Cite this article
Ku D, Na S, Kim J and Lee S (2021)
Real-time taxi demand prediction using recurrent neural network.
Proceedings of the Institution of Civil Engineers – Municipal Engineer **174(2)**: 75–87,
https://doi.org/10.1680/jmuen.20.00005

Research Article
Paper 2000005
Received 16/01/2020;
Accepted 18/12/2020;
Published online 14/04/2021

Keywords: statistical analysis/transport management/transport planning

Municipal Engineer

ice Publishing

Real-time taxi demand prediction using recurrent neural network

Donggyun Ku
PhD candidate, Department of Transportation Engineering, University of Seoul, Seoul, Republic of Korea (Orcid:0000-0002-8106-393X)

Sungyong Na
PhD candidate, Department of Transportation Engineering, University of Seoul, Seoul, Republic of Korea (Orcid:0000-0002-5425-3367)

Jooyoung Kim
Professor, Department of Transportation Planning & Management, Korea National University of Transportation, Gyeonggi, Republic of Korea (Orcid:0000-0002-5425-3367)

Seungjae Lee
Professor, Department of Transportation Engineering, University of Seoul, Seoul, Republic of Korea (Orcid:0000-0001-9081-2835) (corresponding author: sjlee@uos.ac.kr)

The study aims to predict the location of taxi users, based on an algorithm that was built using a map learning method, which is one of the techniques of deep learning. As the location data of taxi riders showed sequential characteristics over time, learning was performed using a recurrent neural network, which is suitable for predicting dynamic changes over time. The main data used in the analysis were the Seoul Metropolitan Government's taxi tachometer data. These data were collected over a span of six months, from February 2018 to July 2018. Seoul Metropolitan Government's building data and Seoul public transportation smart card data were used as secondary data sources to reflect taxi traffic characteristics. Deep learning results were reviewed using different accuracy values based on combinations of the data sources, such as taxi data only, taxi data and building data and taxi data and smart card data. As a result, the algorithm was able to accurately obtain the distribution of taxi passengers' boarding positions compared to the actual taxi riding pattern through statistical analysis. On the basis of these predictions, the asymmetric characteristics of taxi traffic in terms of transport planning and management can be solved.

Notation

$\sim c_t$	tanh function of the input layer
b_c	bias of hidden layer$_t$
b_f	bias of hidden layer$_{t,t-1}$
b_i	bias of hidden layer$_{t-1}$
c_t	cell state$_t$
f_t	sigmoid function of the forget layer
h_{t-1}	hidden layer$_{t-1}$
i_t	sigmoid function of the input layer
W_c	weight of the hidden layer of time$_t$
W_i	weight of the hidden layer of time$_{t-1}$
x_t	taxi occupied time$_t$

1. Introduction

Taxis have asymmetrical traffic characteristics. This means that the time and place in which taxis operate are concentrated. The location data of taxis in Seoul was used to analyse the driving characteristics of taxis, finding that they showed different characteristics depending on the day of the week or the month, time and districts (Choo *et al*., 2013). This shows that unlike public transportation, which has the characteristics of operating efficiency and user equilibrium according to exclusive routes mentioned by Ku *et al*. (2020), taxi traffic characteristics have asymmetry.

To solve this problem, demand–response taxis such as Kakao taxi and call taxi were introduced. However, Cho and Lee (2016) noted that the taxi system should be improved to solve the asymmetry problem faced by Kakao taxi users. Accordingly, the government plans to optimise taxi revenue by applying the deep learning theory to taxi tachometer data, building data and smart card data in order to predict the existing driving behaviour, characteristics of taxis, and location of taxi stations in terms of demand response services, and to establish demand–response taxi services. Currently, several deep learning studies in Korea have been conducted on image recognition techniques used in license plate recognition systems, on speech recognition techniques used in virtual assistant systems, and so on Nevertheless, studies on traffic are judged to be as well meaningful because they are in limited areas, such as signal system optimisation. In addition, taxi passengers can reduce waiting times. From the point of view of taxi drivers, the pattern of taxi passengers' boarding locations can be identified on a daily and hourly basis to maximise taxi profits and minimise the tolerance rate. Furthermore, it could be a justification for controlling the number of taxis in Seoul when they are saturated.

Therefore, this study identifies the characteristics of Seoul taxis operation through deep learning technology and uses Seoul Metropolitan Government's building data and smart card data to further reflect the characteristics of taxi traffic, to predict the location of Seoul taxis' users. By optimising taxi revenues and allowing taxis to operate more efficiently, this research study will pave the way for building semi-public taxi services, and even mobility as a service (MaaS).

2. Literature review

This chapter provides an overview of the state-of-the-art taxi and deep learning research in Korea and elsewhere. Rodrigues *et al.* (2019) predicted future taxi demand by dividing cities into grid format based on geographical coordinates, and by converting taxi request data into heat maps. Liu and Chen (2017) also predicted the flow of public transportation passengers using the methods of map learning and non-map learning. Liao *et al.* (2018) attempted to predict the demand for taxis in New York City by using deep learning techniques. They also drew conclusion implications by systematically comparing two recently developed deep learning techniques, ST-ResNet and FLC-Net, and applying them to predict the taxi demand in New York City. Xie *et al.* (2017) conducted a test on the proper ratio of taxis using taxi data, and Xu *et al.* (2017) conducted a study to predict the destination of a taxi based on the taxi's driving direction and speed, and by applying a recurrent neural network (RNN) to the taxi data. De Brébisson *et al.* (2015) predicted taxi destinations based on learning of the catchment area with an RNN model fed by a taxi dataset, obtaining more accurate prediction of the destinations than other approaches.

In Korea, Kim and Kang (2019) analysed the performance of temporal-guided networks, which is drawing the most attention among deep learning models for taxi demand prediction, as a preliminary study to build digital twins for self-driving taxi services in Seoul. Other studies have suggested that customers should further increase by improving factors such as waiting time, service time and cost. Jung *et al.* (2016) and Lee *et al.* (2016) developed an algorithm to recognise traffic signs using the map learning method. Sameen and Pradhan (2017) developed an algorithm to detect license plates through the map learning method and video analysis. Sameen and Pradhan (2017) conducted a prediction and verification of the severity of traffic accidents using the non-map learning method. Kim *et al.* (2020) developed a deep learning-based prediction model that takes into account time, space and trajectory patterns, by reflecting urban characteristics and implementing them for efficient learning. The application of a short-term traffic speed prediction model was constructed using long short-term memory (LSTM) with 5 min interval speed data.

In Korea, studies have been conducted in various fields using deep learning techniques, but map learning research using big data such as taxi data or buildings or smart card data is still scarce. Therefore, this study aims to ensure the accuracy of taxi demand prediction by applying RNN and LSTM techniques that reflect taxi and other traffic characteristics (building, smart card data), adding to the taxi passenger prediction research already underway overseas.

