WILEY | Hindawi

Research Article

Factors Identification and Prediction for Mind Wandering Driving Using Machine Learning

Ciyun Lin (iD),[1] **Hongli Zhang** (iD),[1] **Bowen Gong** (iD),[1] **and Dayong Wu** (iD)[2]

[1]*Department of Traffic Information and Control Engineering, Jilin University, Changchun 130022, China*
[2]*Texas A&M Transportation Institute, Texas A&M University, College Station, TX 77843, USA*

Correspondence should be addressed to Bowen Gong; gongbowen@jlu.edu.cn

Traffic safety is affected by many complex factors. Mind wandering (MW) is a fatal cause affecting driving safety and is hard to be detected and prevented due to its uncertain and complex occurrence mechanism. The aim of this study was to propose a framework for analyzing and predicting MW based on readily available driving status data. The data used in this study are the single-trip information collected by the questionnaire, which includes drivers' personal characteristics, contextual information in which MW occurs, and in-vehicle environmental factors. After investigating the extent of factors that influence MW, these chosen factors are used to forecast MW. Based on these results, we select factors reliable to be obtained in real life to forecast MW. To verify that the new factors explored are useful in improving the forecast accuracy, the compared analysis is conducted with the results found by our approach and the existing approaches. We compare results obtained by four machine-learning-enabled forecasting approaches on a real-life data set. The result shows that the factors found in this paper can significantly improve forecast accuracy. The confusion matrix, ROC curves, and AUC are conducted, and the performance of the gradient boosting decision tree algorithm is better than other forecast approaches. The importance rankings of most factors obtained by the Gradient Boosting Decision Tree and questionnaire are the same.

1. Introduction

The U.S. National Highway Traffic Safety Administration reports that distracted driving caused 3,142 deaths in 2019 and expects a 10.59% increase per year for fatal traffic accidents [1]. Using a smartphone, tuning a radio, or chatting with passengers, and many other behaviors unrelated to the driving tasks can impact driving performance [2–4]. In addition, there is a kind of distracted driving behavior wherein the driver is thinking unrelated to the driving task at hand without any attention [5]. It is called mind-wandering (MW) in driving, and MW is a situation in which executive control shifts away from a primary task to the processing of personal goals [6]. The episode of MW is linked to risky driving behavior, which would be a potential safety hazard [7].

MW occurs with high frequency and has many negative impacts on driving safety. MW frequently occurs during long-distance driving and under optimal driving conditions and is not easily detectable by the driver [8]. Normally, MWs could occur five times in 15 min commuting trips [9]. When it occurs, the driver would be less attentive to the primary task of driving [10]. MW reduces drivers' cognitive and visual performance and results in their low-perceived-risk status. However, drivers' average driving speed during MW would be higher than that under normal driving. Thus, drivers may not react quickly to dangerous emergency situations [11]. Moreover, the driver should also monitor the current driving scene when operating autonomous vehicles. MW would also be a noticeable safety hazard for automated long-distance trips due to its occurrence mechanism [12, 13]. Thus, the effective analysis and forecast of MW are significant for driving safety.

To respond to the impact of MW on driving behavior, existing research has explored the differences in self-reported mind-wandering according to drivers' demographic characteristics and the context in which MW occurs [7, 14–16]. Some scholars have identified real-time MW

based on eye movements, steering wheel reversal rates, and other driving performance data collected by the driving simulator [17]. These proposed methods have high accuracy in detecting MW. However, these parameters may not be easily available in daily driving. To our best knowledge, there has been no existing research to predict MW based on readily available data without a driving simulator.

Thus, this paper aims to propose a framework for analyzing and predicting MW based on readily available driving status data. The impact of readily available factors on MW should be explored. Based on such factors and their specific impacts on MW, a framework to predict this dangerous driving behavior should be established. To ensure the validity of the forecasting framework, it is also necessary to verify the forecasting results using drivers' real-life MW. To achieve these objectives, we analyze the relationships between factors and MW, and propose a forecasting framework based on our analysis results. First, in this paper, the questionnaire method is used to collect data of MW episodes when driving. The collected data contain drivers' personal characteristics (age, gender, driving experience, and educational background), context information when MW occurs (time, distance, trip purpose, etc.), and the number of MW for the reporting trip. Second, Chi squares of independence are performed on the number of MW occurrences variable, and the extrinsic and intrinsic factors triggering the MW, and the extent of factors that influence MW are unveiled. Third, we set evaluations of predicting MW using the factors proposed in this paper. Then, we compare the forecasting results of the four approaches based on machine-learning algorithms with a real-life data set, and provide the relative importance of the factors.

This paper has three main contributions that distinguish our research on factor identification and forecast for MW from the existing studies. First, we identify the extrinsic and intrinsic factors for MW and indicate the extent of these factors. These results could provide suggestions for developing MW prevention strategies. Second, to our best knowledge, this study is one of the first studies to apply machine learning to predict MW based on readily available data. Third, we compare four forecasting approaches based on machine-learning algorithms for the occurrence of MW to cope with the randomness and uncertainty of MW. The performance advantages of such algorithms are also analyzed, which could provide suggestions for real-life applications.

The rest of this paper is organized as follows: in Section 2, we review existing studies on the analysis and forecast of MW. The data collection and analysis of MW are presented in Section 3. Section 4 describes four machine learning algorithms for predicting MW and evaluating the performance and effect of MW forecast. This paper is discussed in Section 5 with our contributions and perspectives on future work. Finally, Section 6 concludes the paper.

2. Literature Review

To distinguish our research from the existing studies, previous research on the identification of different factors' impact on MW and the detection of MW based on driving behavior are reviewed. We also review the literature for the applications of machine learning in driver behavior analysis and forecast.

To respond to the uncertainty and unconscious behavior of MW, some scholars focused on the identification of scenarios that MW is prone to occur using simulated-based methods. Studies showed that the more times the participants drive, the less attention their play on driving [8, 18]; the participants are more familiar with the scene within the the driving simulator, and more frequency of MW is occurs [19]; the less visual complexity of the oncoming traffic scene that the participants meeting, the less attentive their play on and the more frequency of MW is occuring [20]. It is also noticeable that when drivers pay attention to MW, they become more aware than ever of the occurrence of MW as an unsafe behavior [21]. Although simulation-based studies can obtain comprehensive data of driving behavior characteristics, some identified factors related to MW researched by scholars are even hard to get and quantify in real life. Besides, the data may not reflect the reality as the driver in the exprimental situation always feels nervous or feels like playing a game [19].

The previous research used questionnaires to ask drivers to review their past experiences of MW and analyzed the interactive factors related to the MW. Male drivers and young drivers are reported having aggressive driving behavior more frequently, and when their are tired, they are reported more to occur MW [14, 22]. As for environmental factors, traffic environments like monotonous motorways and by-passes, also under normal weather conditions, can highly impact MW. Besides, MW is common in daily life which is closely correlated with the situation of drivers such as stress, alone in car, and driving time [7, 14, 23]. These researches revealed factors linked to MW from drivers' perspective and environmental factors. However, there are still some relationships between extrinsic intrinsic factors and MW that they did not focus on, such as traffic flow conditions and in-vehicle environments. These factors need to be further explored.

