Thesis for the Degree of Master of Engineering

Analysis of Incident Impact Factors and Development of SMOGN-DNN Model for Prediction of Incident Clearance Time

by

Gyu Ri Yun

Department of Spatial Information Engineering

The Graduate School

Pukyong National University

August, 2021

Analysis of Incident Impact Factors and Development of SMOGN-DNN Model for Prediction of Incident Clearance Time
(유고처리시간 예측을 위한 영향요인분석 및 SMOGN-DNN 모델 개발)

Advisor: Prof. Sang Hoon Bae

by

Gyu Ri Yun

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
in Department of Spatial Information Engineering, The Graduate
School, Pukyong National University

August, 2021

Analysis of Incident Impact Factors and Development of SMOGN-DNN Model for Prediction of Incident Clearance Time

A dissertation

by

Gyu Ri Yun

Approved by:

Baesanghoon

(Chairman)

(Member)　　(Member)

August, 27th, 2021

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AADT	Annual Average Daily Traffic
KNN	K-Nearest Neighbor
MAPE	Mean Average Percentage Error
CTM	Classification Tree Method
SVM	Support Vector Machine
ANN	Artificial Neural Network
DNN	Deep Neural Network
SMOGN	Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error

유고처리시간 예측을 위한 영향요인분석 및 SMOGN-DNN 모델 개발

윤규리

부경대학교 대학원 공간정보시스템공학과

요약

정체현상은 높은 교통비용과 혼잡을 발생시킨다. 교통사고, 기상이변 등의 유고로 발생하는 비반복 정체에 대한 연구는 첨두시 같은 반복정체에 비해 부족한 실정이다. 비반복 정체를 일으키는 유고에 가장 중요한 지표는 유고가 지속되는 시간이다. 비반복 정체로 인한 높은 교통비용과 혼잡을 효과적으로 해소하기 위해서 유고 처리시간을 예측하는 것은 중요하다. 이를 위해 본 연구에서는 서울특별시 도시고속도로인 내부순환로를 공간적 범위로 선정하여 약 5년치 유고 데이터를 수집하였으며, 해당 데이터 중 종속변수는 유고 지속시간이 아닌 처리시간 데이터로 획득되었다. 독립변수에 대해서는 유고 처리시간 영향요인을 분석하고 해당 영향 요인 중 총 22개의 특성을 선정하여 전처리 및 분석을 통해 예측 모델의 학습데이터로 생성하였다. 기존 연구와 본 연구의 학습데이터 생성 및 예측 모델의 평가 과정에서 높은 유고 처리시간에 대한 과소 예측 문제가 발생하였다. 이에 대한 해결방안으로 본 연구에서는 소수데이터는 오버샘플링하고, 대다수의 데이터는 랜덤으로 언더샘플링해주는 SMOGN기법을 적용한 오버샘플링 학습데이터를 생성하였다. 본 연구에서 개발된 예측 모델은 인공신경망 모델인 ANN과 DNN 모델을 활용하였으며 두 가지 모델의 성능을 비교하는 동시에 원본 학습데이터와 오버샘플링 학습데이터를 모델에 적용하여 결과 비교를 시도하였다. 그 결과 ANN모델에 비해 DNN모델을 사용한 경우와 원본 데이터에 비해 오버샘플링 데이터를 사용한 경우에 더 정확한 예측을 하였으며 결과적으로 SMOGN기법을 적용한 오버샘플링 데이터로 구축된 DNN모델이 MAE기준 예측오차 18.3분으로 최적모델로 선정되었다. 이는 기존에 개발된 예측모델의 과소 예측에 대한 한계점을 보완할 수 있을 것으로 기대하지만, 아직 실무에 적용되기엔 여전히 높은 예측오차를 가지며, 다른 도시고속도로의 데이터를 적용해보는 것으로 추가 검증이 향후 연구로 수행되어야 한다.

Analysis of Incident Impact Factors and Development of SMOGN-DNN Model for Prediction of Incident Clearance Time

Gyu Ri Yun

Department of Spatial Information Engineering,
The Graduate School, Pukyong National University

Abstract

Traffic congestion causes high costs and congestion. Studies on non-recurring congestion caused by accidents such as traffic accidents and extreme weather conditions are insufficient compared to recurrent congestion such as peak hours. The most important indicator for non-recurring congestion is the duration of traffic incident. It is important to predict the incident duration in order to effectively solve high traffic costs and congestion caused by non-recurring congestion. Based on the literature review, various studies on the prediction of incident duration were insufficient in Korea compared to other countries, and various studies suitable for domestic road conditions were needed. To this end, in this study, Naebusunhwan-ro, an urban expressway in Seoul, was selected as a spatial scope to collect data on retention for a period of about 5 years. For independent variables, factors affecting the incident clearance time were analyzed and a total of 22 characteristics were selected among the influencing factors and generated as training data of the predictive model through pre-processing and analysis. In the process of generating training data and evaluating the prediction model of the previous study and this study, there was a problem of under-prediction with respect to the high error processing time. As a solution to this problem, in this study, over-sampling training data by applying the SMOGN technique that over-sampling a small number of data and under-sampling the majority of data at random was generated. The prediction model developed in this study utilized the artificial neural network model ANN and DNN model, and at the same time, the performance of the two models was compared, and the original training data and over-sampling training data were applied to the model to compare the results. As a result, more accurate predictions were made when the DNN model was used compared to the ANN model and when the over-sampling data was used compared to the original data. As a result, the DNN model built with the over-sampling data applied with the SMOGN method showed a prediction error of 18.3 minutes of MAE was selected as the optimal model. This is expected to be able to supplement the limitations of the under-prediction of the previously developed prediction model, but it still has a high prediction error to be applied in practice. By applying data from other urban highways, additional verification should be performed as a future study.

Ⅰ. Introduction

1. Background

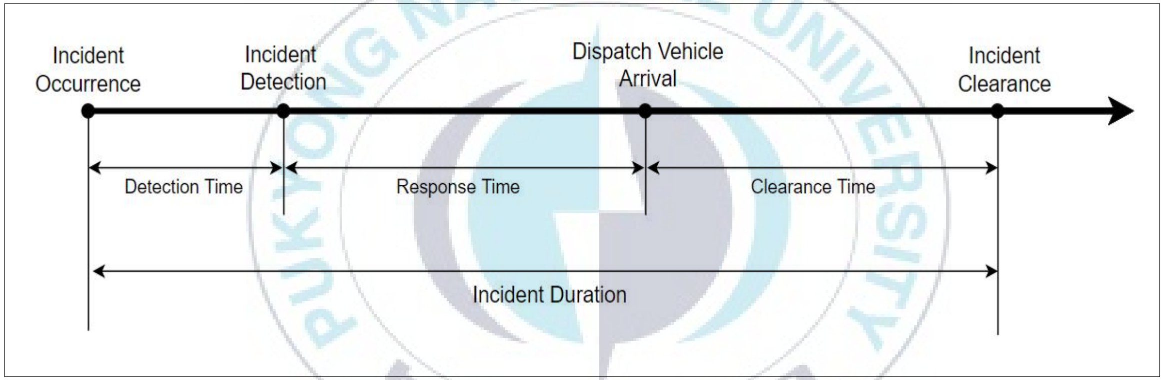
As the number of individual vehicle registrations steadily increases, the demand on limited roads is increasing, and accordingly, the AADT(Annual Average Daily Traffic) is also increasing(MOLIT, 2021). As the average of traffic volume increases, traffic congestion occurs on the road. The main cause of traffic congestion can be divided into recurrent congestion and non-recurring congestion. Recurrent congestion is about the congestion that occurs repeatedly at a certain time such as commuting time, and non-recurring congestion is about the congestion that occurs due to incident, which refers to all unexpected situations that occur on the road such as traffic accidents, construction works, and extreme weather conditions. According to a research by the Korea Transport Institute, as of 2017, about 290 trillion won in annual transportation costs were incurred. Among them, it was reported that the cost due to traffic accidents and traffic congestion accounted for about 35% of the cost (KOTI, 2019).

Related agencies are conducting research on analysis of traffic patterns and prediction traffic conditions to efficient traffic management by providing traffic information and reducing transportation costs. In particular, Previous studies have shown that traffic information is provided with high accuracy by analyzing repeated patterns of traffic condition and predicting them in real time using speed or traffic volume. However, prior research have limited consideration of non-recurring congestion caused by unpredictable incidents.

