IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 September 2023, accepted 8 October 2023, date of publication 12 October 2023, date of current version 18 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3324035

RESEARCH ARTICLE

Traffic Speed Prediction of Urban Road Network Based on High Importance Links Using XGB and SHAP

EUN HAK LEE
Multimodal Planning and Environment Division, Texas A&M Transportation Institute, Bryan, TX 77807, USA
e-mail: e-lee@tti.tamu.edu

ABSTRACT As the intelligent transportation system has been introduced, traffic speed prediction has become one of the foremost challenging tasks within complex urban road networks. The main idea of this study is to identify links that have a significant impact on the target link and develop a high-performance travel speed prediction model using those links. This study proposes the Extreme gradient boosting model with high importance links (HI-XGB) to predict traffic speed in the urban area. High importance links for predicting the target link speed are selected using Shapley additive explanations. With the selected input features, extreme gradient boosting is used to predict traffic speed. The results show that the performance of the HI-XGB model with one- and 12-time steps ahead achieved 98.5% and 90.7% accuracy, respectively. Feature analysis and link classification analysis are performed to identify the impact of the spatial characteristic on predicted speed. Among the eight features, the speed of the target link at t and the speed change of the target link at t−1 have the most impact on the predicted target link speed. In addition, link classification analysis is performed to identify the impact of the spatial characteristic of the input feature on predicted speed. The result indicates that links other than upstream or downstream could have a greater impact on traffic speed prediction.

INDEX TERMS Traffic speed prediction, urban road network, extreme gradient boosting (XGB), shapley additive explanations (SHAP), explainable artificial intelligence (XAI).

I. INTRODUCTION

With the introduction of the intelligent transportation system (ITS), traffic speed prediction has been regarded as one of the key challenging tasks in a complex urban road network. Advanced ITS provides an opportunity for accurate traffic speed prediction by collecting traffic state data. With the availability of big data, many traffic speed prediction methods have been developed based on data-driven statistical and machine-learning models. Accurate traffic speed prediction benefits both road travelers and operators such as road management agencies. Predicted traffic speed information enables travelers to select the best routes and departure times. For road network operators, predicted traffic speed is conducive to efficiently controlling the traffic conditions of urban road networks.

Traffic speed prediction of urban road networks is usually more challenging compared to other local regions [1], [2], [3]. For example, urban roads, with their frequent intersections, mixed road classifications, traffic signals, varying speed limits, and congested traffic patterns influenced by rush hour, events, and road conditions, pose unique challenges for speed prediction. Their road design, characterized by shorter segments, curves, and intersections, further impacts the accuracy of speed prediction. The presence of multiple access points like driveways, side streets, and parking lots adds to the variability in predicting speeds on urban roads. Therefore, previous studies tried to explore the relationship and correlation between links including upstream, downstream, and target links [4], [5], [6], [7]. For example, Dai et al., [4] explored

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

© 2023 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

spatial and temporal trends between the links to predict short-term traffic speed. Similarly, Park et al., [6] investigated the impact of inherent variation and spatio-temporal dependency on urban road networks. These studies implied that the travel speed in the urban area varies greatly and the correlation between the upstream and downstream links and the target link is not consistent.

There have been noteworthy studies on traffic speed prediction using data-driven techniques. The prediction approaches can be divided into three categories, i.e., conventional, parametric, and non-parametric [1]. Conventional approaches refer to the historical average and are widely adopted by practical fields such as the transportation industries. Parametric approaches are also regarded as an important solution to traffic speed prediction. Among them, auto-regressive integrated moving average (ARIMA) and Kalman filtering are widely used in previous studies [2], [3].

With the rise of deep and machine learning techniques in recent years, prediction models have evolved to be more sophisticated and accurate by capturing highly complex data correlations. For example, Lv et al. [8] developed a prediction model based on a deep belief network considering the spatio-temporal correlation of traffic dynamics. Similarly, Yu et al. [9] and Zhang et al. [10] proposed recurrent neural network (RNN) models to predict traffic states by reflecting temporal correlation. Zhang et al. [11] and Yu et al. [12] proposed convolutional neural network (CNN) models to predict traffic states by reflecting latent spatial factors. Long short-term memory (LSTM) and gated recurrent unit (GRU) were also used to predict travel time and speed, considering spatio-temporal characteristics [13], [14], [15], [16]. Moreover, advanced techniques have been used to predict traffic flow characteristics, i.e., the multi-task learning, deep multimodal learning model, attention mechanism, and graph convolution network [17], [18], [19], [20], [21]. Although these models showed notable performance, such as high accuracy, model interpretation remains difficult and is a significant drawback.

The explainable artificial intelligence (XAI) has emerged as a solution to overcome this problem. Extreme gradient boosting (XGB) and Shapley additive explanation (SHAP) have been used to understand the output of machine learning models [22], [23]. XGB is one of the notable machine learning techniques due to its speed and scalability. XGB efficiently predicts target values by reducing computational complexity. Several previous studies predicted traffic conditions with the advantages of XGB, i.e., high accuracy and fast processing time. For example, Dong et al. [24] used XGB to predict short-term traffic flow and revealed that XGB showed better performance than other machine learning models on traffic flow prediction. Similarly, Mei et al. [25] and Sun et al. [26] predicted short-term traffic flow based on the XGB model. XGB has also been used to estimate traffic speed and has been shown to achieve higher performance than other models [27]. Regarding model interpretation, SHAP has been utilized to understand the output of developed models. Proposed by Lundberg and Lee [28], SHAP was developed based on game theory and local explanations. Furthermore, SHAP provides an insightful understanding and nonlinear joint impact of features on the model output. SHAP values provide two interpretability aspects, i.e., global and local interpretability [29], [30]. Global interpretability provides the positive or negative relationship for each feature with the target, and local interpretability provides contributions to the developed model. With the combination of XGB and SHAP values, it is possible to achieve notable model performance and understand the impact of inputs on output.

Overall, previous studies have attempted to predict speeds by exploring relationships and correlations between upstream, downstream, and target links. These studies have indicated that travel speeds in urban areas vary significantly, and the correlations between upstream, downstream, and target links are not always consistent. However, there were limitations in modeling the complex nonlinear spatio-temporal correlations of traffic dynamics. While the advancement of deep learning and machine learning techniques has led to more sophisticated and accurate prediction models, there remains a challenge in interpreting these models.

