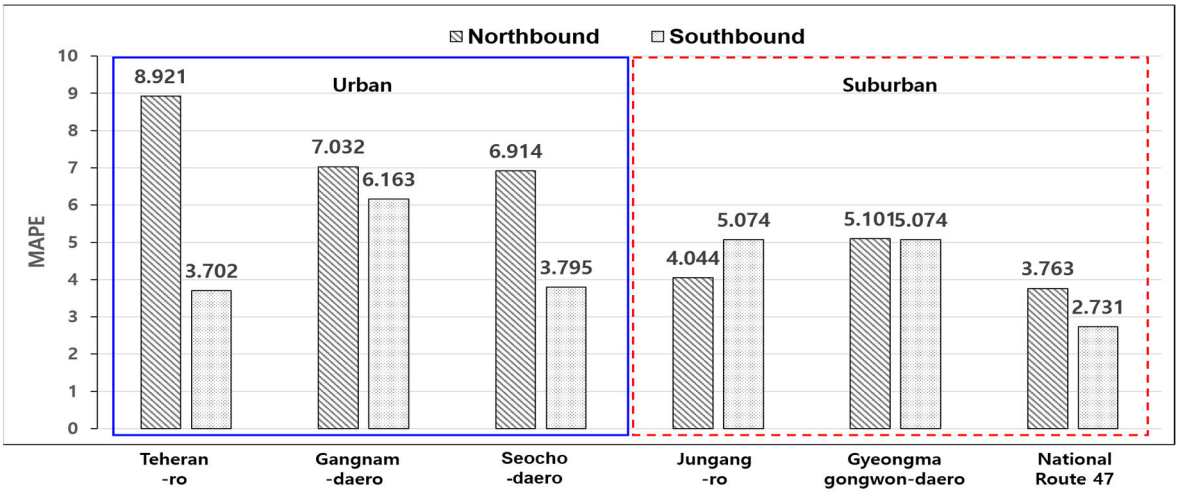

FIGURE 12. MAPE results of suburban and urban areas.