3. Research methodologies

This study was conducted in two stages. The first consisted of primary supervised learning that predicted the location of taxi passengers by learning only from the Seoul Metropolitan Government's taxi tachometer data. This is the first deep learning analysis that applied both CNN and RNN to determine whether each of the neural networks could predict changes over time. The second stage consisted of secondary supervised learning that additionally learned building data and smart card data that can reflect taxi traffic characteristics. The two-step method is shown in Figure 1.

Primary supervised learning conducted an analysis using only the taxi tachometer data to determine whether it could accurately reflect the changing characteristics of taxi traffic over time. The analysis first confirmed the accuracy by fixing the taxi's timing (fixed day and time). Point-in-time analysis is a learning method that embodies a picture, such as image recognition. Next, the accuracy was checked by learning all available taxi data. The accuracy was analysed by applying RNN and LSTM techniques. The input values of RNN and LSTM were set to 50 m × 50 m for the taxi data in coordinate units. The sum of the number of taxi passengers in each catchment area was set as the input value. The output was calculated for the number of passengers and the accuracy of each time of each catchment area. Then, secondary supervised learning was applied to predict the changing characteristics of taxi traffic over time and to check the accuracy of the prediction of the passengers' boarding locations reflecting those taxi traffic characteristics, using Seoul Metropolitan Government's building data, which can reflect taxi traffic characteristics, and Seoul's smart card data, as auxiliary data.

3.1 Taxi tachometer data

In this study, proper processing was carried out on data collected in real time by 75 000 taxis from the Seoul Metropolitan Government's taxi tachometers, which are mandatory for location transmission. The average capacity of taxi

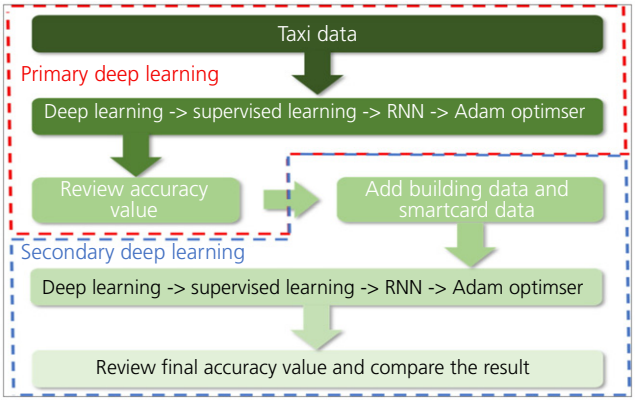

Figure 1. Process flow chart

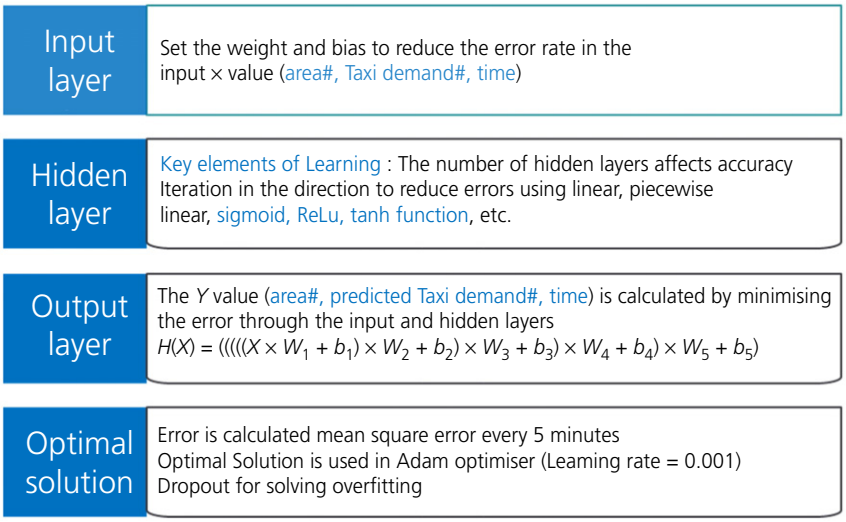

Figure 2. Input and output value for learning

data was approximately 247.3 GB. The purpose of data collection was to identify the taxi passenger's boarding location. The ID and coordinates of each taxi were gauged over 10 s, along with the speed, operation time, altitude, direction and whether the taxi had passengers (tolerance = 0 if the taxi had no passenger, else, tolerance = 1). To use the data for predicting the location of taxi passengers, the analysis was performed by selecting a point with an ID, a set of coordinates, a time and a tolerance that changes from 0 to 1 (Figure 2).

The data used for this study were extracted from the Seoul Metropolitan Government's taxi tachometer data with the geographic information system (GIS) for tolerances and actual differences. Table 1 describes the taxi tachometer data, which is used as the main data for predicting the location of the taxi passengers. The data were collected from February to July 2018. The application of data set the taxi passenger entry point to 1(0→1), extracting by changing the busy (1) to extract points with vacant (0). Figure 3 shows the location of the taxi data vacancy or busy expressed in GIS and Figure 4 shows how to make passengers' boarding location.

Seoul taxi tachometer data can track the location and time of each taxi ID. It can also estimate the taxi driver's income and the taxi's tolerance time. Traffic patterns can also be checked by region and time and have asymmetry. The research study by Lim and Kwon (2019), based on taxi data, compared and analysed the locations of taxi rides and taxi stands in the micro-space range according to the density of taxi rides by each link. As a result, the taxi ride space in Seoul had a high degree of space autocorrelation, and the taxi ride density area was distributed in the widest range.

Table 1. Seoul taxi tachometer data

Data purpose	To identify the taxi passenger's destination
Data list	Points that change from ID, coordinates, time, vacant (0) to busy (1)
Data date	February 2018–July 2018
Data day	Weekday (Mon, Tues, Wed, Thu), weekend (Fri, Sat, Sun)
Data size	247.3 GB
Data application	Set taxi passenger entry point to 1(0→1), Extracting by changing the busy (1) to vacant (0). Extract point with vacant (0)

In addition, Hwang and Yun (2006) compared and analysed the average driving distance, average business hours, operating distance, tolerance distance and distance difference rate, with the one-month tachometer data of the corporate and private taxis. The operating distance was defined as the distance the vehicle moves while the passenger was on board during working hours, while the tolerance distance was defined as the distance travelled without passengers on board. The distance change rate was also defined as the proportion of the distance travelled while the passenger was on board. As a result, it was revealed that there was a difference in the operation status between the two taxis, highlighting the need for an accurate sample survey of private taxis. Kim *et al.* (2017) evaluated the potential implementation of electric taxis by analysing the operating and charging behaviour of electric taxis based on actual Seoul tachometer data of electric taxis, and by carrying out a feasibility study and environmental analysis for implementing electric taxis. The analysis of the boarding frequency by time of day and passenger boarding distance for each day showed financial feasibility and high implementation feasibility of converting private taxis to electric