In the event of an incident on a road that is a normal flow, lanes that the incident occur have been blocked until the end of the incident, causing serious congestion and disappearing traffic pattern. For these reason, it is difficult to predict the traffic situation in which non-recurring congestion has occurred by using the prediction method for recurring congestion. Therefore, in order to more effectively relieve traffic congestion, it is necessary to consider the case of non-repetitive congestion as well as repeated congestion.

Previous studies have demonstrated that there are many influencing factors that affect traffic condition when an incident happened and incident duration which is the most important factor could be predicted by analyzing those factors. Incident duration affects the duration of congestion, accordingly, it could be significant information for predicting real-time traffic condition and providing optimal path and travel time when non-recurring congestion occurs.

As shown in <Fig.1>, incident duration refers to the time from when the incident occurred to the time when the incident was actually handled after it was detected and vehicles to deal with it were dispatched. Since it is difficult to know the exact time when the incident occurred, the period from the detection of the incident to the time the incident is completed is recorded. Therefore, instead of incident duration, the incident clearance time was used in this study which refers to the time from when an incident is detected to the time when incident handling is completed, instead of incident duration.



<Fig. 1> Definition of Incident Duration

2. Previous Study

As an early domestic study on prediction of incident duration, independent variables were selected through correlation analysis based on data from Washington State, USA, and an optimal model was calculated using the SPSS statistical program. This study suggested the necessity of accumulating domestic traffic data as there was a limit to reflect the model developed due to the difficulty of acquiring domestic data, and mentioned that a model suitable for domestic road conditions should be developed in future research(Han WG, 2001).

Afterwards, based on the domestic highway traffic accident breaking news data of the Suwon branch of the Korea Expressway Corporation, a statistical method, Multiple Regression Model, was applied to predict the incident duration. A model was constructed by analyzing the variables of traffic accident data from various viewpoints, and it was confirmed that there was statistical significance between the variables through correlation analysis(Shin C. H., 2002).

As the domestic transportation system has become more advanced than before, a decision tree model was developed according to the type of accident using the Seoul Urban Expressway incident data of the Transportation Management Center. The incident severity divided into three stages was calculated using the number of closed lanes that had a higher correlation compared to other variables. It was judged that in order to improve the reliability of the model, it was necessary to clarify the end of incident time and to accumulate relatively insufficient data for different types of incidents(Kim J. W., 2005).

In accordance with the case of systematic analysis abroad, they attempted to systematically analyze the factors affecting the incident duration by using the incident data of the Korea Expressway Corporation. Exponential linear regression equations were applied to derive explanatory variables(Lee K. Y., 2012).

A non-parametric model using the KNN Algorithm was presented based on 7 years of traffic accident data on national highways in order to develop a higher accuracy prediction model for the incident duration. In addition, as it is analyzed that each incident class has a very large effect on the incident clearance time, the model was constructed by classifying the incident clearance time by incident class, and it was confirmed that the MAPE was lower than the results of the existing models, thus supplementing the limitations of the existing models(Lee SB, 2015).

Overseas, research on the prediction of incident duration has been carried out in a variety of ways from the past to the present. Regarding the research methods, studies have been carried out using Regression Model, Fuzzy System, CTM, Neural Network, Bayesian Networks, Hazard-based duration model, and SVM. Recently, researches using the Combined/Hybrid Model that fused several models have been conducted. As the evaluation method, MAPE was mainly used(Li, 2018).

Through literature review of domestic and foreign studies on the prediction of incident duration, it was confirmed that there are several challenges to domestic research. Accordingly, the number of domestic studies performed and diversity in the prediction models used in the studies was relatively insufficient compared to overseas studies.

In addition, in domestic studies conducted in the past, there were limitations in carrying out research suitable for domestic road conditions and conducting detailed and systematic analysis due to difficulties in acquiring domestic data. Recently, as domestic incident data has been accumulated due to the advancement of the transportation system, studies could have been conducted to predict the incident duration through correlation analysis between variables and analysis of the factors affecting the incident duration. However, the spatial scope of the domestic study was mainly limited to highways, and the most of research had used the statistical method rather than the latest technology. Considering that the traffic conditions are different due to the different road types in each country, it seems that various studies suitable for the domestic road conditions are needed separately from overseas studies.

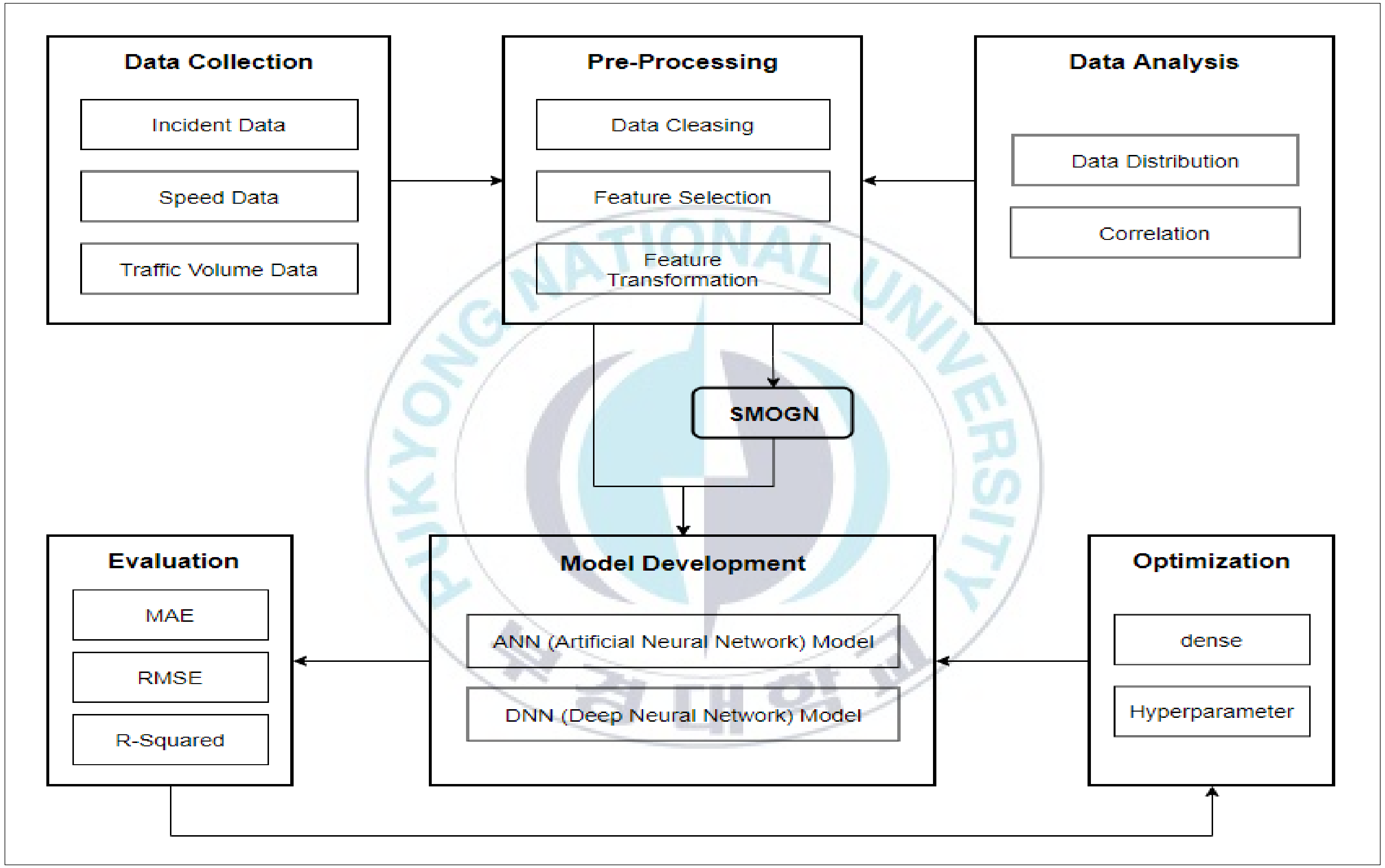
As a common challenge in domestic and foreign studies, there was a problem of under-prediction as some incident in the case of a high value were treated as a fractional data of the distribution of the data. Furthermore, the developed prediction models have a high error value to be applied to the field, so methodological research is needed to further increase the accuracy.

3. Objective

As shown in <Fig. 2>, this study aims to develop a predictive model for the incident clearance time suitable for domestic road conditions. First, domestic incident data was obtained as a data collection process. In order to select a variable to be used in a model for prediction, important factors affecting the incident clearance time were identified, and then each factor was analyzed. Data created through analysis is preprocessed to build input dataset and apply it to predictive models utilizing artificial neural networks.