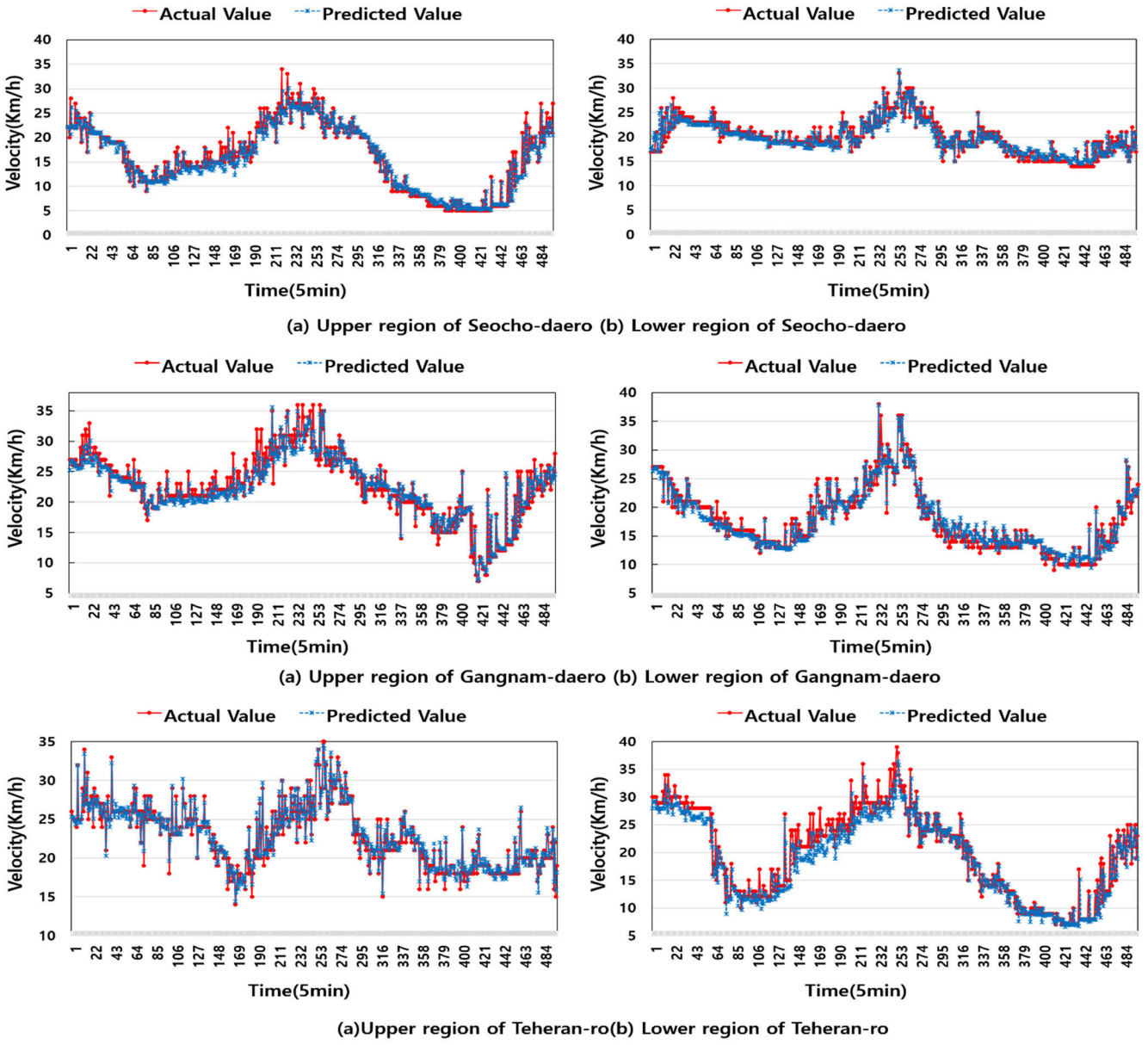

FIGURE 13. Analysis of prediction results for urban areas.

In addition, the data used in the experiment is the traffic data of a day. The data includes data on an urban area with high congestion and a suburban area with relatively low congestion. The performance of the LSTM model for congestion prediction was evaluated using uninterrupted and interrupted flow regions. An uninterrupted flow region has no external influences that control the traffic flow. An interrupted flow region refers to a region with interrupted traffic flow that has crossroads and trunk lines that cause interruptions due to traffic signals or traffic control facilities. An example of an uninterrupted flow region is a suburban area with highways, while an example of an interrupted flow region is an urban area with traffic signals and traffic control facilities. Fig. 12 shows a graph of the MAPE results for suburban and urban areas. For the suburban areas, three regions were extracted, and the northbound and southbound speeds were predicted. As shown in the graphs of the prediction results, the average MAPE was approximately 4.297%. As for the suburban areas, three regions were extracted from the urban areas, and the northbound and southbound speeds were predicted. The average MAPE for the urban areas was approximately 6.087%. The urban areas showed a somewhat lower accuracy than the suburban areas, and the reasons for this were analyzed. The suburban areas included fewer surrounding buildings and no traffic signals, and the speed limit within these regions was higher than in the urban areas. By contrast, the urban areas included numerous buildings, the large influence of a floating population other than drivers, traffic signals