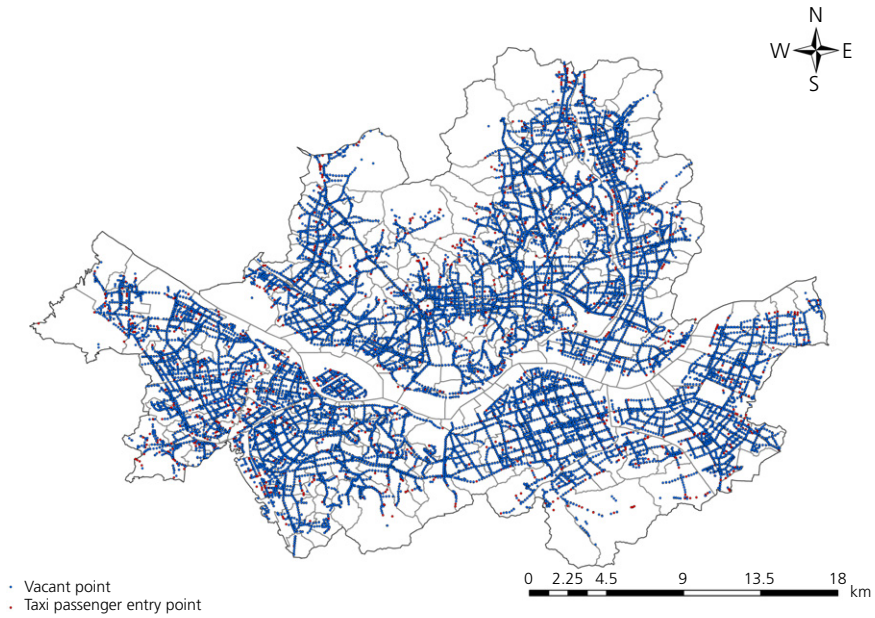

Figure 3. Location of taxi data vacancy using GIS

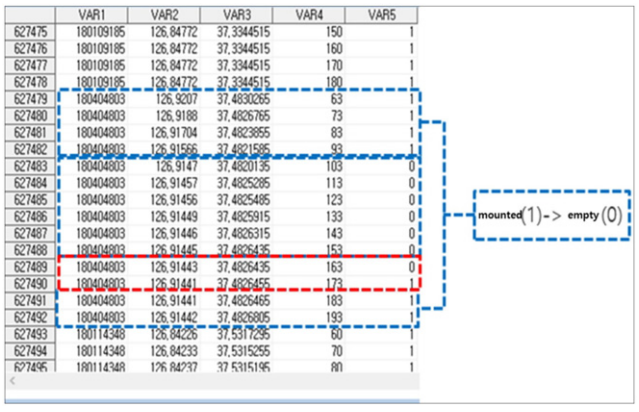

Figure 4. Taxi data establishment

taxis, which would further increase the asymmetry of the taxi traffic parameter such as journey duration and taxis space zones.

3.2 Seoul building and smart card data

To guide building data, the analysis was conducted by extracting traffic volume according to the building's identity, coordinates, area and use. The processed and learned building data are described in Table 2.

To guide and guide smart card data, the analysis was performed by extracting the number of people on and off the station ID, coordinates, time and station. The processed and learned smart card data are described in Table 3.

Table 2. Seoul building data

Data purpose	To identify the taxi passenger's destination
Data list	ID, coordinates, time, building area, building use
Data year	2018
Data day	—
Data size	16.6 GB
Data application	Reflects the population generated for each building use and assumes taxi passengers

Table 3. Seoul smart card data

Data purpose	To identify the taxi passenger's destination
Data list	ID, coordinates, time, number of passengers, number of people getting off
Data date	February 2018–July 2018
Data day	Weekday (Mon, Tues, Wed, Thu), weekend (Fri, Sat, Sun)
Data size	216.8 GB
Data application	Assume conversion rate to taxi

3.3 Recurrent neural network

This study applied an RNN, a deep learning technique for learning data that changes over time, such as time-series data. RNN was used to reflect taxi time characteristics because learning by time is possible. The circular neural network is an implemented neural network theory that allows the network to

be connected at the reference point (t) and the next point ($t+1$), as represented in Figure 1, to reflect the characteristics of the previous situation. In an RNN, the connections between units have a cyclical structure, and it is known to be suitable for handling data containing a sequence or time flow, such as voice, music and video. RNN architectures have the advantage of being able to link previous information to the present and varying lengths of input and output. In addition, as the network structure allows the acceptance of inputs and outputs regardless of the length of the sequence, it has the advantage of being able to assume a variety of flexible structures depending on the situation.

RNN can be used not only to predict stock prices and time-series data but also to develop a model (Jung *et al.*, 2017) that predicts traffic and trade conditions in urban areas.

However, if a network of neural networks is connected at all levels when performing iterations, the learning accuracy may be compromised or inefficient due to the accumulation or loss of bias values from long-standing data. Furthermore, a wide gap in characteristics over time can lead to a long-term dependency problem, which can lead to greater errors as learning of historical information accumulates. In other words, RNN includes long-term information that is not needed (long-term dependencies). In addition, the backpropagation is multiplication [−1, 1]→Convergence weights 0. Each neural network's weights receive an update proportional to the partial derivative of the error function with respect to the current weight in each iteration of the training. This can be a disadvantage when learning other taxi data with clear traffic characteristics over time. Thus, LSTM network architectures were implemented to address these issues.

3.4 Long short-term memory

LSTM is a type of RNN that is known to be particularly appropriate for long-term learning. LSTM can adjust the layer at which the input data are learned cyclically, to control the impact of previous information on current information, add related information, and to additionally control the level at which the output is affected again. This technique creates a forget gate for problems that occur when analysing over a long period of time, thus erasing data from prior periods that are less relevant at the time of analysis. Therefore, the learning time is reduced, and the accuracy is improved (Figure 5).

In LSTM, the cell state is divided into two vectors 'h_t', 'c_t'. The 'h_t' means short-term state and the 'c_t' means long-term state. The core of LSTM is to learn what the network remembers, deletes and what to read from the long term (c_t). The long-term status (c_t) is learned through the forget gate (f_t), input gate (i_t, g_t) and output (o_t). The structure is shown in Figure 6.

