2024 National Freeway Intelligent Transportation Management Innovation Competition

Prediction of Accident-Induced Congestion Length on National Freeway No.1 (Mileposts 10-99 km)

Department: Department of Statistics

Student Name:

郭依璇 (Yi-Hsuan Kuo)

李博業 (Bo-Ye, Li)

林瑋珈

陳皓鈞

Advisor: Ping-Yang Chen, Ph.D.

October, 2024

Table of Contents

ABSTRACT

This study applies machine learning techniques to predict post-accident traffic queue length, with the aim of reducing prediction errors and the underestimation rate. In addition to the data provided by the competition organizer, this research integrates real-time traffic information from the Freeway Bureau database and external weather databases, incorporating weather conditions to increase the complexity of explanatory variables. These additional real-time factors enable the models to better reflect actual traffic conditions and improve prediction performance.

Based on model architectures such as CatBoost and multilayer perceptrons (MLP), a two-stage traffic queue length prediction framework is proposed. The overall research process is illustrated in Figure 1. Using information related to weather conditions, real-time traffic status, accidents, and construction activities, the first stage determines whether an accident is likely to cause traffic congestion. For accidents predicted to result in congestion, the second stage further estimates the length of the traffic queue along the affected road segment. Experimental results on the test set show a root mean square error (RMSE) of 1.45 kilometers, with an underestimation rate of approximately 24.94%.

Based on the findings of this study, it is recommended to integrate the proposed model with the “1968” app developed by the Freeway Bureau. By incorporating real-time accident-induced congestion impact information into various app functions, the system can leverage its high update frequency to provide early warnings for potential traffic congestion. Furthermore, by utilizing location-based push notifications, voice

alerts can be sent to drivers approaching congested areas, along with alternative route suggestions, thereby helping to alleviate traffic congestion in the Greater Taipei Area and surrounding regions.

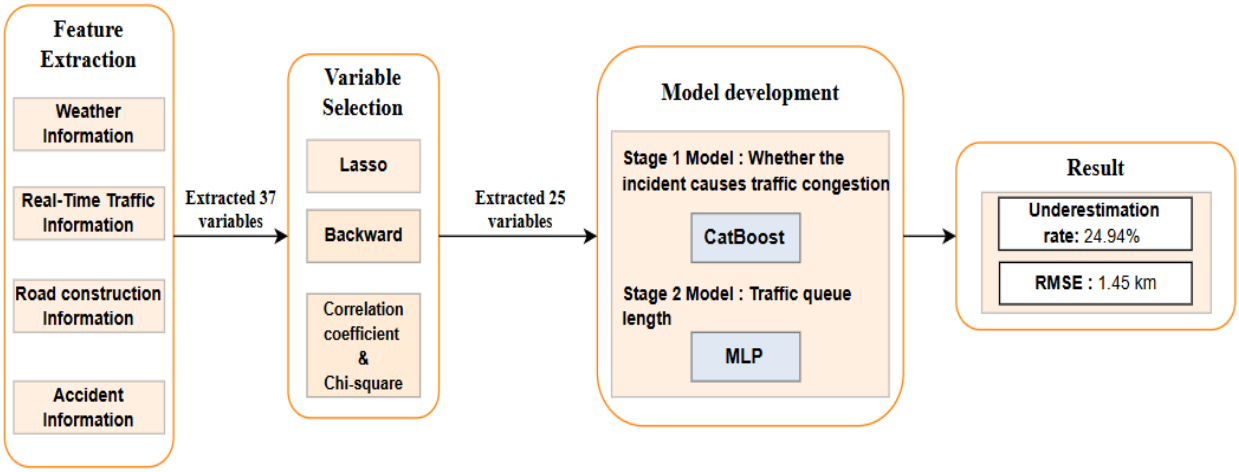


Fig 1: Research Framework Diagram

1 Introduction

Taiwan's national highway system covers all major cities and key economic regions, providing fast and convenient transportation services. However, with the increasing number of vehicles and the intensification of economic activities, the highway system is facing increasingly severe congestion problems.

The total length of Taiwan’s national highways is approximately 1,000 kilometers, encompassing several major expressways, including National Highway No. 1, No. 3, and No. 5. These highways connect the northern, central, and southern economic zones, carrying millions of vehicles daily, with National Highway No. 1 serving as the most important north–south transportation corridor. Traffic congestion on the highways mainly occurs during peak commuting hours, weekends, and holidays,

particularly near major cities and industrial areas. According to data from the Ministry of Transportation and Communications, National Highway No. 1 passes through many densely populated areas, such as Neihu in Taipei, Linkou in New Taipei, Nankan in Taoyuan, and Hsinchu Science Park, where severe congestion frequently occurs.

Based on the congestion-prone sections compiled by the Ministry of Transportation and Communications [1], this study focuses primarily on the 10–99 km segment of National Highway No. 1, covering the above-mentioned densely populated and congestion-prone areas. Traffic congestion along this segment is mainly caused by the following factors:

1. **High traffic volume:** During peak commuting hours, traffic volume rises sharply, far exceeding the road segment's designed capacity.

2. **Frequent entrances and exits:** This segment has multiple on- and off-ramps, causing vehicles to frequently change lanes, which often leads to traffic bottlenecks.

3. **Frequent accidents:** The combination of high traffic volume and complex traffic conditions results in frequent accidents, each of which further exacerbates congestion.

This study aims to predict the congestion caused by traffic accidents along the 10–99 km segment of National Highway No. 1. Through precise data analysis and model construction, it seeks to provide effective measures and recommendations to help alleviate traffic congestion.

2 Literature Review

1. **A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data [2]:** This study classifies traffic congestion levels into five grades and models the process in two stages based on the time when the police arrive at the accident scene. The aim is to predict how long it takes for congestion to return to its pre-accident level.

   The features primarily include real-time traffic factors, event-related factors, and weather factors. In the first-stage modeling, since the police have not yet arrived, only real-time traffic and weather factors are available. Binary classification and multiclass classification are used to predict whether congestion will worsen. The accuracy of the test set ranges from 0.73 to 0.83, depending on the original congestion level of the road. The most important factors in the first stage are real-time traffic variables, including the speed difference one minute before and after the accident, traffic volume from five minutes before the event to the time of the event, rainfall at the time of the event, and the proportion of heavy vehicles from five minutes before the event to the time of the accident.

   If the first-stage prediction indicates worsening congestion, the model proceeds to the second stage after the police finish their investigation. The second-stage model primarily predicts the time needed to return to the original congestion level. In the test set, the absolute difference (AD) is mostly below

10 minutes. Important factors in the second stage focus on accident-related variables, such as whether there were injuries, whether the cause was a rear-end collision, the average speed one minute after police arrival, the total number of vehicles involved, and general weather conditions.

This study's approach of first predicting congestion categories and then modeling continuous recovery time is a solid idea. However, the study does not discuss whether the second-stage time predictions tend to overestimate or underestimate, and there is still room for improvement in the accuracy of the first stage.

2. **Congestion Prediction With Big Data for Real-Time Highway Traffic [3]:** This study conducts an in-depth investigation of real-time highway congestion prediction. In selecting feature variables, multiple key factors are considered, including road speed, segment density, traffic volume, rainfall, and real-time traffic incident reports. Our study also references this work to incorporate these features, which not only reflect current traffic conditions but also help the model more accurately predict future traffic situations.