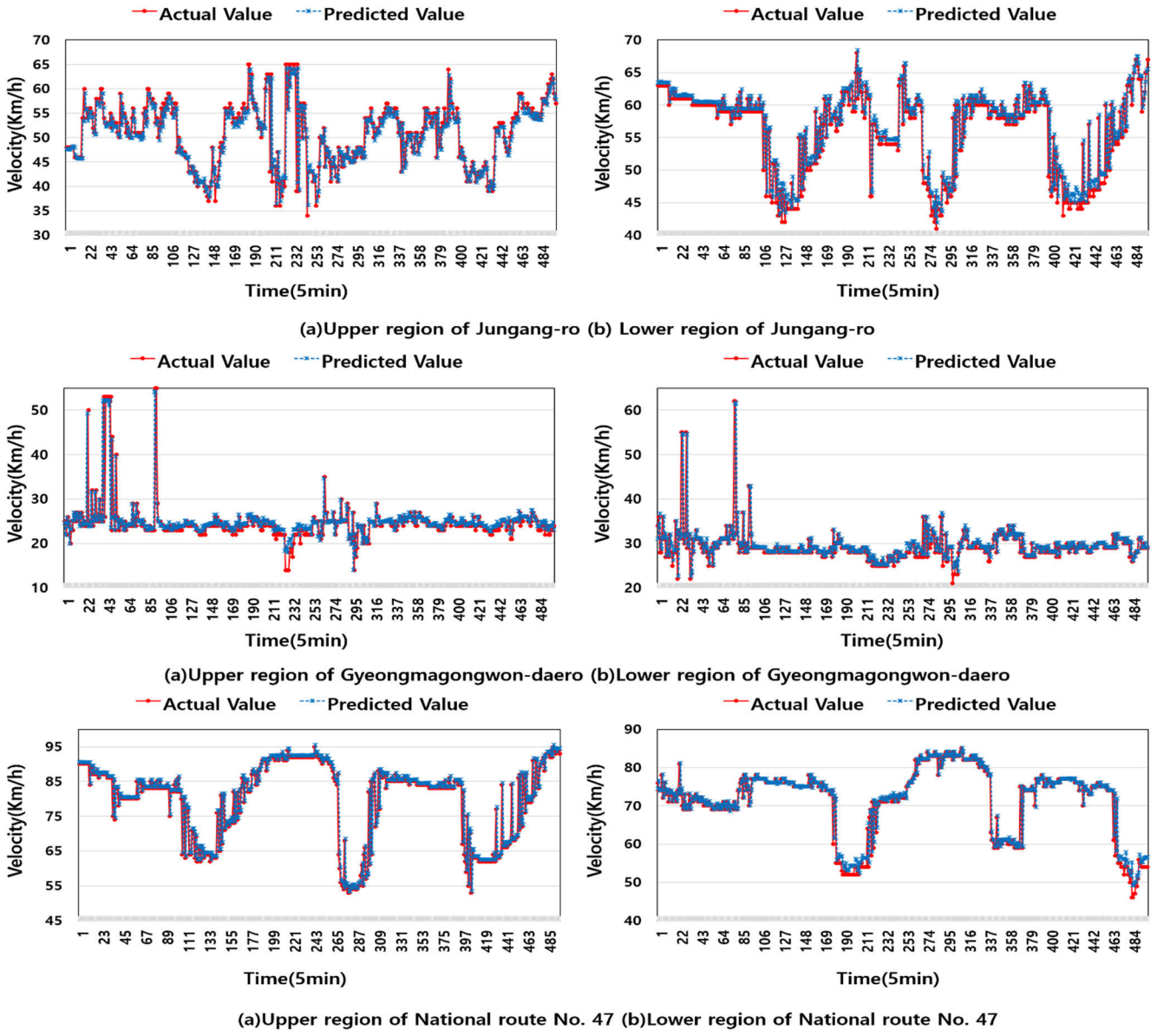

FIGURE 14. Analysis of prediction results for suburban areas.

at crossroads, and numerous variables interrupting the traffic flow. For these reasons, it is more difficult to predict the traffic flow in urban areas. In addition, Fig. 13 shows the results of a comparative analysis of the actual and predicted values in terms of the MAPE for three sections of the city center, while Fig. 14 shows the same comparison for three sections on the outskirts of the city. However, the MAPE reduces the denominator as the actual measurement approaches 0. This results in a significant increase in Absolute Percentage Error (APE) even if the absolute error value is small, resulting in a biased value when the average is taken. Therefore, RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error) is used for measuring performance in order to prevent the distortion of overall prediction performance. MAE calculates results through identical standards in different circumstances. Also, RMSE reduces distortion through route about errors dependent on size, which is the problem of MSE (Mean Squared Error), and displays the average of errors themselves intuitively.

In this study, the performances are compared between urban area and suburban area to evaluate the performances of prediction. Figure 13 and 14 show the results of performance evaluation through RMSE and MAE of urban and suburban areas. In the results of performance evaluation through RMSE in Figure 13, Southbound of Seocho-daero shows the best performance, which is 1.543. The Northbound of National