To address the aforementioned challenges in predicting urban traffic speed, this study aims to incorporate highly influential links in addition to upstream and downstream links when predicting traffic speed on urban roads. The main idea of this study is to identify links that have a significant impact on the target link and develop a high-performance travel speed prediction model using those links. Specifically, the spatiotemporal impact of urban roads is understood from a network perspective, without being limited to upstream and downstream considerations. The link speed data is obtained from Transport Operation & Information Service (TOPIS) in Seoul. The XGB with high importance links (HI-XGB) model is proposed to predict traffic speed in urban areas. The input features that impact target link speed are selected based on the importance of SHAP, and the XGB model is developed to predict traffic speed. With the results of the HI-XGB model, feature analysis based on SHAP is performed to identify high importance links for predicting target link speed.

II. DATA DESCRIPTION

Gangnam district is the most popular area in the city of Seoul, Korea, and is a modern center, attracting the largest floating population among the city's twenty-five districts. Gangnam district experiences steady traffic congestion due to its geographical location. Gangnam district is the main gateway for entering Seoul from the southeastern suburban area, and thus constant flow of external traffic is present. Additionally, Gangnam district is densely developed as a commercial and residential area and suffers from many traffic jams. The road network in Gangnam district consists of 67 intersections, and its length stretches to about 203.0 km. The daily average traffic speed is about 30.1 km/h.

The government of Seoul has been operating TOPIS since 2004. TOPIS is a transportation management and information system that monitors and records the overall traffic situation in Seoul. TOPIS collects link speeds of the entire road network in Seoul using GPS information of 70,000 taxis. Seoul's traffic speed data can be obtained from Open Data Portal (data.seoul.go.kr) in Seoul, and it includes link speed information that is updated every five minutes. The road network data is obtained from the Open Data Portal (www.data.go.kr), and it includes node and link data. TOPIS data from 20 weekdays from June 4, 2018, to June 29, 2018, are used as the dataset, including 731 links in the road network of Gangnam district.

The data were preprocessed using the link IDs recorded in the traffic speed data and road network data. Furthermore, the upstream and downstream of the target link were labeled based on the starting and ending points of the link. The overall missing rate is approximately 2%, and these missing values were interpolated using the moving average technique.

The road network data of Gangnam district and an example of traffic speed data are shown in Table 1 and Fig. 1, respectively.

TABLE 1. Example of traffic speed data.

Time (2018/06 /04)	Link ID	Link speed (km/h)	Upstream link ID	Upstream link speed (km/h)	Downstream link ID	Downstream link speed (km/h)
00:05:00	1	30.1	2	33.7	5	30.1
00:05:00	2	33.7	3	32.5	1	30.1
00:05:00	3	32.5	4	32.1	2	33.7
…	…	…	…	…	…	…
23:55:00	1014	32.1	1012	45.2	1011	45.1

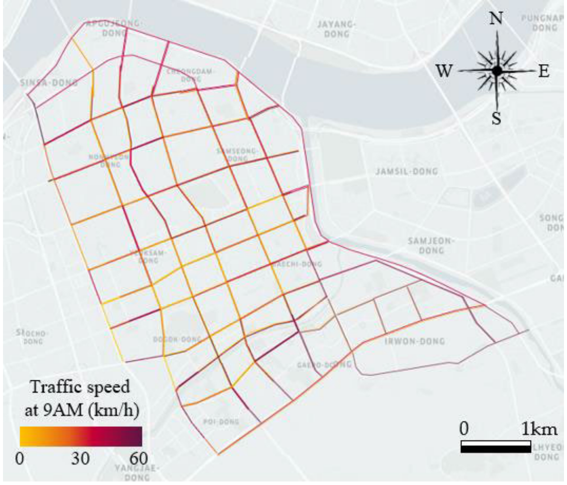


FIGURE 1. Road network of gangnam district.

III. METHODOLOGY

A. TRAFFIC SPEED PREDICTION MODELING STRATEGY

Accurate traffic speed prediction requires an understanding of the links that impact highly on the target link speed [3], [4], [5], [6]. The traffic speed of uninterrupted flow, primarily on freeways or expressways, changes depending on the time series trends of the target link and the traffic conditions of upstream or downstream links [31], [32]. However, the traffic speed of road networks in urban areas, such as signalized intersections and mixed road classifications, can be affected by distant links rather than upstream and downstream links. This indicates that finding links with a high impact, such as similar spatiotemporal characteristics, on the target link is essential for urban traffic speed prediction. Specifically, links other than upstream or downstream could have a greater impact on traffic speed prediction. In this regard, this study proposed a HI-XGB, as shown in Fig. 2.

There are many factors in influence the traffic flow pattern, for example, link speed, link speed change, day of week, holiday or not, and weather [33]. Among these factors, link speed and link speed change are the most representative factors in traffic characteristic predictions [34]. For an intuitive implication of spatial impact on target link speed, only two factors, i.e., link speed and link speed change, were used in this study.

The conventional approach in the traffic speed prediction field usually uses upstream and downstream link speed as input features, as shown in Fig. 2(a). Regarding the temporal characteristics, the speed of the previous time periods, i.e., t, t−1, . . . , and t−5, were used to predict traffic speed at t+1. Five conventional models were compared to validate the performance of the HI-XGB model.

The framework of the HI-XGB model consisted of two models, i.e., the feature selection model and the prediction model, as shown in Fig. 2(b). The feature selection model was to select the input features that had a high impact on the target link speed. The feature importance was estimated by SHAP and used to determine the impact of speed and speed change on each target link speed. The inputs were set as the speed and speed change of all links from the previous time periods (30 minutes), such as t, t−1, . . . , and t−5, to predict traffic speed at t+1. This is because the transition between the free flow state (80th percentile of traffic speed) and congestion (20th percentile of traffic speed) is 28.3 minutes on average. Based on the results of the feature selection model, six features scored high importance, i.e., three speed features and three speed change features, were selected as input features for the prediction model. With the selected input features, the XGB model was developed to predict the traffic speed of the target link. With the proposed model, the traffic speeds of the road network of Gangnam district were predicted, and the performance of each model was compared to select the best model. Additionally, the impact of spatio-temporal factors on the predicted output was interpreted using the SHAP value.