Recently, models using artificial neural network are well known to produce higher accuracy results than models using existing statistical techniques. Among those, Artificial Neural Network(ANN) Model, which is a representative artificial neural network model and was confirmed to have high accuracy on average in Li's study introduced in the literature review. Since the model has not been applied to domestic studies, this study aims to predict the incident clearance time by applying the ANN model and the Deep Neural Network (DNN) model, an artificial neural network more advanced from ANN.

Furthermore, the problem of under-prediction existed as a challenge to the previous research due to the lack of historical information about the high value of the incident clearance time. As additional methodological research was needed to solve this problem, in this study, Synthetic Minority Over-Sampling Technique for Regression with Gaussian Noise (SMOGN) Method, which over-sampling the minority data, was applied.



<Fig. 2> Research Flow

Ⅱ. Methodology

In this study, incident data including incident clearance time and transportation information were collected as basic data for developing a predictive model and variables necessary for training the model were selected from the collected data. Accordingly, analysis of both each variable and incident clearance time was performed, and data pre-processing was performed to generate training data in a form suitable for the neural network model. In the model building process, in order to solve the problem of under-prediction due to asymmetric data distribution, an input dataset built by using the SMOGN technique was applied to the models, and then prediction evaluation was performed between neural network models with different configurations.

1. Dataset

1) Data Collection

This study used the data provided by the Seoul Facilities Corporation, an institution that manages the transportation system of Seoul Metropolitan City. The spatial scope of the research data is Naebu-Soonhwan-road, which is an urban highway that runs from Mapo-gu to Seongdong-gu, Seoul, and is a continuous flow road consisting of 10 ICs with a total length of about 42 km based on a round trip road. In terms of the temporal range, incident data from September 4, 2014 to September 23, 2019, which is a period consistent with the obtained two data, was used as research data.

Factors influencing the incident duration can be summarized in <Table 1>. The incident information obtained from the Seoul Facilities Corporation includes the clearance time of the incident, the start date and time of the incident, the end date and time of the incident, the section where the incident occurred, the type of the incident, the blocked lanes, the type and number of vehicles to be dispatched. Transportation information includes average speed and traffic volume information. In addition, information that could not be obtained from the data, such as the geometry characteristics of the road and weather condition, was additionally collected using satellite maps and public data from the Korea Meteorological Administration.

<Table 1> Factors and their significant contributions to traffic incident duration (Li, 2018)

Types of Factors	Factors
Incident characteristics	Incident severity, incident type, towing requirements, type of involved vehicles, number of casualties, number of lanes blocked and incident location
Environmental conditions	Rain, snow, dry, or wet
Temporal factors	Time of day, day of week, season, month of year
Roadway geometry	Street, intersection, road layout, horizontal/vertical alignment, bottlenecks, roadway type
Traffic flow conditions	Flow, speed, occupancy, queue length
Operational factors	Lane closures, freeway courtesy service characteristics
Vehicle characteristics	Large trucks, trucks with trailers, taxis, special vehicles, compact trucks, number of vehicles involved
Others	Driver, special events, time that a police officer reaches the site, police response time, report mechanism, accident characteristics reported at accident notification

2) Data Understanding and Pre-processing

In the collected data, the column on the type of incident, which is a factor corresponding to the characteristics of the incident, consists of breakdown, construction, traffic accident, weather, falling object, event/training. Among these, the incident for construction and events/training were removed in the study data because, unlike other types, they were notified of a pre-planned schedule to the drivers so that they could know the duration before the incident occurred. Also, it is impossible to predict duration of the incident for weather and in the case of falling objects, there is an incident that does not block the road and thus does not affect road traffic. Therefore, in terms of the incident type, classes of the column was divided into detailed types of incident including breakdowns and falling, rollover, fire, clash, and crash.

In addition, when it comes to other factors that correspond to the characteristics of the incident, the information of the blocked lanes and the location of occurrences of the incident was also collected. Classes of the column regarding the blocked lane was divided into 1-lane blocking, 2-lane blocking, 3-lane blocking, and 2 lanes blocking because the spatial scope of study is a road with the maximum of 3 lanes except for the IC section. For the location where the incident occurred, only the road section name was recorded, and additional information such as latitude and longitude could not be obtained.

Among the collected data, there are six columns for the temporal factor include the start date and time, the end date and time, which are string type and datetime format such as '0000-00-00 00:00'. Since time is ordinal, it is necessary to convert it into meaningful information. For this reason, the values for day and hour were extracted from the time information, and then the values of the day column were converted into weekdays or weekends, and the values of the time column were converted into peak hours and non-peak hours.

Furthermore, there are seven columns of vehicle dispatched to incident situations and those include police, patrol, ambulance, fire truck, towing car, cleaning car, etc and the value of each column is the number of dispatched vehicles. Additionally, a column which is the total number of vehicles is added for better training model.

The average speed and traffic volume information for the road section that the incident occurred and the upstream road section were extracted from the transportation information dataset which was separately collected and then those were combined with the incident dataset. Among them, a column for Congestion Status was created using the speed data. If the average of the speed of the upstream road and the speed of the road in which the incident occurred is more than 50km/h, it is a value of 'smooth', if it is more than 30km/h and less than 50km/h, the value is 'delay', and if it is less than 30km/h, it is 'Congestion' was assigned a value.

Information that could not be obtained from the data provided by the Seoul Facilities Corporation was additionally collected using external data. As for environmental factors, data from the Korea Meteorological Administration including precipitation and snow cover information were collected for a period consistent with the temporal range of the research data.

Accordingly, for road geometry information, a few columns that contains the road characteristic which has the junction such as confluence or fractionation and whether the road is a general road or tunnel, and the number of sections separated from the access road were added.

Moreover, a weather column for environmental factor was created, and when the incident occurred, if there is snowfall, the value is assigned as 'snow', if there is a value in precipitation, it is assigned as 'rain' and if neither of those, it is set as 'sunny’.

The incident data contains the string values using natural language and the columns that are categorical types rather than numeric types. When these columns are applied to the prediction model, it is difficult to learn the actual meaning of the data, and in particular, since the computer cannot understand natural language, string values must be converted into numbers.

When converting a categorical value to a numeric value, the most used techniques are Label Encoding and One-Hot Encoding. First, Label Encoding is a method of encoding the values in categories in a column with ordinal numbers. This method is often used when values can be ranked. Contrary to this, One-Hot Encoding is a Boolean method in which unique values within one column become each column and are expressed as 0,1 values. The data to which this method is applied has a number of zero values, and as the number of zeros increases, it becomes sparse data, which has a disadvantage in training model. An appropriate method was applied to each column by considering the characteristics of the encoding method and the data characteristics at the same time. <Table 2> is a table that summarizes each variable of the collected data and the data type of each column, and <Table 3> is a table that includes the applied encoding method.

While One-Hot Encoding is applied to the column for geometric factors, the values of the 'confluence' and 'fractionation' columns are 0 and 1 for whether the section is affected by the access road or exit road, and 'tennel' column are expressed as values divided by 0 and 1 for whether the road is a tunnel or not. Label encoding was applied to other variables such as ‘blocked lane’, ‘day of the week’, ‘time of day’, ‘congestion level’, and ‘weather’ column because if one-hot encoding is applied excessively, it becomes sparse data and may interfere with learning.

Unlike other columns, Target Encoding was applied to incident type column. As shown in the tables in <Table 4> and <Table 5>, when it is confirmed through the distribution and correlation analysis of the incident clearance time for each variable, the incident type and the type of vehicle to be dispatched are statistically significant with the incident clearance time compared to other columns. Hence, in the column of the incident type, the Target Encoding method was applied to help calculate more significant weights. Unlike label encoding and one-hot encoding, target encoding is not a simple classification method, but a method in which the probability of each type in a column is calculated by considering the relationship with the dependent variable to be predicted.

This is the second most frequently used encoding method after the above two methods, and has the advantage of helping the neural network model to more easily learn relationships between variables without increasing the number of data dimensions.

As the final step of the preprocessing process, if there is a missing value in the data, an error occurs when training the neural network model. Therefore, Interpolation is generally applied to the value or the data is deleted. Since there were few missing values in the data of this study, the row with missing data was deleted from the dataset.