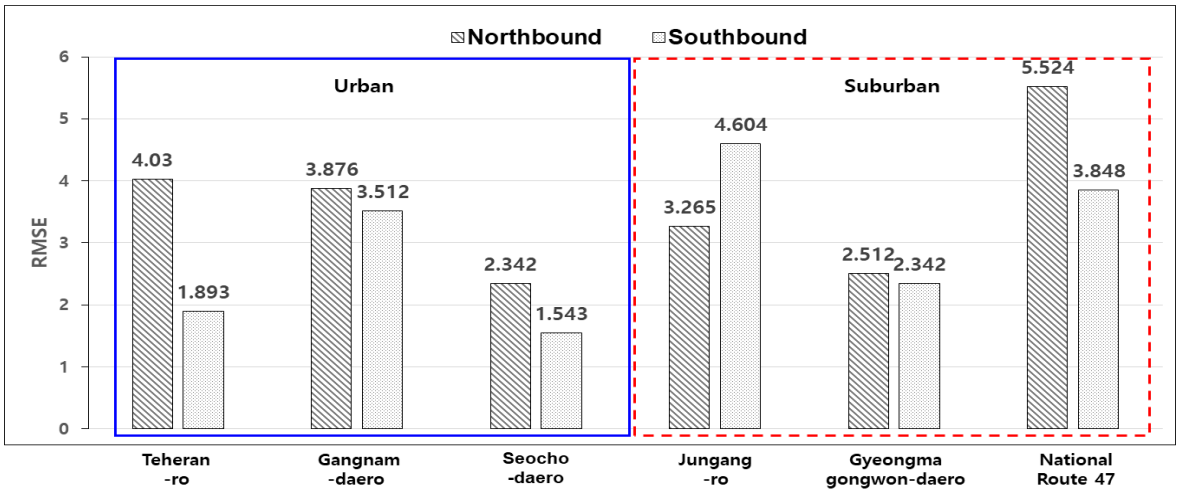

FIGURE 15. RMSE results of suburban and urban areas.

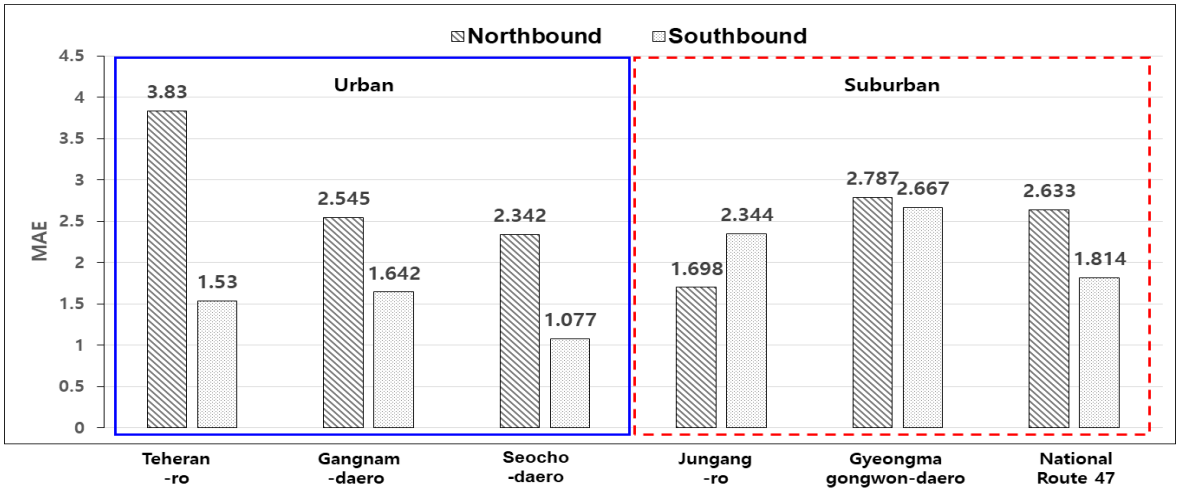

FIGURE 16. MAE results of suburban and urban areas.

Route 47 shows relatively low performance, which is 5.524. The results of 12 routes of MAE show 3.27 in average. MAE in Figure 14 shows the best performance in the Southbound of Seocho-daero like RMSE, and the Northbound of Teheran-ro shows the lowest performance, which is 3.83. The results of 12 routes of MAE show 2.24 in average. The results through MAPE show that the congestion in urban areas has poor prediction performance. But, the results of RMSE and MAE show that the performance of some suburban areas is poorer than the prediction of urban areas' congestion. This is because MAE and RMSE do not depend on the speed values or situation changes in urban areas with high congestion level and suburban areas where congestion level is not high, but are the results of calculation through identical standards. Therefore, when the three performance evaluation indexes are comprehensively analyzed, the prediction performance of urban areas except the Northbound of Teheran-ro is mostly better than that of suburban areas.