Highly distracting MW thoughts are related to a higher risk of being involved in an accident [5]. As for the characteristics of off-task thoughts while driving, most off-task thoughts reported by drivers were present- and future-oriented, of neutral emotional valence, principally for the solution of personal and professional problems, and occurring without any contextual cue [7]. Over half of drivers' MW reports were related to things they saw or heard [9]. Although inner thoughts during MW were investigated, the factors inherent in these thoughts' differences were not discussed further.

To our best knowledge, while some scholars present methods for real-time MW detection, only a few studies focus on predicting MW. Most scholars used driving performance information and eye movement information collected by the driver-in-loop simulator to detect the drivers' cognitive status for MW detection. After feature extraction, important features of MW were set as classification features, and the support vector machine classifier was trained and cross-validated within subjects [24]. Osman

et al. [25] presented a bi-level hierarchical classification methodology with the Decision Tree algorithm to distinct types of secondary tasks. The inputs to the model were five driving behavior parameters and their standard deviations. Besides, research showed steering wheel reversal rate could also be an effective identification factor of MW [17]. Some of these methods had high accuracy and instantaneity in MW detection and provided knowledge on risky driving behavior interruption. However, these parameters were collected in the simulator study, which is not available readily in daily driving. Further research is needed to predict MW during actual driving and use data easily available, to our best knowledge.

Machine learning has recently received attention due to the emergence of big data generated by multiple sources and the availability of computational power [26]. It performs well in solving several complex and nonlinear problems [27, 28] and is widely applied in the prediction of drivers' behavior and some traffic-safety-related researches [29]. Chung et al. [30] proposed an ensemble machine-learning-based algorithm for electric vehicle user behavior prediction, including stay duration and energy consumption based on historical charging records. Deng et al. [31] predicted the performance of lane-changing behaviors combined with environmental and eye-tracking data based on four different machine learning approaches. Saeed et al. [32] carried out an empirical assessment of uncorrelated and correlated random parameter count models on multilane highways considering two crash severities for predicting road crash frequencies. These researches can provide a reference for constructing a framework for MW predictive models that analyze the relationship between forecast objectives and related variables, and make these related variables as inputs for the forecast of MW based on different machine-learning algorithms. While various studies have been conducted on the applications of machine learning, there has been no existing research applying such methods to predict MW.

3. Data Source and Factors Analysis

3.1. The Questionnaire. The questionnaire was published on an open-source questionnaire platform (https://www.wenjuan.com), and people could fill it out online by clicking on the link on the website. The questionnaire was distributed and propagated on social media like WeChat groups, QQ groups, and other personal media platforms such as WeChat Moments and microblogging to enlarge the occupation and age structure cover scale. Participants were likely a mix of employed and unemployed people, students, and retired people from cities, smaller towns, and rural areas [23]. The snapshot of the questionnaire is shown in Figure 1. Finally, we collected 201 questionnaires in about two months and distributed and propagated them. The mean response time for the questionnaire was 6.3 minutes (SD = 5.02), and the questionnaire was collected between December 10, 2019, and February 24, 2020.

To verify the questionnaires' availability, we just set the time threshold to judge whether the questionnaire is acceptable. The threshold is set as 2 min according to the

Figure 1: The snapshot of the questionnaire.

question number, reading speed, and comprehension based on the experiences of contactable participants. If the time for filling the questionnaire is less than 2 min, we considered that the participants did not take the questionnaire seriously and classified it as an invalid questionnaire. We found that only 11 questionnaires are invalid in this survey.

The detailed content of the questionnaire in http://wjw.com (Shanghai Zhongyan Network Technology Co., Ltd, Shanghai, China) can be viewed at https://www.wenjuan.com/s/UnauInE. The questionnaire contained three sections with 25 questions and could be answered based on previous driving experiences. The first part of the questionnaire collected personal information about the driver's age, gender, educational background, and years of driving experience. Besides, to investigate the influence of MW on driving safety, traffic accidents caused by MW of the participants and their opinion about the intervention with MW were asked. The second part of the questionnaire asked the participants to report specific information about the trip in which MW occurred. The report content includes the MW occurrence frequency, time, duration, distance, trip purpose, traffic flow condition, road pavement condition, location, weather, and mood of the trip. The third part of the questionnaire was added to explore the influence of in-vehicle environments

such as temperature, multimedia, and passenger on the occurrence of MW. Besides, the content of the mind wanding was also concerned and added in the questionnaire.

3.2. The Participant. Only automobile drivers could participate in the survey for this study, and this statement was put out before asking volunteers to fill the questionnaire. The questionnaire requested that the participants are required to have a driving license and have driving experience. In the questionnaire, we stated the objective of data collection and promised to keep the personal information confidential.

The participants included 129 males and 61 females. They were aged between 19 and 60 years based on valid questionnaires. The participants' minimum, average, and maximum driving years were 1, 5.65, and 20, respectively. 78.9% of the participants had been involved in risk traffic behavior due to MW. Although missing road signs and markings, panic braking occurrence most, there are still 21.05% of the drivers who reported they had occurred traffic accident due to MW. Table 1 summarizes the participants' information in detail.

3.3. Data Analysis

3.3.1. Forecast Variables. A significance analysis between the potential influence factors and MW was performed. The potential influence factors were divided into two groups: personal characteristics and contextual information in which MW occurs. The personal characteristics information includes gender, age, years of driving experience, and educational background. The contextual information covers the time, duration, distance of driving, trip purpose, traffic flow condition, road pavement condition, location, weather, and mood along with the driving trip. The results of the significance analysis with the chi-square test are shown in Table 2.

It can be seen from Table 2 that there is no significant evidence to prove gender, time, trip purpose, weather, and road pavement condition being conspicuously related to the occurrence and frequency of MW. For factors related to MW, driving experience, educational background, age, duration, mood, distance, traffic flow condition, and location have a relatively significant effect on the occurrence of MW. According to the questionnaire, drivers with five to ten years of driving experience, with higher education, under 40 years old are prone to MW. Within the driving influence factors, driving with long duration and distance, driving in emotionally unstable state (excited or frustrated), driving in free flow condition, and driving in the city are prone to MW.

Age is a drivers' attribute that has the most significant impact on MW. Thus, it is necessary to clarify the relationship between age and MW. To explore how drivers' age influences participants' frequency of MW, we divided the participants into three groups based on the frequency of MW in a single trip: less than one, between two and three,

and more than three. They were defined as low frequency, middle frequency, and high frequency of MW, respectively. The distribution of drivers' age in different MW frequencies is shown in Table 3.

As shown in Table 3, the age distribution in the low frequency and middle frequency of MW has consistency. Within low frequency and middle frequency groups, the drivers under 30 years old make up the largest proportion (44.07%). Then, the middle-aged drivers aged between 40 and 50 years make up the second largest proportion (34.75%). On the contrary, in the high frequency of MW, the participants aged between 30 and 40 years make up the major proportion, about 38.89%.

Furthermore, different drivers have variations in MW occurrence frequency. Figure 2 shows the MW occurrence frequency for drivers. Almost all drivers seldom report high-frequency MW, excluding drivers aged 30–40 years. 20% of drivers aged 30–40 have a high frequency of MW, which is significantly higher than drivers of other ages. Older drivers also tend to have high-frequency MW and injuries [33]. There is research showing that drivers about 35 years old are likely to be distracted [34], and teens reported more distracting behavior while driving compared to their parents [35], which is coincident with our findings.