<Table 2> Columns for Predicting Incident Clearance Time

Feature	Column	Type	Unit
Clearance Time (Target Variable)	Clearance Time	Numeric	minute
Vehicles reach the site	police, patrol, ambulance, fire_truck, tow_truck, cleaning_car, etc	Numeric	Number of vehicles
	total_cars	Numeric	Number of vehicles
Distance from Approach Road	access	Numeric	Number of Section
Traffic Flow Conditions	speed	Numeric	km/h
	speed_upstream	Numeric	km/h
	volume	Numeric	vehicles per hour
	volume_upstream	Numeric	vehicles per hour

<Table 3> Encoded Columns for Predicting Incident Clearance Time

Feature	Column	Type	Unit	Encoding Method
Lane Closures	Closure_Lane	Categorical	1~4	Label Encoding
Day of Week	day_cat	Categorical	1 = weekday 2 = weekend	Label Encoding
Time of Day	hour_cat	Categorical	1 = Non-Congestion 2 = Congestion	Label Encoding
Incident Type	incident_type_target	Numeric	Average of Incident Clearance Time of Incident Type	Target Encoding
roadway type	confluence, fractionation, tennel	Categorical	0, 1	One-Hot Encoding
Traffic Flow Conditions	traffic_status	Categorical	1 = smooth 2 = Delay 3 = Congestion	Label Encoding
Environmental conditions	weather	Categorical	1 = sunny 2 = rain 3 = snow	Label Encoding

<Table 4> Summary of Incident Clearance Time by Features

Column	Features	count	sum	mean	std
Incident_Type	broken	3995	92389	23.126	18.333
	crash	3815	74214	19.453	15.387
	clash	133	4477	33.662	23.339
	falling	28	1838	65.643	55.465
	fire	16	772	48.25	35.311
	rollover	16	747	46.688	26.361
Lane Closures	Lane 3	4282	97545	22.780	18.465
	Lane 2	1315	25561	19.438	15.223
	Lane 1	2201	44131	20.050	15.950
	2 Lanes	205	7200	35.122	29.100
traffic_status	Congestion	1456	28599	19.642	15.013
	Delay	1969	40053	20.342	16.884
	smooth	4578	105785	23.107	18.949
day_cat	weekday	7141	156037	21.851	17.920
	weekend	862	18400	21.356	17.318
hour_cat	Non-Congestion	4360	99037	22.715	19.011
	Congestion	3643	75400	20.697	16.300
confluence	0	6194	130669	21.096	17.179
	1	1809	43768	24.195	19.818
fractionation	0	4113	85181	20.710	17.929
	1	3890	89256	22.945	17.708
tennel	0	4999	116515	23.308	18.318
	1	3004	57922	19.282	16.761
weather	sunny	7376	160646	21.780	17.660
	rain	335	7598	22.681	20.291
	snow	292	6193	21.209	19.737

<Table 5> Correlation of Independent Variables with Target Variable

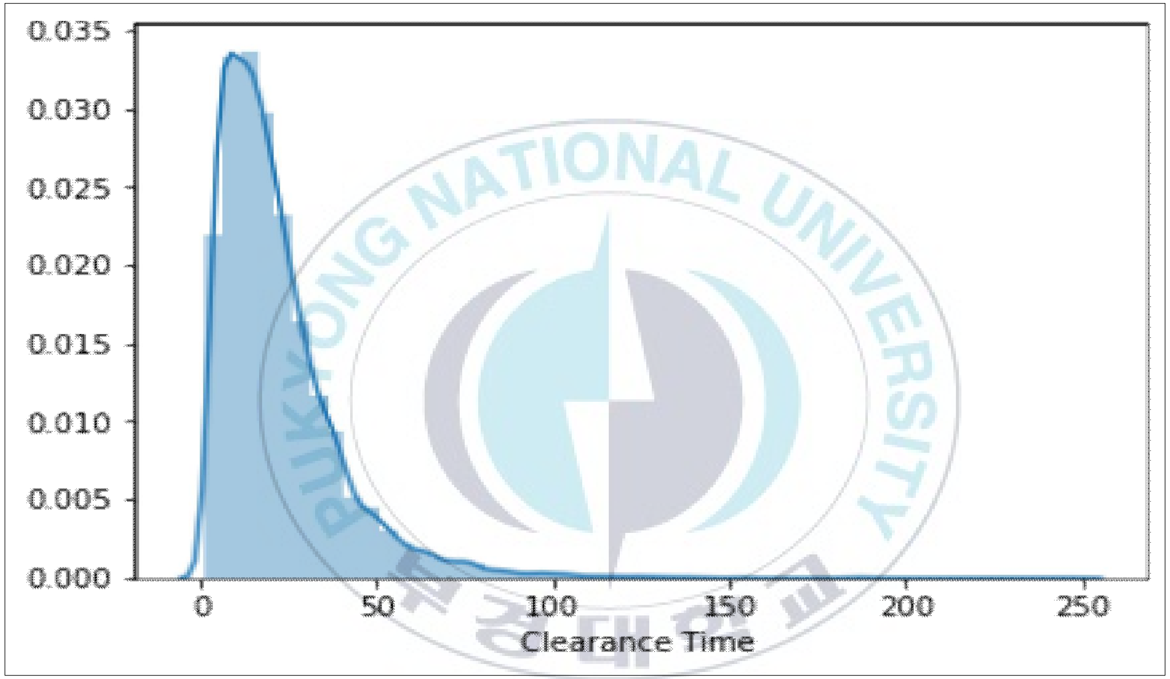
	total _cars	police	tow_truck	incident_ type_target	ambulance	cleaning_ car	fire_truck
Correlation	0.339	0.277	0.241	0.218	0.203	0.201	0.189
	falling	speed	clash	traffic_ status	speed_ upstream	broken	confluence
Correlation	0.146	0.094	0.086	0.082	0.078	0.074	0.072
	fire	fractionation	rollover	patrol	etc	incident_ type	weather
Correlation	0.066	0.063	0.062	0.055	0.048	0.018	0
	access	day_cat	day	Closure_ Lane	hour	hour_cat	volume_ upstream
Correlation	-0.005	-0.009	-0.013	-0.019	-0.044	-0.056	-0.107
	tennel	crash	volume				
	-0.109	-0.125	-0.136				

3) Data Over-Sampling

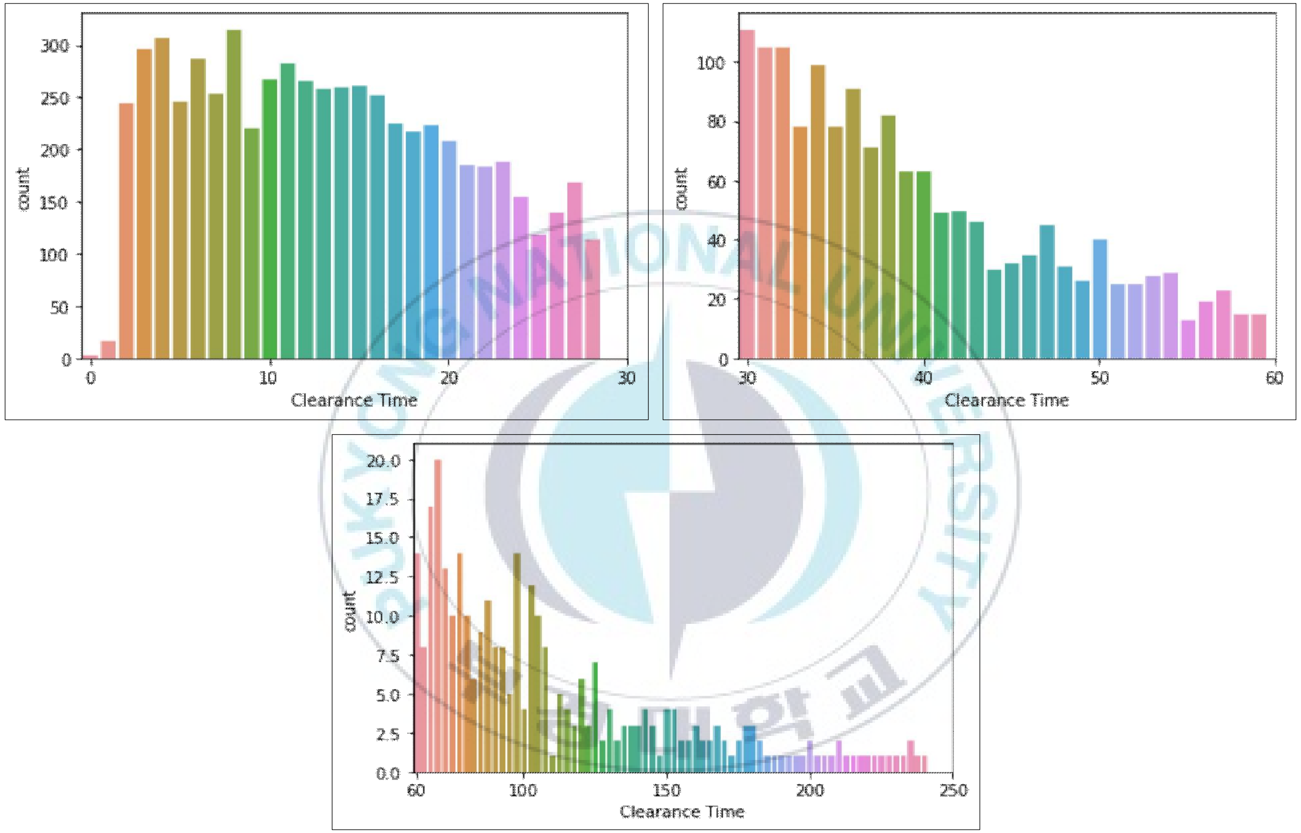
<Fig. 3> is a graph showing the distribution of the incident clearance time, which is a predictor variable. It is known that the ideal input data of a machine learning model for predicting a continuous number value should follow a normal distribution. However, It can be seen that most incidents of the data are distributed in the range of 40 minutes or less.

