Research Article

Fatigue Driving Prediction on Commercial Dangerous Goods Truck Using Location Data: The Relationship between Fatigue Driving and Driving Environment

Shifeng Niu [ID] and Guiqiang Li [ID]

Automotive Transportation Safety Assurance Technology Key Laboratory of Transportation Industry, Chang'an University, Xi'an, Shaanxi 710064, China

Correspondence should be addressed to Shifeng Niu; nsf530@chd.edu.cn and Guiqiang Li; 1119084992@qq.com

The approaches monitoring fatigue driving are studied because of the fact that traffic accidents caused by fatigue driving often have fatal consequences. This paper proposes a new approach to predict driving fatigue using location data of commercial dangerous goods truck (CDT) and driver's yawn data. The proposed location data are from an existing dataset of a transportation company that was collected from 166 vehicles and drivers in an actual driving environment. Six different categories of the predictor set are considered as fatigue-related indexes including travel time, day of week, road type, continuous driving time, average velocity, and overall mileage. The driver's yawn data are used as a proxy for ground truth for the classification algorithm. From the six different categories of the predictor set, we obtain a set of 17 predictor variables to train logistic regression, neural network, and random forest classifiers. Then, we evaluate the predictive performance of the classifiers based on three indexes: accuracy, F1-measure, and area under the ROC curve (AUROC). The results show that the random forest is more suitable for predicting fatigue driving using location data according to its best accuracy (74.18%), F1-measure (62.02%), and AUROC (0.8059). Finally, we analyze the relationship between fatigue driving and driving environment according to variable importance described by random forest. In summary, our results obviously exhibit the potential of location data for reducing the accident rate caused by fatigue driving in practice.

1. Introduction

The transportation volume of the CDT continues to rise throughout the world with the rapid development of the modern manufacturing and logistics industries [1]. Dangerous goods transportation has a high potential risk which refers to the possibility of incurring traffic accidents with disastrous consequences [2, 3]. For example, explosions in densely populated areas or the release of toxic chemicals can lead to casualties directly or indirectly through environmental degradation [4]. Dangerous goods usually have characteristics, such as flammable, explosive, volatile, easy-corrosive, and so on. Thus, the transportation accidents involving dangerous goods usually show the following features: unpredictability, severe losses, and sudden and long-term effects [5]. When the

catastrophic accidents occur, the consequences cannot be often controlled and lowered [6]. Therefore, the safety of dangerous goods transportation has caught the attention of the public, transportation companies of dangerous goods, and decision makers and researchers within governmental and nongovernmental safety organization [7, 8].

Fatigue driving is one of the main reasons for fatal traffic accidents according to the causality analysis of traffic accidents [9, 10]. Up to 20 percent of traffic accidents are caused by fatigue driving [11–13]. Commercial truck drivers have relatively long driving time and are more prone to fatigue. Studies show that fatigue driving has been a major reason for commercial truck accidents [14–16]. Fatigue driving of commercial truck drivers increases the accident rate and leads to severe property loss, injuries, and fatalities [17–20].

Many previous researches have focused on the fatigue driving problem of commercial truck drivers [21–25]. Various sources and types of real-time data have been used in detecting driving fatigue. Physiological signals, being continuously available, objective, and fairly direct indicators, were often used to detect fatigue [26]. The electroencephalogram (EEG) and electrooculogram (EOG) are often used as a medium for detection [27–29]. However, EEG signals are very susceptible to noise and movements of the body [30]. EOG detection removes some problems of EEG; it only gives reasons for a certain aspect of the degree of human fatigue [28]. In addition, most physiological signal acquisition devices need to contact the driver's body, which may interfere with the normal operation of the driver and affect the driving safety. Thus, the alternative approaches without contacting the driver's body were developed using camera and other driving data. Fatigue may affect driver's behaviors including face and body activities [31]. The ocular and eyelid movements are often used to detect fatigue [32]. However, the image acquisition device is expensive and easy to be affected by the light. So, some other relative detection information was used to detect driving fatigue. The standard deviation of lane position (SDLP) or steering wheel movements are also often measured to detect the drivers' fatigue [33–35].

