International Journal of Intelligent Transportation Systems Research (2022) 20:309–319
https://doi.org/10.1007/s13177-021-00293-y

Improving the Accuracy of Traffic Accident Prediction Models on Expressways by Considering Additional Information

Yuki Wakatsuki[1] · Jumpei Tatebe[1] · Jian Xing[2]

Received: 8 June 2021 / Revised: 16 December 2021 / Accepted: 22 December 2021 / Published online: 16 February 2022

Abstract
This study aims to improve the accuracy of a convolutional neural network (CNN) based model. That predicts the likelihood of accidents on a specific road section from the present to 2 h in the future using a wide range of temporal and spatial sensor information developed in previous studies as input to reduce accidents. In addition to previous studies that only used traffic data (i.e., speed, traffic volume, time occupancy, etc.), we considered time data (i.e., day of the week, time of day, etc.) and weather data as additional explanatory variables. Then, using the chi-square test, we selected the information that contributed to improving the accuracy of accident occurrence prediction and added it as input to the CNN-based model. Compared with the base model, the average F1-score of the proposed model was improved by 19.1%.

Keywords Traffic accident · Prediction · Convolutional neural network · Chi-square test

1 Introduction

Road traffic injuries and deaths are serious problems globally and are anticipated to increase in the future [1]; as of 2016, the annual global mortality number due to road accidents were reported to be approximately 1.35 million [1]. Real-time traffic accident prediction is expected to reduce traffic accidents and improve road management capabilities [2].

Many studies on traffic accident prediction have been conducted based on this background. Recently, many neural network-based models for predicting traffic accidents have been developed [3–7]. Tsubota et al. [7] modeled a fixed section of an expressway using 2D input data, taking into account time and space up to the previous hour at multiple locations. They created input data for elements related to three channels of traffic information—speed, traffic volume, and time occupancy rate (OCC)—to develop a convolutional neural network (CNN)-based accident prediction model that considers vehicle interaction. Consequently, they suggested that the output values of the CNN-based model might represent accident risks. However, the selection and preparation methods of input data were not fully considered in the model. Generally, the input data and target time for prediction significantly influence the model's accuracy, and there is room for improving the accuracy. Thus, various studies have been conducted on the selection and addition of input data. In studies, such as [4–6, 8, 9], attempts have been made to use information other than traffic information for accident prediction. In these studies, weather information, day of the week information, time information, aerial photographs, road images, locations where accidents have occurred, and the number of accidents in the past were considered information other than traffic information. Research has been conducted on how much information additional than traffic information contributes to accidents [10]. Accidents occur at different rates depending on the weather, the day of the week, and the time of day. In other words, traffic information and other information (i.e., from now on referred to as additional information) influence the likelihood of accidents. Although there have been examples of accident risk prediction models that consider additional information, no measures evaluate which information is effective in predicting accidents.

✉ Yuki Wakatsuki
wakatsuki.yuki@gridpredict.co.jp

Jumpei Tatebe
tatebe.junpei@gridpredict.co.jp

Jian Xing
xing@ri-nexco.co.jp

1 GRID Inc., Ao Bldg. 6F., 3-11-7, Kitaaoyama, Minato-ku, Tokyo 107-0061, Japan

2 Nippon Expressway Research Institute Co., Ltd., 1-4-1, Tadao, Machida, Tokyo 194-8508, Japan

This study examines the selection and preparation methods of input data for the CNN-based model developed by Tsubota et al. [7] to improve accuracy further.

2 Subject Road and Data

The target road for this study was the inbound section of the Tomei Expressway in Japan from Gotemba to Tokyo (Fig. 1), approximately 84 km. There were 44 vehicle detectors in the target section, and they were divided into four sections of 12, 8, 12, and 12 detectors each. Each vehicle detector counts speed, traffic volume, and OCC at a 5-min interval. The period covered was from 2009 to 2018; we used data from 2009 to 2017 as training data and 2018 as validation data to build a CNN-based model. For details, refer to "2. STUDY SITE AND DATA" [7]. This section explains the geographical relationship between bottlenecks and four sections and the traffic accident data. To add to the explanation, each section has the following traffic characteristics.

Section 1 tends to be congested in the morning near the Tokyo Interchange. Section 2 tends to become congested in the afternoon and evening near the Yamato Tunnel. Section 3 tends to become congested as the congestion occurred in Section 2 is extended upstream. Section 4 tends to become congested in the evening, although not as frequently as Section 3.

The additional candidate data items were weather information, calendar information, time, and accident intervals (details are explained in 3.2.2). The weather data include weather event data obtained and managed independently by Central Nippon Expressway Co., Ltd., and public data (Japan Meteorological Agency [JMA]; https://www.data.jma.go.jp/gmd/risk/obsdl/index.php, Ministry of Land Infrastructure Transport and Tourism [MLIT]; http://www1.river.go.jp/).

For calendar information, we used the days of the week (i.e., Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday) and national and non-national holidays. Twenty-four categories ranging from 0 to 23 h were used for time of the day. For accident intervals, the number of accidents was calculated for each elapsed time since the last accident, and the ratio of the number of accidents was calculated. The time indicated in the accident report calculated the elapsed time.

3 Methodology

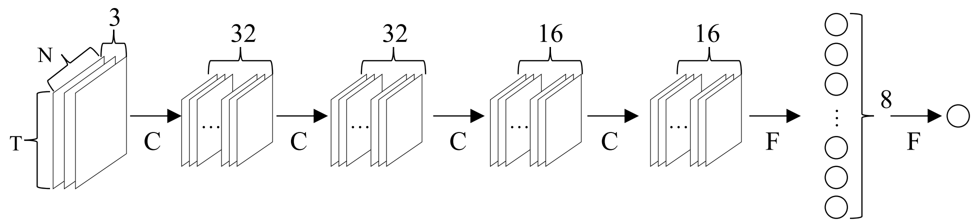
3.1 Problem Setting

The proposed accident occurrence prediction model extends the CNN-based model by Tsubota et al. [7] (from now on, the base model). To begin with, we will explain the structure of the CNN-based model.