The driving experience has shown a relatively significant relationship with MW. Table 4 shows the distribution of drivers' driving experience in different MW frequencies. It can be concluded that drivers with less than 5 years of driving experience are the main components for low frequency and middle frequency. However, for the high-frequency group, drivers with 5–10 years of driving experience constitute the major part.

Figure 3 shows the occurrence of MW among drivers with different driving experiences. The group of drivers with 5–10 years of driving experience has the highest rate of high-frequency MW. It can also be concluded that drivers with many years of driving experience seldom have MW.

The previous study has not researched the relationship between educational background and MW. To find how the educational background influences the frequency of MW, we analyzed the distribution of drivers' educational backgrounds in different MW frequencies in a single trip, as shown in Table 5.

As shown in Table 5, drivers with bachelor's and specialist degrees majorly constitute low-frequency and middle-frequency groups. For the high-frequency group, participants with a masters' degree or Ph.D. degree made up the major component (56%). Previous research showed that people with high working memory capacity are more prone to MW than those with low working memory capacity [36]. This study, to some extent, validates the findings above.

Figure 4 shows the occurrence of MW among drivers of different educational backgrounds. We can find drivers with lower education are less likely to experience a high frequency of MW. There is a noticeable phenomenon that the well-educated ones are most prone to have high-frequency MW (19.61%).

TABLE 1: The participants' attribute distribution.

Factors	Variate	Percentage (%)
Gender	Male	67.9
	Female	32.1
Age	20–30	43.7
	30–40	18.4
	40–50	34.2
	More than 50	3.7
Education	Primary school	1.1
	Junior secondary school	14.2
	Senior secondary school	12.1
	Bachelor and specialist degree	45.8
	Master's degree or above	27.9
Traffic accident experience	Yes	21.1
	No	78.9
Driving years	0–5	60.0
	5–10	25.8
	10–15	7.9
	More than 15	6.3
Daily driving times	0–1	34.2
	1–2	24.2
	2–3	23.2
	3–4	10.0
	More than 4	8.4

TABLE 2: The significance of the potential influence factors on MW.

	Factors	χ^2	P
Personal attribute information	Age	14.505	0.070
	Driving experience	2.520	0.112
	Educational background	27.410	0.124
	Gender	1.630	0.803
Driving situation information	Duration	40.441	0.019
	Mood	12.447	0.132
	Distance	20.253	0.209
	Traffic flow condition	5.716	0.221
	Location	4.704	0.319
	Time	11.040	0.525
	Trip purpose	4.410	0.818
	Weather	12.745	0.888
	Road pavement condition	6.486	0.890

FIGURE 2: MW occurrence in drivers of different ages.

TABLE 3: The distribution of drivers' age in different MW frequencies.

	Under 30	30–40	40–50	50 or older	Total
Low frequency	52	22	41	3	118
Middle frequency	24	6	21	3	54
High frequency	7	7	3	1	18

TABLE 4: The distribution of drivers' driving experience in different MW frequencies.

	0–5	5–10	10–15	15 and above	Total
Low frequency	76	26	8	8	118
Middle frequency	30	14	6	4	54
High frequency	8	9	1	0	18

3.3.2. In-Vehicle Influence Factors. The in-vehicle environments also affect driving behavior. However, how the in-vehicle environments impact the driving behavior is still unclear, as it is difficult to collect the field-based in-vehicle information in previous works [7]. The research found that in-vehicle environments are highly related to traffic crashes [37]. Therefore, this section tries to find the potential relationship between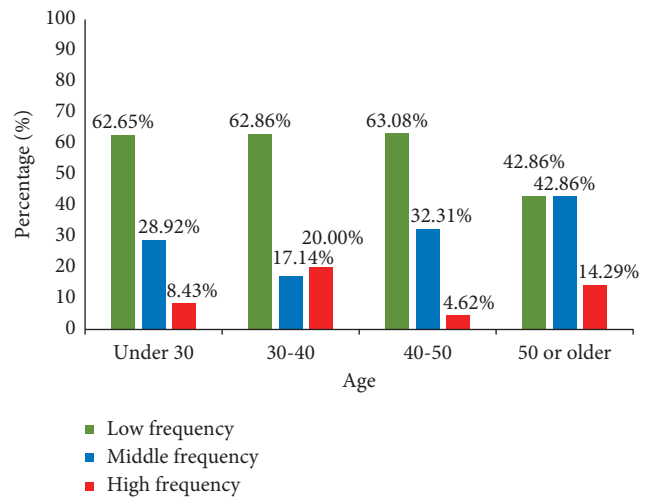 the in-vehicle environments and MW. In the questionnaire, the possible influence factors: passenger, temperature, and multimedia were surveyed. Table 6 shows the survey results of three potential influence factors related to MW by listing the percentage of drivers' choice of the factors most prone to MW.
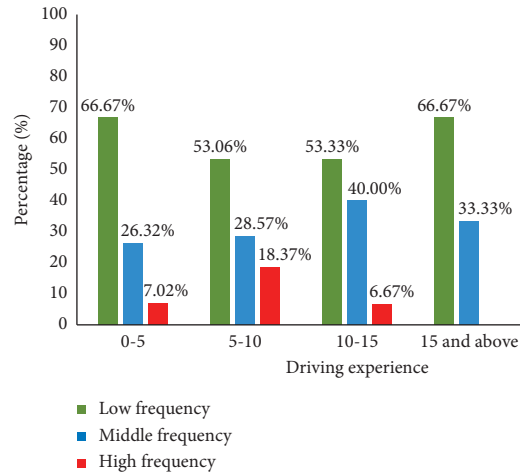
FIGURE 3: MW occurrence in drivers of different driving experiences.

TABLE 5: The distribution of drivers' educational backgrounds in different MW frequencies.

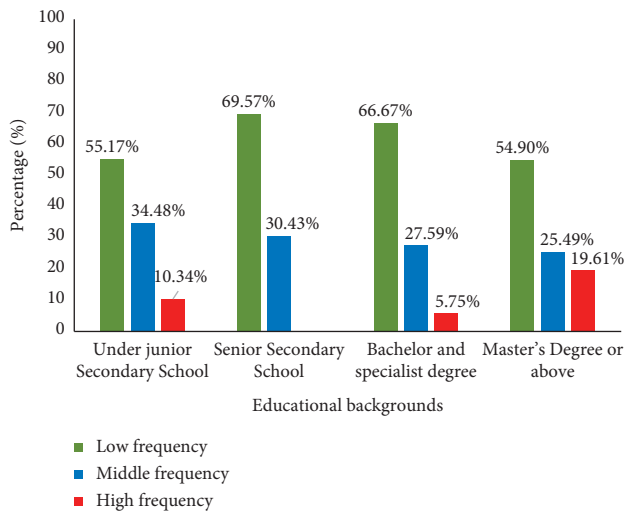|	Under junior secondary school	Senior secondary school	Bachelor and specialist degree	Master's degree or above	Total
Low frequency	16	16	58	28	118
Middle frequency	10	7	24	13	54
High frequency	3	0	5	10	18

FIGURE 4: MW occurrence in drivers of different educational backgrounds.