   The study uses a Support Vector Machine (SVM) to build a real-time highway traffic congestion prediction model (SRHTCP). Results show that the SRHTCP model significantly outperforms prediction methods based on weighted exponential moving averages, with a 25.6% improvement in prediction accuracy. The model also performs well in terms of Mean Absolute Relative Error (MARE). SRHTCP can not only predict vehicle speed in the next time interval in real time but also effectively analyze highway congestion.

By integrating traffic, weather, and social media data, the model provides a comprehensive reflection of traffic conditions and enhances prediction reliability.

3 Research Methods and Steps

3.1 Data Source

This study utilized weather factors from the Timeanddate website [4], including temperature (°C), wind speed (km/h), humidity (%), and atmospheric pressure (mbar); static Etag information (v2.0), M03A (traffic volume statistics by vehicle type), and M06A (raw travel path data) from the Traffic Data Bank of the Taiwan Ministry of Transportation and Communications [5]; road construction data from January–October 2023 and January–February 2024 [6]; traffic accident data on National Highways A1, A2, and A3 from January–October 2023 and January–February 2024; and traffic accident notification reports from the same periods [6].

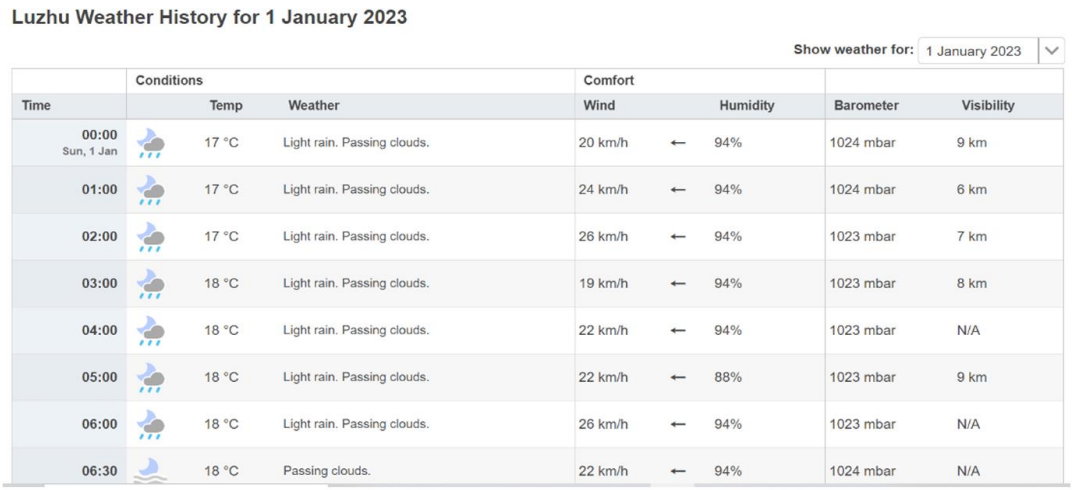
After filtering for National Highway No. 1 and the 10–99 km segment, a total of 5,702 accident events were extracted as the training set, and 962 accident events were used as the test set.

3.2 Feature Extraction

3.2.1 Weather Information

The Timeanddate website provides detailed hourly weather information for various

regions in Taiwan (Figure 2)



Luzhu Weather History for 1 January 2023

Show weather for: 1 January 2023

	Conditions		Comfort			
Time	Temp	Weather	Wind	Humidity	Barometer	Visibility
00:00 Sun, 1 Jan	17 °C	Light rain. Passing clouds.	20 km/h ←	94%	1024 mbar	9 km
01:00	17 °C	Light rain. Passing clouds.	24 km/h ←	94%	1024 mbar	6 km
02:00	17 °C	Light rain. Passing clouds.	26 km/h ←	94%	1023 mbar	7 km
03:00	18 °C	Light rain. Passing clouds.	19 km/h ←	94%	1023 mbar	8 km
04:00	18 °C	Light rain. Passing clouds.	22 km/h ←	94%	1023 mbar	N/A
05:00	18 °C	Light rain. Passing clouds.	22 km/h ←	88%	1023 mbar	9 km
06:00	18 °C	Light rain. Passing clouds.	26 km/h ←	94%	1023 mbar	N/A
06:30	18 °C	Passing clouds.	22 km/h ←	94%	1024 mbar	N/A

Fig 2. Data presentation from the Timeanddate website

Web scraping was used to extract key weather factors for each accident, including temperature (°C), wind speed (km/h), humidity (%), and atmospheric pressure (mbar).Additionally, since the website's meteorological data is organized by region and does not provide detailed information about weather stations (e.g., latitude and longitude), this study can only classify each road segment according to the corresponding queryable region, as shown in the table below.

Mile Markers of National Highway No. 1	Regions
10-15	Xizhi District
15-17	Neihu District
17-25	Zhongshan District
25-27	Sanchong District
27-35	Wugu District
35-41	Linkou District
41-49	Luzhu District

Mile Markers of National Highway No. 1	Regions
49-52	Dayuan District
52-62	Zhongli District
62-71	Pingzhen District
71-86	Hukou District
86-91	Zhubei District
91-95	East District, Hsinchu
95-99	Baoshan, Hsinchu County

Table 1: Kilometer Sections and Corresponding Regions of National Highway No. 1

3.2.2 Extraction of Traffic Volume and Average Speed Features on National Highways

Static Etag information (v2.0) [4] from the Freeway Bureau database was used, focusing on stations along National Highway No. 1. In addition, to account for potential station malfunctions, stations outside the 10–99 km segment were also included. The data format is shown in Figure 3.

	ETagGantryID	RoadName	RoadDirection	RoadSection/Start	RoadSection/End	LocationMile
0	01F0061S	國道1號	S	大華系統	五堵	6.1
1	01F0061N	國道1號	N	五堵	大華系統	6.1
2	01F0099S	國道1號	S	五堵	汐止&汐止系統	9.9
3	01F0099N	國道1號	N	汐止&汐止系統	五堵	9.9
4	01F0147N	國道1號	N	東湖	汐止&汐止系統	14.7
...	...	...	...	...	...	...
66	01F0980S	國道1號	S	新竹(科學園區)	新竹系統	98.0
67	01F1045N	國道1號	N	頭份	新竹系統	104.5
68	01F1045S	國道1號	S	新竹系統	頭份	104.5
69	01F1123N	國道1號	N	頭屋	頭份	112.3
70	01F1123S	國道1號	S	頭份	頭屋	112.3

Fig 3: Example of the Data Format

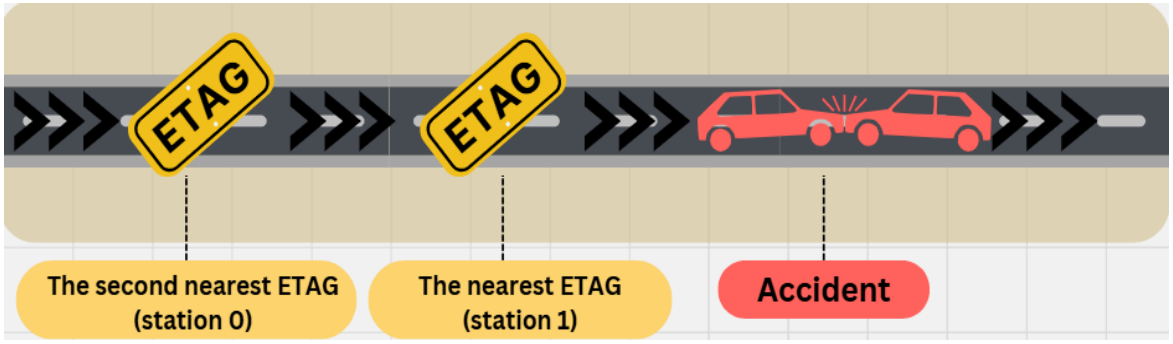


Fig 4: Definition of ETAG location

Based on Figure 4, two types of stations are defined as follows:

- **Station 0:** the second nearest station to the incident that the vehicle has already

passed.