Skewness is an index indicating how skewed the data distribution is compared to the normal distribution. A negative value is shown in the case of a distribution skewed to the right, and a positive value in the case of a distribution skewed to the left. The skewness of the normal distribution is 0, and when the absolute value of skewness exceeds 0.5, it can be judged that the data has some degree of bias, and when it exceeds 1, the data is highly biased. Based on this, as a result of measuring the skewness of the incident clearance time distribution, a very high value was derived with a value of 2.52.

As shown in <Fig. 4>, when the incident clearance time is divided into 'low' for less than 30 minutes, 'median' for 30 minutes or more and less than 60 minutes, and 'high' for more than 60 minutes, the higher the incident clearance time, the smaller the number of data. In the case of the incident, which has a high value, there is a possibility of under-prediction due to poor learning.



<Fig. 3> Distribution of Incident Clearance Time



<Fig. 4> Distribution of Incident Clearance Time by Level

As a method applicable to data having an unbalanced data distribution for a continuous variable, it is common to take a log function to the variable. When the logarithm of the variable value is taken, the calculation is faster because it makes a large number very small, and the difference between the smallest and largest values is very small, creating a distribution close to a normal distribution and reducing skewness. However, this method has a disadvantage in that it may change the value itself, making it difficult to interpret the model. Therefore, in this study, we tried to solve the under-prediction problem by using the over-sampling method that generates new values based on the original data values rather than the method of converting the data values.

SMOGN is the most advanced over-sampling method to date for dealing with imbalanced regression problems. The method is named SMOGN and combines random under-sampling with two oversampling techniques: SmoteR and introduction of Gaussian Noise. The key idea of SMOGN algorithm is to combine both strategies for generating synthetic examples with the goal of simultaneously limiting the risks that SmoteR can incur into by using the more conservative strategy of introducing Gaussian Noise, and allow an increase of the diversity in examples generation, which is not possible to achieve using only the introduction of Gaussian Noise(Paula Branco, 2017).

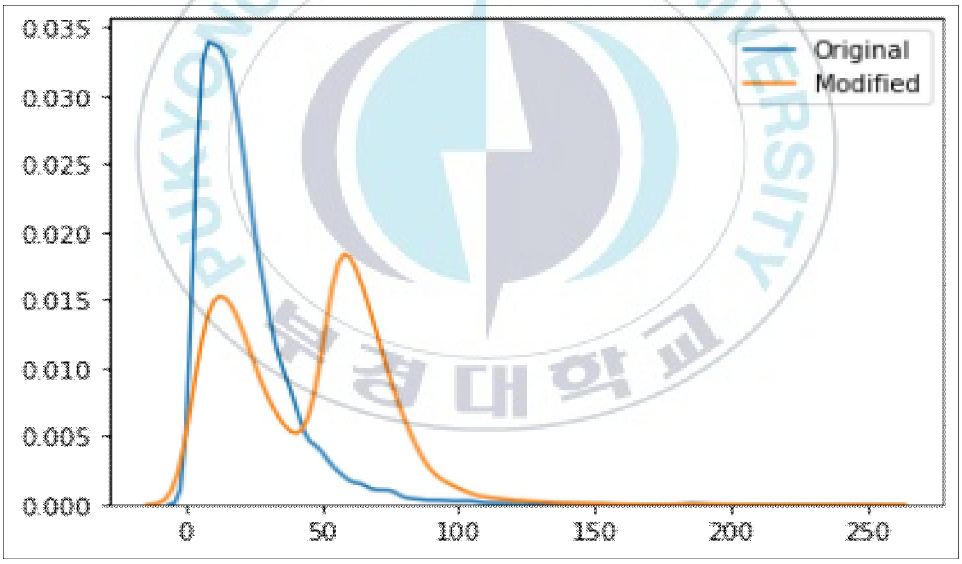
This technique is provided as a Python library, and the library provides three modes: beginner, mediate, and advanced. From beginner to advanced, detailed parameter settings are possible, so in this study, advanced mode was used for detailed parameter control.

There are a total of 9 parameters used for detailed factors including 'data', 'y', 'k', 'pert', 'samp_method', 'drop_na_row', 'rel_thres', 'rel_method', 'rel_ctrl_pts_rg'. Between them, ‘data’ and ‘y’ are parameters that receive the data to be oversampled and the dependent variable. 'k' and 'pert' are about how close to the actual value when creating a new example, and 'samp_method' and 'rel_thres' are about how large the sampling range will be. The rel_method argument takes a string, either 'auto' or 'manual'. If 'auto' is specified, "minority" values are automatically determined by box plot extremes. If 'manual' is specified, "minority" values are determined by the user. In this study, a 2d array (matrix) value was entered in the rel_ctrl_pts_rg argument to manually determine the "minority" values. Lastly, The drop_na_row argument specifies whether or not to automatically remove observations(rows) that contain missing values. The input value of each argument was selected as the value when the skewness of the over-sampling data generated through trial and error was derived the lowest as 0.346. The values entered in each argument are summarized in <Table 6>.

<Table 6> The arguments for manual operation with SMOGN

Arguments	Input Value
data	Incident Dataset after preprocessing
y	'incident clearance time'
k	5
pert	0.03
samp_method	'balance'
rel_thres	0.1
rel_method	'manual'
rel_ctrl_pts_rg	[[18, 1, 0], [60, 0, 0], [120, 0, 0], [160, 0, 0]]
drop_na_row	True

<Fig. 5> is a comparison graph of the distribution of over-sampling data to which SMOGN is applied and the original data. As a result, after applying the SMOGN method to the origin data length of 7048 cases, the oversampling data was reduced to 6263 cases. As shown in the graph, most of the data were distributed in about 40 minutes or less before applying SMOGN, however, after applying SMOGN, it was confirmed that it was distributed evenly throughout and comparing the boxplot summary numbers, it goes from [1, 10, 17, 28, 55] to [1, 18, 52, 64, 133].



<Fig. 5> Comparison of Distribution of Incident Clearance Time

2. Model for Predicting Incident Clearance Time

In this study, two dataset were created with the original data and the over-sampling data applied with the SMOGN technique, and the model built with the data without over-sampling and the model applied with the over-sampling data were compared to check the effect of improving the accuracy.

In addition, artificial neural network models such as ANN (Artificial Neural Network) and DNN (Deep Neural Network) models were built, and MAE (Mean Absolute Error), RMSE (Root Mean Square Error), and R-Squared values were used for model evaluation. An optimal model with high prediction accuracy was selected through evaluation while adjusting various hyperparameters of the model.

The programming language used for this study is Python 3.8.3 version, and the Keras API of the Tensorflow 2.4.1 version framework was used to build the model.