TABLE 6: In-vehicle environments related to MW.

(a) Drivers' choice about the situation of passenger presence most prone MW

The situation of passenger presence	Percentage (%)
Driving alone	39.47
Driving with passengers and chatting	53.16
Driving with passengers but not chatting	7.37

(b) Drivers' choice about the temperature in-vehicle most prone MW

Temperature	Percentage (%)
Warm temperature with air conditioner	53.68
Temperature with windows closed	33.68
Temperature of open windows	8.95
Cold temperature with air conditioner	3.68

(c) Drivers' choice about the multimedia in-vehicle most prone MW

Multimedia	Percentage (%)
None	46.32
Playing music	22.63
Playing radio	8.95
Other video & audio	22.11

It can be seen from Table 6 that the drivers in the situation of chatting with passengers, warm temperature, and no playing multimedia in the vehicle are prone to MW ($P < 0.001$). Furthermore, different gender of drivers has significantly different choices in situations that are most prone to MW ($P < 0.05$). Male drivers are more prone to MW when they drive alone. According to our investigation, driving with passengers and chatting is highly related to MW, especially for female drivers. 47.29% of the male drivers and 65.57% of the female drivers think driving with passengers and chatting makes one most vulnerable to MW. 42.64% of the male drivers and 32.79% of the female drivers think driving alone can lead to MW.

3.3.3. Off-Task Thoughts. It is interesting to know drivers' thoughts on MW, and what kinds of thoughts make them susceptible to MW. For this purpose, in the questionnaire,

we added the following question, "What kind of thoughts are you most likely to have when MW occurs?" Before separating the questionnaire, we classified the off-task thoughts into 7 classes. When participants answer the questionnaire, they need to choose the thought that appeared during the trip. The frequency of daily routines caused MW is shown in Figure 5.

It can be seen from Figure 5 that 54 drivers (about 28.4%) confirmed that they are most likely to think about work-related routines when they have MW; "personal emotional problems" and "family-related affairs" account for 23.2% and 18.9%, respectively. Therefore, it can be deduced that "work-related problems," "personal emotional problems," and "family-related affairs" are the main intrinsic factors of MW induction.

To explore the character of daily routine-related wandering further, significance analysis between drivers' main attributes and daily-routine-related wandering is conducted using the chi-square test. The results are shown in Table 7. Only age has a significant relationship with daily-routine-related wandering.

The data also show that daily-routine-related wanderings during MW episodes have a statistically significant difference for drivers of diverse ages, with $\chi^2 = 18$, $p = 0.003$. The questionnaire results about the daily-routine-related wandering of different ages of drivers are shown in Table 8.

It can be seen from Table 8 that drivers under 30 years old are prone to fall into personal-related issues when experiencing MW. For drivers between 30 and 40 years old, about 40% of them report they are easily caught up in work-related problems. The main wandering issue turns to be family-related affairs for drivers more than 40 years old. It can be concluded that the main daily-routine-related wandering is different for various ages of drivers. Adolescent drivers focus on personal emotional problems. With increasing age, the drivers' issues will be adjusted to undertaking and family. Furthermore, when the issues focus on turn out in life, it will be easier to be caught in MW.

In summary, we explored new factors related to MW. We also analyzed the internal reasons for the differences in the thoughts of MW as these reasons have not been studied. Previous studies have focused on how drivers' gender, age, years of driving experience, and characteristics of driving tasks influence MW. In this section, based on the existing conclusions, drivers' educational background, trip purpose, traffic flow condition, location, mood, and their relationships with MW were taken into consideration. These factors are easy to be collected and can be used to predict MW occurrence. Besides, whether passenger presence, in-vehicle temperature, and in-vehicle multimedia lead to MW was also investigated.

4. Forecasting Approaches Based on Machine Learning

Effective forecasting of MW can improve driving safety greatly; it is also an urgent need for the driver currently. According to our survey, 93.16% of the drivers report that they are willing to follow an approach to prevent MW. This study predicts the occurrence of MW based on readily available driving status data that no researcher has done, to our best knowledge. The factors associated with MW are identified in the previous section. First, we choose the early selected factors as input variables for forecasting MW, as the other impact factors found by our approach could not be reliable to be obtained by the driving simulator. Second, since there is no crucial need to predict the exact value of the number of occurrences of MW in terms of early warning of actual unsafe driving behaviors, we have further processed the data. Here, trips with less than two single-trip MWs were defined as low-risk trips, and the remaining were defined as high-risk trips; then, the forecast of MW is transformed into a classification process [38, 39]. Third, we compare four forecasting approaches based on machine-learning algorithms for the occurrence of MW with a real-life data set and provide the relative importance of the factors.

4.1. Forecasting Method

4.1.1. Random Forest Classifier. Random forest is an ensemble learning methodology. Liking other ensemble learning techniques, its performance is boosted via a voting scheme [40]. The method combines Brieman's bagging idea and Ho's "random subspace method" to construct a collection of decision trees with controlled variations [41]. The main idea of the random forest is to grow a large population of unpruned decision trees to bootstrap a sample of training data by randomly selecting features at each segmentation, and then choosing the best split among these features [42]. Compared with other algorithms, random forest training needs less time, and it can handle high-dimensional data and does not have to make feature selection [43]. After training, it can inform which features are more important.

Based on these advantages, the random forest algorithm was chosen in this study to predict the MW in the trip. Those variables significantly associated with MW were chosen as input, then it was ascertained whether the current trip belonged to the high-risk trip or low-risk trip. Besides, the importance ranking of these features could also be given.

4.1.2. Gradient Boosting Decision Tree. Gradient Boosting Decision Tree (GBDT) is a supervised learning algorithm. GBDT repeats the selection of an ordinary model and adjusts it based on the previous models' performance. It uses gradient boosting to improve model performance [44]. GBDT combines regression trees with a gradient boosting technique and has been widely applied in various disciplines, such as credit risk assessment, transport crash prediction, and fault prognosis in electronic circuits [45]. The method has high prediction accuracy and can flexibly handle various data types, including continuous and discrete values.

As the factors associated with MW collected in this experiment were discrete and continuous variables, the GBDT method was also chosen to predict the number of single-trip MW occurrences. Similar to the random forest approach, variables significantly associated with MW were used as input variables for the forecast, and the classification of the trip belonging was outputted.
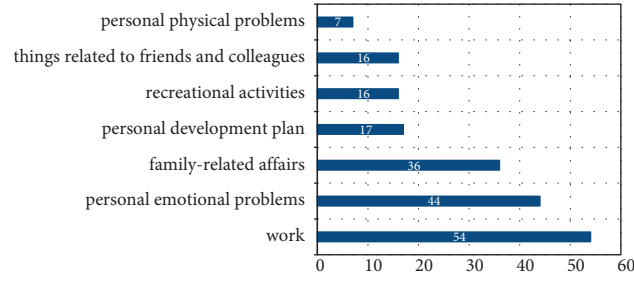
FIGURE 5: The frequency of daily routines caused MW.

TABLE 7: The distribution of daily routine-related wandering according to the age of drivers.

	χ^2	P
Age	39.314	0.003
Gender	3.726	0.714
Education	23.658	0.167
Driving years	10.551	0.103

TABLE 8: The distribution of daily routine-related wandering according to the age of drivers.