The CNN-based model consists of an input layer, convolutional layers, a fully-connected layer, and an output layer, as shown in Fig. 2. The input layer is based on speed, traffic volume, and OCC data aggregated every 5 min. Here N in the figure is the number of detectors installed in the study section. T is the number of hours in a fixed interval; T uses the traffic data for 5 min in the past hour, so T = 12. The three matrices of speed, traffic volume, and OCC become a tensor of dimensions N × T × 3, which is the input for the first layer. In the convolutional layers, the features contained in the input data are extracted. In the fully-connected layer, the data is classified based on the features extracted in the

Fig. 1 Illustration of Study sections

Fig. 2 CNN-based model network



convolutional layer. Input 3D traffic data is classified in this model into those that result in accidents and those that do not.

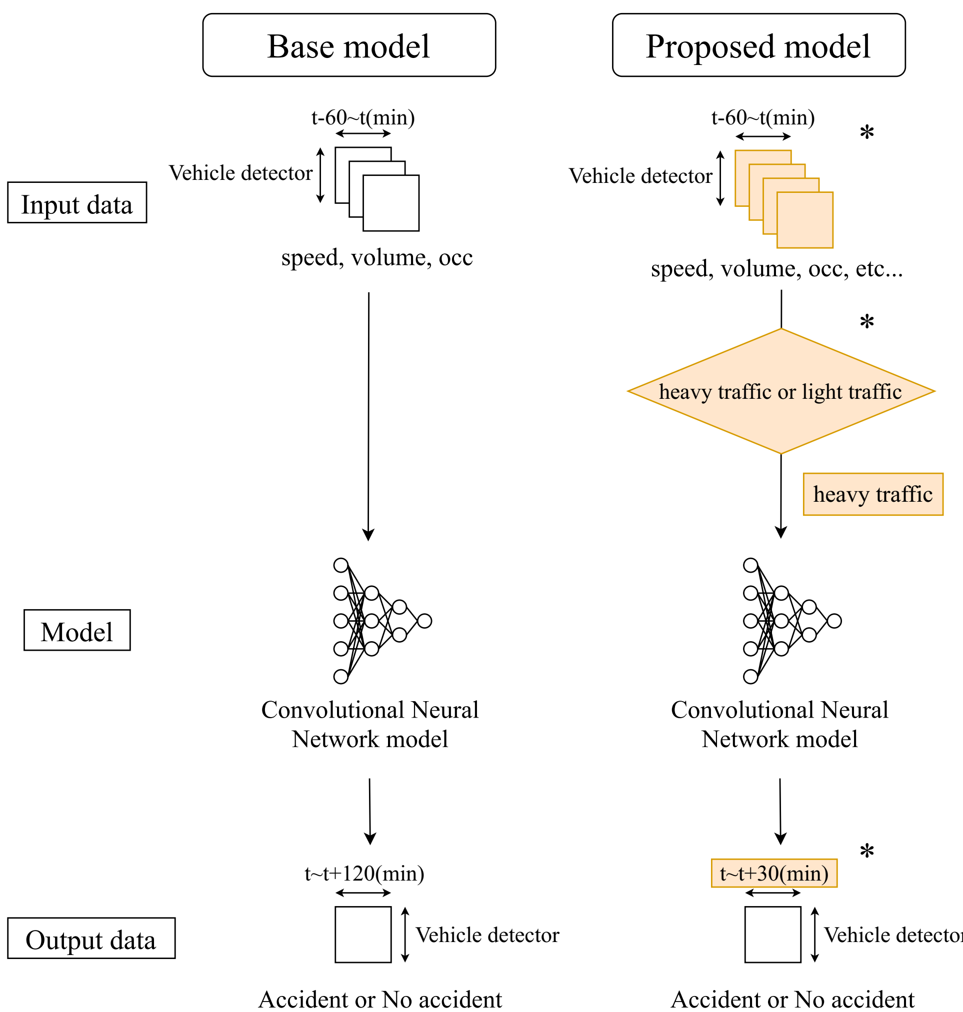
Next, we describe the structure of the proposed model. The conceptual diagrams of the base and proposed model are shown in Fig. 3. The flow of the CNN-based models from input to prediction is shown. The modifications made in this study are indicated in orange. In a previous study, they input the past hour of the three-channel data (speed, traffic volume, and OCC) into the CNN-based model to predict accidents from the present to 2 h in the future. Meanwhile, the proposed model in this study changes three things: the input data creation method, classification of the input data according to the traffic condition, and the target time for a prediction. Each of them will be explained in detail in the next section.

3.2 Input Data Preparation Method

3.2.1 Evaluation of Additional Candidate Data Items by Chi-Square Test

There are three reasons why we determined if weather, day of the week, holidays, time of the day, and accident intervals should be added as inputs to the model. They are described below.

Fig. 3 Schematic of the base and proposed models, classification, prediction time) are indicated by asterisk showing the outline of the CNN-based model from input to prediction. Differences between both methods (data overview, data)



The first is that it is necessary to represent the accident characteristics immediately before an indirectly occurs. Many traffic accidents are due to driver's carelessness or speeding; it is easy to imagine that each driver and vehicle's information is necessary to predict such accidents. However, there is no such data available for study, so it is essential to indirectly express the characteristics before an accident from data that may affect a traffic situation, such as vehicle detectors and weather data.