D. EVALUATION OF MODEL GOODNESS-OF-FIT IN COMPARISON WITH DIFFERENT MODELS

To demonstrate the reliability of the model proposed in this study, the goodness-of-fit of the model was evaluated. The proposed model was compared with other models presented in relevant studies. The data used for the comparison was preprocessed by the method proposed in this study. The performance index used for the comparison was the MAPE. The models used for comparison are RNN, LSTM, and STGCN models. Table 2 presents the prediction results for the different data and models in the comparison. As shown in Table 2, in terms of the MAPE, the proposed method had better goodness-of-fit than the other methods. The RNN performed worse than the LSTM models. This is because the RNN has the problem of long-term dependency. According to the comparison, there is performance improvement of 0.97 for the proposed model over that of Mou *et al.* [27]. The LSTM model used in this study is therefore good for

TABLE 2. Evaluation of model goodness-of-fit in comparison with different models.

Reference	Model	MAPE
Sun et al. [26]	RNN	7.22
Mou et al. [27]	LSTM	6.09
Yu et al. [28]	STGCN	6.43
Ranjan et al. [47]	LSTM	6.81
Zheng et al. [48]	LSTM	6.72
Zhao et al. [49]	LSTM	9.70
Current work	LSTM	5.12

traffic congestion prediction, since it accounts for temporal features.

V. CONCLUSION

In this study, an LSTM-based traffic congestion prediction method using a correction for missing temporal and spatial data was proposed. Based on experimental results, outliers and missing values in the traffic data influenced the prediction results. To improve the model performance, the outliers were removed, and the data were pre-processed using spatial and temporal trends and pattern data. As a predictive model, LSTM was applied. It is derived from the RNN model and solves the problem of long-term dependency. In the LSTM model, the result of a hidden layer is passed into the same hidden layer as an input. Because the model considers sequential or temporal aspects, it can be applied to learn the time-series features of traffic data. In an experiment to evaluate the model performance, suburban areas were used as an example of uninterrupted flow regions and urban areas as an example of interrupted flow regions. The suburban areas were less influenced by the traffic flows with external interference than the urban areas, and therefore had fewer variables at the time of prediction. The model thus demonstrated higher prediction accuracy for suburban areas. In comparison with relevant models, the proposed method was found to achieve better performance with a difference in the MAPE of 3%–17%. As a future study, we plan to increase the accuracy of the traffic congestion prediction in low-speed regions and urban areas and to establish a model with better user performance.

REFERENCES

[1] M. Chung and J. Kim, "The Internet information and technology research directions based on the fourth industrial revolution," *KSII Trans. Internet Inf. Syst.*, vol. 10, no. 3, pp. 1311–1320, 2016.

[2] R. Coppola and M. Morisio, "Connected Car: Technologies Issues Future Trends," *ACM Comput. Surveys*, vol. 49, no. 3, p. 46, 2016.

[3] D. Gunnarsson, S. Kuntz, G. Farrall, A. Iwai, and R. Ernst, "Trends in automotive embedded systems," in *Proc. Int. Conf. Compil., Architectures Synth. Embedded Syst. (CASES)*, 2012, pp. 9–10.

[4] S. Oh and K. Chung, "Performance evaluation of silence-feature normalization model using cepstrum features of noise signals," *Wireless Pers. Commun.*, vol. 98, no. 4, pp. 3287–3297, Feb. 2018.

[5] G. Dimitrakopoulos and P. Demestichas, "Intelligent transportation systems," *IEEE Veh. Technol. Mag.*, vol. 5, no. 1, pp. 77–84, Mar. 2010.

[6] S.-H. An, B.-H. Lee, and D.-R. Shin, "A survey of intelligent transportation systems," in *Proc. 3rd Int. Conf. Comput. Intell., Commun. Syst. Netw.*, Jul. 2011, pp. 332–337.

[7] M. Böhm, S. Fuchs, R. Pfliegl, and R. Kölbl, "Driver behavior and user acceptance of cooperative systems based on infrastructure-to-vehicle communication," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2129, no. 1, pp. 136–144, Jan. 2009.

[8] W. Ge, Y. Cao, Z. Ding, and L. Guo, "Forecasting model of traffic flow prediction model based on multi-resolution SVR," in *Proc. 3rd Int. Conf. Innov. Artif. Intell. (ICIAI)*, 2019, pp. 1–5.