	Under 30	30–40	40 or older	Total
Work	21	14	19	54
Personal emotional problems	25	7	12	44
Family-related affairs	7	4	25	36
Personal development plan	8	4	5	17
Recreational activities	12	1	3	16
Things related to friends and colleagues	6	5	5	16
Personal physical problems	4	0	3	7

4.1.3. Naïve Bayes. A Naïve Bayes classifier estimates the posterior probability of Y given X, $P(Y|X)$, there are m samples, each sample has n features, and the feature output has K categories defined as C_1, C_2, ..., C_k. Applying Bayes' theorem and the assumption of conditional independence, the posterior probability can be represented as [46]:

$$P(Y|X_1,\ldots,X_n) = \frac{P(Y)P(X_1,\ldots,X_n|Y)}{P(X_1,\ldots,X_n)} \propto P(Y)\prod_{i=1}^{n} P(X_i|Y). \tag{1}$$

The posterior probability maximization is needed to determine the classification, calculate all K conditional probabilities $P(Y = C_k|X = X^{(\text{test})})$, and then find the category corresponding to the maximum conditional probability, which is the plain Bayesian prediction. Using the independence assumption of the plain Bayesian, one can obtain the plain Bayesian inference formula in the usual sense:

$$C_{\text{result}} = \underset{C_k}{\arg\max} \; P(Y = C_k)\prod_{j=1}^{n} P\left(X_j = X_j^{(\text{test})}|Y = C_k\right). \tag{2}$$

Although the Naïve Bayes classifier is based on the "naïve" conditional independent assumption, compared to other classification algorithms, it still demonstrates preferable performance in analyzing many real data sets that do not strictly follow the conditional independent assumption [47]. Therefore, as a classical classification method, it was chosen here to predict the occurrence of MW. The event code with the largest estimated probability was then chosen as the forecast for the MW.

4.1.4. Multiple Linear Regression. Multiple linear regression can deal with the relationship between a dependent variable and an independent variable or multiple independent variables. Regression analysis generally includes the following steps: first, determine the independent and dependent variables in the data; second, calculate the coefficients of the multiple linear regression model and standardize the regression coefficients; third, use the determined multiple linear regression model to make predictions. The general representation of the model is as follows:

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon, \tag{3}$$

$\beta_0, \beta_1, \ldots, \beta_p$ are fixed and unknown, where the random variables X_j, $j = 1, \ldots, p$, denote p regressor variables. The random variable ε denotes an error term. It is uncorrelated with the regressors and has expectation 0 and variance $\sigma^2 > 0$.

Although multiple linear regression is weaker than the previous machine-learning methods for classification problems, it has a better explanation and can quantitatively express the relationship between the probability of occurrence of MW and each factor, so it was chosen to predict the occurrence of MW. A linear equation was fitted to predict MW using historical data on the occurrence of MW from the questionnaire.

4.2. Evaluation. To verify whether the new factors we explored are useful in improving MW's forecasting accuracy, we conducted three evaluation procedures. The input of compared evaluations are factors associated with MW found by our approach, associated factors found by existing approaches, and factors that only have a significant relationship with MW, separately.

The data used for training were the MW driving obtained from a questionnaire and driver demographics and relevant factors for that trip. 70% of the data were randomly chosen for training models, and the remaining 30% were used for evaluating performance. To score the algorithms' performance, we provided the accuracy and confusion matrix of the algorithms' forecast on the test data. Moreover, ROC curves, the Area under a ROC Curve (AUC), and the ranking of risk factors were shown to analyze the results.

4.2.1. Comparative Analysis. The first evaluation procedure was to select factors found in previous studies related to MW driving (driving experience, gender, age, time, duration, distance, and road pavement condition) as input variables for the forecast. The second evaluation procedure was to select the factors we investigated (age, gender, driving experience, educational background, duration, mood, distance, traffic flow conditions, location, time, trip purpose, weather, and road pavement condition) as input to make the forecast. The third evaluation forecasts MW using factors that have a relatively significant relationship with MW (age, driving experience, duration, and mood).

For these three evaluations, the comparative results of random forest, Gradient Boosting Decision Tree, Naïve Bayes, and multiple linear regression are shown in the Table 9. We conducted forecasting 10 times for these experiments and calculated the average forecasting accuracy.

From Table 9, it can be seen that the variables we aided have a significant effect on the improvement of forecasting accuracy. Although some of these factors did not show a statistically significant relation with MW driving, they were convenient to collect and could improve forecasting accuracy. Therefore, we analyzed the results of the MW forecast with the factors found by our approach in Section 3.

By comparing the results of four machine-learning algorithms, the accuracy of the GBDT algorithm is 73.69%. The random forest algorithm, Naïve Bayes algorithm, and Multiple Linear regression algorithm have an accuracy of 70.28%, 68.42%, and 65.42%, respectively. It can be observed that the GBDT algorithm has better performance in the forecast of MW. It is a meaningful result to achieve

TABLE 9: The accuracy of MW forecast.

Algorithms	Accuracy		
	1 (%)	2 (%)	3 (%)
Random forest	49.12	70.28	68.18
GBDT	59.65	73.69	70.40
Naïve Bayes	57.89	68.42	67.25
Multiple linear regression	60.32	65.42	65.80

acceptable forecasting results only with the driver's personal information and context information when experiencing MW.

Furthermore, we analyze the forecasting results of these four algorithms in detail. The forecast confusion matrix [48] of the algorithms is shown in Figure 6.

The results show that the forecast value of the random forest algorithm, GBDT algorithm, and Naïve Bayes algorithm is lower than the actual value of MW. The Naïve Bayes elicits the most inferior performance. In contrast, the forecast value of the linear regression algorithm is higher than the real value of MW.

4.2.2. Receiver Operating Characteristic. ROC is widely used for evaluating the performance of binary classifiers. The ROC curve shows how the number of correctly classified positive cases varies with the number of incorrectly classified negative cases [46]. It consists of the x-axis of false positive (FP) rate and the y-axis of true positive (TP) rate. FP defines the number of the estimated instances incorrectly classified as positive when they were negative. TP rate is used to measure the proportion of instances correctly predicted as positive in all actual positive instances. Random forest algorithm, GBDT algorithm, and Naïve Bayes algorithm were constructed from the training data set, and the ROC performance was evaluated using the test data set. ROC curves are shown in Figure 7.

If a classifier is perfect for predicting all cases correctly, the ROC curve of the classifier will be in the upper-left corner. As our sample size is not very large, the ROC curves are not smooth enough. To show the differences between these curves, we smoothed the curve with a polynomial fit. It can be seen from Figure 7 that the ROC curves of the GBDT algorithm are higher than the random forest algorithm and the GNB algorithm. It can be indicated that the GBDT algorithm performs better in the MW forecast. The MW forecasting results using the random forest algorithm and the GNB algorithm show no significant difference.