The second reason is that accident data is scarce. Because traffic conditions are constantly changing, each accident occurs under a unique set of circumstances. In order to build an accident occurrence prediction model from such accident data, it is necessary to extrapolate the characteristics from a massive amount of data, which necessitates a massive amount of accident data. However, the frequency of accidents is approximately 100 annually in each section, making this difficult.

The third reason is that while there are various accident factors (i.e., candidate explanatory variables), it is difficult to consider them all. It has been shown that the frequency of accidents varies depending on the weather, day of the week, and time of the day [10]. Typically, we aim to build an accident occurrence prediction model that considers all factors. Still, it is assumed that increasing the number of explanatory variables reduces the number of accidents per pattern (i.e., the number of training data) and complicates the learning process.

Therefore, it is necessary to express the characteristics immediately before an accident with a limited number of explanatory variables in this study. Thus, from the data format of the descriptive and objective variables in this study, we selected data items to be employed as explanatory variables using the chi-square test.

It is determined that there is a statistically significant difference among specific data item categories. In that case, it is assumed that the difference in categories affects the target event. As a result, the data points with significant differences were explanatory variables that contributed to the prediction of accident occurrence. However, the chi-square test is more likely to show statistically significant results when the sample size is too large [11, 12]. Therefore, we must consider the sample size when conducting the test. In this study, the sample size of the dataset used for the test was very large—more than one million records. As a result, the test must be conducted by sampling portions of the total data. The number of samples in this study was adjusted using the expected frequency value obtained when calculating the chi-square value. The procedure is shown in Fig. 4. First, we randomly sampled the training data (Fig. 4①). Next, we cross-tabulated the sampled data (Fig. 4②). Afterward, we created an expectation table from the cross-tabulation (Fig. 4③). Finally, we checked whether the minimum value of the expectation table exceeded the threshold value, *Th*. If not, we returned to ①.

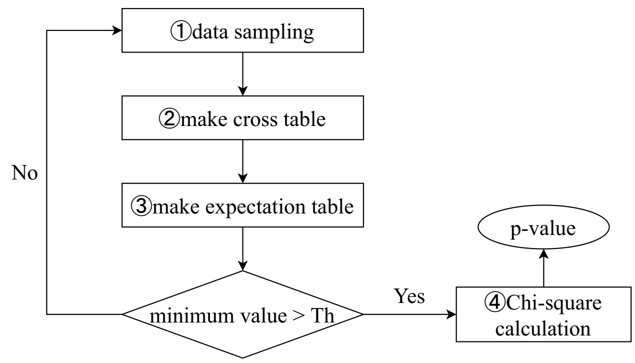


Fig. 4 Flow of the chi-square test conducted in this study

Meanwhile, if it did, the sampling was terminated, and the chi-square value was calculated (Fig. 4④). Experiment ①–④ were conducted T times. The average of the p values in the T experiment was tested for the target p value of a certain data item (in this study, $Th = 10$, $T = 20$, and the significance level was set to $p = 0.05$.).

In the chi-square test, Cochran's rule [13] is known as a rule for setting an appropriate threshold (*Th*), and *Th* was examined following Cochran's rule, with the value of *Th* being 5 or more so that the value of the expected frequency would not exceed 20%. Although not shown here, when *Th* was changed to 5, 10, or 15, the chi-square test results showed that the relationship between the variables was not affected. Two variables, time and accident occurrence interval, showed smaller p-values than the other variables.

The above procedure was done for each of Sections 1–4. In this study, the averages of the p-values obtained through 20 experiments that met the significance level of 0.05 were selected as the explanatory variables to be added.

3.2.2 How to Create Input Data for Explanatory Variables

The four explanatory variables—weather, day of the week, holiday, and time of the day—were encoded to be input into the proposed model. Meanwhile, the explanatory variables were created for accident intervals based on values obtained by aggregating the number of accidents based on the elapsed time since the last accident occurrence.

Qualitative data, such as weather, days of the week, holidays, and time of the day, need to be encoded and converted into numerical values when inputting them into the CNN-based model. In this study, Target encoding was adopted among multiple encoding methods to express the relationship between the objective variable and the explanatory variables that are effective for accident occurrence prediction. The calculation method of target encoding is as follows:

$$e_C = \frac{N_{acc}^C}{Q^C}, \tag{1}$$

where C is the number of categories, N_{acc}^C is the number of accidents, Q^C is the total traffic volume (100 million vehicles) every 5 min in category C, and e_C is the number of accidents relative to the total traffic volume every 5 min in category C; this normalizes the likelihood of an accident by traffic volume.

The accident interval is an explanatory variable that incorporates the ideas of survival time analysis based on the hypothesis that accidents may occur more frequently in situations where accidents are more likely to occur. The term "survival time analysis" refers to studying the time required for an event to occur [14]. For example, suppose we consider a survival function that indicates the probability of survival up to a survival time t. In that case, the survival function decays and approaches zero as t increases [14]. This study conducted a survival time analysis until a traffic accident occurred. First, the time between the occurrence of one accident and the next accident was tabulated. Next, the range of time elapsed since the accident was set to 30 min, and the number of accidents occurring in each range of time, such as 0–30 min, 30–60 min, was tabulated.

The calculation method for the accident interval is as follows:

$$r_C = \frac{N_{acc}^C}{N_{acc}}, \tag{2}$$

where C is the number of categories (the range of the elapsed time from the accident every 30 min), N_{acc}^C is the number of accidents in category C, and N_{acc} is the number of all accidents. The accident occurrence rate r_C indicates the number of accidents in a certain category N_{acc}^C relative to the overall number of accidents N_{acc}. r_C was calculated by setting N_{acc}^C to 0 for accidents that occurred more than 24 h ago, as it was considered that these accidents would not affect accidents that happened 30 min ago. This means that the total number of categories was 49. The r_C calculated in this way was named the explanatory variable of accident intervals.