1) Input Dataset

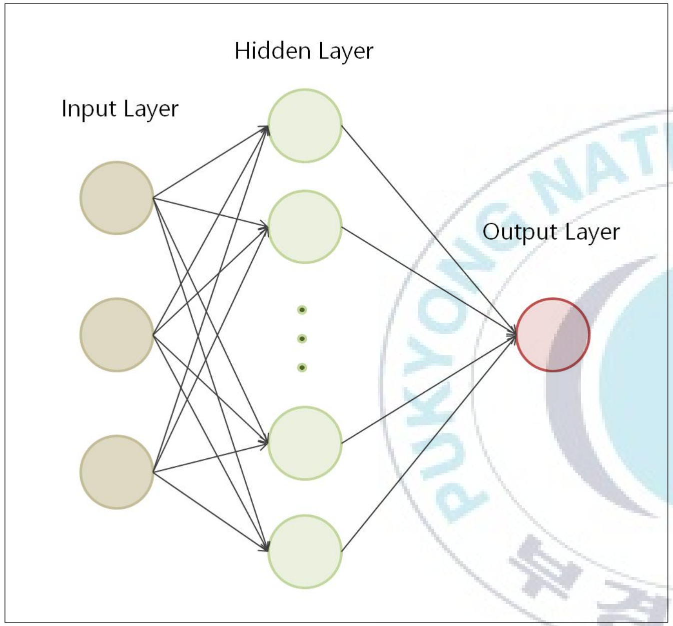
In the input data construction process of this study, in order to prevent imbalanced data from lowering the reliability of the model, as shown in <Table 7>, less than 30 minutes was classified as low, between 30 and 60 minutes as medium, and over 60 minutes as high. By combining the data sampled at a constant rate in each step, training, validation, and test data were generated.

<Table 7> Shape of Dataset

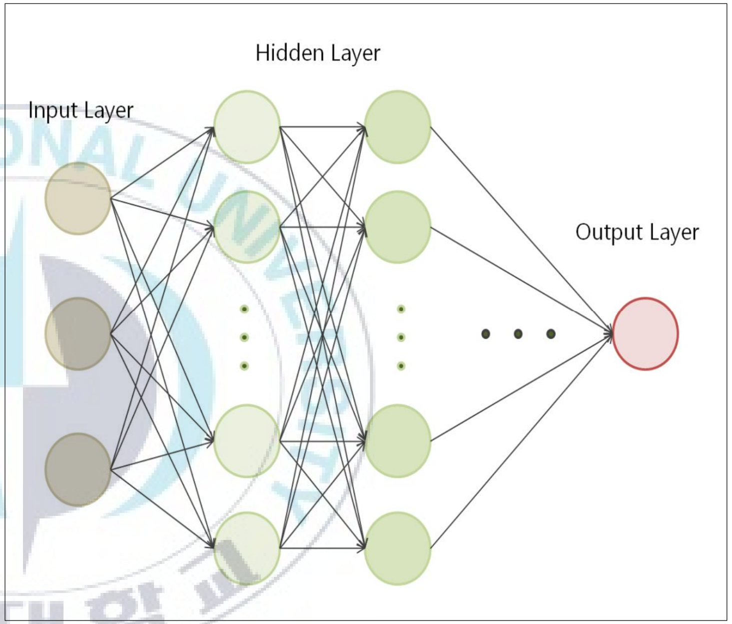
dataset	clearance time level	raw dataset	dataset applied SMOGN
train dataset	low	5493	2362
	median	1301	1874
	high	316	2027
validation dataset	low	610	267
	median	145	210
	high	28	225
test dataset	low	62	62
	median	76	76
	high	32	32

2) Neural Network Model

In this study, an artificial neural network model was used to predict incident clearance time. The artificial neural network is known as the most basic machine learning model by mimicking the network formed by synaptic bonding like the structure of a human neuron and applying it to an algorithm. The basic principle is that a certain signal or stimulus is sent to the neurons of the human brain, and the resulting signal is transmitted again according to the threshold of the signal. This process is the same as the process of sending each column value of data as a signal to the input layer of the artificial neural network, calculating each weight in the hidden layer, and delivering the result to the final output layer. As shown in <Fig. 6> and <Fig. 7>, the ANN model is a basic model with one hidden layer, and the DNN model is a model that has two or more hidden layers and enables more complex learning.



<Fig. 6> ANN Model Architecture



<Fig. 7> DNN Model Architecture

Furthermore, in order to verify the effect by applying the previously described over-sampling method and artificial neural network model, a Simple-ANN model and a Simple-DNN model applied with the original data were constructed and the SMOGN-ANN model and SMOGN-DNN model applied over-sampling data were constructed.

There are several hyperparameters that need to be manually set to calculate the weights of the artificial neural network model, including the number of neurons in the hidden layer, activation function, optimizer, learning rate, epoch, and batch size. It is important to select the hyperparameter of the model suitable for the dataset because very different results appear even if the same hyperparameter is used depending on which data is used for the model. Therefore, in this study, the model was optimized by variously adjusting the hyperparameters and checking the results.

For the ANN model, the optimal hyperparameter was selected through the number of nodes in [32, 64, 128, 256, 512, 1028, 2056], epochs in [500, 1000, 2000, 3000], and the learning rate of [0.01, 0.001, 0.0001]. Also, for the DNN model, the number of nodes in [[32,32], [64,64], [128, 128], [256, 256], [512, 512], [1028, 1028]], epochs in [500, 1000, 2000], the learning rate of [0.01, 0.001, 0.0001], and dropout of [0.01, 0.1, 0.2]. Besides, the Adam optimization function and ReLu(Rectified linear unit) activation function were used for the prediction model, and the batch size was set to 300.

In order to increase the reliability of the model, not only the Best Model but also the Worst Model were selected and the results were analyzed. The optimal input value for constructing Simple-ANN is built with Epoch 1000 in Dense [32] and learning rate of 0.0001, and in case of SMOGN-ANN, the best result is obtained with input values of

Epoch 500 in Dense [128] and learning rate 0.0001. The optimal hyperparameter of Simple-DNN model is Dense [64,64], Epoch 500, Learning Rate 0.001, Dropout 0.1, and SMOGN-DNN model gives the best result with input values of Epoch 500, Learning Rate 0.0001, Dropout 0.01 in Dense[64,64].

Among the models predicted with the worst accuracy, the worst input value for constructing Simple-ANN was built with Epoch 3000 and Learning Rate 0.01 in Dense [512]. The hyperparameter of the worst SMOGN-ANN model was predicted with the lowest accuracy with the input value of Epoch 2000 and Learning Rate 0.01 in Dense [128]. In the Simple-DNN model, the hyperparameters input as Dense[128, 128], Epoch 2000, Learning Rate 0.01, Dropout 0.01 were the worst, and SMOGN-DNN input Dense[512,512] with Epoch 500, Learning Rate 0.01, Dropout 0.01. value showed the worst results. The results of each model are summarized in <Table 8>.

<Table 8> Comparison of Worst&Best Models with Hyperparameter

Model		Dense	Epoch	Learning Rate	Dropout	MAE	RMSE	R-Squared
Worst	Simple-ANN	[512]	3000	0.01	-	36.75	46.39	-2.36
	Simple-DNN	[128, 128]	2000	0.01	0.01	26.25	35.11	-0.93
	SMOGN-ANN	[128]	2000	0.01	-	31.33	41.08	-1.64
	SMOGN-DNN	[512, 512]	500	0.01	0.01	23.86	31.83	-0.58
Best	Simple-ANN	[32]	1000	0.0001	-	18.79	25.57	-0.02
	Simple-DNN	[64, 64]	500	0.001	0.1	17.97	24.53	0.06
	SMOGN-ANN	[128]	500	0.0001	-	18.72	24.29	0.08
	SMOGN-DNN	[64, 64]	500	0.0001	0.01	18.3	**23.52**	**0.14**

Ⅲ. Results and Discussion

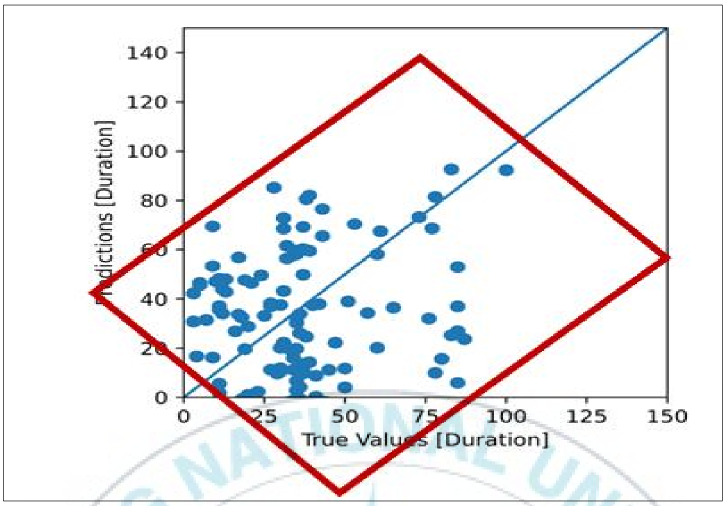
For quantitative evaluation of the predictive model, MAE(Mean Absolute Error), RMSE(Root Mean Squared Error), and R-Squared value were used. Since it was difficult to clearly verify with only one evaluation index, all three indicators were used to compare and evaluate each model.