The Area under ROC Curve (AUC) is defined to evaluate the overall performance of a classifier quantitatively. The AUC is the area enclosed by the ROC curve, the horizontal axis, and the right boundary of the ROC space, with a maximum value of 1 indicating a perfect classification result. A value of 0.5 indicates that the classifier produced a random classification result [47]. As shown in Figure 7, the GBDT algorithm achieves the best performance in the MW forecast, with an AUC of 0.71. The random forest algorithm and the GNB algorithm both achieve performance with an AUC of 0.62.
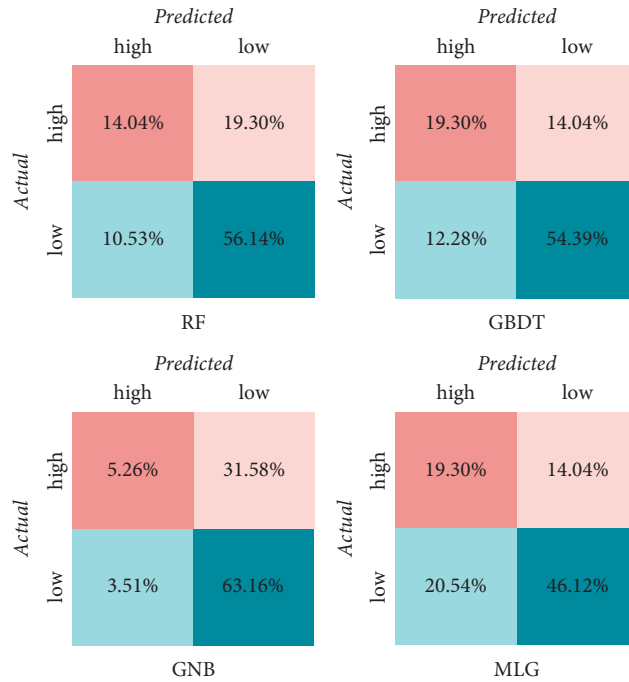
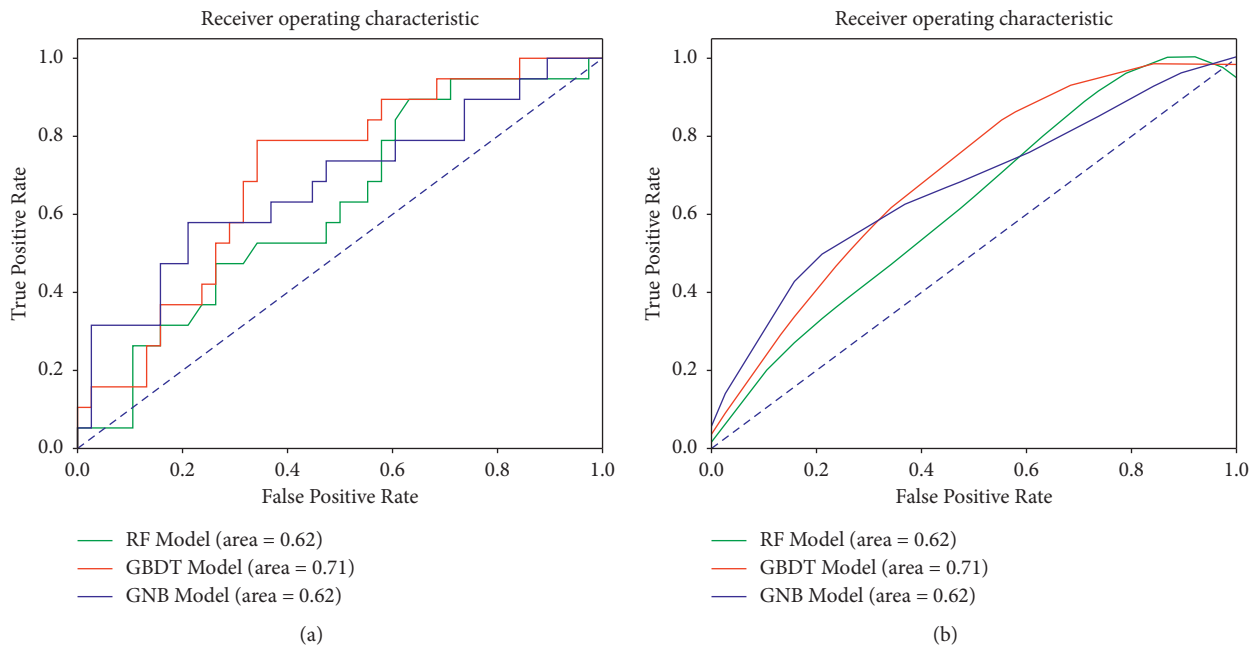Figure 6: The forecast confusion matrix of algorithms.

Figure 7: The ROC curve of different algorithms. (a) The original ROC curve of the algorithms; (b) The fitting ROC curve of the algorithms.

In summary, it can be proved that the GBDT algorithm has the advantages for the MW forecast with higher forecasting accuracy and produces an acceptable performance based on the comparison. Thus, the GBDT algorithm has the potential to be applied in real-life MW forecasts.

4.3. Ranking of Risk Factors. To evaluate the contribution of different factors to the forecast of the frequency of MW, we extract the variables' weight distribution. Variables' rank is compared within the GBDT algorithm and data analysis. The compared results are shown in Table 10 and Figure 8.

The variables' weight distribution ranking is explored in Table 10. In Table 10, the age, duration, and driving experience rank in the top three for the GBDT algorithm and data analysis. It can be deduced that these three factors can represent highly influential factors of MW driving. Besides, educational background, mood, distance, time, location, and

TABLE 10: Variables' weight distribution ranking.

| Factors	GBDT		Data analysis	
	Rank	Importance	Rank	P
Age	1	0.233	2	0.070
Duration	3	0.109	1	0.019
Driving experience	2	0.170	3	0.112
Educational background	4	0.107	4	0.124
Mood	5	0.077	5	0.132
Distance	6	0.075	6	0.209
Time	8	0.049	9	0.525
Location	10	0.030	8	0.319
Gender	9	0.035	10	0.803
Weather	7	0.054	12	0.888
Traffic flow condition	12	0.024	7	0.221
Trip purpose	13	0.029	11	0.818
Road pavement condition	11	0.006	13	0.890
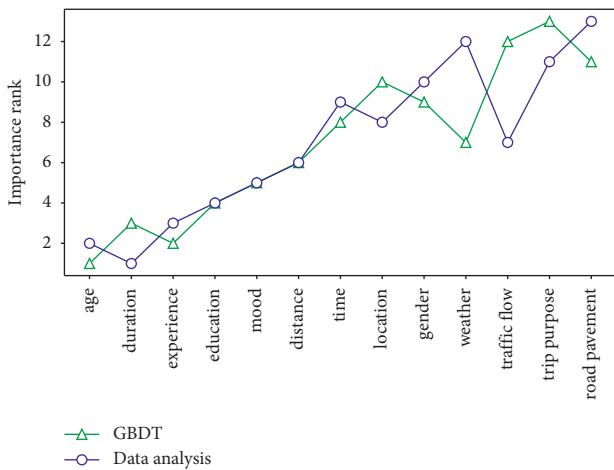
FIGURE 8: Comparison of different factors' weight ranking.

gender have a relatively significant impact on the occurrence of MW. Trip purpose and road pavement conditions do not significantly affect the occurrence of MW. It also can be seen from Figure 8 that there is the same consistent trend in the ranking of influence factors within the two analyzes. The ranking of influence factors also has some individual differences. The importance ranks of the weather and traffic flow conditions are different from these two analyzes. Although these two factors have different important ranks, they hardly impact MW significantly.