3.2.3 Classification of Data According to Traffic Conditions

As can be imagined, the factors that contribute to accidents differ greatly depending on the traffic condition. The factors that contribute to accidents differ in light and heavy traffic conditions. As a result, it is thought that the input data suitable for accident risk prediction may also vary. Therefore, we considered categorizing traffic conditions into two categories (light/heavy traffic) based on traffic conditions. We assumed different accident factors and aimed to develop accident occurrence prediction models using different approaches for each.

This study focuses on accidents during restraint driving in heavy traffic conditions (i.e., a traffic condition in which the driver is restrained by surrounding vehicles). The accidents in heavy traffic are mainly rear-ended collisions or sideswipes, which are quite different from single-vehicle crashes with roadside facilities in light traffic conditions (i.e., especially solo driving). The severity of accidents is not taken into account in the study. Due to their low relative speed, vehicle-to-vehicle collisions in heavy traffic are generally less severe than single-vehicle collisions in light traffic.

The light traffic state is defined in this study as the area of $V > V_b$ and $Q < Q_b$ in the Q-V diagram as shown in Fig. 5 while the remaining area is defined as the heavy traffic state. Here V_b was set to 60 (km/h), close to the critical speed at which the traffic capacity is reached. Q_b is the boundary between light and heavy traffic conditions that here is assumed to be the boundary of LOS B and LOS C. According to the German Highway Capacity Manual (HBS) [15], the traffic volume at this boundary is 55% of the traffic capacity. Considering that the average traffic capacity of the three-lane section of the Tomei Expressway is approximately 5135 veh/h, Q_b is calculated as $Q_b = 5135 \times 0.55 = 2825$ veh/h.

The space mean speed V and average traffic volume Q of a long section are calculated from individual detector sections as follows:

$$V = \frac{\sum_i \sum_j q_{ij} L_i}{\sum_i \sum_j \frac{q_{ij} L_i}{v_{ij}}} = \frac{\sum_i \sum_j k_{ij} v_{ij} L_i}{\sum_i \sum_j K_{ij} L_i}, \tag{3}$$

$$Q = \frac{\sum_i \sum_j q_{ij} L_i}{\sum_i L_i}, \tag{4}$$

where i is the number of the spatial division unit in space-time (approximately 2 km), j is the number of the

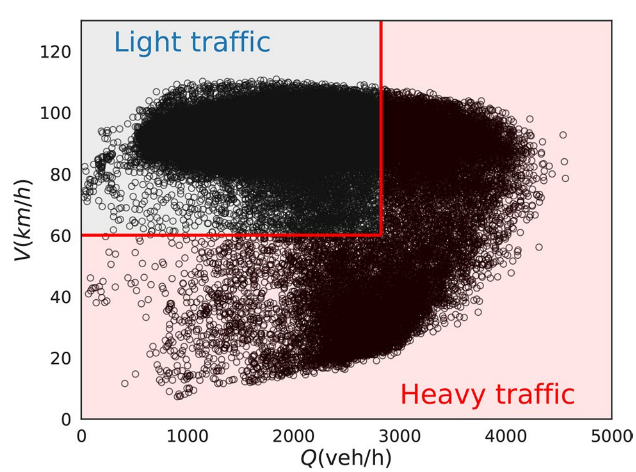


Fig. 5 Scatterplot of average traffic volume and space mean speed (Q-V diagram). Classify the area of the Q-V diagram into two categories: light traffic and heavy traffic

time-division unit in space-time (5 min), q_{ij} is the traffic volume (vehicles), L_i is the section of the detector (km), k_{ij} is the density $= \frac{q_{ij}}{v_{ij}}$(veh/km), and v_{ij} is speed (km/h). Thus, when Q < 2825 veh/h in free flow, the vehicle will be considered in solo driving, and when Q ≥ 2825 veh/h in free flow and when V < 60 km/h in congested flow, the vehicle is restraint driving. Although the Q-V diagram of each section varies depending on the section characteristics, the definition of the boundary of light and heavy traffic states is fixed for each section in the study. Since the purpose of dividing traffic conditions by V and Q is to roughly distinguish between light and heavy traffic states, the differences in V and Q for each section are not considered. Figure 5 shows as an example the Q-V diagram for a section of the Tomei Expressway. Table 1 shows the number of accident data in heavy traffic and section length for each section. Each section has three lanes. For example, there were a total of 3703 heavy traffic periods, of which 89 periods had accidents, and 3614 periods did not have accidents in Section 1.

3.3 Model Development

3.3.1 How to Input to the CNN-Based Model

The base model accepts three-channel input data: speed, traffic volume, and OCC. This study uses this three-channel CNN-based model as a base. It adds the explanatory variables selected in 3.2.1 in two strategies (light/heavy traffic). The conceptual diagram of the modified CNN-based model is shown in Fig. 6. Weather, day of the week, holiday, and time of the day was added to the input layer [Fig. 6 (1)], and accident interval was added to fully-connected layers by one unit [Fig. 6 (2)]. Extracting the features distributed on the spatiotemporal data was possible by adding one channel to the input layer. Meanwhile, adding one unit to fully-connected layers allow us to input the values extracted from the spatiotemporal features. The accident interval is a value that expresses the risk of an accident occurring at a specific time interval, so this addition method was chosen.

3.3.2 Predicted Target Time

Tsubota et al. [7] predicted the presence or absence of accidents occurring between the current time and 2 h ahead; to predict accidents 2 h in advance, some correlation between the traffic conditions in the target section in the past hour (the input data) and accidents 2 h later is considered necessary. Assuming that a given vehicle is traveling at 50 km/h, the maximum length of the section is approximately 25 km (Table 1, thus the vehicle will have finished passing through any section in 30 min. Therefore, we thought there would be little correlation between the accident 2 h later and the input data.