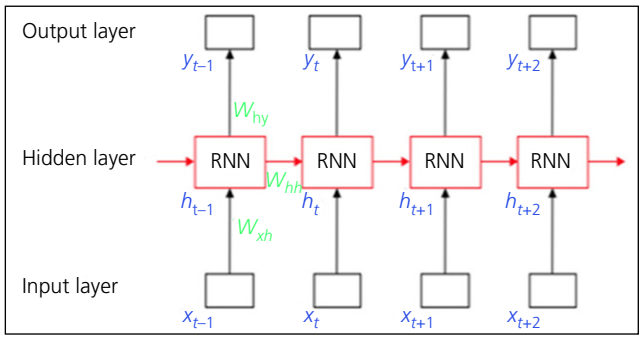

Figure 5. Schematic diagram of RNN

The forget gate (f_t) is a gate to forget information from the past. Receive previous short-term memory (h_{t-1}) and current input information (x_t). It is activated through the sigmoid function and exported to the current long-term state (c_t). At this time, the output range of the sigmoid function is either 0 or 1: if the value is 0, the model forgets the information of the previous state, and if it is 1, it remembers the information of the previous state completely. The forget gate (f_t) is shown in the following expression:

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right)$$

$$\sigma = \sum\left(W_f[h_{t-1}, x_t] + b_f\right)$$

The input gate ($i_t \odot \sim c_t$) is a gate for remembering current information. The values enabled by the hyperbolic tangent function (tanh), such as those enabled by the short-term memory(h_{t-1}) and the current input information (x_t) sigmoid function, are calculated by calculating the matrix product (Hadamard product) and added to the long-term state (c_t) that passed the forget gate (f_t). i_t ranges from 0 to 1 and g_t ranges from −1 to 1, which indicates the strength and direction of information. The input gate ($i_t \odot \sim c_t$) is shown in the following expression:

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

$$\sim c_t = \tanh(W_c[h_{t-1}, x_t] + b_c)$$

$$\sigma = \sum(W_i[h_{t-1}, x_t] + b_i)$$

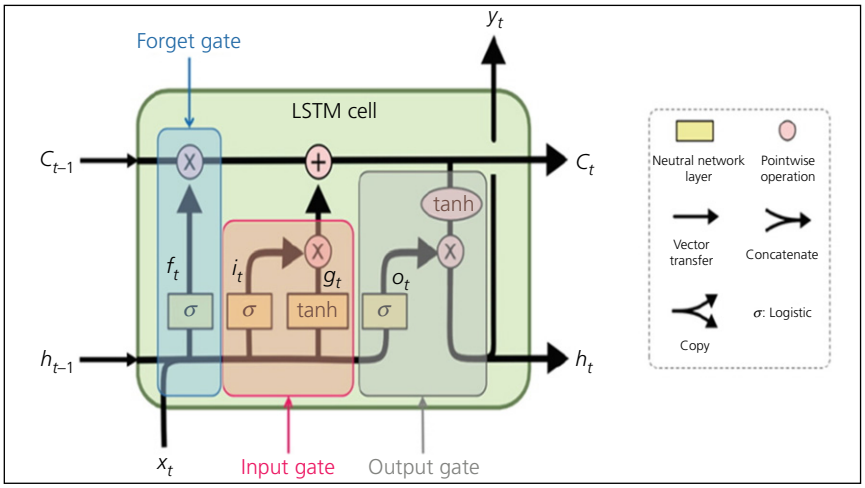

Figure 6. Structure of LSTM cell (source: http://colah.github.io)

Update the long-term state (c_t) in the t time through the values calculated from the forget gate (f_t) and input gate ($i_t \odot \sim gt$), and follow this expression

$$c_t = f_t \odot c_{t-1} + i_t \odot \sim c_t$$

The output gate (o_t) is the step in determining the output value (y_t) and the short-term state (h_t) in t time, and the output value and the short-term state have the same value. The output gate (o_t) outputs the value enabled through the short-term memory (h_{t-1}) and the current input (x_t) signature function, and the output value (y_t, h_t) is determined by the matrix multiplying the output value and the long-term state (c_t) renewed in the preceding stage with the hyperbolic tangent function. Output gate (o_t) and output value (y_t, h_t) are shown in the following expression:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$

$$y_t,\ h_t = o_t \odot \tanh(c_t)$$

$$\sigma = \sum(W_i[h_{t-1}, x_t] + b_i)$$

4. Taxi supervised learning with building and smartcard data

The taxi tachometer data were the main data for predicting the individual boarding location of taxi users. In addition, the Seoul

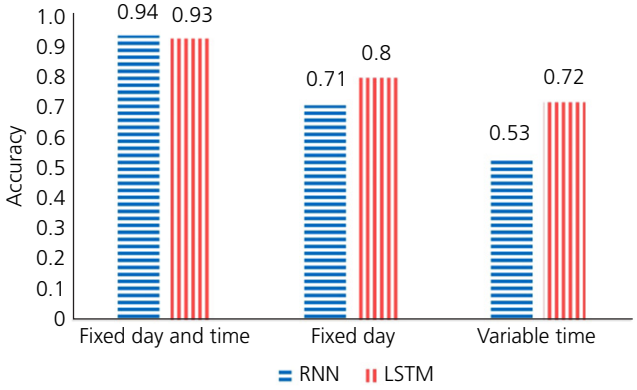

Figure 7. Accuracy variation of each technique (graph)

Metropolitan Government's building data, which could reflect taxi-traffic characteristics, were added to the input data to improve the accuracy. This study used the building data to apply traffic volume, area characteristics, coordinates and buildings' values in the map learning. Like the building data, the city's public transportation data, which have built-in taxi traffic characteristics, were added to the input data in order to enhance the accuracy of predicting the individual ride location. In this study, taxi data, station ID, coordinates (X = iteration, Y = accuracy), time and number of people travelling to and from each station were extracted and used for analysis (Figures 7 and 8).

4.1 Primary supervised learning

This analysis was performed using RNN and LSTM to find algorithms that reflect changing taxi characteristics over time.

In addition, the accuracy of each algorithm was verified by comparing and analysing the learning of data with fixed and not fixed time.

When the day of the week and the time were fixed, an algorithm-specific analysis was performed on short-term forecasts, such as recognising a single image. According to the analysis by algorithm, the accuracy was 0.94 when using RNN and 0.93 when using LSTM. High accuracy means that in short-term situations, the taxi ride was learned and predicted accurately. However, as a result of removing the essential characteristics of taxi data, which dynamically change over time, additional supplementary learning is needed (Table 4).

In Figure 9, the connection of white dots represents the taxi's boarding location in the learned data, and the shaded areas represent the concentration of the actual boarding locations. The reason for high accuracy of the RNN, 0.9 or higher, is that when taxi data are learned with a fixed time, there is not much difference in the algorithm changes of the map learning hidden layer. This means that accurate forecasts can be obtained by applying any algorithm in the time fixation analysis. The results of learning by algorithm are shown in Table 5. However, learning taxi data with fixed time can secure high accuracy, but it cannot reflect the changing characteristics of taxi traffic over time, so additional learning is needed according to changes in time.