• **Station 1:** the nearest station to the incident that the vehicle has already passed.

These definitions facilitate the explanation of subsequent feature extraction.

(1) Traffic volume and large-vehicle ratio ten minutes before the incident (Pre_TrafficVolume and Pre_LargeVehicleRatio):

This feature is designed to evaluate traffic conditions prior to the occurrence of an incident. The data are primarily obtained from the Freeway Bureau's M03A dataset. As shown in Figure 5, the data are aggregated in five-minute intervals; for example, the time label *2023-08-17 06:00* represents the total traffic volume during the period from *06:00 to 06:05* on *2023-08-17*.

	Time Interval	ETAG ID	Direction	Car Types	Vehicle counts
0	2023-08-17 06:00	01F0005N	N	小客車	19
1	2023-08-17 06:00	01F0005N	N	小貨車	7
2	2023-08-17 06:00	01F0005N	N	大客車	3
3	2023-08-17 06:00	01F0005N	N	大貨車	0
4	2023-08-17 06:00	01F0005N	N	聯結車	0

Fig 5: Total traffic volume per five-minute interval

After identifying Station 1 for a given incident, this study extracts the traffic volume data from the gantry associated with that station during the ten minutes preceding the incident time (as defined in Equations (1) and (2)), excluding the time interval in which the incident occurred.

Specifically:

• If the incident occurred at **03:04**, data from **02:50** and **02:55** are extracted.

• If the incident occurred at **03:05**, data from **02:55** and **03:00** are extracted.

• If the incident occurred at **03:00**, data from **02:50** and **02:55** are extracted.

$$\textbf{Total Traffic Volume} = V_{pc} + V_{lt} + V_{bus} + V_{ht} + V_{av} \quad (1)$$

$$\textbf{Large-Vehicle Ratio} = \frac{V_{bus} + V_{ht} + V_{av}}{\text{Total Traffic Volume}} \quad (2)$$

- V_{pc}：traffic volume of passenger cars
- V_{lt}：traffic volume of light trucks
- V_{bus}：traffic volume of buses
- V_{ht}：traffic volume of heavy trucks
- V_{av}：traffic volume of articulated vehicles

(2) Post-incident traffic volume within a two-minute interval:

This feature aims to quantify the traffic inflow into the incident zone immediately after an incident occurs, which is expected to be informative for estimating traffic congestion length. Given the limited temporal resolution of available traffic volume data, traffic counts are obtained from the Freeway Bureau's M03A dataset. Since the dataset is aggregated at five-minute intervals, a weighted average method is employed to approximate the traffic volume during the two-minute period following the incident time.

For example:

• An incident at **03:32** uses **2/5** of the traffic volume from **03:30–03:35**.

• An incident at **03:29** uses **1/5** of the traffic volumes from **03:25–03:30** and **03:30–03:35**, respectively.

• An incident at **03:35** uses **2/5** of the traffic volume from **03:35–03:40**.

(3) Pre-incident average vehicle speed (Pre_AverageCarSpeed):

This feature aims to assess roadway conditions prior to an incident. Vehicle trajectory data are obtained from the Freeway Bureau's M06A dataset, which logs the times at which vehicles pass through gantries. Data are aggregated in hourly intervals; for example, the timestamp *2023-12-23 04:00* corresponds to all vehicles passing the initial gantry between *04:00:00* and *04:59:59* on that date.

Each record contains sequential gantry passage times, for instance:

'2023-12-23 04:34:39+01F0928N; 2023-12-23 04:37:38+01F0880N; 2023-12-23 04:45:52+01F0750N; 2023-12-23 04:54:54+01H0608N; 2023-12-23 04:56:25+01F0584N',

indicating that the vehicle passed gantry **01F0928N** (km 92.8) at *04:34:39*, **01F0880N** (km 88.0) at *04:37:38*, etc.

For a given incident, the complete set of vehicle trajectories passing **Station 1** during the ten minutes preceding the event is extracted, resulting in **n** records. Individual vehicle speeds prior to the incident are computed using the **distance between stations** and the corresponding time differences, as specified in Equation (3).

$$\text{speed}_j = \frac{\text{Distance between ETAG (km)}}{\text{Time difference (s)}},\ j = 1 \ldots . . n \qquad (3)$$

Based on the above example, for **Station 1** identified as **01F0928N**, if the timestamp **04:34:39** falls within the ten-minute interval preceding the incident, the vehicle's

speed is computed as (92.8 - 88)/(179) = 0.0268(km/s).

Subsequently, as specified in Equation (4), the average of the speeds from the **n** vehicles is calculated to represent the pre-incident average speed for the ten minutes prior to the event.

$$\text{Pre\_AverageCarSpeed}_i = \frac{\sum_{j=0}^{n} \text{speed}_j}{n},\ i = 1 \ldots .5702 \quad (4)$$

Although the M06A dataset allows for more precise extraction of vehicle speed, several challenges remain. The issues encountered and the strategies employed in this study are as follows:

1. **Selecting passenger cars:**

   Each M06A file contains a large volume of data, with at least 100,000 records per hour. Since the goal is to compute average vehicle speed and the majority of vehicles on the freeway are passenger cars, data for this vehicle type are less likely to be missing and can be quickly filtered. Therefore, the speed calculation for this feature is performed exclusively for passenger cars.

2. **Gantry malfunctions:**

   ETAG gantries occasionally fail due to technical issues, leading to missing data for **Station 1** during certain periods. In such cases, the feature is calculated using the preceding gantry, i.e., **Station 0**.

3. **Large data volume for individual events:**

   M06A records are organized based on the timestamp at which a vehicle passes

the initial gantry. Consequently, files corresponding only to the hour of the incident may be insufficient and sometimes do not contain the relevant data. For example, a vehicle may pass the initial gantry at **12:00**, but reach the target gantry only at **15:00**. To account for this, the study extracts data from the four hours preceding the incident. For instance, if an incident occurs at **04:16**, vehicle records from **12:00:00, 01:00:00, 02:00:00, 03:00:00,** and **04:00:00** are collected for speed calculation.

3.2.3 Road construction Information

The data used in this study were obtained from the National Freeway Intelligent Transportation Management Innovation Competition website [6], specifically the files *Traffic Accident SMS Notification Data.xlsx* for January–October 2023 and January–February 2024.

Duplicate records were first removed. Since some fields in the dataset differed from the content in the **"SMS Content"** column, the records in this column were adopted as the authoritative source. Each SMS was then split by keywords and processed according to the following categories:

(1) **Time-related features (TimeCategory, DateCategory, ProcessingMinutes):**

The year, month, day, hour, and minute of each incident, as well as the incident occurrence and clearance times, were extracted from the SMS. The hour of the incident was stored as the **TimeCategory**, while the **DateCategory** was assigned based on whether the date was a weekday, regular weekend, or public holiday. Finally, the time required to clear the

incident was calculated as **ProcessingMinutes**.