[9] F. G. Habtemichael and M. Cetin, "Short-term traffic flow rate forecasting based on identifying similar traffic patterns," *Transp. Res. C, Emerg. Technol.*, vol. 66, pp. 61–78, May 2016.

[10] W. Zhang, Y. Yu, Y. Qi, F. Shu, and Y. Wang, "Short-term traffic flow prediction based on spatio-temporal analysis and CNN deep learning," *Transportmetrica A, Transp. Sci.*, vol. 15, no. 2, pp. 1688–1711, Nov. 2019.

[11] D. H. Hong and C. Hwang, "Support vector fuzzy regression machines," *Fuzzy Sets Syst.*, vol. 138, no. 2, pp. 271–281, Sep. 2003.

[12] G. Santamaría-Bonfil, A. Reyes-Ballesteros, and C. Gershenson, "Wind speed forecasting for wind farms: A method based on support vector regression," *Renew. Energy*, vol. 85, pp. 790–809, Jan. 2016.

[13] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, 2012, pp. 9–10.

[14] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2013, pp. 8614–8618.

[15] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.

[16] M. Mejdoub and C. B. Amar, "Classification improvement of local feature vectors over the KNN algorithm," *Multimedia Tools Appl.*, vol. 64, no. 1, pp. 197–218, May 2013.

[17] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[18] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2014, pp. 338–342.

[19] 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.

[20] J. C. Kim and K. Chung, "Prediction model of user physical activity using data characteristics-based long short-term memory recurrent neural networks," *KSII Trans. Internet Inf. Syst.*, vol. 13, no. 4, pp. 2060–2077, Apr. 2019.

[21] J.-C. Kim and K. Chung, "Associative feature information extraction using text mining from health big data," *Wireless Pers. Commun.*, vol. 105, no. 2, pp. 691–707, Mar. 2019.

[22] D. Ni, J. D. Leonard, A. Guin, and C. Feng, "Multiple imputation scheme for overcoming the missing values and variability issues in ITS data," *J. Transp. Eng.*, vol. 131, no. 12, pp. 931–938, Dec. 2005.

[23] X. Luo, X. Meng, W. Gan, and Y. Chen, "Traffic data imputation algorithm based on improved low-rank matrix decomposition," *J. Sensors*, vol. 2019, pp. 1–11, Jul. 2019.

[24] J. Chen and J. Shao, "Nearest neighbour imputation for survey data," *J. Official Statist.*, vol. 16, no. 2, pp. 113–131, 2000.

[25] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation," *BMC Med. Informat. Decis. Making*, vol. 16, no. S3, p. 74, Jul. 2016.

[26] S. Sun, J. Chen, and J. Sun, "Traffic congestion prediction based on GPS trajectory data," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 5, May 2019, Art. no. 155014771984744.

[27] L. Mou, P. Zhao, H. Xie, and Y. Chen, "T-LSTM: A long short-term memory neural network enhanced by temporal information for traffic flow prediction," *IEEE Access*, vol. 7, pp. 98053–98060, 2019.

[28] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 3634–3640.

[29] *USC Infolab*. Accessed: Jun. 12, 2020. [Online]. Available: https://infolab.usc.edu/

[30] *Blue Signal*. Accessed: Jun. 12, 2020. [Online]. Available: https://www.bluesignal.co.kr/

[31] A. Famili, W.-M. Shen, R. Weber, and E. Simoudis, "Data preprocessing and intelligent data analysis," *Intell. Data Anal.*, vol. 1, no. 1, pp. 3–23, Jan. 1997.

[32] S. Garcia, J. Luengo, and F. Herrera, *Data Preprocessing in Data Mining*. Cham, Switzerland: Springer, 2015.