Based on the above description, we decided that the target time for prediction was from the current time to 30 min.

3.4 Evaluation

The confusion matrix and F1-score—commonly used as evaluation indicators for classification models—were used

Table 1 Number of accidents per section in heavy traffic

Section No.	Section	# of periods with accident	# of periods without accident	Length of section (km)
1	Yokohama-Machida – Tokyo	89	3614	19.7
2	Atsugi – Yokohama-Machida	147	5274	15.2
3	Oi-Matsuda – Atsugi	82	2318	22.8
4	Gotemba – Oi-Matsuda	77	1225	25.7

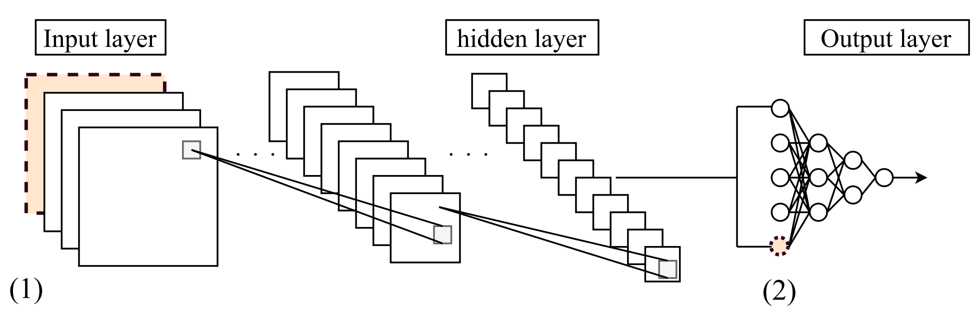


Fig. 6 Conceptual diagram of the proposed model. The explanatory variables to be added are indicated by dashed square and circle. (1) shows the position of one channel added to the input layer of the existing CNN model, and (2) shows the position of unit added fully connected layers

as model evaluation indicators. The confusion matrix is the aggregation of the prediction results of the model by the actual classification class (Table 2).

The F1-score can be calculated using the TP, FN, and FP of the confusion matrix, which is given by

$$F1\ score = \frac{2\times Precision\times Recall}{Precision+Recall} \tag{5}$$

where $Precision = \frac{TP}{TP+FP}$, $Recall = \frac{TP}{TP+FN}$. The F1-score is the score obtained by harmonizing Precision and Recall. F1-score produces values in the range of 0–1, and the closer it is to 1, the better the model's prediction performance.

Table 2 Confusion matrix

		Accident alert (predicted condition)	
		Positive	Negative
Accident occurrence (True condition)	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

4 Results

4.1 Chi-Square Test Results

The chi-square test results are shown in Fig. 7, which shows the distribution of the *p*-values obtained from 20 chi-square tests for each data item by section. The.

The *p*-values for the time of the day and accident interval were low in all experiments, with *p*-values <0.05, indicating a significant difference in the distribution of the presence or absence of accidents between the categories of time and accident interval. For days of the week and holidays, differences in the distribution of *p*-values could be seen for each section. The *p*-values for Section 3 for days of the week and Sections 3 and 4 for holidays were small; significant differences were observed. The day and accident intervals were selected as explanatory variables based on these results since significant differences were found for all sections.

As indicated in 3.2.2, the number of categories for day and accident interval time is larger than other explanatory variables. According to Berman, E., Wang, X. [16], "The degrees of freedom address the practical, statistical problem that the magnitude of most test statistics is affected by the