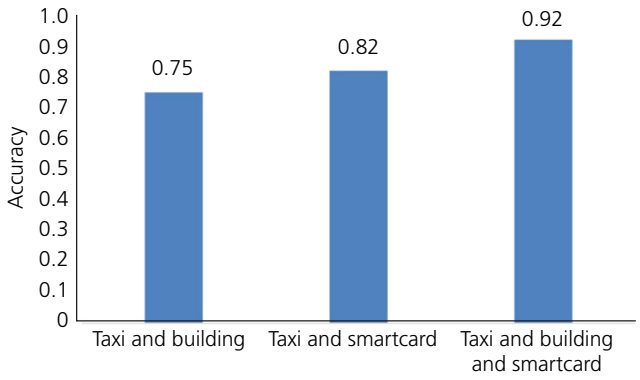

Figure 8. Accuracy variation of each technique (graph)

The analysis was then performed without fixing the time of taxi data in order to analyse the traffic characteristics of the taxi that changed over time. This is a learning method that predicts the ride position according to the time change of the taxi, such as the implementation of animation by completing a series of photographs. First, the day of the week was fixed in the taxi data of 17 days (three Saturdays) and the analysis was performed according to the application of RNN and LSTM. The accuracy of the RNN decreased from 0.93 to 0.71, even though the days of the week were fixed to minimise changes in taxi characteristics due to time changes. The LSTM was also estimated at 0.8. The days of the week were fixed to control the width of changes in taxi traffic characteristics over time. As the time changed, the taxi usage pattern became more complicated, resulting in a relatively lower accuracy.

Finally, all taxi data without a fixed day of the week and time were analysed. Taxi data collected over six months were learned in chronological order. This was done to determine whether the techniques could really reflect changes in taxi traffic characteristics over time. The analysis showed that for RNN, the accuracy was 0.53, and for LSTM, it was 0.72. Hence, it is analysed that additional data learning is needed to fully learn the traffic characteristics of taxis over time and reflect it in the algorithm. The results of the accuracy analysis are categorised as: both the time and days of the week being fixed prior to analysis, and only the days of the week being fixed.

4.2 Secondary supervised learning

To ensure the accuracy of taxi ride demand, the input value was added considering the number of passengers by subway time of smart card data, type of building according to the shape of the building and the amount of traffic generated according to the floor area. Although it does not match one-to-one with taxi passengers, it is based on the experience that areas with high demand for taxis occur basically around subway stations and in areas where commercial districts are concentrated.

Additional data learning was carried out in order to learn the dynamic changes in the taxi traffic characteristics over time more effectively. This was conducted using LSTM, which showed superior accuracy compared to using RNN.

Table 4. Data set example for learning

Taxi data	**ID** 180109185	**X** 126.84772	**Y** 37.3344515	**Time** 150	**Empty/Mounted** 1
Building data	ID	X	Y	Usage	Area
	589022	126.968732	37.590591	Multi-unit dwelling	6572.1
Smart card data	ID	X	Y	Time	Drop-out
	24511800	126.92234	37.49977	00 : 00	4

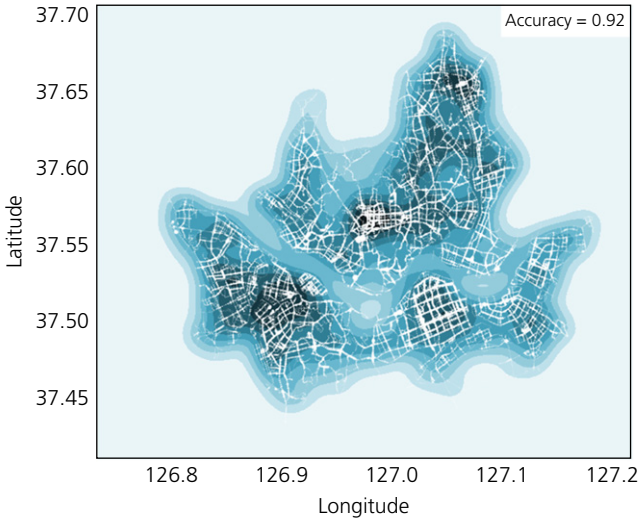

Figure 9. Compare learning data and actual data when fixed on hours and days of the week

Table 5. Accuracy variation of each technique

Technique	Fixed date and time	Fixed date	Variable time
RNN	0.94	0.71	0.53
LSTM	0.93	0.80	0.72

The Seoul Metropolitan Government's building data and smart card data were added to the auxiliary data to enhancing learning accuracy. The accuracy before and after further analysis was then compared.

Building data could reflect taxi-traffic characteristics and were included as additional learning data. The value of the building could be evaluated, thus estimating the amount of traffic generated by each building. The estimated amount of traffic flow caused by different buildings could be predicted in areas where the taxi demand was high, thus complementing and reflecting taxi-traffic characteristics. This was carried out under the assumption that if the traffic generated due to the building was high, the taxi demand would be high in the surrounding area. Further study and analysis of building data showed an accuracy value of 0.75. The accuracy did not increase significantly because the building data provided a daily traffic volume for each building and its area, which could not reflect the time-based characteristics. If the building data existed temporally and would be learned, the learning effect could be increased by increasing the taxi traffic characteristics reflected by the building data.

Following that, the LSTM model learned additional smart card data provided by the Seoul Metropolitan Government with the aim of improving the accuracy of predicting boarding locations of taxi passenger. This was because it could reflect the taxi-traffic characteristics similar to the building data. The smart card data were analysed based on the assumption that the demand for taxis would be high in proportion to the number of people arriving to and leaving from the public transport service at that time and in a location.

The accuracy value considering additional smart card data was found to be 0.82. This was 0.1 units higher than when considering only taxi data. It is believed that smart card data reflect traffic characteristics more realistically, based on the time of travel as compared to building data. However, the accuracy is still lower than 0.92, when the days and hours of the week are fixed, and the smart card data does not fully reflect changes in the traffic characteristics of taxis. Therefore, additional data that can reflect taxi-traffic characteristics, such as weather and average income, should be learned.

To improve the accuracy of predicting boarding locations of taxi passengers, both building data and smart card data were studied. After learning taxi data from the main data, and learning the building data and smart card data from the auxiliary data, the accuracy value considering the fixed day of the week and time improved to 0.89, which is close to 0.92.

4.3 A study on the learning of taxi data

Further, it was attempted to understand why learning using taxi data only showed an accuracy value of 0.72, which increased when learning additional data. First, the fact that only the taxi data were processed to obtain an accuracy of 0.72 did not properly reflect changes in the traffic characteristics over time. This can be adjudged to have occurred due to the lack of taxi data points. Comparing the average number of taxi rides per hour on Saturdays (26 days), which was analysed, with the accuracy of each time, a fairly similar trend was observed.