(2) **Direction (Direction):**

The SMS was used to extract the direction of traffic, classified as northbound, southbound, eastbound, westbound, or bidirectional. As this study focuses only on a segment of National Freeway No. 1, only **northbound** and **southbound** directions are included.

(3) **Mileage-related features (Mileage, CongestionMileage):**

The incident location was extracted from the SMS as **Mileage**. Depending on whether congestion occurred, **CongestionMileage** was recorded: 0 km if no congestion, or the reported congestion distance (km) if congestion occurred.

(4) **Casualties (Deaths, Injuries):**

The number of fatalities (**Deaths**) and injuries (**Injuries**) for each accident was extracted from the SMS.

(5) **Lane involvement (InnerShoulder, InnerLane, InnerMiddleLane, MiddleLane, OuterMiddleLane, OuterLane, OuterShoulder, Ramp):**

As the SMS does not provide detailed lane-level information, the original eight lane columns in the dataset were used to indicate which lanes were affected by the incident.

(6) **Vehicle-related features (AccidentVehicle, ConstructionVehicle, Car, MediumLargeBus, MediumSmallTruck, LargeTruck_Trailer, OtherVehicles):**

Vehicle types involved in or affected by each incident were extracted from the

SMS. Since different SMS records use varying terms for the same vehicle type (e.g., “小客”, “小客車”, “小車”, “小自客” for passenger cars), terms were standardized and classified into six categories: engineering vehicles, passenger cars, medium-to-large buses, medium-to-small trucks, large trucks & trailers, and other vehicles. The number of vehicles in each category was then counted per incident, and the total number of involved vehicles was calculated.

(7) **Accident cause (RearEndCollision, SelfCollision, Fire, Overturn, OtherCause):**

The cause of each accident was extracted from the SMS and initially categorized into eight types: rear-end collision, self-collision, side-swipe, slipping, overturn, debris, fire, and others. Some incidents involved multiple causes simultaneously (e.g., debris and self-collision); these were consolidated into five categories for analysis: **rear-end collision, self-collision, overturn, fire, and other causes**.

3.2.4 Accident Information

The construction data used in this study were obtained from the National Freeway Intelligent Transportation Management Innovation Competition website [6], specifically the files *Roadwork Segment Data.xlsx* for January–October 2023 and January–February 2024.

Construction records corresponding to the selected freeway segment were first filtered based on **freeway number** and **mile markers**. The records were then separated by **southbound** and **northbound lanes**, and the number of lanes occupied by each construction project (**LaneOccupancy**) was recorded, including lanes, shoulders, and

median barriers. The start and completion times of each construction project were also recorded.

To combine construction data with accident records, an accident was considered potentially influenced by construction if:

1. The accident location fell within ±300 meters of the construction site, **and**
2. The construction was still ongoing at least 20 minutes before the time of the latest accident in that segment.

In such cases, the total number of lanes occupied by construction was recorded as the **LaneOccupancy due to construction**.

Following the above data preprocessing, the study's dataset features are summarized in Table 2, along with the results of feature selection using **Lasso**, **Backward Selection**, **Chi-square test**, and **correlation analysis**.

Column Name	**Column Definition**	**Variable Selection Results**		
		Lasso	Backward	Chi-square & correlation analysis
Accident Description				
DateCategory	Date Category of the Accident	V	V	V
TimeCategory	Time Category of the Accident	V	V	V
Direction	Road Direction of the Accident	V	V	V
Mileage	Mileage of the Accident	V	V	V
ProcessingMinutes	Incident clearance time in minutes	V	V	V
Deaths	Number of deaths			V
Injuries	Number of injuries	V	V	V
CongestionMileage	Resulting congestion mileage due to the accident	Target variable is no need to be selected		
Traffic real-time information				

Pre_LargeVehicleRatio	Ratio of buses, heavy trucks and articulated vehicles ten minutes prior to the accident	V	V	V
Pre_AverageCarSpeed	Average vehicle speed ten minutes prior to the accident	V	V	V
Pre_TrafficVolume	Total traffic volume ten minutes prior to the accident	V	V	V
Post_TrafficVolume	Total traffic volume two minutes after the accident		V	V
Road construction information				
LaneOccupancy	Number of lanes occupied by construction			V
Road situation				
InnerShoulder	Whether the accident influenced the inner shoulder (Yes/No)			V
InnerLane	Whether the accident influenced the inner lane (Yes/No)	V	V	V
InnerMiddleLane	Whether the accident influenced the inner middle lane (Yes/No)	V	V	V
MiddleLane	Whether the accident influenced the middle lane (Yes/No)	V	V	V
OuterMiddleLane	Whether the accident influenced the outer middle lane (Yes/No)	V	V	V
OuterLane	Whether the accident influenced the outer lane (Yes/No)		V	
OuterShoulder	Whether the accident influenced the outer shoulder (Yes/No)			V
Ramp	Whether the accident influenced the ramp (Yes/No)	V	V	V
Vehicle Types				
AccidentVehicle	Total number of vehicles involved in the accident	V		V
ConstructionVehicle	Number of construction vehicle involved in the accident			V
Car	Number of cars involved in the accident		V	V
MediumLargeBus	Number of medium-large			V

	buses involved in the accident			
MediumSmallTruck	Number of medium-small buses involved in the accident	V	V	V
LargeTruck_Trailer	Number of large trucks and articulated vehicles involved in the accident	V	V	V
OtherVehicles	Number of other vehicle types involved in the accident			
Accident Causes				
RearEndCollision	Whether the accident cause included a rear-end collision. (Yes/No)			V
SelfCollision	Whether the accident cause included a self-collision. (Yes/No)			
Fire	Whether the accident cause included a fire. (Yes/No)			V
Overturn	Whether the accident cause included an overturn. (Yes/No)	V	V	V
OtherCause	Whether other causes were involved in the accident. (Yes/No)			
Weather information				
Temp	Temperature at the time of the accident			V
WindSpeed	Wind speed at the time of the accident		V	V
Humidity	Humidity at the time of the accident			V
atm	Atmospheric pressure at the time of the accident			V

Table 2: Dataset variables table

3.3 Variable Selection

Based on the results of the variable significance tests, the following adjustments were made:

(1) **Casualties**

The test results for injuries were acceptable, whereas those for fatalities were

considerably poor. Given that the number of accidents involving fatalities was extremely small, which may have affected the statistical test results, fatalities and injuries were therefore merged into a single variable representing the total number of casualties.

(2) **Accident Vehicles and Large Vehicles**

Since the majority of accident-involved vehicles were passenger cars, the number of passenger cars showed little distinction from the total number of accident vehicles. As a result, the passenger car variable was removed. Moreover, most of the remaining vehicle types were classified as large vehicles; thus, these categories were combined into a single variable representing the number of large vehicles.

(3) **Rear-End Collision**

Among the accident cause variables, the rollover variable demonstrated better test performance compared to the others and was therefore retained. Considering that most traffic accidents were rear-end collisions, the variables representing fire, self-collision, and other causes were merged into the rear-end collision variable.

Finally, temperature, humidity, and atmospheric pressure were removed. Consequently, the original set of 37 variables was reduced to 25 variables.

3.4 Model Development

3.4.1 Evaluation Index

The evaluation indicators in this study are divided into two aspects: statistical

indicators and error-based metrics.

(1) Statistical indicators

Based on the comparison between the actual conditions and the predicted results, four states can be identified, as shown in Table 3, which are then used to construct Table 4.