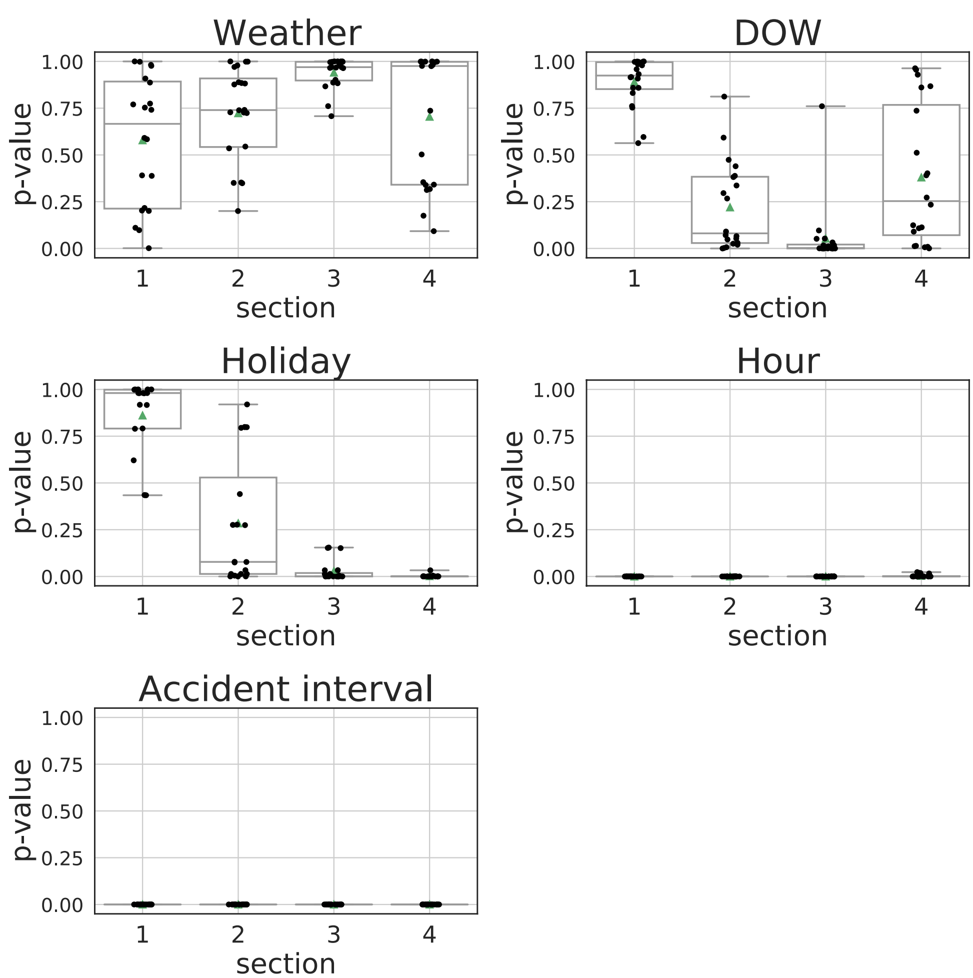


Fig. 7 Results of the chi-square test. Distributions of p-values obtained from 20 sampling experiments for weather, day of the week, holiday, time of the day, and accident interval are shown for each section

number of observations or categories." (p.184). Thus, it can be seen that the difference in the number of categories has little effect on this experiment. Thus, it can be seen that the difference in the number of categories has little effect on this experiment.

Table 3 shows the resultant explanatory variables for the base model and the proposed model constructed based on the chi-square test results.

4.2 Base Model Construction Results

The proposed model was constructed as shown in Fig. 6. Table 4 displays the F1 scores of the base model with speed, traffic volume, and OCC as inputs, with the F1-score for each section listed for 2018.

Results of the construction of the model with the explanatory variables selected by the chi-square test added to the base model.

Table 4 shows the F1-score of the proposed model with speed, traffic volume, OCC, time of the day, and accident interval as inputs. To compare the base model with the proposed model, the time predicted in the base model was changed from 2 h to 30 min. The F1-score for Sections 1, 3, and 4 was higher than the base model's. The section averages showed an improvement of approximately 19.1%. The selection of explanatory variables by the chi-square test improved the F1-score above.

5 Discussions

Our results showed that it was possible to improve the accuracy of the section averages by adding the explanatory variables selected by the chi-square test to the base model. In previous studies, additional information, such as weather, day of the week, time of the day, aerial photographs, road images, locations where accidents have occurred, and the number of past accident occurrences have been used for accident occurrence prediction [4–6, 8, 9]. This study examined statistically additional candidate explanatory variables that previous studies had not considered. As a result, we confirmed that adding the time of day and accident interval contributed to improved accuracy among the available information, such as weather, day of the week, holidays, time of day, and accident interval. The accident interval is a new data item that previous studies have not considered.

We will discuss the difference in prediction characteristics and accuracy between the proposed model in the following.

Table 3 Explanatory variables of base model and proposed model

	Base model	Proposed model
Explanatory variables	speed, traffic volume, OCC	speed, traffic volume, OCC, time of the day, and accident interval

Table 4 Comparison of F1 scores in each section between base model and proposed model

	Base model	Proposed model
Section 1	0.080	0.083
Section 2	0.189	0.172
Section 3	0.066	0.157
Section 4	0.207	0.235
Section average	0.136	0.162
Improvement rate	–	19.1(%)

5.1 Analysis of Model Characteristics

Table 5 compares the Precision, Recall, and F1-score of the base and proposed models. It shows that the proposed model has higher Precision and lower Recall except for Section 2. In other words, there are fewer false alarms than the base model.

Section 1 Comparing the base model and the proposed model in Section 1 (Table 6), the proposed model had higher F1-score and Precision values between 6:00 and 11:59 than the base model. Section 1 is characterized by many accidents during traffic congestion in the morning rush hour to Tokyo. We believe that the model captured this characteristic by adding time.

Section 2 Comparing the base model and the proposed model in Section 2 (Table 7), the values of the base model were generally higher than those of the proposed model, or the results were the same. These results are because the speed, traffic volume, and OCC included in both models may

Table 5 Comparison of F1, Precision and Recall values in each section between base model and proposed model

Time period	Base model/ Proposed model	F1	Precision	Recall
0:00–5:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000
6:00–11:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000
12:00–17:59	Base model	0.225	0.152	0.427
	Proposed model	0.211	0.155	0.333
18:00–23:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000

Table 6 Comparison of F1, Precision and Recall values in each time period in Section 1

Time period	Base model/ Proposed model	F1	Precision	Recall
0:00–5:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000
6:00–11:59	Base model	0.084	0.045	0.732
	Proposed model	0.093	0.073	0.127
12:00–17:59	Base model	0.039	0.024	0.118
	Proposed model	0.000	0.000	0.000
18:00–23:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000

Table 7 Comparison of F1, Precision and Recall values in each time period in Section 2

Time period	Base model/ Proposed model	F1	Precision	Recall
0:00–5:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000
6:00–11:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000
12:00–17:59	Base model	0.225	0.152	0.427
	Proposed model	0.211	0.155	0.333
18:00–23:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000

Table 8 Comparison of F1, Precision and Recall values in each time period in Section 3

Time period	Base model/ Proposed model	F1	Precision	Recall
0:00–5:59	Base model	0.051	0.026	1.000
	Proposed model	0.000	0.000	0.000
6:00–11:59	Base model	0.113	0.060	1.000
	Proposed model	0.000	0.000	0.000
12:00–17:59	Base model	0.077	0.040	1.000
	Proposed model	0.163	0.103	0.387
18:00–23:59	Base model	0.036	0.019	1.000
	Proposed model	0.148	0.088	0.462

Table 9 Comparison of F1, Precision and Recall values in each time period in Section 4

Time period	Base model/ Proposed model	F1	Precision	Recall
0:00–5:59	Base model	0.000	0.000	0.000
	Proposed model	0.000	0.000	0.000
6:00–11:59	Base model	0.533	0.444	0.667
	Proposed model	0.615	0.571	0.667
12:00–17:59	Base model	0.199	0.138	0.355
	Proposed model	0.222	0.169	0.323
18:00–23:59	Base model	0.125	0.080	0.286
	Proposed model	0.000	0.000	0.000

have a greater contribution to the accident than the time of day and the accident interval that were selected from the chi-square test. Section 2 has the Yamato Tunnel where traffic congestion occurs frequently. This congestion causes traffic accidents. Since congestion can be described by only speed, traffic volume, and OCC, we think that the time of day and the accident interval did not improve the F1-score.