B. EXTREME GRADIENT BOOSTING

XGB is an ensemble machine-learning method using a sequence of decision trees. It has three main advantages, i.e., predictive accuracy, fast computation, and interpretability [28], [29], [30]. The idea of XGB is to correct the

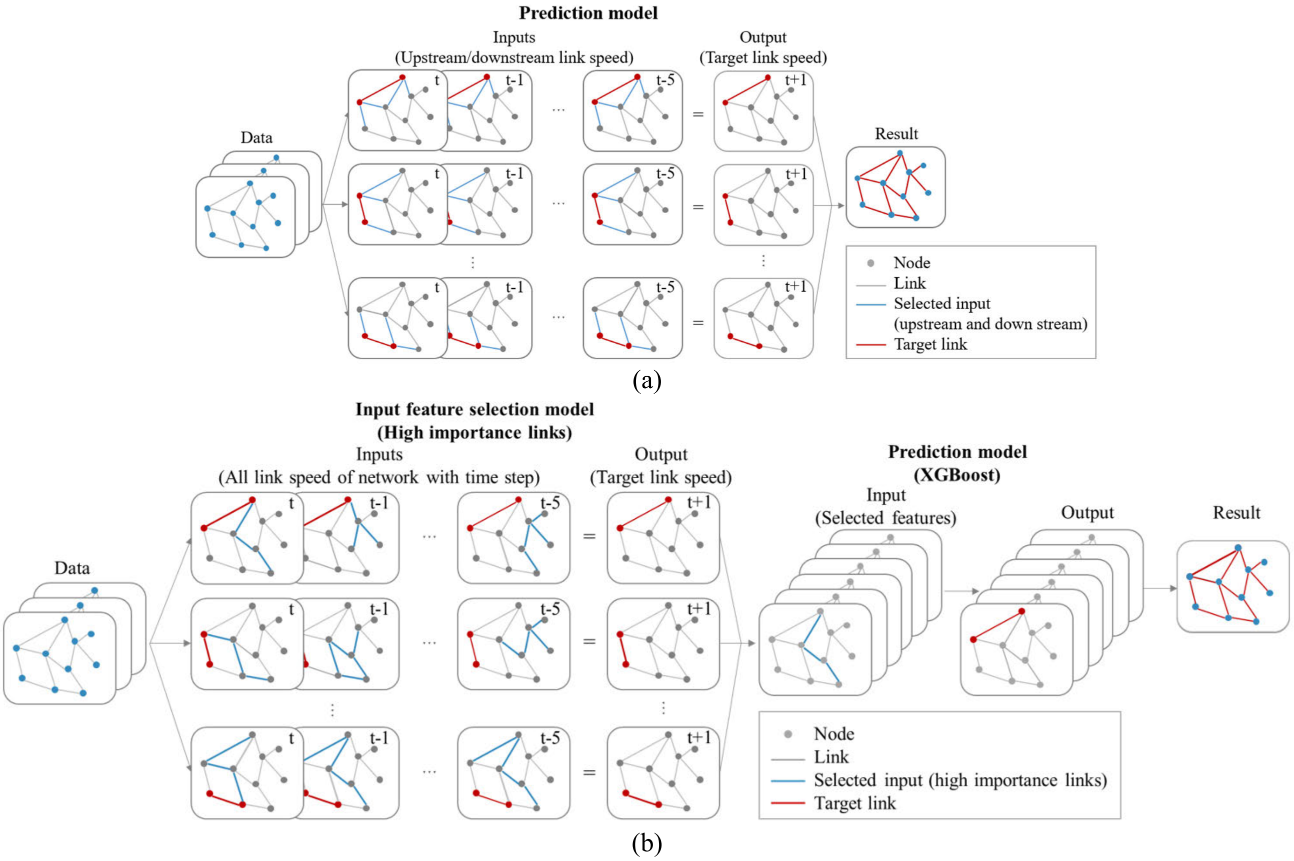


FIGURE 2. Concept of prediction model: (a) Conventional model with upstream and downstream links; (b) HI-XGB model.

performance of prior models by adding trees to the ensemble. Thus, XGB continuously adds trees and splits features to learn an improved function to fit the last predicted residuals.

In this study, XGB is used to predict the traffic speed of an urban road network. The pre-processed dataset with 4,210,560 samples (731 links × 288 time steps × 20 days) includes independent variables x_i and dependent variables y_i $(D = \{(x_i, y_i)\}, |D| = 4,210,560)$. Each x_i has m features of traffic speed $(m = 1, 2, \ldots, 8)$, and these features have corresponding dependent variables $(x_i \in \mathcal{R}^m, y_i \in \mathcal{R})$. The tree ensemble model predicts the target value $(\hat{y}_i)$ using K additive functions, as shown in Equation (1):

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i), F = \left\{f(x) = w_{q(x)}\right\}, \left(q : \mathcal{R}^m \rightarrow T, w \in \mathcal{R}^T\right) \quad (1)$$

where $\hat{y}_i$ are the target values, y_i are the dependent variables such as the traffic speed of the target link at the next time step, x_i are the independent variables, m are the features (m = from 1 to 8 in order of highest importance), q is a tree structure, w is the weight of leaf node, K is the number of tree functions, F is the space of trees, and f_k is an independent tree structure with leaf scores.

The objective function consists of the loss function and regularization term. The loss function measures how well the model fits the data, and the regularization term controls the complexity of the model to prevent overfitting. The objective is to minimize $\mathcal{L}(\phi)$, and the formulation is shown in Equation (2):

$$\mathcal{L}(\phi) = \sum_i l\left(y_i, \hat{y}_i\right) + \sum_k \Omega(f_k) \quad (2)$$

where Ω is a regularization term of the model complexity, and l is a loss function.

The regularization term includes two penalty terms, i.e., the penalty for the number of trees and the penalty for the leaf weights. The penalty terms are determined by optimizing the objective function using gradient descent and second-order Taylor approximation. The Ω is calculated as in Equation (3):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \|w_i\|^2 \quad (3)$$

where γ is a parameter that controls the minimum loss reduction required to make a further partition on a leaf node, T is the number of trees, λ is a parameter that controls the L2 regularization on the leaf weights, and w_i is the score of the leaf i.

The leaf weights are the values assigned to the terminal nodes of each tree, which represent the predicted output for the samples that fall into that leaf. The optimal weights are calculated by minimizing a loss function that measures the difference between the true and predicted outputs, as well as a regularization term that penalizes the complexity of the model. The weight calculations for the leaf nodes using gradients and Hessians contribute to the overall model's performance by finding the best-split points and leaf values that minimize the objective function. This way, XGB learns a more accurate and robust model that generalizes well to data. The optimal weight w_i^* of the leaf, j is calculated as in Equation 4, and the corresponding optimal value is estimated by Equations (5) to (7):

$$w_i^* = -\frac{\sum_{i\in I_j} \partial_{\hat{y}^{t-1}} l(y_i, \hat{y}^{t-1})}{\sum_{i\in I_j} \partial^2_{\hat{y}^{t-1}} l\left(y_i, \hat{y}^{t-1}\right) + \lambda} \tag{4}$$

$$g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{t-1}\right) \tag{5}$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{t-1}\right) \tag{6}$$

$$\tilde{\mathcal{L}}^t(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \tag{7}$$

Normally, it is difficult to enumerate all possible tree structures q. Therefore, a greedy algorithm, which extends a single leaf to many branches iteratively, is used to calculate the optimal value. This algorithm is usually employed to evaluate spilled candidates. The formulation of the greedy algorithm is shown in Equation (8):

$$\mathcal{L}_s = -\frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \tag{8}$$

where $I = I_L \cup I_R$, I_L is the instance set of left nodes after the split and I_R is the instance set of right nodes after the split.