When the building data and smart card data were learned from auxiliary data along with taxi data, more accurate results (0.89) were obtained if the day and time of the week were fixed. The problem of lack of sample numbers that occurred during the time when the number of taxi rides was low appears to have been solved through secondary data learning. Therefore, it is deemed necessary to secure additional taxi data to predict the taxi boarding location, which is the aim of this study. It is also inferred that the key is to secure additional data that reflects the taxi-traffic characteristics to the extent that the accuracy is supplemented at a time when the average number of taxis is low (Table 6).

4.4 Visualisation of predicted taxi passengers boarding location

The asymmetrical phenomenon of taxi traffic was discovered through the identification of taxi status and learning of taxi

tachometer data. To overcome this phenomenon, the predicted results from the learned data were provided to the driver and the user, and visualisations were conducted to make it easier to understand. Through this, the purpose is to implement taxi traffic patterns that can satisfy both users, drivers and mitigate asymmetry. For visualisation, taxi ride locations distributed by coordinates were divided into 20 m^2 of catchment area. The demand for taxis over time was visualised on Fridays at 0 : 00–01 : 00, 12 : 00–13 : 00 and 21 : 00–22 : 00 based on the learned data. The number of taxi passengers learned by time was classified into a catchment area, before being sorted into the administrative dong to which the catchment area belongs (Figure 9). The result showed that Yeouido-dong had the highest total based on training data, and the highest accuracy was found between 0 : 00 and 01 : 00, when taxi demand was the highest during the day. This is judged to be due to the difference in the amount of data on taxi rides over time, and if a lot of data would be obtained, accuracy would be improved. The specific training data, test data and accuracy are described in Table 7.

Table 6. Accuracy variation of each technique

Technique	Taxi and building	Taxi and smartcard	Taxi and building and smartcard
LSTM	0.75	0.82	0.92

In addition, the Seoul subway network (smart card data) and building data were used to compare the relationship between taxi passengers and vehicles. The smart card data were analysed by extracting the number of people on and off the station ID, coordinates, time and station. The building data were obtained by extracting traffic volume according to the building's identity, coordinates, area and use. Figure 10 shows the relationship between subway and taxi passenger vehicles using the Seoul subway network (smart card data). The figure shows that the demand for taxis is concentrated around the subway station. Therefore, it was found that many taxi rides were predicted near the main transit stations with high connectivity in the subway network. Figure 11 shows the relationship between the purpose of the building and the taxi ride using Seoul building data. If one looks closely, the more sales facilities there are in the area (such as shopping malls and restaurants), the more the demand for taxis is concentrated. As a result, many taxi rides were predicted in areas where sales facilities were concentrated (Figure 12).

As mentioned earlier, the results of predicting Seoul city data and taxi rides by time and space show asymmetry in taxi traffic in dense areas.

5. Conclusion

5.1 Summary and conclusion

The steps for deep learning were carried out in two stages. First, one looked at the changes in the accuracy values of deep

Table 7. Train data, test data, accuracy for time interval

		Time interval			
Sortation	**Dong name**	**0 : 00–1 : 00**	**12 : 00–13 : 00**	**20 : 00–21 : 00**	**Total**
Train data	Yeoui-dong	56 789	63 610	40 440	938 796
	Sinwol 7-dong	30 771	53 570	35 098	812 678
	Jayang 2-dong	63 239	13 961	27 298	728 991
	Hangangno-dong	38 075	32 909	42 625	680 326
	Haengsin 3-dong	65 115	16 623	42 582	677 642
	⋮	⋮	⋮	⋮	⋮
	Sum	14 252 832	10 203 342	11 990 981	246 425 040
Test data	Yeoui-dong	61 822	78 502	40 177	981 080
	Sinwol 7-dong	32 497	57 854	36 516	891 471
	Jayang 2-dong	68 501	17 541	26 994	771 610
	Hangangno-dong	42 163	28 970	44 020	714 845
	Haengsin 3-dong	72 098	19 096	41 438	740 668
	⋮	⋮	⋮	⋮	⋮
	Sum	14 876 994	11 098 471	12 713 817	261 459 719
Accuracy	Yeoui-dong	0.92	0.81	0.99	0.96
	Sinwol 7-dong	0.95	0.93	0.96	0.91
	Jayang 2-dong	0.92	0.80	0.99	0.94
	Hangangno-dong	0.90	0.86	0.97	0.95
	Haengsin 3-dong	0.90	0.87	0.97	0.91
	⋮	⋮	⋮	⋮	⋮
	Average	0.94	0.88	0.92	0.92