Actual situation Predicted result	+ (Has traffic congestion)	− (Has no traffic congestion)
+ (Has traffic congestion)	TP (True positive)	FP (False positive)
− (Has no traffic congestion)	FN (False negative)	TN (True negative)

Table 3: Model prediction outcomes

Evaluation Indicators	Definition	Formula	Meaning
Recall	Ratio of predicted positives among the true positives	$\frac{TP}{TP + FN}$	The ability of correctly identify cases that are actually congested.
Precision	Ratio of true positive among the predicted positives	$\frac{TP}{TP + FP}$	The ability to correctly identify cases that will result in congestion.
Accuracy	Ratio of correctly predicted cases among all cases	$\frac{TP + TN}{TP + FP + TN + FN}$	The ability to correctly identify all situations.

Table 4: Statistical evaluation indicators

(2) Error-based metrics

Evaluation Indicators	Definition	Formula	Meaning

RMSE	The square root of the mean of the squared differences between predicted and actual values	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	The difference between the predicted congestion mileage and the actual congestion mileage
Ratio of underrated cases	The proportion of test cases in which the predicted congestion mileage is less than the actual congestion mileage	$\frac{n_{(\hat{y}<y)}}{n} * 100\%$	The proportion of cases in which the model underestimates accident severity

Table 5: Error-based metrics

3.4.2 Model Training

Based on the boxplot (Figure 6), it was observed that the variability of congestion length is high, making it difficult to directly predict the congestion length. Therefore, this study aims to leverage the monotonic property of the logarithmic function to reduce the range and smooth the data. However, for events without congestion, the logarithm cannot be directly applied. Following previous studies, this research adopts a two-stage prediction approach. In the first stage, the objective is to predict whether the accident will cause congestion. In the second stage, the congestion length is predicted only for the events identified as congested in the first stage.

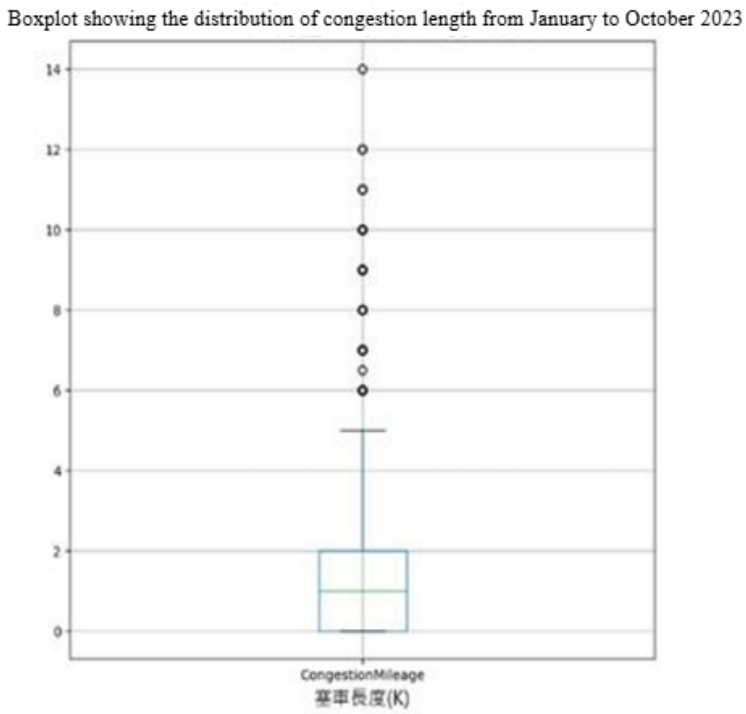


Fig 6: Boxplot showing the distribution of congestion length: The data exhibits high variability, making it difficult to directly predict congestion values.

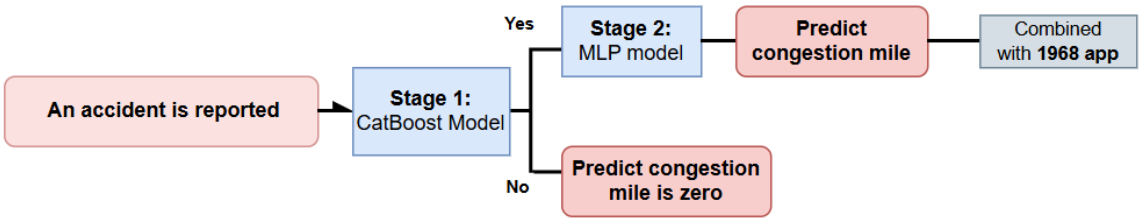


Fig 7: Prediction flow chart

Stage 1: Predict whether an unseen accident will cause congestion

In this stage, the objective is to classify whether an unseen event will result in congestion. The training set consists of events from January to October 2023 (5,702 records), while the test set includes events from January to February 2024 (962 records). The models evaluated in this stage include Support Vector Machines (SVM),

Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, AdaBoost, XGBoost, and CatBoost. The initial results are shown in Figure 9, indicating that tree-based models achieve higher accuracy, with CatBoost demonstrating the best overall performance.

For events from January to October 2023, incidents with a **Congestion Mileage** greater than 0 are classified as congested, while those with a value equal to 0 are considered non-congested. As shown in the pie chart (Figure 8), the dataset exhibits a relatively balanced distribution between congested and non-congested events. Furthermore, for events with congestion, applying a logarithmic transformation reduces the data range and variability, as illustrated by the boxplot.

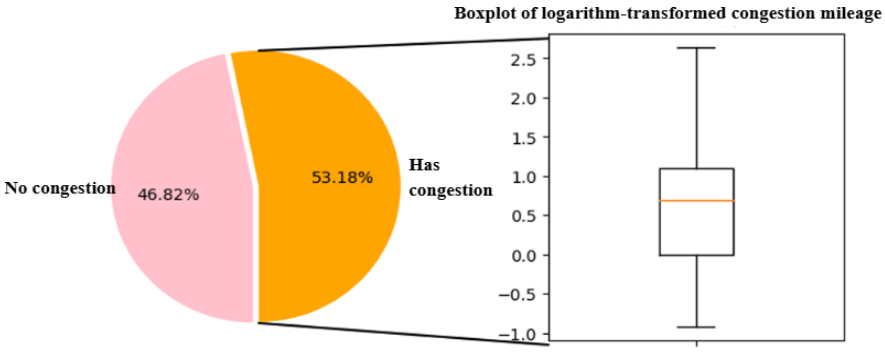


Fig 8: Identify whether events caused congestion and log-transformation to reduce data variability.

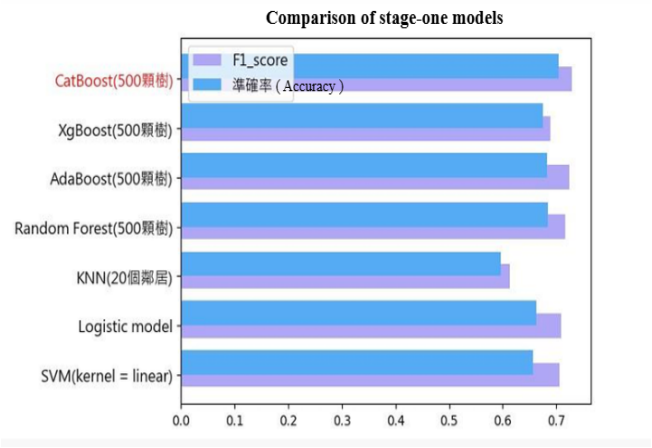


Fig 9: Among the first-stage models, tree-based models demonstrate superior performance, with CatBoost achieving the best results

(1) CatBoost Model Overview

The name **CatBoost** is derived from the words *Category* and *Boosting*. Traditional boosting-based models typically require categorical variables—such as *TimeCategory* and *DateCategory*—to be transformed using one-hot encoding. This transformation increases data dimensionality, leading to a higher risk of overfitting and increased computational cost, while also potentially neglecting the relationship between categorical variables and the target variable. To address these issues, CatBoost was introduced to handle categorical features more effectively.