C. HYPERPARAMETER TUNING FOR XGB

In this study, 85% of the dataset, selected randomly, was used as the training set, and the remaining 15% was used as the test set to validate model performance. Eight hyperparameters were tuned to maximize model performance. Hyperparameter tuning of XGB is requisite to prevent the overfitting problem and to minimize the complexity of the model. Optimal hyperparameters were selected by performing a 10-fold cross-validation. The number of iterations refers to the maximum number of boosting iterations. The learning rate is the scale of the weights of each tree, and it makes a robust model by changing the impact of each tree [23]. The max_depth parameter is the maximum depth of the trees. The subsample and colsample_bytree are tuned to prevent the overfitting problem of the model. The subsample parameter is the ratio of selected random observations for training instances. The colsample_bytree parameter stands for the ratio of columns when building each tree. Alpha and lambda stand for L1 and L2 regulation terms on the weights, respectively. Gamma is the minimum loss reduction required to make an additional partition on a leaf node. The hyperparameters of XGB in this study were selected as 2000 for the number of iterations, 0.1 for learning rate, 16 for max_depth, 0.8 for subsample, 0.9 for colsample_bytree, 0.3 for an alpha, and 0.3 for lambda.

Additionally, the robustness of the model performance was evaluated based on the toy networks with 8000 links using optimal hyperparameters. The results showed that the performance of the model remained robust, as the features with high importance were consistently selected for each of the 8000 links.

D. SHAPLEY ADDITIVE EXPLANATIONS FOR MODEL INTERPRETATION

SHAP is used to interpret the impact of input features on the model output. Specifically, SHAP presents the order of feature importance and relative importance of input features. SHAP repeatedly asks about the impact of the feature on each predicted value, and the SHAP value is estimated as the answer. The SHAP was used to identify the impact of individualized features on model output, and the results were illustrated using the SHAP summary plot function. The formulation of the SHAP is shown in Equation (9). For the linear function, $g\left(z'\right)$ is defined by the additive feature function shown in Equation (10):

$$\theta_i = \sum_{S\subseteq N\{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}\left[v\left(S\cup\{i\}\right) - v(S)\right] \tag{9}$$

$$g\left(z'\right) = f\left(h_x\left(z'\right)\right) \tag{10}$$

where z' is the simplified input vector, h_x is a function from the simplified to the original input.

IV. APPLICATION

A. RESULTS OF THE HI-XGB MODEL FOR TRAFFIC SPEED PREDICTION

The performance of the HI-XGB model was evaluated by comparing five naïve models, i.e., ARIMA, SVM, LSTM, GRU, and XGB. Three performance measures, i.e., mean absolute percentage error (MAPE), root mean square error (RMSE), and mean absolute error (MAE), were used to evaluate the performance of the prediction models [12]. These measures are well-known metrics for evaluating the difference between the precision of a prediction and its bias. The formulations of MAPE, RMSE, and MAE are shown in Equations (11)-(13):

$$MAPE = \frac{100}{n}\sqrt{\sum\nolimits_n \frac{|\mu_{t+1} - \tilde{\mu}_{t+1}|}{\mu_{t+1}}} \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_n (\mu_{t+1} - \tilde{\mu}_{t+1})^2}{n}} \tag{12}$$

$$MAE = \frac{1}{n}\sum\nolimits_n |\mu_{t+1} - \tilde{\mu}_{t+1}| \tag{13}$$

where n is the number of predicted traffic speeds, μ_{t+1} is the actual traffic speed at time $t+1$, and $\tilde{\mu}_{t+1}$ is the traffic speed at time $t+1$.

The training was done with 85% of the dataset, and the remaining 15% was used to test data. Train and test data were randomly selected. The train data comprised 3,578,976 of 4,210,560 link speeds, and the test data included 631,584. The link traffic speeds of the road network of Gangnam district were predicted by the six prediction models with 1-, 6-, 12-, and 24-time steps, i.e., conventional ARIMA, SVM, LSTM, GRU, XGB, and proposed HI-XGB models. The results of the traffic speed prediction by the proposed XGB model are shown in Table 2 and Fig. 3.

TABLE 2. Results of traffic speed prediction with XGB.

Model		MAPE	RMSE (km/h)	MAE (km/h)
ARIMA with UD links	1-time step	0.106	2.78	4.34
	6-time steps	0.139	3.54	5.63
	12-time steps	0.170	4.23	6.68
	24-time steps	0.218	5.32	8.26
SVM with UD links	1-time step	0.089	2.25	3.61
	6-time steps	0.115	2.97	4.68
	12-time steps	0.153	4.04	6.44
	24-time steps	0.182	4.88	7.74
LSTM with UD links	1-time step	0.056	1.39	2.22
	6-time steps	0.095	2.37	3.77
	12-time steps	0.132	3.26	5.24
	24-time steps	0.153	3.85	6.17
GRU with UD links	1-time step	0.057	1.49	2.38
	6-time steps	0.085	2.21	3.49
	12-time steps	0.134	3.49	5.55
	24-time steps	0.155	4.15	6.66
XGB with UD links	1-time step	0.045	1.19	1.89
	6-time steps	0.081	2.13	3.39
	12-time steps	0.088	2.31	3.68
	24-time steps	0.110	2.73	4.34
XGB with HI links (HI-XGB)	1-time step	0.015	1.01	0.78
	6-time steps	0.074	1.95	3.22
	12-time steps	0.083	2.17	3.56
	24-time steps	0.093	2.41	3.90

UD links: upstream and downstream links
HI links: high importance links

Among the six prediction models, the HI-XGB model showed the highest performance in all time steps. Specifically, the results of MAPEs for the HI-XGB model with 1-, 6-, 12-, and 24-time steps ahead were estimated to be 0.015, 0.074, 0.083, and 0.093, respectively. The results of MAPEs for the conventional ARIMA, SVM, LSTM, GRU, and XGB models with a one-time step ahead were estimated to be 0.106, 0.089, 0.056, 0.057, and 0.045, respectively. Interestingly, the HI-XGB model shows particularly robust performance in long-term prediction, i.e., 12-time steps ahead and 24-time steps ahead. For example, the error of the conventional ARIMA model was increased by about 0.05 when the time step was increased from 12 to 24. However, the error of the HI-XGB model was increased by about 0.01 in the same condition. Among the conventional models, it showed high performance in the order of the conventional XGB model and LSTM model. These results implied that the XGB model could show better performance than LSTM, which is a type of time-series model, in an urban area.