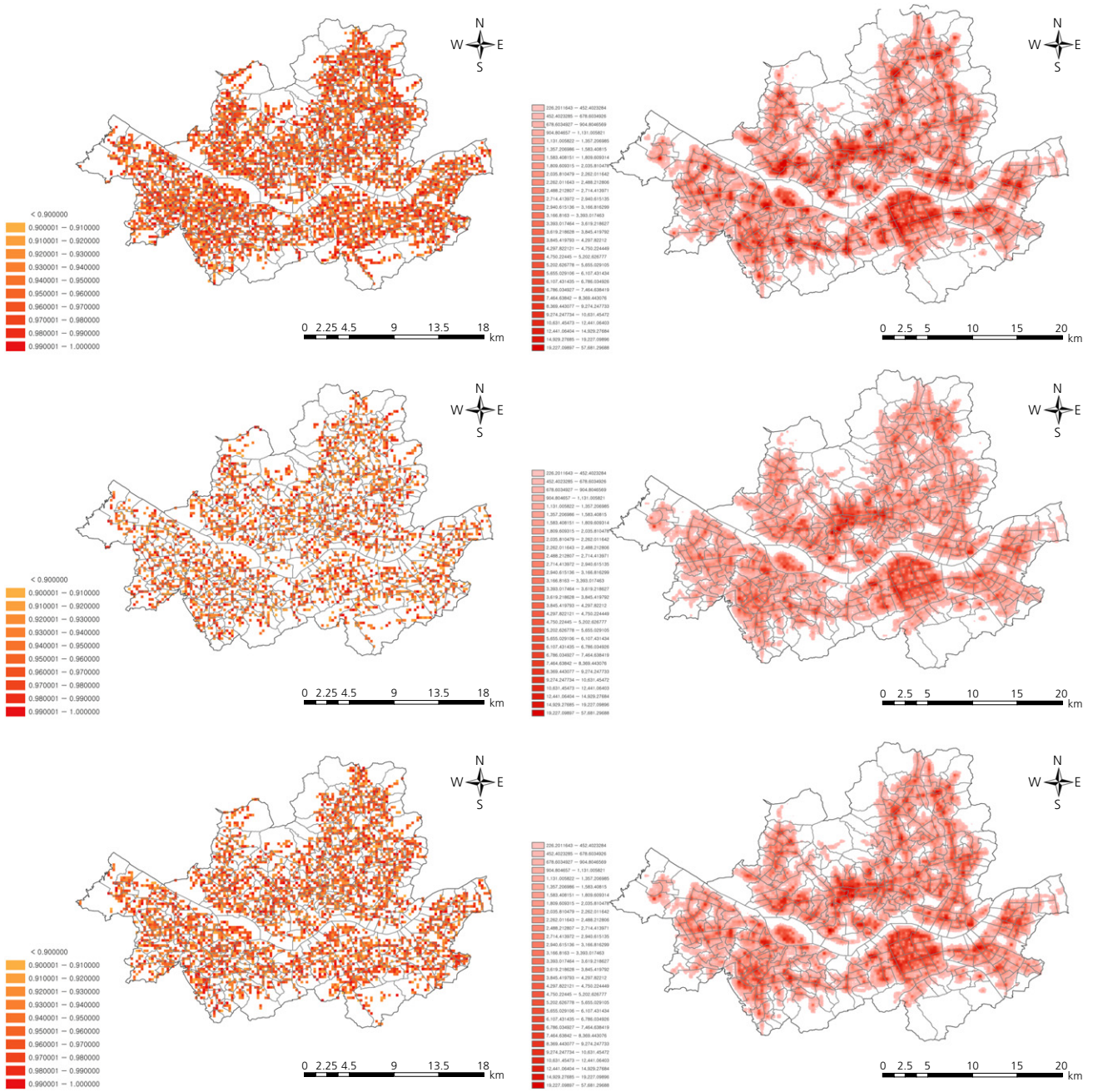

Figure 10. Spatial distribution of Seoul metropolitan taxi

learning when the day and time were fixed using taxi data only, and when the day and time were not fixed. The analysis showed that the accuracy was 0.94 for RNN and 0.93 for LSTM when the day of the week and the time were fixed. The accuracy was 0.71 for RNN and 0.80 for LSTM when the day and time were fixed, and the accuracy when changing both day of the week and time was 0.53 for RNN and 0.72 for LSTM. Subsequently, additional learning was conducted for the deep learning of Seoul building data and smart card data, with auxiliary data to reflect the traffic characteristics of taxis over time through the LSTM technique. When only additional building data were considered, the accuracy was 0.75, which was an improvement but did not have a significant effect. If only additional smart card data were considered, the accuracy

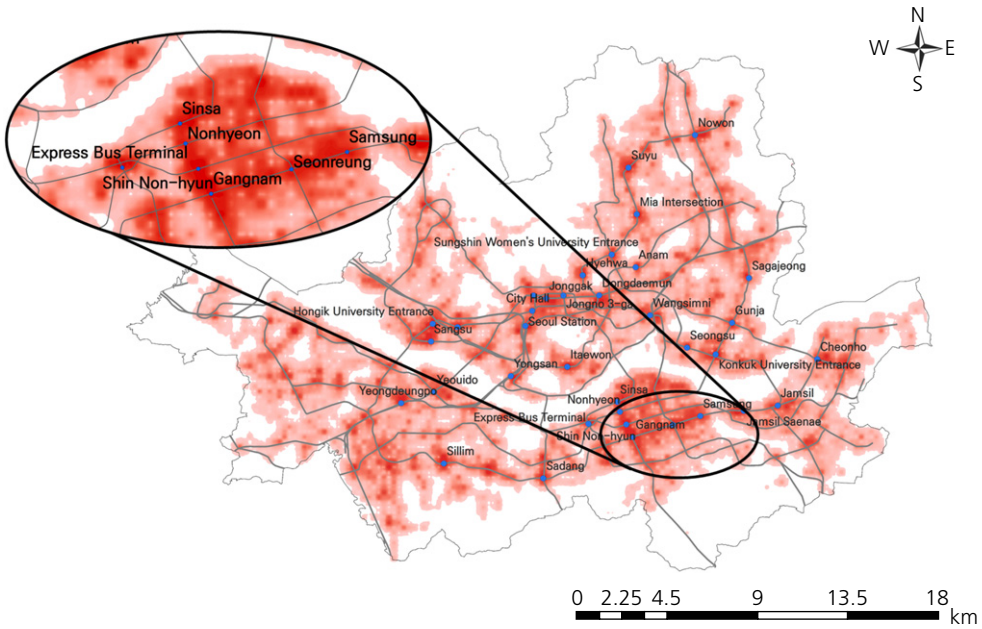

Figure 11. Spatial distribution of Seoul metropolitan taxi and subway network

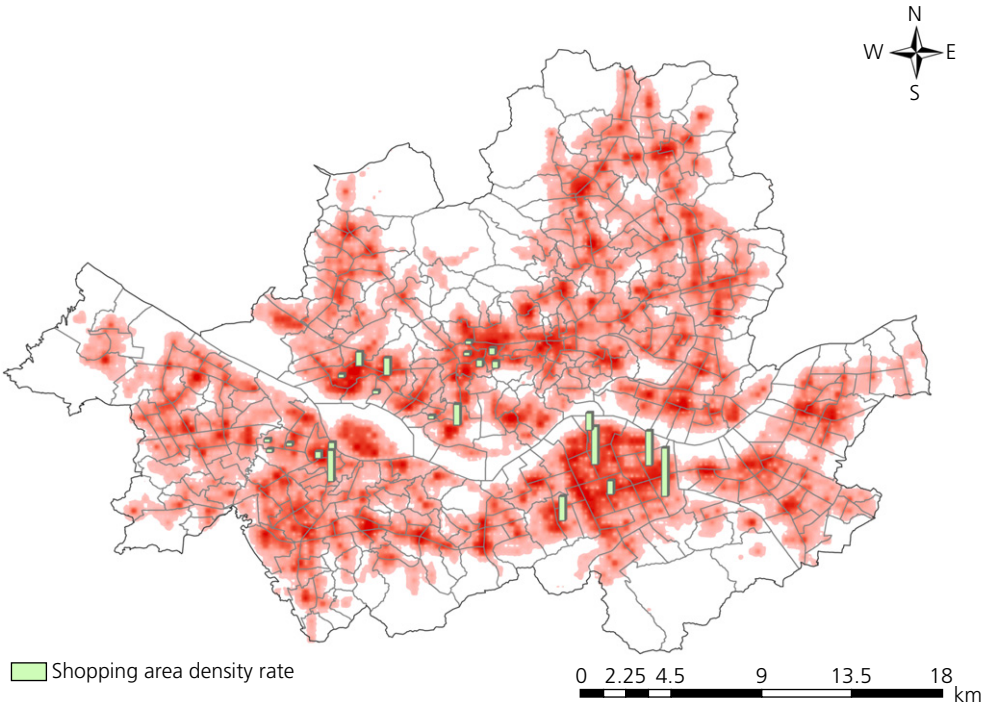

Figure 12. Spatial distribution of Seoul metropolitan taxi and shopping area density rate

was 0.82, and if both building data and smart card data were learned, the accuracy reached 0.92 for both the day of the week and time being fixed in the previous step.