The model employs an alternative approach for handling categorical variables, which preserves the inherent order within each variable and captures their relationship with the target variable. This not only reduces data dimensionality but also makes more effective use of categorical features. Moreover, the tree-based structure provides greater flexibility in capturing feature interactions.

(2) Model Parameter Selection

After experimenting with different model parameters, the results are presented in Figure 10. As shown, the accuracy remains around 0.7 across most settings, with the highest accuracy of approximately 0.713 achieved when using 950 trees with a depth of 4, corresponding to the white region in the figure.

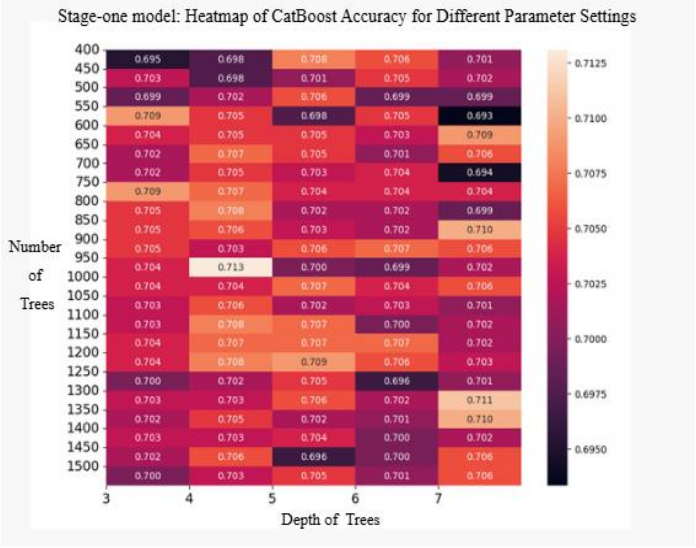


Fig 10: Heatmap of CatBoost Parameter Tuning

(3) Discussion of First-Stage Model Results and Adjustments

After selecting the optimal model parameters, the model was trained using the combination that achieved the highest accuracy, and the resulting confusion matrix is presented in **Figure 11**. From a public perspective, among prediction errors, False Negatives—cases in which an event causes congestion but is predicted as non-congested—are more likely to provoke negative reactions than False Positives. Therefore, the model is adjusted to capture as many truly congested events as possible, making the improvement of **Recall** the primary objective.

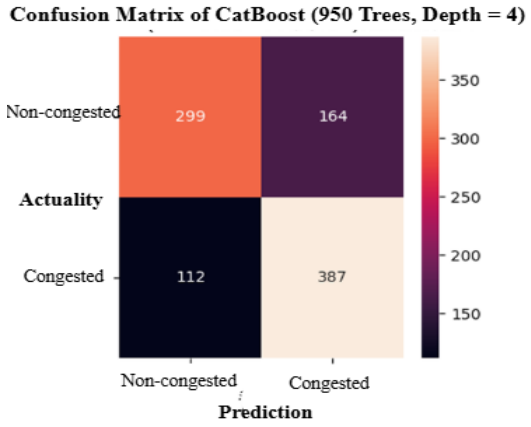


Fig 11:Confusion matrix of stage-one model

This study achieves the objective by adjusting the proportion of initial training samples in CatBoost. In addition to **Recall** and **Accuracy**, the **F-score** is also used as a reference metric (Equation 5), as it balances the trade-off between **Precision** and **Recall**. When B = 2 , the metric is referred to as the **F2-score**, which places more weight on Recall. Therefore, the F2-score is selected as an additional evaluation metric.

$$F_B = \frac{(1+B^2)(\text{Precision}\cdot\text{Recall})}{(B^2\cdot\text{Precision})+\text{Recall}} \quad (5)$$

From **Figure 13**, the model performance under different sample weights can be observed. The sample proportion corresponding to the **maximum F2-score** is selected, with congested events set to approximately 0.87 and non-congested events to 0.13. This configuration achieves a **Recall** value above 0.9. As shown in **Figure 12**, the number of False Negatives is reduced to 7, indicating a high probability of correctly predicting events that will cause congestion.

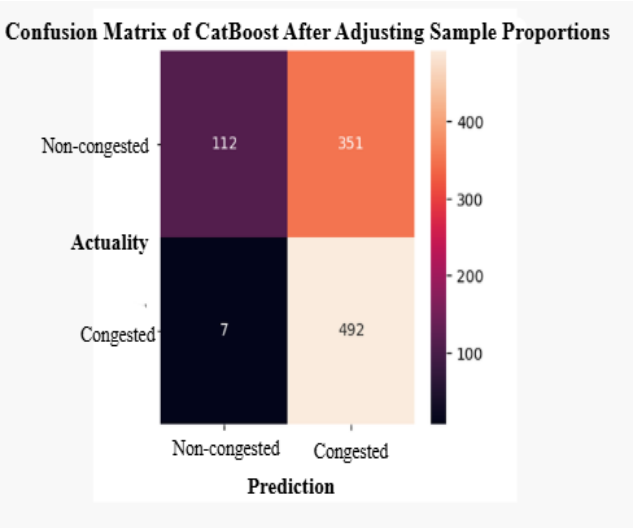


Fig 12: Confusion Matrix of stage-one model after adjustment

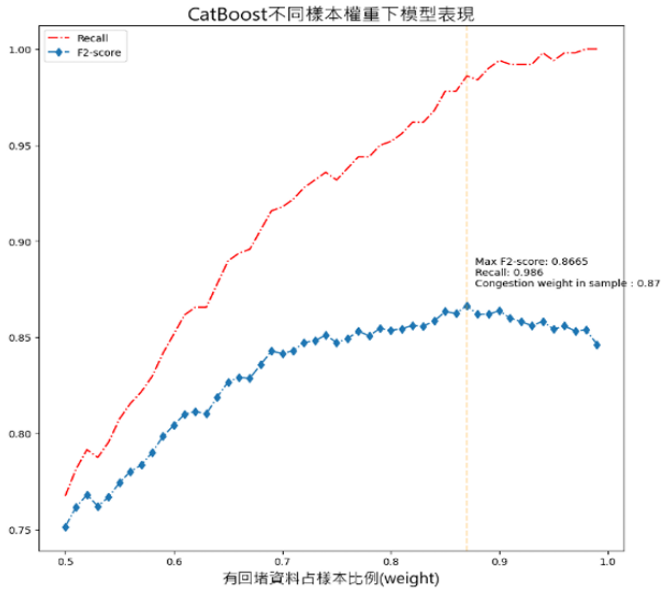


Fig 13: Line plots of the two training metrics under different sample proportions. The selection criterion was to maximize the **F2-score** while maintaining a high **Recall**.

(4) Discussion of important features

Shapley values were used to quantify the contribution of each feature to the model's

predictions. Based on **Figure 14**, the most important features for the first-stage model are:

1. **Pre-accident average speed (Pre_AverageCarSpeed):** Lower values indicate longer congestion lengths.