Section 3 Comparing the base model and the proposed model in Section 3 (Table 8), the proposed model had higher F1-score and Precision values between 12:00 and 17:59 and between 18:00 and 23:59 than the base model. Section 3 is characterized by the number of vehicles returning to the Tokyo area in the evening and night of the busy season and holidays. The resulting traffic congestion causes many accidents. The proposed model captured its characteristics by adding time and accident occurrence intervals from the above.

Section 4 Comparing the base model and the proposed model in Section 4 (Table 9), the proposed model had higher F1 scores and Precision values between 6:00 and 11:59 and between 12:00 and 17:59 than the base model. Since there was less traffic congestion and fewer accidents between 6:00 and 11:59 even if it was heavy traffic, it was challenging to predict accidents using only speed, traffic volume, and OCC. Therefore, the time and accident interval may have contributed to the improvement of the F1-score and precision.. From 12:00 to 17:59, the traffic jam that occurred on Section 3 was extended to Section 4, which is often congested during the busy season, weekends, and holidays, and this leads to accidents, so the time and the accident interval were considered to be explanatory factors for the accidents.

5.1.1 Summary of each section's Model Characteristics

From the trend of the difference between the base model and the proposed model in the F1-score, Precision, and Recall of each section, it can be said that the proposed model was able to predict accidents with fewer false alarms in a specific time but more misses occurred. This can be attributed to the fact that the proposed model appropriately trained past accident occurrences and the situations in which accidents frequently occurred by adding the time and accident intervals. On the other hand, the proposed model can not be trained for accidents in situations where traffic jams are not frequent.

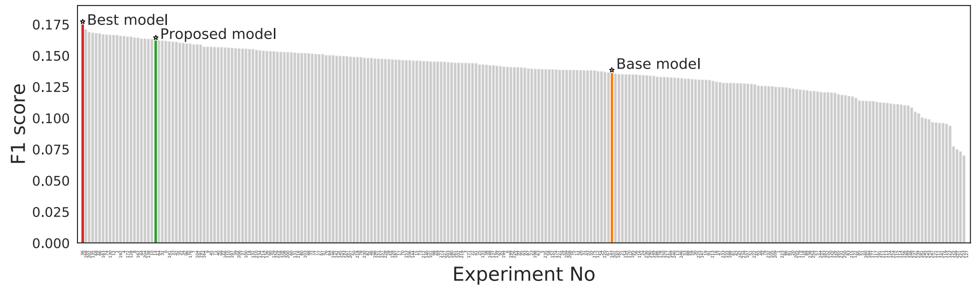


Fig. 8 Results of the F1-score from the all-explanatory variables search in descending order, from left to right, for each model. Figure shows the score of Best model (the model with the highest score), Proposed model, Base model. Unannotated bars indicate other models

However, there were cases where the accuracy of the base model was higher than that of the model for other than specific periods. This suggests that other factors contribute to accidents other than the data items targeted in this study. To find these factors is what we need to do in the future.

5.2 Analysis of the Predictive Accuracy

In this section, we compare the accuracy of the proposed model with that of a model using other combinations of explanatory variables as input to confirm whether the descriptive variable selection method using the chi-square test is effective. This allows us to see if the explanatory variables that show statistically significant differences in their contribution to accidents across categories help to improve the CNN-based model's prediction accuracy. Specifically, we constructed a CNN-based model with all eight combinations of explanatory variables (254 ways) and compared their F1-score for the 2018 data to the proposed model's F1-score.

The accuracies of the proposed model and the model using other combinations of explanatory variables are shown in Fig. 8, which shows the F1-score of the section averages of the models with the 254 combinations of explanatory variables in descending order, from left to right. The model with the highest F1-score (referred to as the best model) was the model with the addition of speed, traffic volume, day of the week, holidays, time of the day, and accident interval. The middle section F1-score of this model was approximately 0.175, indicating that the best model slightly exceeded the F1-score of the proposed model (Table 4). However, the time and accident intervals in the proposed model were included as inputs in the best model. The fact that the explanatory variables selected by the chi-square test were also included in the best model obtained by searching for combinations of all explanatory variables suggested that this selection method was effective.

The chi-square test performed well as an explanatory variable selection method in this study. However, because of the characteristic of being naive to sample size [11, 12], the process for adjusting the sample size should be investigated in the future.

6 Conclusion

This study examined input data's selection and creation methods and prediction time for a CNN-based model using input from a previous study using spatiotemporal 2D data of speed, traffic volume, and OCC [7]. For selecting input data, time of the day and accident interval were selected as additional explanatory variables using the chi-square test. Target encoding was used to create the input data. The explanatory variables were developed considering the correlation with the target variable. The target time for prediction was shortened. These measures resulted in results that exceeded the F1-score of existing models.

After analyzing the characteristics of the chi-square test model constructed by the above measures, it can be said that the proposed model was able to predict accidents with fewer false alarms during specific periods. This can be attributed to the fact that the model appropriately trained the past accident occurrence periods and situations where accidents frequently occurred by adding the time and accident interval. However, there were cases where the precision of the base model was higher than that of the model for other than specific periods. This suggests that other factors contribute to accidents other than the data items targeted in this study.