The results show that the accuracy of predicting taxi passenger boarding locations can be increased by learning additional data that can indicate the traffic characteristics of taxis, thereby implying that the accuracy can be further improved through data that realistically reflect the traffic characteristics rather than building data or smart card data.

Finally, the expected boarding location of taxi users predicted through deep learning was set up, aggregated and visualised in an area of 20 m^2 and compared with the visualisation results of the actual data. Comparisons have shown that a similar distribution of taxi demand has been observed, which proves the validity of the prediction of the corresponding algorithm.

In this study, the location of taxi passengers through deep learning techniques is predicted, which are currently limited in the domain of transportation engineering. In addition, if existing taxi-related research focuses on building demand–response services that respond flexibly to the already generated demand, this study can be used to predict taxi-riding demand. When previous studies on taxi demand–response and demand prediction research through deep learning are combined, studies leading to small enhancements in accuracy are expected to emerge as milestones in taxi-related research that will help maximise profits for taxi drivers. Additionally, taxi operations will be streamlined and will ultimately contribute to the establishment of an integrated transportation system (MaaS) through the quasi-public taxi transportation systems.

5.2 Limitations and future scope

Until now, this research study has made real-time predictions on the number of taxi passengers to make taxi demand–response services. The limitations of this study are explained in the following paragraphs.

This study did not reflect the nature of the traffic because it only used Seoul city's taxi data, and the actual taxi traffic directed towards Incheon and Gyunggi. This is deemed to be the most significant factor responsible for the failure to achieve a higher accuracy. In addition, the data itself had limitations, which used secondary learning to reflect taxi-traffic characteristics. Building data were used as auxiliary data, which was collected daily, and thus features were characterised according to the day of the week, but not over time. A study using building data built over time would have produced better results.

In addition to building data and smart card data, in the future, additional data that can reflect taxi-traffic characteristics should be obtained and processed to further increase accuracy. To perform a more accurate analysis, it is necessary to previously set the boarding area of the individual taxi rides rather than aggregating the boarding areas after the analysis. Such complementary measures will enable the development of a maximum fare calculation system for taxi drivers using the reinforced learning method if an accuracy of at least 0.95 in the prediction of the taxi ride location can be obtained. In fact, when the service is introduced, the traffic characteristics of taxis will change because they will be provided with learned data. Therefore, the effect of applying taxi services in the future should be analysed.

Acknowledgements

This study was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019K1A4A7A03112460).

REFERENCES

Cho MJ and Lee YJ (2016) A study on the impacts of use motivation and dissatisfying factors on the continuous intention to use Kakao taxi service. *The e-Business Studies* **17(3)**: 93–116.

Choo SH, Lee HS, Lee TK and Kim JY (2013) Analyzing travel behavior of taxi in Seoul using GPS data. *Journal of Korean Society of Transportation* **68**: 261–266.

De Brébisson A, Simon É, Auvolat A, Vincent Pand Bengio Y (2015) Artificial neural networks applied to taxi destination prediction. arXiv preprint arXiv:150800021.

Hwang JH and Yun DS (2016) A comparative analysis of operational condition of corporation-owned and owner-driver taxies using tachometer output data. *Journal of Korean Society of Transportation* **24(6)**: 45–54.

Jung SW, Lee UH, Jung JW and Shim HC (2016) Development of traffic sign recognition algorithm with deep convolutional neural network. In *KSAE Spring Conference Proceedings*. The Korean Society of Automotive Engineers, Seoul, South Korea, pp. 1100–1103.

Jung HJ, Yoon JS and Bae SH (2017) Traffic congestion estimation by adopting recurrent neural network. *The Korea Institute of Intelligent Transport Systems* **16(6)**: 67–78.

Kim SS and Kang NW (2019) Preliminary study for digital twin design of autonomous taxi service in Seoul: prediction of passenger demand using deep learning and big data. In *Proceedings of The Korean Society of Mechanical Engineers Spring Conference*. The Korean Society of Mechanical Engineers, Seoul, Korea, pp. 2273–2274.

Kim J, Lee S and Kim KS (2017) A study on the activation plan of electric taxi in Seoul. *Journal of Cleaner Production* **146**: 83–93.

Kim Y, Kim J, Han Y, Kim J and Hwang J (2020) Development of traffic speed prediction model reflecting spatio-temporal impact based on deep neural network. *The Journal of the Korea Institute of Intelligent Transport Systems* **19(1)**: 1–16.

Ku D, Na S, Kim J and Lee S (2020) Interpretations of Downs–Thomson paradox with median bus lane operations. *Research in Transportation Economics* **83**: 100909.

Lee D, Yoon S, Lee J and Park DS (2016) Real-time license plate detection based on faster R-CNN. *KIPS Transactions on Software and Data Engineering* **5(11)**: 511–520.

Liao S, Zhou L, Di X, Yuan B and Xiong J (2018) Large-scale short-term urban taxi demand forecasting using deep learning. *2018 23rd Asia and South Pacific Design Automation Conference (ASP-DAC), Jeju, South Korea*, pp. 428–433.

Lim DH and Kwon YH (2019) The comparison on the spatial dense area of taxi ridings and the location of taxi stand using big data analysis. *Journal of the Urban Design Institute of Korea Urban Design* **20(5)**: 99–116.

Liu L and Chen RC (2017) A novel passenger flow prediction model using deep learning methods. *Transportation Research Part C: Emerging Technologies* **84**: 74–91.

Rodrigues F, Markou I and Pereira FC (2019) Combining time-series and textual data for taxi demand prediction in event areas: a deep learning approach. *Information Fusion* **49**: 120–129.

Sameen MI and Pradhan B (2017) Severity prediction of traffic accidents with recurrent neural networks. *Applied Sciences* **7(6)**: 476.

Xie J, Nie YM and Liu X (2017) Testing the proportionality condition with taxi trajectory data. *Transportation Research Part B: Methodological* **104**: 583–601.

Xu J, Rahmatizadeh R, Bölöni L and Turgut D (2017) Real-time prediction of taxi demand using recurrent neural networks. *IEEE Transactions on Intelligent Transportation Systems* **19(8)**: 2572–2581.