2. **Ramp involvement (Ramp):** Accidents occurring on ramps are less likely to cause congestion.

3. **Time category (TimeCategory):** As a categorical variable, certain time periods contribute more significantly to the target prediction.

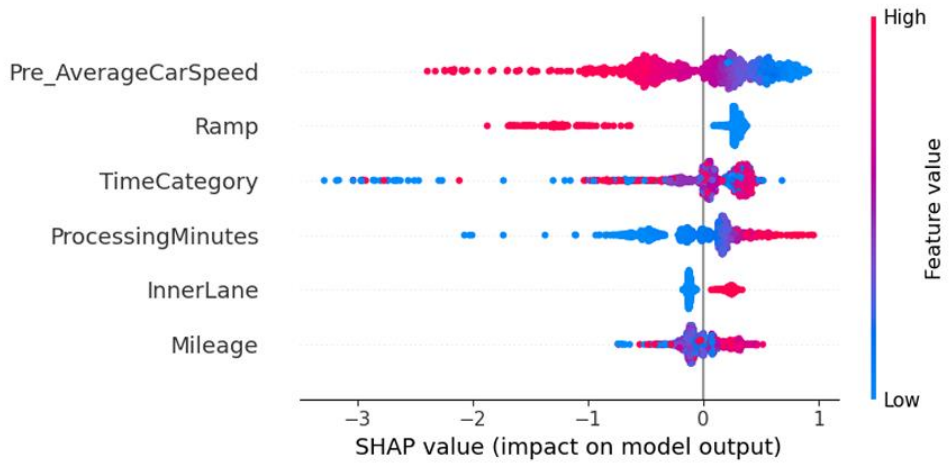


Fig 14: Feature Importance Ranking Using Shapley Values

Stage 2: Predict congestion miles (Congestion Mileage)

Based on the previous stage, the **"Congestion Mileage"** variable is log-transformed, and the features are selected from the 25 variables obtained through prior dimensionality reduction. The training set for this stage consists of congested events (Congestion = 1) from January to October 2023 (112th year) with a total of 3,032

records. The test set includes events predicted as congested by the first-stage CatBoost model from January to February 2024 (113th year), totaling 843 records.

In this stage, we trained four models: **Linear Regression**, **Random Forest**, **Support Vector Regression (SVR)**, and **Multilayer Perceptron (MLP)**. **Figure 13** presents the comparison of their performance. The objective of parameter tuning in this study was to:

- **Minimize RMSE:** The goal is to minimize the difference between the predicted congestion mileage and the actual congestion mileage.
- **Minimize underestimation rate:** If the predicted congestion mileage is lower than the actual value, the model provides limited utility for drivers (the primary users of the highway). Therefore, the most effective way to improve model performance is to reduce the proportion of underestimated congestion mileage.

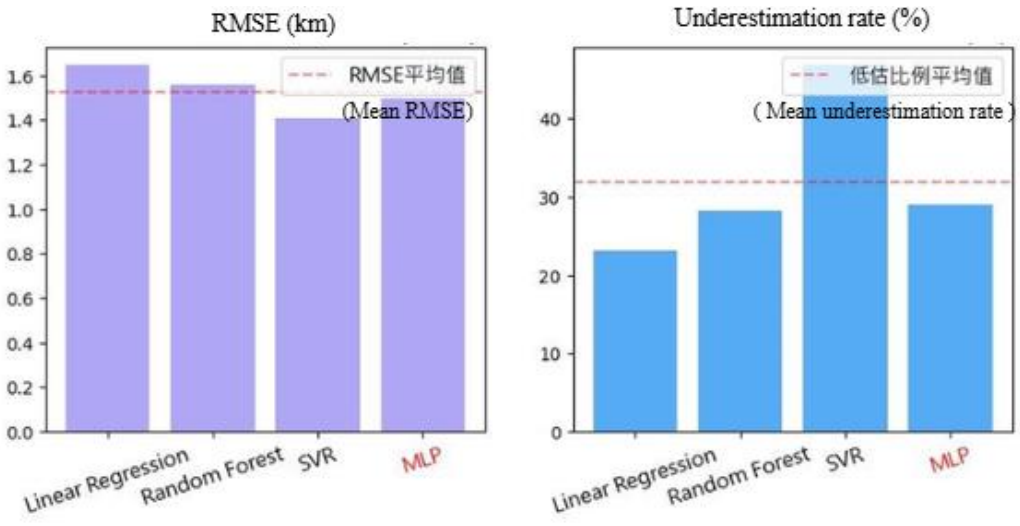


Fig 15: **Comparison of Second-Stage Models:** Among the four models, the **MLP** outperforms the others in terms of **RMSE** and **underestimation rate**, both being better than the average performance.

As shown in **Figure 15**, the **MLP** model performs below the average line for both metrics, indicating smaller prediction errors and lower underestimation rates. Therefore, the **MLP** was selected as the final model for the second stage.

The **Multilayer Perceptron (MLP)** consists of multiple layers of neurons, including an input layer, several hidden layers, and an output layer. Each neuron in a layer is connected to all neurons in the next layer, allowing the MLP to learn and model complex nonlinear relationships.

For this study, the MLP was constructed with an input layer, six hidden layers containing 256, 128, 128, 64, 64, and 32 neurons respectively, and an output layer. The **Rectified Linear Unit (ReLU)** activation function was used to introduce nonlinearity, enabling the network to learn complex patterns. **Dropout layers** were added between hidden layers to randomly deactivate some neurons, reducing the risk of overfitting. The **Adam optimizer** was employed to accelerate learning and adjust the learning rate. During training, the model was trained for 150 epochs with a batch size of 128 for weight updates. After training, the log-transformed predictions were converted back to the original scale for comparison with the actual values.

Based on the results, the preliminary hypothesis for the superior performance of the **MLP** model compared to the other three models is as follows:

1. **Nonlinear modeling capability:** The MLP can capture complex nonlinear relationships between input features and the target variable. When dealing with heterogeneous and nonlinear data, including traffic factors, construction factors, and weather conditions, the MLP is able to produce more accurate predictions.

2. **Deep feature learning:** Through multiple hidden layers, the MLP learns deep feature representations from the data, a capability far exceeding that of single-layer models such as linear regression.

3. **Overfitting mitigation:** Properly incorporating **Dropout layers** in the network effectively reduces the risk of overfitting.

4. **Adam optimizer:** With its adaptive learning rate strategy, the Adam optimizer efficiently finds optimal weights, greatly facilitating fast convergence and improving model stability.

For the second-stage MLP model, when applied to the test set consisting of events predicted as congested in the first stage, approximately **29.024%** of the predicted congestion mileage was lower than the actual value, indicating that the model slightly underestimated congestion severity. The **RMSE**, representing the square root of the mean squared error between predicted and actual congestion mileage, was approximately **1.496 km**.

4 Research Result and Discussion

4.1 Research Result

4.1.1 Two-Stage model result

After combining the two trained models, the test set achieved an **RMSE** of **1.45 km** and an underestimation rate of **24.94%**. We consider an error of approximately 1.45 km acceptable for highway applications. However, there remains room for improvement in reducing underestimation, as the first-stage model already achieves a sufficiently high **Recall**. The main source of underestimation is therefore attributed to the second-stage model. Future work should focus on more refined modeling of this stage to further decrease the underestimation rate, while currently minimizing overall prediction error remains the guiding principle.