5. Discussion and Conclusion

As a high risk of driving behavior, MW can give rise to traffic crashes, a potentially safety hazard. This study identified the extrinsic and intrinsic factors that triggered the MW that had not been explored and indicated the extent of factors' influence on MW. We proposed a framework for predicting MW to cope with the randomness and uncertainty of MW. Then, we compared four forecasting approaches for the occurrence of MW based on machine-learning algorithms. Our research can provide suggestions for developing MW driving prevention strategies and can be applied to in-vehicle

information systems or mapping software to give drivers early warning regarding MW.

MW is widespread, and its occurrence is not completely untraceable, linked to certain external factors. It can be concluded from the data analyses and variables' importance of machine-learning algorithms that duration, age, and driving experience have significant impacts on the occurrence of MW. Educational background, mood, distance, traffic flow conditions, and location also have relatively obvious influences on MW. In contrast, gender, time, trip purpose, weather, and road pavement conditions have inconsiderable implications on MW. For in-vehicle environmental factors, chatting with passengers when driving, comfortable temperature, and playing multimedia in the vehicle are prone to MW driving. With the presence of passengers, the mean proportion of time having elevated gravitational-force events in curves is significantly higher than when there was no passenger [4]. Driving with passengers or people they have met has an impact on driving behavior. As to playing multimedia in the vehicle, however, a previous study showed that radio tuning tasks might seem reasonably distracting as drivers get inattentive in no-task driving [48]. Thus, it is common for drivers that their minds do not always focus on the driving task.

For off-task thoughts, 28.4% of drivers think that they are most likely to wander about work-related routines when they have MW; "personal emotional problems" and "family-related affairs" account for 23.2% and 18.9% respectively. It can also be concluded that the main daily routine-related wandering is different for various age groups. Adolescent drivers focus on personal emotional problems, and with increasing age, the drivers focusing on them will be adjusted to undertaking and family. Furthermore, when the issues focus on turn out in life, it will be easier to be caught in MW driving.

As the factors associated with MW are identified, these related factors are used to predict MW. The Random forest algorithm, Gradient Boosting Decision Tree algorithm, Naive Bayes algorithm, and Multiple Linear regression model are chosen to predict the MW. By comparing the results of the four machine-learning algorithms, the

accuracy of the GBDT algorithm is 73.69%, and the random forest algorithm, the Naïve Bayes algorithm, and the Multiple Linear regression algorithm have an accuracy of 70.28%, 68.42%, and 65.42%, respectively. The GBDT algorithm achieves the best performance in MW forecast, with an AUC of 0.71, while the random forest algorithm and the Naïve Bayes algorithm both achieve performance with an AUC of 0.62.

Among the above factors used to predict MW, some factors are considered for the first time. These factors proved by us include drivers' educational background, trip purpose, traffic flow condition, location, and mood. The results of the comparison experiments illustrate that the addition of these factors can effectively improve the forecasting accuracy of MW.

In the future, it is advised that under the premise of ensuring the authenticity of survey samples, more drivers should be surveyed to enlarge the data set. In addition, when more dimensions of data related to MW are collected, some factors that are more detailed such as the type of vehicle can be considered to improve the forecasting accuracy of MW.

MW is a high-risk driving behavior, and it is hard to be detected. However, its occurrence is not untraceable. This paper reported some factors related to MW (e.g., educational background, and mood). Based on these findings, machine-learning algorithms can be used to forecast the occurrence of MW, and the factors we explored can improve forecasting accuracy. The evaluation shows that the GBDT algorithm performs better in the MW forecast.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] NHTSA, "Distracted driving," 2020, https://www.nhtsa.gov/risky-driving/distracted-driving.

[2] O. Oviedo-Trespalacios, A. P. Afghari, and M. M. Haque, "A hierarchical Bayesian multivariate ordered model of distracted drivers' decision to initiate risk-compensating behaviour," *Analytic Methods in Accident Research*, vol. 26, Article ID 100121, 2020.

[3] D. Stavrinos, J. L. Jones, A. A. Garner et al., "Impact of distracted driving on safety and traffic flow," *Accident Analysis & Prevention*, vol. 61, pp. 63–70, 2013.

[4] F. Zhang, S. Mehrotra, and S. C. Roberts, "Driving distracted with friends: effect of passengers and driver distraction on young drivers' behavior," *Accident Analysis & Prevention*, vol. 132, Article ID 105246, 2019.

[5] C. Galera, L. Orriols, K. M'Bailara et al., "Mind wandering and driving: responsibility case-control study," *BMJ: British Medical Journal*, vol. 345, Article ID 10e8105, 11e8105 pages, 2012.

[6] J. Smallwood and J. W. Schooler, "The restless mind," *Psychological Bulletin*, vol. 132, pp. 946–958, 2006.

[7] G. BerthieBerthié, C. Lemercier, P.-V.P. V. Paubel et al., "The restless mind while driving: drivers' thoughts behind the wheel," *Accident Analysis & Prevention*, vol. 76, pp. 159–65165, 2015.

[8] M. R. Yanko and T. M. Spalek, "Route familiarity breeds inattention: a driving simulator study," *Accident Analysis & Prevention*, vol. 57, pp. 80–686, 2013.

[9] B. R. D. Burdett, S. G. Charlton, and N. J. Starkey, "Inside the commuting driver's wandering mindInside the commuting driver's wandering mind," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 57, pp. 59–74, 2018.

[10] M. H. Martens and R. F. T. Brouwer, "Measuring being lost in thought: an exploratory driving simulator study," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 20, pp. 17–28, 2013.

[11] T. Ito and K. Yamazaki, "A study of a driver model using unconscious driving behaviors," *Procedia Computer Science*, vol. 96, pp. 754–61761, 2016.

[12] D. Schnebelen, C. Charron, and F. Mars, "Estimating the out-of-the-loop phenomenon from visual strategies during highly automated driving," *Accident Analysis & Prevention*, vol. 148, Article ID 105776, 2020.

[13] B. R. D. Burdett, N. J. Starkey, and S. G. Charlton, "The close to home effect in road crashes," *Safety Science*, vol. 98, pp. 1–8, 2017.

[14] W. Qu, Y. Ge, Y. Xiong, R. Carciofo, W. Zhao, and K. Zhang, "The relationship between mind wandering and dangerous driving behavior among Chinese drivers," *Safety Science*, vol. 78, pp. 41–48, 2015.

[15] D. A. Albert, M. C. Ouimet, J. Jarret et al., "Linking mind wandering tendency to risky driving in young male drivers," *Accident Analysis & Prevention*, vol. 111, pp. 125–32132, 2018.

[16] C. Lemercier, C. Pêcher, G. Berthié et al., "Inattention behind the wheel: hHow factual internal thoughts impact attentional control while driving," *Safety Science*, vol. 62, pp. 279–85285, 2014.

[17] G. K. Kountouriotis, P. Spyridakos, O. M. J. Carsten, and N. Merat, "Identifying cognitive distraction using steering wheel reversal rates," *Accident Analysis & Prevention*, vol. 96, pp. 39–45, 2016.