First, among the worst models, in the case of the Simple-ANN model with one hidden layer without applying over-sampling data, the MAE was 36.75 minutes, which resulted in the highest error. On the contrary, in the case of the SMOGN-DNN model with two hidden layers with applying over-sampling data, the MAE of 23.86 min was calculated, which is much lower than Simple-ANN model. As a result of examining the RMSE and R-Squared values as well as the MAE of the worst model, the accuracy of the DNN model was significantly higher and a much lower error value was derived than that of the ANN model.

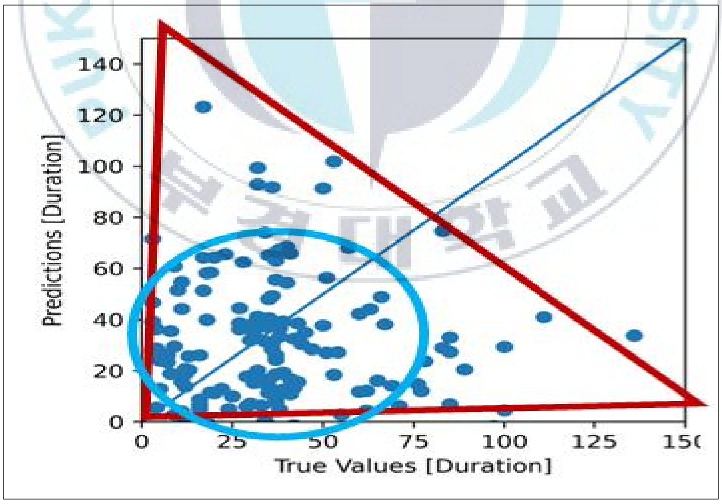
As summarized in <Table 8>, the Best Models showed significantly higher accuracy in all three indicators than the Worst Model. As a result of comparing the accuracy between the hyperparameters of the worst model and the best model, it was observed that when the input values for the number of nodes, epoch, and learning rate of the hidden layer are high, the error value is rather lower.

Also, examining the evaluation index for Best Model, it can be determined that the SMOGN-DNN Model, which is the model with the highest R-Squared value among each model, is the most accurate model. However, in the case of the MAE and RMSE values of the Best

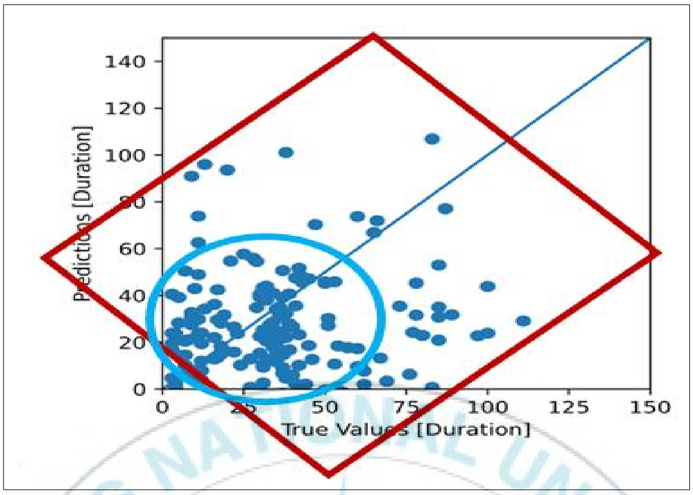
Model, it can be seen that there is little difference in the evaluation index values. For more reliable model validation, as shown in <Fig. 8>~<Fig. 15>, qualitative evaluation was attempted by analyzing the graph comparing the actual data and the predicted data for the worst model and the best model.



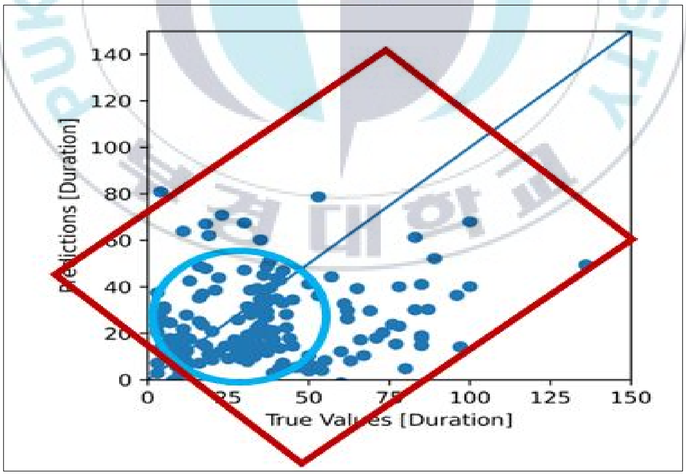
<Fig. 8> Prediction of Incident Clearance Time with Worst Simple-ANN Model



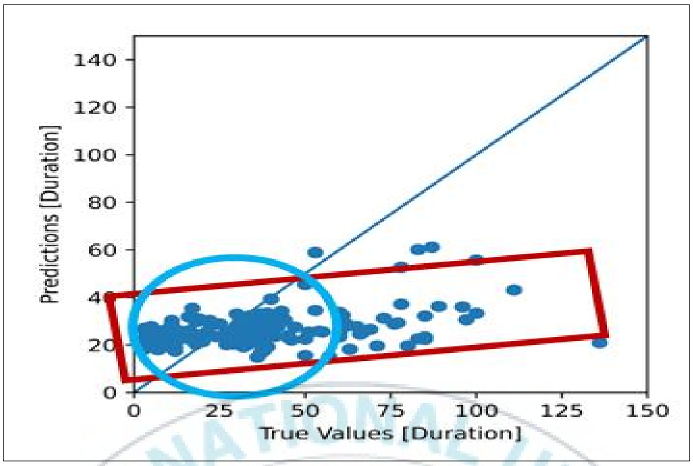
<Fig. 9> Prediction of Incident Clearance Time with Worst SMOGN-ANN Model



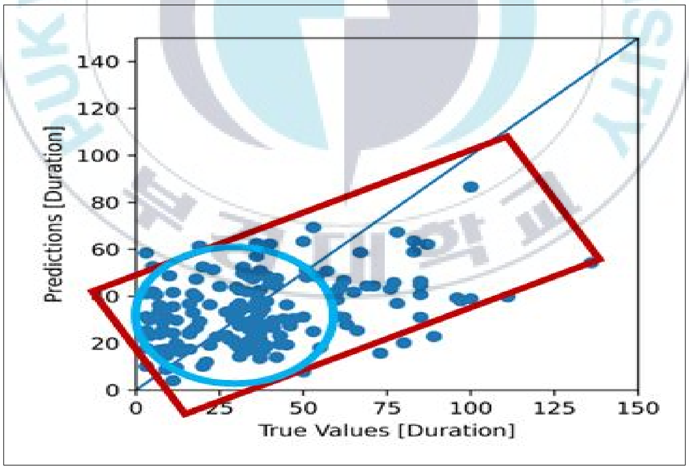
<Fig. 10> Prediction of Incident Clearance Time with Worst Simple-DNN Model



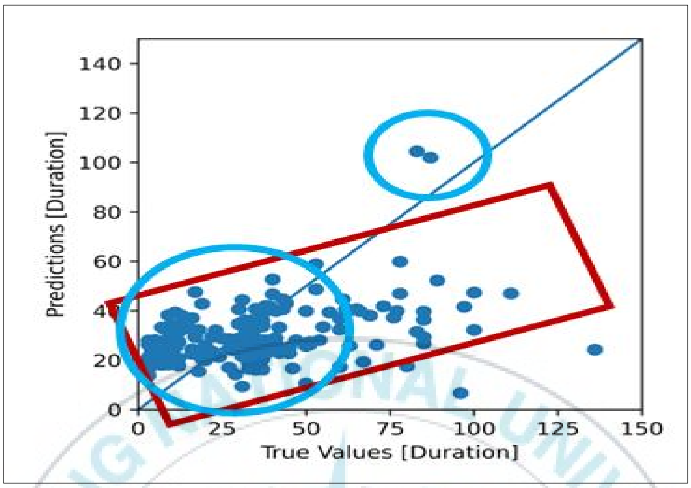
<Fig. 11> Prediction of Incident Clearance Time with Worst SMOGN-DNN Model