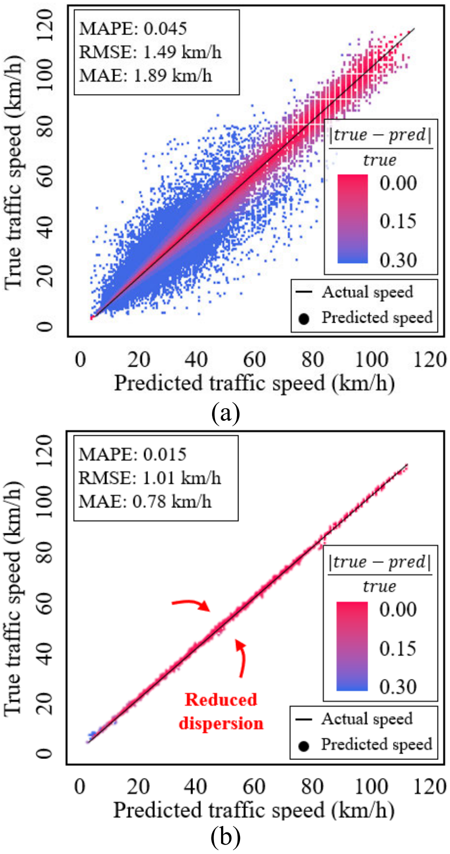


FIGURE 3. Result of traffic speed prediction: (a) Conventional XGB model with upstream and downstream links; (b) HI-XGB model with important features.

Since the road network showed interrupted flow characteristics due to signalized intersections, the effect of temporal characteristics on traffic speed was smaller than that of networks with an uninterrupted flow, such as highways. Similarly, the effect of spatial characteristics on traffic speed was also small in urban areas. In general, it is known that an upstream link and a downstream link have a large impact on the target link. However, in urban areas, the effect was small due to the intermittent flow characteristics. These results suggest that the effect of spatially or temporally distant links could be greater than that of upstream or downstream links on the target link.

Overall, the accuracy of the HI-XGB model with a 1-time step ahead showed the highest performance with a MAPE of 0.015. Since the characteristics of the upstream link, downstream link, and right-before time zone had little effect on the target link in the urban area, it was crucial to find the link and time that affected the target link.

B. FEATURE ANALYSIS

The importance scores, such as SHAP values of eight features of the HI-XGB model, were summarized in Fig. 4.

The features were ordered by their importance in predicting traffic speed. The SHAP value implied the impact of the input feature on the output. It was interpreted that the larger the SHAP value, the greater the impact. The average SHAP value of target link speed at t, target link speed change at t−1, 1st high important link speed, 1st high important link speed change, 2nd high important link speed, 3rd high important link speed, 2nd high important link speed change, and 3rd high important link speed change were estimated to be 8.54, 3.41, 0.04, 0.02, 0.01, 0.01, 0.01, 0.01, on average, respectively.

The SHAP values of target link speed at t and target link speed change at t−1 significantly impacted the predicted speed, such as target link speed at t+1. These results implied that the traffic speed of the target link at the previous time step has a significant impact on speed prediction.

Regarding the impact of excluding target links on predicted speed, six features had a significant impact on the predicted traffic speed. All six features had a positive correlation with the predicted speed, such as target link speed at t+1. Specifically, the predicted speed of the target link increased as the three important link speeds increased. Also, the predicted speed of the target link has increased as the three important link speed changes increased. These results indicated that the speed and speed change of links with similar spatiotemporal characteristics to the target link speed was reasonably selected, even if they were not close in time or near distance.

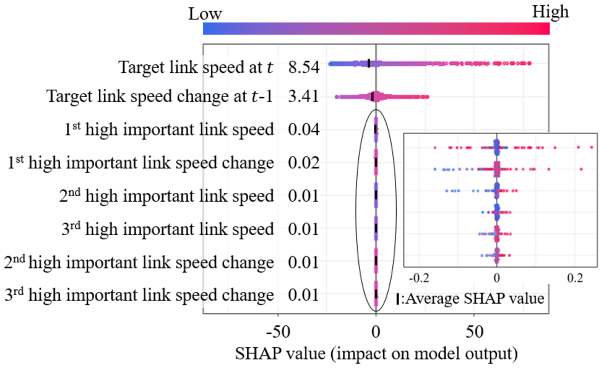


FIGURE 4. Results of SHAP values for eight input features.

C. FEATURE DEPENDENCY ANALYSIS

The results of feature dependency analysis were performed to identify the impact of speeds and speed changes on the output of the HI-XGB, as shown in Fig. 5.

The impact of selected link speed which has high importance on predicted speed and relationships between selected links, is shown in Fig. 5(a). The results showed that the three important link speeds, i.e., link speeds with high importance scores (1st to 3rd), had a positive correlation with each other. The SHAP value of the most important link speed increased when the 2nd and 3rd important link speeds decreased. These results indicated that the high importance score from SHAP was estimated as link speed had similar spatio-temporal characteristics to the predicted speed.

The impacts of target link speed at t, target link speed change at t−1, and important link speed are shown in Fig. 5(b). The results showed that the SHAP value of the target link speed at t decreased when the target link speed changed at t−1, and the important link speed decreased. Specifically, SHAP values of target link speed at t were decreased from 30 to 0 when speed changes were −20 to 0. Then, SHAP values of target link speed at t were slightly increased from 0 to 5 when speed changes were 0 to 20. Regarding the important link speed, the SHAP value of the target link speed at t decreased linearly.

Overall, target link speed at t, target link speed change at t−1, and important link had a significant impact on predicted speed. The selected link speed features had a positively correlated impact, and the selected speed change had a greater impact on the prediction speed as the value decreased.