[18] S. G. Charlton and N. J. Starkey, "Driving without awareness: the effects of practice and automaticity on attention and driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 6, pp. 456–71471, 2011.

[19] S. G. Charlton and N. J. Starkey, "Driving on familiar roads: aAutomaticity and inattention blindness," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 19, pp. 121–33133, 2013.

[20] M. Geden, A.-M. Staicu, and J. Feng, "The impacts of perceptual load and driving duration on mind wandering in

driving," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 57, pp. 75–83, 2018.

[21] C. L. Baldwin, D. M. Roberts, D. Barragan, J. D. Lee, N. Lerner, and J. S. Higgins, "Detecting and quantifying mind wandering during simulated driving," *Frontiers in Human Neuroscience*, vol. 11, p. 406, 2017.

[22] H. E. K. Heather Walker and L. M. Trick, "Mind-wandering while driving: the impact of fatigue, task length, and sustained attention abilities," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 59, pp. 81–97, 2018.

[23] B. R. D. Burdett, S. G. Charlton, and N. J. Starkey, "Not all minds wander equally: the influence of traits, states and road environment factors on self-reported mind wandering during everyday driving," *Accident Analysis & Prevention*, vol. 95, pp. 1–7, 2016.

[24] Y.S. E. L. Yuan LiaoLiao, S. E. Li, W. Wang, Y. Wang, G. Li, and B. Cheng, "Detection of driver cognitive distraction: a comparison study of stop-controlled intersection and speed-limited highway," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 6, pp. 1628–371637, 2016.

[25] O. A. Osman, M. Hajij, S. Karbalaieali, and S. Ishak, "A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data," *Accident Analysis & Prevention*, vol. 123, pp. 274–81281, 2019.

[26] C. Y. Lin, D. Y. Wu, H. C. Liu, X. T. Xia, and N. Bhattarai, "Factor identification and prediction for teen driver crash severity using machine learning: a case study," *Applied Sciences-Basel*, vol. 10, no. 5, p. 1675, 2020.

[27] J. Tang, L. Zheng, C. Han et al., "Statistical and machine-learning methods for clearance time prediction of road incidents: a methodology review," *Analytic Methods in Accident Research*, vol. 27, Article ID 100123, 2020.

[28] X. Chen, Z. Li, Y. Yang, L. Qi, and R. Ke, "High-resolution vehicle trajectory extraction and denoising from aerial videos," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 3190–2023202, 2021.

[29] J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, "Crash injury severity analysis using a two-layer stacking framework," *Accident Analysis & Prevention*, vol. 122, pp. 226–38238, 2019.

[30] Y.-W. Chung, B. Khaki, T. Li, C. Chu, and R. Gadh, "Ensemble machine learning-based algorithm for electric vehicle user behavior prediction," *Applied Energy*, vol. 254, Article ID 113732, 2019.

[31] Q. Deng, J. Wang, K. Hillebrand, C. R. Benjamin, and D. Söffker, "Prediction performance of lane changing behaviors: a study of combining environmental and eye-tracking data in a driving simulator," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3561–703570, 2020.

[32] T. U. Saeed, T. Hall, H. Baroud, and M. J. Volovski, "Analyzing road crash frequencies with uncorrelated and correlated random-parameters count models: an empirical assessment of multilane highways," *Analytic Methods in Accident Research*, vol. 23, Article ID 100101, 2019.

[33] B. Donmez and Z. Liu, "Associations of distraction involvement and age with driver injury severities," *Journal of Safety Research*, vol. 52, pp. 23–828, 2015.

[34] C. Lyon, S. Brown, W. Vanlaar, and R. Robertson, "Prevalence and trends of distracted driving in Ccanada," *Journal of Safety Research*, vol. 76, pp. 118–26126, 2021.

[35] C. Raymond Bingham, J. S. Zakrajsek, F. Almani, J. T. Shope, and T. B. Sayer, "Do as i say, not as i do: distracted driving behavior of teens and their parents," *Journal of Safety Research*, vol. 55, pp. 21–929, 2015.

[36] J. Smallwood, K. Brown, B. Baird, and J. W. Schooler, "Cooperation between the default mode network and the frontal–parietal network in the production of an internal train of thought," *Brain Rresearch*, vol. 1428, pp. 60–70, 20112012.

[37] J. K. Caird, S. M. Simmons, K. Wiley, K. A. Johnston, and W. J. Horrey, "Does talking on a cell phone, with a passenger, or dialing affect driving performance? An updated systematic review and meta-analysis of experimental studies," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 60, no. 1, pp. 101–33133, 2018.

[38] B. Wu, J. Zhang, T. L. Yip, and C. Guedes Soares, "A quantitative decision-making model for emergency response to oil spill from ships," *Maritime Policy & Management*, vol. 48, no. 3, pp. 299–315, 2021.

[39] F. Zong, M. Zeng, J. Lv, and C. Wang, "A credit charging scheme incorporating carpool and carbon emissions," *Transportation Research Part D: Transport and Environment*, vol. 94, Article ID 102711, 2021.

[40] M. W. Ahmad, M. Mourshed, and Y. Rezgui, "Trees vs. neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption," *Energy and Buildings*, vol. 147, pp. 77–89, 2017.

[41] R. Harb, X. Yan, E. Radwan, and X. Su, "Exploring precrash maneuvers using classification trees and random forests," *Accident Analysis & Prevention*, vol. 41, no. 1, pp. 98–107, 2009.

[42] Y. Xu, S. Bao, and A. K. Pradhan, "Modeling drivers' reaction when being tailgated: A random forests method," *Journal of Safety Research*, 2021, In press.

[43] F. Zong, M. Zeng, Z. He, and Y. Yuan, "Bus-car mode identification: Traffic condition–based random-forests methodBus-cCar mMode iIdentification: tTraffic cCondition-bBased rRandom-fForests mMethod," *Journal of Transportation Engineering, Part A: Systems*, vol. 146, no. 10, Article ID 04020113, 2020.

[44] Z. Zheng, P. Lu, and B. Lantz, "Commercial truck crash injury severity analysis using gradient boosting data mining model," *Journal of Safety Research*, vol. 65, pp. 115–24124, 2018.

[45] R. Sun, G. Wang, W. Zhang, L.-T. Hsu, and W. Y. Ochieng, "A gradient boosting decision tree based GPS signal reception classification algorithm," *Applied Soft Computing*, vol. 86, Article ID 105942, 2020.

[46] O. H.W. R. Oh Hoon KwonaKwon, W.Yoonjin Yoon.Rhee, and Y. Yoon, "Application of classification algorithms for analysis of road safety risk factor dependencies," *Accident Analysis & Prevention*, vol. 75, pp. 1–15, 2015.

[47] C. Chen, G. Zhang, J. Yang, J. C. Milton, and A. D. AlcantaraAlcántara, "An explanatory analysis of driver injury severity in rear-end crashes using a decision table/Naïve Bayes (DTNB) hybrid classifier," *Accid Anal PrevAccident Analysis & Prevention*, vol. 90, pp. 95–107, 2016.

[48] R. Arvin, A. J. Khattak, and H. Qi, "Safety critical event prediction through unified analysis of driver and vehicle volatilities: application of deep learning methods," *Accident Analysis & Prevention*, vol. 151, Article ID 105949, 2020-2021.