4.1.2 Significant variables

As shown in **Figure 14**, the most important factors are:

1. **Event-related factors:** the location and time of the accident, the incident handling time, and the mileage.

2. **Pre-accident traffic conditions:** the average vehicle speed before the accident.

4.2 Discussion of Error Causes

1. Lack of feature precision:

On highways, due to high vehicle speeds, missing even one minute of data can

result in significant information loss, which negatively impacts model performance. For example:

- As shown in **Figure 14**, the most important feature is the **pre-accident average vehicle speed**, which requires integrating all gate data from the four hours prior to a single event and accurately extracting data from the ten minutes before the accident to calculate speed. Calculating this for each event takes at least 3 minutes. However, another feature, **traffic volume in the ten minutes before the accident**, is derived from data aggregated in five-minute intervals, reducing precision and significantly limiting its predictive utility.

- Regarding weather data, prior studies suggest that pre-accident rainfall could be an important feature. However, historical and real-time rainfall data are difficult to obtain, and there are no dedicated weather stations on the highway. This study had to rely on foreign weather websites and select nearby observation stations via maps, some of which lacked rainfall data entirely. Therefore, for future accident prediction, collecting more granular data, such as **per-minute traffic volume and per-minute rainfall**, is recommended.

2. Highway gate malfunctions:

During feature extraction, there is a risk of gate failures, which can result in missing data for extended periods. It is recommended to design contingency plans to mitigate the impact of gate malfunctions.

3. Accident report content:

- When parsing accident reports, we found many unclear definitions or inconsistent vehicle type naming. For example, construction vehicles might be labeled as "工工程車" and "施工車," cement trucks as "水泥車" and "預拌混凝土車," or small cars as "小自客_x000D_\n" and "小自客\n," requiring manual reclassification.

- The goal of this study is to predict congestion length, but for at least 100 events, congestion data were missing. It is recommended that future accident reports follow a standardized format and have clear vehicle type classification to reduce feature engineering complexity and improve model accuracy.

5 Practical Application and Feasibility

Based on the current highway monitoring and notification systems provided by the **Highway Bureau**, road monitoring includes weather conditions, accidents, and vehicle detection, while notification systems comprise the **1968 Highway App and website**, roadside variable message signs, real-time traffic radio, and police broadcast channels. Most of the data used in this study were obtained directly from the Highway Bureau. Only a small portion of weather data was sourced from other websites, due to the unavailability of real-time weather at the time of accidents from the Meteorological Bureau; however, highway weather monitoring systems can adequately complement these missing data. Therefore, for future applications—whether constructing and predicting models for other highway sections, regularly updating predictive models, or expanding datasets and variables to improve model

performance—the process is straightforward and feasible, as it mainly involves utilizing existing data. This demonstrates the high practicability of the proposed approach.

1. **Real-time traffic conditions and network maps for northern, central, and southern regions (see Figure 16):**

   Currently, this function provides real-time vehicle speeds for various highway segments with an update rate of one minute. If the system could predict potential congestion in affected segments immediately after an accident and issue early warnings for areas expected to experience severe delays, including the estimated congestion mileage, it would offer instant visual notifications. This would allow co-drivers or drivers yet to enter the road to make informed decisions about whether to change routes, reducing additional traffic inflow to congested areas.

2. **Route travel time estimation and subscribed segment notifications (see Figure 17):**

   These two features, similar to Google Maps' travel time estimation, allow users to calculate the estimated travel time between selected entry and exit ramps based on current traffic conditions or pre-subscribed time periods. During accidents, integrating the predictive models from this study can provide real-time updates on the expected congestion levels for each location. This enhances the accuracy of travel time predictions, reduces forecasting errors, and helps prevent drivers from misjudging travel durations.

3. **Traffic event notifications (see Figure 18):**

When enabled, this feature provides audio notifications of incidents within a specified distance ahead based on the vehicle's current location. Drivers can select the types of notifications they wish to receive, such as traffic controls, construction, or accidents. For accident-related notifications, the predicted congestion mileage and the locations of entry ramps preceding the congestion can be included. This provides drivers with options to bypass congested segments, helping to reduce traffic flow in affected areas and prevent further congestion escalation.

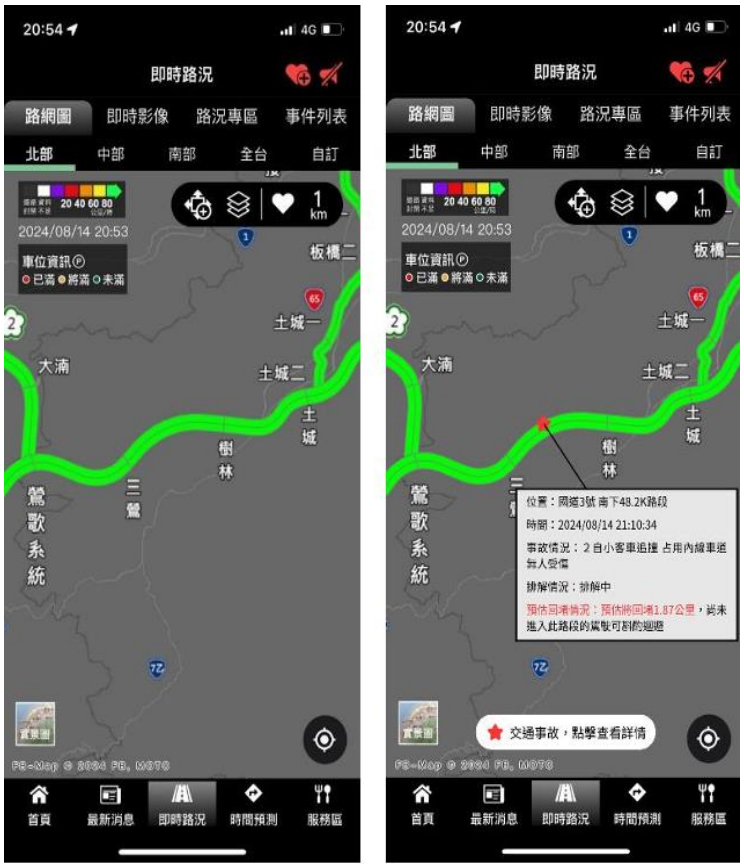


Fig 16 **Real-time traffic conditions and network maps:**

At the moment an accident occurs, the system can mark the location, display clearance status, and provide predicted

Fig 17 **Segment travel time prediction:**

The system can incorporate the presence of current accidents when predicting travel times and provide

Fig 18 **Traffic event notifications:**

The system can provide voice alerts of upcoming accidents based on the user's current location, including predicted congestion information for the affected segments.

6 References

1. 國道易壅塞路段(https://data.gov.tw/dataset/33191)
2. Abdi, A., Seyedabrishami, S., & O'Hern, S. (2023). A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data. Journal of advanced transportation, 2023(1), 8737185.
3. Tseng, F. H., Hsueh, J. H., Tseng, C. W., Yang, Y. T., Chao, H. C., & Chou, L. D. (2018). Congestion prediction with big data for real-time highway traffic. IEEE Access, 6, 57311 57323.
4. Timeanddate (https://www.timeanddate.com/)
5. 高工局交通資料庫 (https://tisvcloud.freeway.gov.tw/history-list.php)

6. 113年國道智慧交通管理創意競賽資料下載

   (https://freeway2024.tw/links#links)