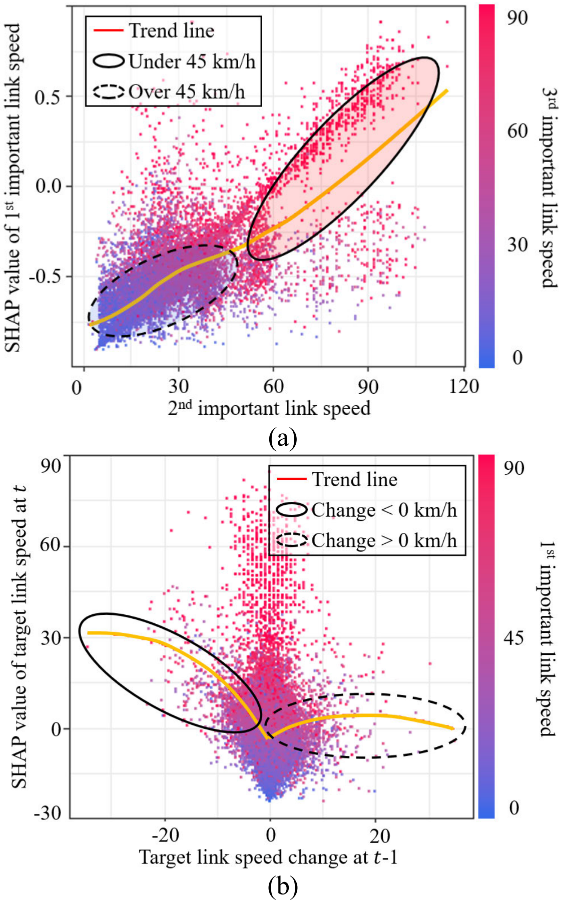
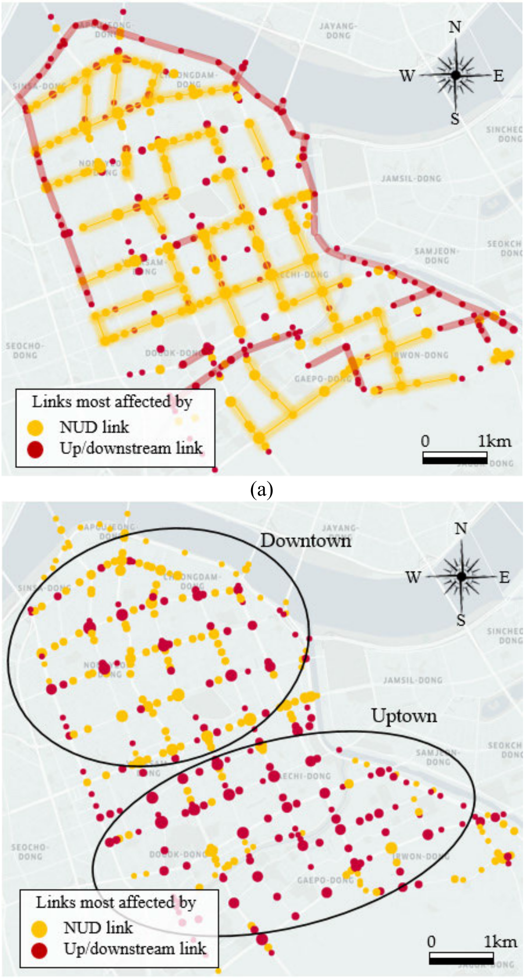


FIGURE 5. Result of SHAP dependency analysis: (a) Impact of three important link speeds on model output, (b) Impact of target link speed, link speed change, and important link speed on model output.

D. SPATIAL CLASSIFICATION OF IMPORTANT LINK

The spatial classification analysis was performed to identify the impact of the link speed and speed change on the target link from a spatial perspective. The road network of Gangnam district was classified into two groups based on the SHAP value. Group 1 was defined as the set of links that were most impacted by links other than up/downstream (NUD) links. Group 2 was defined as the set of links that were most impacted by upstream and downstream (UD) links. Specifically, if the NUD link had the highest SHAP value the target link was classified into Group 1 and if the UD link had the highest SHAP value the target link was classified into Group 2. The impact of the input link speed and speed change on the predicted target link speed is shown in Fig. 6 and Table 3.



(a)

(b)

FIGURE 6. Impact of the input link speed and speed change on the predicted target link speed: (a) speed feature; (b) Speed change feature.

Regarding the speed feature, the numbers of links in Groups 1 and 2 were 425 and 306, respectively. The road links in Groups 1 and 2 were mostly collectors and arterials, respectively. The average speeds of Groups 1 and 2 were 28.56 and 34.41, respectively. SHAP values of Groups 1 and 2 showed 0.24 and 0.36. These results indicated that more NUD links impacted the target link than UD links. Also, the collector link was more affected by the NUD links than the UD links. Regarding the speed change feature, the numbers of links in Groups 1 and 2 were 304 and 427, respectively. The road links in Groups 1 and 2 are mostly located in the downtown and uptown areas, respectively. SHAP values of Groups 1 and 2 showed 0.23 and 0.34. These results indicated that the speed change of UD links impacts more than that of NUD links.

TABLE 3. Impact of the input link speed and speed change on the predicted target link speed.

Group	Speed feature		Speed change feature	
	1	2	1	2
Link	Most affected by NUD links[1]	Most affected by UD links[2]	Most affected by NUD links[1]	Most affected by UD links[2]
Road class	Collector (Intersection)	Arterial	-	-
Location	-	-	Downtown	Uptown
Number of links	425	306	304	427
Average SHAP	0.24	0.36	0.23	0.34
Average true speed (km/h)	28.56	34.41	29.02	29.96
Average predicted speed (km/h)	28.51	34.12	28.22	29.97

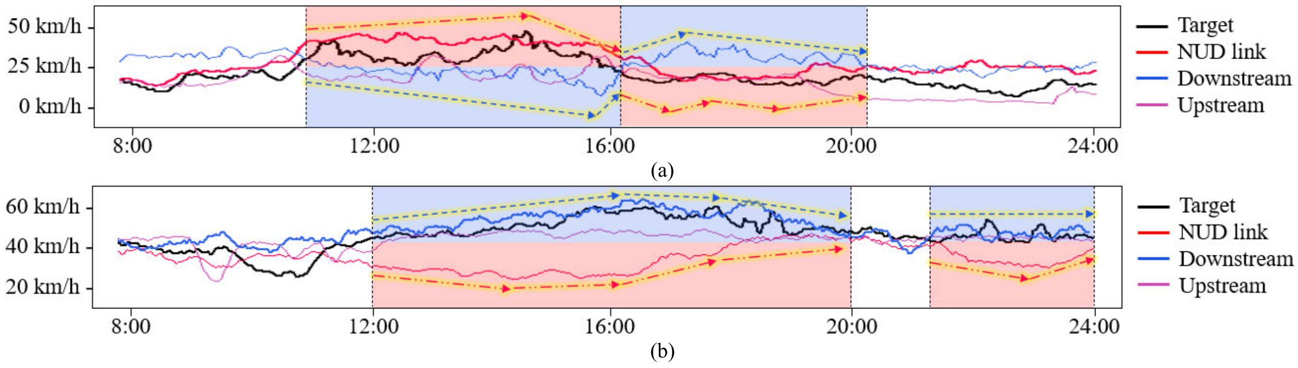
[1] NUD link: links that were not upstream or downstream links
[2] UD link: links that were upstream or downstream links

Overall, the road network in Gangnam was affected by a large number of NUD links, but it was greatly affected by the UD links. In terms of speed change, links in the downtown area were heavily impacted by NUD links, and links in the uptown area, which is relatively less congested, were impacted by UD links.