[33] T. Pham-Gia and T. L. Hung, "The mean and median absolute deviations," *Math. Comput. Model.*, vol. 34, nos. 7–8, pp. 921–936, Oct. 2001.
[34] P. J. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *J. Amer. Stat. Assoc.*, vol. 88, no. 424, pp. 1273–1283, Dec. 1993.
[35] Z. Zhao, W. Chen, X. Wu, P. C. Y. Chen, and J. Liu, "LSTM network: A deep learning approach for short-term traffic forecast," *IET Intell. Transp. Syst.*, vol. 11, no. 2, pp. 68–75, Mar. 2017.
[36] L. Xianglong, J. Qinqin, and N. Liyao, "Short-term traffic flow prediction based on deep learning," *Comput. Appl. Res.*, vol. 34, no. 1, pp. 91–97, Jan. 2017.
[37] R. A. Jacobs, "Increased rates of convergence through learning rate adaptation," *Neural Netw.*, vol. 1, no. 4, pp. 295–308, 1988.
[38] Y. Gal, "A theoretically grounded application of dropout in recurrent neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 1019–1027.
[39] Q. V. Le, J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, and A. Y. Ng, "On optimization methods for deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 265–272.
[40] M. Denil, B. Shakibi, L. Dinh, N. de Freitas, and M. Ranzato, "Predicting parameters in deep learning," *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2148–2156.
[41] *Ministry of Land, Infrastructure and Transport*. Accessed: Jun. 12, 2020. [Online]. Available: http://openapi.its.go.kr/
[42] *Intelligent Traffic System Standard Node Link Management System*. Accessed: Jun. 12, 2020. [Online]. Available: http://nodelink.its.go.kr/
[43] J.-C. Kim and K. Chung, "Emerging risk forecast system using associative index mining analysis," *Cluster Comput.*, vol. 20, no. 1, pp. 547–558, Mar. 2017.
[44] K. Chung and R. C. Park, "Cloud based U-healthcare network with QoS guarantee for mobile health service," *Cluster Comput.*, vol. 22, no. S1, pp. 2001–2015, Jan. 2019.
[45] A. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, Jun. 2016.
[46] U. Khair, H. Fahmi, S. A. Hakim, and R. Rahim, "Forecasting error calculation with mean absolute deviation and mean absolute percentage error," *J. Phys., Conf. Ser.*, vol. 930, no. 1, pp. 25–26 Aug. 2017.
[47] N. Ranjan, S. Bhandari, H. P. Zhao, H. Kim, and P. Khan, "City-wide traffic congestion prediction based on CNN, LSTM and transpose CNN," *IEEE Access*, vol. 8, pp. 81606–81620, 2020.
[48] Y. Zheng, L. Liao, F. Zou, M. Xu, and Z. Chen, "PLSTM: Long short-term memory neural networks for propagatable traffic congested states prediction," in *Proc. Int. Conf. Genetic Evol. Comput.*, 2019, pp. 399–406.
[49] J. Zhao, Y. Gao, Z. Bai, H. Wang, and S. Lu, "Traffic speed prediction under non-recurrent congestion: Based on LSTM method and BeiDou navigation satellite system data," *IEEE Intell. Transp. Syst. Mag.*, vol. 11, no. 2, pp. 70–81, Mar. 2019.

DONG-HOON SHIN received the B.S. degree from the Department of Computer Engineering, Dongseo University, South Korea, in 2019. He is currently pursuing the master's degree with the Department of Computer Science, Kyonggi University, Suwon, South Korea. He has been a Researcher with the Data Mining Laboratory, Kyonggi University. His research interests include data mining, artificial intelligent, healthcare, biomedical and health informatics, knowledge system, VR/AR, and deep learning.

KYUNGYONG CHUNG received the B.S., M.S., and Ph.D. degrees from the Department of Computer Information Engineering, Inha University, South Korea, in 2000, 2002, and 2005, respectively. He has worked with the Software Technology Leading Department, Korea IT Industry Promotion Agency (KIPA). From 2006 to 2016, he was a Professor with the School of Computer Information Engineering, Sangji University, South Korea. Since 2017, he has been a Professor with the Division of Computer Science and Engineering, Kyonggi University, Suwon, South Korea. His research interests include data mining, artificial intelligent, healthcare, biomedical and health informatics, knowledge systems, HCI, and recommendation systems.

ROY C. PARK received the B.S. degree from the Department of Industry Engineering and the M.S. and Ph.D. degrees from the Department of Computer Information Engineering, Sangji University, South Korea, in 2010 and 2015, respectively. From 2015 to 2018, he was a Professor with the Division of Computing Engineering, Dongseo University, South Korea. Since 2019, he has been a Professor with the Department of Information Communication Software Engineering, Sangji University, Wonju, South Korea. His research interests include WLAN systems, heterogeneous networks, ubiquitous network service, human-inspired artificial intelligent and computing, health informatics, knowledge systems, peer-to-peer, and cloud networks.