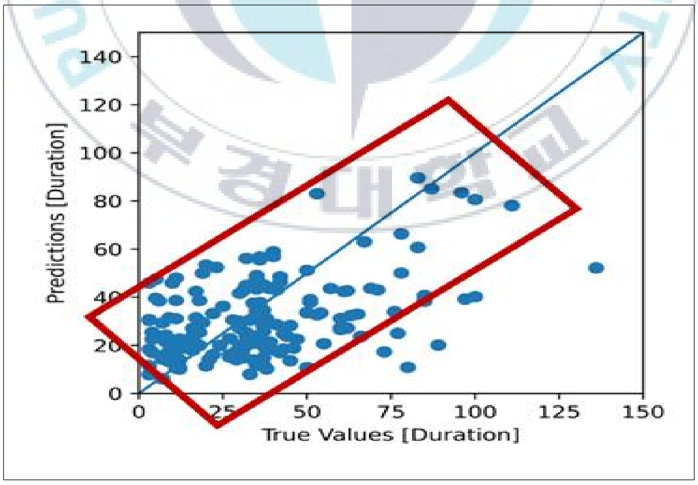
<Fig. 12> Prediction of Incident Clearance Time with Best Simple-ANN Model



<Fig. 13> Prediction of Incident Clearance Time with Best SMOGN-ANN Model



<Fig. 14> Prediction of Incident Clearance Time with Best Simple-DNN Model



<Fig. 15> Prediction of Incident Clearance Time with Best SMOGN-DNN Model

Analyzing the prediction graph of the Worst Model shown in <Fig. 8>~<Fig. 11>, it can be seen that the predicted values partially coincided with the correct answer without being biased to either side, but had a very wide error range. In the prediction of the high values of incident clearance time of the Worst Simple-ANN Model, the points of prediction was made with a somewhat low error value, however, in the prediction of the low values, there are very high error values. For these reason, in Simple-ANN model, the highest MAE value of 36.75 minutes was calculated. In the Worst SMOGN-ANN model, the prediction accuracy is higher for the low values of incident clearance time than the Worst Simple-ANN model, but it seems to have a very large error for the high values. Also, the Simple-DNN and SMOGN-DNN models showed results that were not visually different from the previous models, but the points distributed in the range of low values appeared to have higher density than other models. Therefore, it can be demonstrated that when a DNN model or SMOGN technique is applied, the accuracy of prediction for low values of incident clearance time is improved, and it is also possible to explain the increase in prediction accuracy in the quantitative evaluation of <Table 8>.

As results of analysis of the best models shown in <Fig. 12>~<Fig. 15>, it was definitely found that predicting incident clearance time with a narrower error range compared to the Worst Model. However, the prediction graph of the Simple-ANN model shows a poor prediction for a higher error than the worst model. Seeing that all predicted values of the model were predicted to be less than or equal to 60 minutes, it is inferred that the model was overfitted for the low values of incident clearance time. Accordingly, it could be judged that the overfitting was caused by applying the input dataset, in which most of clearance time data was distributed to the low side. On the other hand, when it comes to the SMOGN-ANN Model and the Simple-DNN Model, the overfitting for the low values seems to decrease as the predictions were made with a slightly higher slope than the former model and the prediction accuracy for high values is improved. Nevertheless, it still has a large error range.

Finally, in the graph of the SMOGN-DNN Model, it is confirmed that most of the points are distributed close to the regression line from low to high values while predicting values with a narrow error range. As a result, the prediction of the SMOGN-DNN model does not overfit at low values and shows high accuracy even at a high values compared to the case where over-sampling data is not applied or the ANN model is used. Even if the analyses of quantitative and qualitative evaluation were consistent, however, additional consideration was performed on the fact that there was no difference in MAE and RMSE values between Best Models in quantitative evaluation.

First, as the basic characteristics of evaluation indicators, MAE and RMSE may have very low values or very high values depending on the range of actual values. However, R-Squared can evaluate the model as a relative value to some extent because the number of samples and the average of observations affect the R-Squared value. According to this fact, it can be inferred that the Best Model predicted the low values much more accurately than the Worst Model even though the error range of the high incident clearance time was large so that the MAE and RMSE values were low. In addition, it is judged that the R-Squared value was higher in the Best Model because the overall error range was narrower than the Worst Model.

Taking a look at the graph between the Best Models, the SMOGN-DNN Model has the highest prediction accuracy among the models in this study even though the prediction accuracy is slightly lower for the low values compared to other models. As for other features, when comparing the number of data with low, medium, and high dependent variable in the test dataset, the number of low or medium values was larger than the number of high values, so it can be deduced that the SMOGN-DNN Model has disadvantage of predicting the low values compared to other models that overfit to the low values. As a result, in the evaluation of the three models except for the SMOGN-DNN Model, the prediction accuracy for the low values was supplemented by overfitting the majority of the data, although it failed to fit the small number of data with high values, and the SMOGN-DNN model predicted the values with a similar error in the overall range without significant deviation. For this reason, it is presumed that the MAE and RMSE values of the four best models came out with similar results.

Ⅳ. Conclusion

In this study, to develop a neural network model for the incident clearance time prediction, the input dataset of the models were constructed through a total of 22 factors by analyzing the factors affecting the incident duration. In addition, as a solution to the problem of under-prediction for high incident clearance time, the over-sampling data applied with the SMOGN method was applied to the existing model, and comparison and evaluation between models were performed, and the verification process was performed. Accordingly, a total of four models were constructed: the Simple-ANN model and the Simple-DNN model to which the basic data was applied, and the SMOGN-ANN model and the SMOGN-DNN model to which the over-sampling data was applied. The best and worst models were selected to analyze the prediction accuracy of each model while adjusting the hyperparameters for each model. the three evaluation indicators of MAE, RMSE, and R-Squared were used for quantitative verification of the prediction model, and the prediction accuracy was qualitatively analyzed and verified through a graph that can compare the actual value with the predicted value.

As a result, the DNN model predicted the incident clearance time with the lower error on average than the ANN model. and it was confirmed that the lower the number of nodes, the number of epochs, and the learning rate, the higher prediction accuracy. In addition, it was found that the model trained with the over-sampling data using the SMOGN method was predicted with much higher accuracy than the general model. Therefore, according to the results of the study, it is expected that the accuracy of incident clearance time prediction will be higher and the problem of under-prediction for the minority data will be solved.

However, as a limitation of the study, this model is a model built using only the data of Korean road among urban highways, and additional verification is needed by applying other domestic urban highway data for the reliability of the prediction model. In addition, it has an average error of about 18 minutes based on the MAE standard, and it is still considered a large error range for practical application, so additional research is needed to further reduce the error range and develop a model that can make more high-accuracy predictions.

REFERENCES

MOLIT STATISTICS SYSTEM, https://stat.molit.go.kr/portal/main/portalMain.do, 2021.03.30.

The Korea Transport Institute (2019), Transportation cost calculation status and improvement plan

Li, Ruimin et al.(May 2018), "Overview of traffic incident duration analysis and prediction." European Transport Research Review 10 : 22

Han W. G. (May 2001) "A Model For Estimating Incident Duration" Korean Society of Civil Engineers, v.21 n.3-D

Shin C. H. and Kim J. H. (2002) "Development of Freeway Incident Duration Prediction Models" Journal of Korean Society of Transportation v.20 no.3, pp.17 - 30

Kim J. W. (2005) "Development of a model for estimating incident duration : Focusing on seoul urban expressway traffic management systems" University of Seoul

Lee K. Y., Seo I., K. et al.(2012) "A Study on the Influencing Factors for Incident Duration Time by Expressway Accident" International journal of highway engineering v.14 no.1 = no.51, pp.85 - 94

Lee S. B., Han D. H., Lee Y. I. (2015) "Development of Freeway Traffic Incident Cearance Time Prediction Model by Accident Level" J. Korean Soc. Transp. Vol.33, No.5, pp.497-507

Paula Branco, Luís Torgo, Rita P. Ribeiro (2017) "SMOGN: a Pre-processing Approach for Imbalanced Regression" Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR 74:36-50