Fig. 7. shows the time-series pattern of speeds and speed changes of the target link, NUD link, and UD link from 8:00 to 24:00. The traffic speed in the downtown area is shown in Fig. 7(a). The speed of the NUD link showed a pattern more similar to the target link speed than the downstream link. Specifically, both the target link and NUD link speeds showed a tendency to peak high from 12:00 to 16:00 and low from 16:00 to 20:00. However, the downstream link displayed an opposite trend, with the speed peaking low from 12:00 to 16:00 and high from 16:00 to 20:00.

The traffic speed in the uptown area is shown in Fig. 7(b). The speed of the UD links showed a pattern more similar to the target link speed than the NUD link. Specifically, the speeds of the target link and UD links showed a tendency towards free flow speed from 16:00 to 20:00, while the NUD link was congested. Similarly, the speeds of the target link and UD links showed a tendency towards free flow speed from 21:00 to 24:00, but the NUD link showed a low peak.

These results imply that the speed patterns of the target link can be more similar, even if the links are not physically connected to the target link. This similarity is particularly noticeable in road networks in the downtown area with numerous signals and frequent congestion. On the other hand, road



IPS* stands for importance score estimated by SHAP

FIGURE 7. Time-series pattern of traffic speed: (a) Downtown area, (b) Uptown area.

networks within uptown speed patterns of target links can be more similar to UD links than NUD links. This is because the roads in uptown are less crowded, have fewer traffic signals, and have more arterial routes compared to downtown. Among the UD links, the pattern was more similar to downstream links than to upstream links. This is because the traffic flow of the target link is significantly influenced by the downstream area, for example, signal crossings, sudden accidents, or bad weather [15].

Overall, the findings suggest that road conditions in urban areas such as Gangnam district are shaped not solely by factors related to upstream or downstream links but also by other factors, i.e., frequent intersections, diverse road categories, traffic signals, varying speed limits, rush hours, events, and overall road conditions.

V. CONCLUSION

With the introduction of ITS, traffic speed prediction is a key challenge in a complex urban road network. develop a high-performance travel speed prediction model using those links. To improve model performance and interpret the impact of input features on target links, the HI-XGB model was developed in this study. The proposed model framework was classified into two stages, i.e., the feature selection model and the prediction model. Specifically, the high-importance links that impact target link speed are selected based on SHAP, and XGB is used to predict traffic speed using selected high-importance links as inputs. In addition, the impact of high-importance links' speeds on target link speeds was explored with the Shapley score. Traffic speed data for five weekdays from June 4, 2018, to June 8, 2018, was used as the dataset, including 731 links of the road network of Gangnam district. The model was trained with 85% of the dataset, selected randomly, and tested with the remaining 15%. The HI-XGB model was proposed to accurately predict traffic speed and to understand the impact of spatio-temporal features on predicted speed in an urban area. The results showed that the HI-XGB model with a time step ahead achieved the highest performance with a MAPE of 0.015.

Feature analysis was performed based on the importance score estimated from the SHAP analysis. Among the eight features of the HI-XGB model, target link speeds at t, and target link speed change at t-1 significantly impacted the predicted speed. Also, the six remaining features, such as selected features based on high importance scores, had a significant impact on predicted speed. Specifically, the selected important link speed impacted on predicted speed, and the selected important speed change had a greater impact on the prediction speed as the value decreased. Spatial classification analysis was also performed to identify the spatial impact on predicted speed. The results indicated that the road network in Gangnam was affected by a large number of links that were not up/downstream links, but it was greatly affected by the up/downstream links. This result implies that the speed and speed change patterns could be more similar even if the links were not physically connected to the target link, i.e., upstream or downstream link.

The noteworthy performance of the HI-XGB model supported its ability to predict complex urban road network speeds. SHAP provided an insightful understanding of the predicted speed. Specifically, SHAP evaluated feature importance and nonlinear joint impacts of features and was used for the feature selection and model interpretation. Interesting information came from this study, such as the spatio-temporal impact on target link speed in the urban area, which has not been established by other machine learning techniques. The results indicate that roads in complex urban areas such as Gangnam district are influenced not only by upstream or downstream factors but also by frequent intersections, mixed road classifications, traffic signals, various speed limits, commuting hours, events, and road conditions. From an urban traffic management perspective, it also suggests that roads need to be classified with spatio-temporal characteristics, i.e., UD or NUD links, to improve road performance.

Although the proposed HI-XGB showed notable performance and interpretability in predicting link speeds, there are still opportunities to improve the performance of the prediction model. It would be desirable to consider other factors in the prediction model, i.e., yearly traffic pattern, seasonal traffic pattern, the day of the week, and weather conditions. Moreover, it is beneficial to subdivide and include road conditions such as geometric structure, environment, and surrounding land use in the model to improve understanding of the classified UD and NUD in this study. The insights found in this study could be applicable to other advanced prediction techniques and other urban areas with similar traffic characteristics.

REFERENCES

[1] J. Q. James, C. Markos, and S. Zhang, "Long-term urban traffic speed prediction with deep learning on graphs," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7359–7370, Jul. 2022.

[2] P. Jiao, R. Li, T. Sun, Z. Hou, and A. Ibrahim, "Three revised Kalman filtering models for short-term rail transit passenger flow prediction," *Math. Problems Eng.*, vol. 2016, Mar. 2016, Art. no. 9717582.

[3] M. Van Der Voort, M. Dougherty, and S. Watson, "Combining Kohonen maps with ARIMA time series models to forecast traffic flow," *Transp. Res. Part C, Emerg. Technol.*, vol. 4, no. 5, pp. 307–318, Oct. 1996.

[4] G. Dai, C. Ma, and X. Xu, "Short-term traffic flow prediction method for urban road sections based on space-time analysis and GRU," *IEEE Access*, vol. 7, pp. 143025–143035, 2019.

[5] C. Ma, G. Dai, and J. Zhou, "Short-term traffic flow prediction for urban road sections based on time series analysis and LSTM_BILSTM method," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5615–5624, Jun. 2022.