When the proposed model's accuracy was compared to a CNN model with other combinations of explanatory variables, the F1-score was relatively high. The CNN-based model with the highest F1-score among the descriptive variable varieties included explanatory variables chosen using the chi-square test, indicating that the proposed method is effective. We will search for other factors that contribute to accidents in the future (e.g., information on traffic conditions for each lane, effects of upstream and downstream congestion, and information for each vehicle using ETC2.0 probe data). Then they will be added into the model as explanatory variables to improve accuracy further. As for selecting explanatory variables, we will study how to adjust the number of samples when conducting a chi-square test.

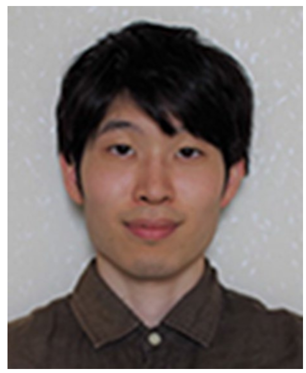
Acknowledgments We thank Prof. Toshio Yoshii and Dr. Takahiro Tsubota of the Faculty of Engineering, Ehime University, for their advisory, Dr. Michael Dziomba of GRID Inc. for his cooperation in this research, and Dr. Katsunori Tanaka of GRID Inc. for his advisory in proofreading.

Declarations

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. World Health Organization: "The global burden of road traffic deaths", Global status report on road safety 2018, pp. 2–13. Geneva (2018)
2. Zhu, L., Yu, F.R., Wang, Y., Ning, B., Tang, T.: Big data analytics in intelligent transportation systems: A survey. IEEE Transactions on Intelligent Transportation Systems **20**(1), 383–398 (2019). https://doi.org/10.1109/TITS.2018.2815678
3. Wang, G., Kim, J.: A large scale neural network model for crash prediction in urban road networks. In Proceedings of Australasian Transport Research Forum 2017, Auckland, New Zealand, 27–29 Nov 2017
4. Ozbayoglu, M., Kucukayan, G., Dogdu, E.: A real-time autonomous highway accident detection model based on big data processing and computational intelligence. In Proceedings of 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, 5–8 Dec 2016. https://doi.org/10.1109/BigData.2016.7840798
5. Ren, H., Song, Y., Wang, J., Hu, Y., Lei, J.: A deep learning approach to the citywide traffic accident risk prediction. In Proceedings of 2018 IEEE International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, 4–7 Nov 2018. https://doi.org/10.1109/ITSC.2018.8569437
6. Li, P.: A Deep, Learning Approach for Real-Time Crash Risk Prediction at Urban Arterials Urban Arterials. University of Central Florida. https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1088&context=etd2020. Accessed 3 Sep 2021
7. Tsubota, T., Yoshii, Y., Xing, J.: Prediction of traffic accident likelihood on intercity expressway by convolutional neural network. Intelligence, Informatics and Infrastructure. **1–1**, 11–17 (2020). https://doi.org/10.11532/jsceiii.1.1_11
8. Hébert, A., Guédon, T., Glatard, T., Jaumard, B.: High-resolution road vehicle collision predictionfor the city of montreal. In Proceedings of 2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA 9–12 Dec 2019. https://doi.org/10.1109/BigData47090.2019.9006009
9. Yuan, Z., Zhou, X., Yang, T.: Hetero-ConvLSTM: A deep learning approach to traffic accident prediction on heterogeneous spatiotemporal Data. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom, 19-23 Aug 2018. https://doi.org/10.1145/3219819.3219922
10. Lin, L., Wang, Q., Sadek, A.W.: Data mining and complex network algorithms for traffic accident analysis. Transp. Res. Rec. **2460**(1), 128–136 (2014). https://doi.org/10.3141/2460-14
11. Armstrong, R.A.: Is there a large sample size problem? Ophthalmic Physiol. Opt. **39**(3), 129–130 (2019). https://doi.org/10.1111/opo.12618
12. Lin, M., Lucas, H.C., Shmueli, G.: Research commentary—too big to fail: large samples and the p-value problem. Inf. Syst. Res. **24**(4), 906–917 (2013). https://doi.org/10.1287/isre.2013.0480
13. Rayson, P., Berridge, D., Francis, B.J.: Extending the cochran rule for the comparison of word frequencies between corpora. In Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT). 4, 1–12 (2004)
14. Kleinbaum, D.G., Klein, M.: Introduction to Survival Analysis, pp. 1–54. Survival Analysis A Self-Learning Text Third Edition, New York, Springer (2011)
15. Geistefeldt, J.: Assessment of Basic Freeway Segments in the German Highway Capacity Manual HBS 2015 and Beyond. Transportation Research Procedia. 15, 417–425 (2016). https://doi.org/10.1016/j.trpro.2016.06.035
16. Berman, E., Wang, X.: Hypothesis Testing with Chi-Square. Essential Statistics for Public Managers and Policy Analysts 4th Edition, pp. 178–201. SAGE Publications, Inc., Thousand Oaks, California (2016)



Yuki Wakatsuki is a data scientist in GRID Inc. He received M. Eng in 2020 from Kochi University of Technology. His research interests include traffic data analysis, traffic accident risk modeling, and developing neural network architecture.

Jumpei Tatebe is a project manager at GRID Inc. He received M. Eng in 2016 from the University of Nagoya. His research interests include traffic data analysis, prediction of future traffic conditions, and application of AI to motorway traffic operations.

Jian Xing is a senior researcher at Nippon Expressway Research Institute Co., Ltd. He received B.Eng. from the University of Tsinghua in 1987 and a D.Eng. in 1992 from the University of Tokyo. His research interests include the big data analysis, application, and evaluation of AI and ITS to motorway traffic operations and management.