[6] H.-C. Park, S. Kang, S.-Y. Kho, and D.-K. Kim, "Investigation of effects of inherent variation and spatiotemporal dependency on urban travel-speed prediction," *J. Transp. Eng., Part A, Syst.*, vol. 146, no. 5, May 2020, Art. no. 04020027.

[7] B. Yu, Y. Lee, and K. Sohn, "Forecasting road traffic speeds by considering area-wide spatio-temporal dependencies based on a graph convolutional neural network (GCN)," *Transp. Res. Part C, Emerg. Technol.*, vol. 114, pp. 189–204, May 2020.

[8] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[9] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks," *Sensors*, vol. 17, no. 7, p. 1501, Jun. 2017.

[10] C. Zhang, J. J. Q. Yu, and Y. Liu, "Spatial–temporal graph attention networks: A deep learning approach for traffic forecasting," *IEEE Access*, vol. 7, pp. 166246–166256, 2019.

[11] J. Zhang, Y. Zheng, D. Qi, R. Li, and X. Yi, "DNN-based prediction model for spatio-temporal data," in *Proc. 24th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, 2016, pp. 1–4.

[12] B. Yu, H. Yin, and Z. Zhu, "Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting," 2017, *arXiv:1709.04875*.

[13] E. H. Lee, S.-Y. Kho, D.-K. Kim, and S.-H. Cho, "Travel time prediction using gated recurrent unit and spatio-temporal algorithm," *Proc. Inst. Civil Eng. Municipal Engineer*, vol. 174, no. 2, pp. 88–96, Jun. 2021.

[14] J. Zhao, Y. Gao, Y. Qu, H. Yin, Y. Liu, and H. Sun, "Travel time prediction: Based on gated recurrent unit method and data fusion," *IEEE Access*, vol. 6, pp. 70463–70472, 2018.

[15] 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, May 2019.

[16] J. Zhao, Y. Gao, Z. Yang, J. Li, Y. Feng, Z. Qin, and Z. Bai, "Truck traffic speed prediction under non-recurrent congestion: Based on optimized deep learning algorithms and GPS data," *IEEE Access*, vol. 7, pp. 9116–9127, 2019.

[17] D. Ma, J. Zhu, X. B. Song, and X. Wang, "Traffic flow and speed forecasting through a Bayesian deep multi-linear relationship network," *Exp. Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 119161.

[18] J. H. Min, S. W. Ham, D.-K. Kim, and E. H. Lee, "Deep multimodal learning for traffic speed estimation combining dedicated short-range communication and vehicle detection system data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2677, no. 5, pp. 247–259, May 2023.

[19] A. Abdelraouf, M. Abdel-Aty, and J. Yuan, "Utilizing attention-based multi-encoder–decoder neural networks for freeway traffic speed prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11960–11969, Aug. 2022.

[20] S. Zhang, Y. Guo, P. Zhao, C. Zheng, and X. Chen, "A graph-based temporal attention framework for multi-sensor traffic flow forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 7743–7758, Jul. 2022.

[21] Y. Zhou, J. Li, H. Chen, Y. Wu, J. Wu, and L. Chen, "A spatiotemporal hierarchical attention mechanism-based model for multi-step station-level crowd flow prediction," *Inf. Sci.*, vol. 544, pp. 308–324, Jan. 2021.

[22] A. B. Parsa, A. Movahedi, H. Taghipour, S. Derrible, and A. K. Mohammadian, "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis," *Accident Anal. Prevention*, vol. 136, Mar. 2020, Art. no. 105405.

[23] E. Hak Lee, K. Kim, S.-Y. Kho, D.-K. Kim, and S.-H. Cho, "Estimating express train preference of urban railway passengers based on extreme gradient boosting (XGBoost) using smart card data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2675, no. 11, pp. 64–76, Nov. 2021.

[24] X. Dong, T. Lei, S. Jin, and Z. Hou, "Short-term traffic flow prediction based on XGBoost," in *Proc. IEEE 7th Data Driven Control Learn. Syst. Conf. (DDCLS)*, May 2018, pp. 854–859.

[25] Z. Mei, F. Xiang, and L. Zhen-hui, "Short-term traffic flow prediction based on combination model of XGBoost-LightGBM," in *Proc. Int. Conf. Sensor Netw. Signal Process. (SNSP)*, Oct. 2018, pp. 322–327.

[26] B. Sun, T. Sun, and P. Jiao, "Spatio-temporal segmented traffic flow prediction with ANPRS data based on improved XGBoost," *J. Adv. Transp.*, vol. 2021, pp. 1–24, May 2021.

[27] S. Bouktif, A. Fiaz, A. Ouni, B. Alnaqbi, F. S. Alsereidi, and F. A. Alsereidi, "Bayesian optimized XGBoost model for traffic speed prediction incorporating weather effects," in *Proc. 4th Int. Conf. Intell. Comput. Data Sci. (ICDS)*, Oct. 2020, pp. 1–7.

[28] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 4765–4774.

[29] E. H. Lee, "Exploring transit use during COVID-19 based on XGB and SHAP using smart card data," *J. Adv. Transp.*, vol. 2022, pp. 1–12, Sep. 2022.

[30] L. S. Shapley, "A value for N-person games," *Contributions to the Theory of Games Annals of Mathematics Studies*, vol. 2, no. 28. Princeton, NJ, USA: Princeton Univ. Press, 1953, pp. 307–317.

[31] T. Cheng, J. Haworth, and J. Wang, "Spatio-temporal autocorrelation of road network data," *J. Geographical Syst.*, vol. 14, no. 4, pp. 389–413, Oct. 2012.

[32] D. Zang, J. Ling, Z. Wei, K. Tang, and J. Cheng, "Long-term traffic speed prediction based on multiscale spatio-temporal feature learning network," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3700–3709, Oct. 2019.

[33] D. Ma, X. Song, and P. Li, "Daily traffic flow forecasting through a contextual convolutional recurrent neural network modeling inter- and intra-day traffic patterns," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 5, pp. 2627–2636, May 2021.

[34] G. Lee, Z. Ding, and J. Laval, "Effects of loop detector position on the macroscopic fundamental diagram," *Transp. Res. Part C, Emerg. Technol.*, vol. 154, Sep. 2023, Art. no. 104239.

EUN HAK LEE received the Ph.D. degree in civil and environmental engineering from Seoul National University, Seoul, South Korea, in 2021. He is currently an Assistant Research Scientist with the Texas A&M Transportation Institute. His research interests include intelligent transportation systems, traffic engineering, machine learning, and mobility data analysis.