The above studies mainly focus on real-time detection of fatigue, which is a good approach to reduce the effects of fatigue driving. However, it may be not enough. When the fatigue is detected, the commercial driver is already on a transport mission and is difficult to abandon the mission or recover from a short rest [36]. If we can use historical data to predict the fatigue status of drivers before a new transportation mission, managers can select the drivers who are not prone to fatigue to undertake the more heavy transportation task by adjusting transportation plans. Fatigue driving is not only related to the driver's current driving, but also related to the driver's previous driving task intensity [37–39]. Long and hectic work schedules will increase the odds of driver fatigue [17, 21, 23, 25, 40–42]. Studies showed that the odds of driver fatigue increased heavily as the continuous days of driving increased [43, 44]. This might be a result of "accumulated fatigue" among the drivers due to long and continuous days of driving [44, 45]. The driver's recent driving tasks and driving environment can be used to predict the possibility of driver fatigue. The results can provide accurate information for driving tasks arrangement and early driver fatigue intervention.

The primary objective of this study is to propose an approach for predicting driver fatigue using characteristics of driver's recent driving task and driving environment extracted from location data and then use these characteristics to predict the possibility of driver fatigue. At present, studies focused on prediction of fatigue driving have emerged [26, 46, 47]. However, to the best knowledge of the authors, the approaches to predict fatigue using drivers' recent driving task and driving environment characteristics extracted from location data are yet to come. The contributions of this paper can be summarized as follows:

(1) Previous studies mainly used real-time data from drivers or vehicles to detect fatigue, but few studies used the historical data. In addition, previous studies have suggested that fatigue driving is intimately related to the driver's previous driving task intensity [37–39]. The proposed approach predicts fatigue driving using the drivers' recent driving task and driving environment characteristics

(2) There have been studies on the prediction of fatigue driving [26, 46, 47]; however, most studies focus on short-time forecasting. Few studies research on long-term prediction methods, specifically on commercial dangerous goods truck (CDT). At present, there is no research on prediction of fatigue driving of CDT within the scope of our literature review. The proposed approach can use the location data of CDT to predict fatigue driving of CDT

(3) At present, most studies on the prediction of fatigue driving mainly use physiological and behavioral indicators. However, physiological and behavioral measurements may interfere with the driver's normal driving, and the corresponding detection devices are relatively expensive and inconvenient to carry, which brings some difficulties to the future popularization and application of real driving conditions. To the best of our knowledge, there has not been a solution that is noninvasive and accurate. The proposed approach uses the location data of CDT to predict fatigue. The location data of CDT are available in many countries, so using location data makes the approach very scalable. In China, all CDTs are equipped with satellite positioning system and the data are uploaded to the national management system. However, we have not found any research on predicting fatigue driving using location data. The proposed approach is established based on six different categories of the predictor set only using raw location data

The paper is organized as follows. Section 2 details the study dataset. The overview for the methodology is described in Section 3. The obtained results are presented and discussed in Sections 4 and 5. Finally, conclusions to the paper are provided in Section 6.

2. Data

2.1. Data Description. We obtained data from the database of a transportation company in the south of China that currently comprises more than 200 CDTs. It has more than 580 drivers and more than 250 managers. The registered capital of the company is about 8 million dollars, with total assets of 32 million dollars. Each vehicle was equipped with devices which contains a GPS sensor, yawn detecting camera, and wireless transmission system. Because of the privacy restrictions of the database, we only took the location data and yawn data from the company's 166 CDT for 11 months in 2017. The location data were updated every 10 seconds, containing vehicle's plate number, speed, latitude, longitude, direction and location address, and time stamp. The yawn data included the vehicle's plate number and the specific time of yawning.

The mileage can be calculated from latitude and longitude data. Continuous driving time can be obtained using the time stamp and speed which is used to judge whether the vehicle is driving or not. The road type containing urban roads, highway except freeway, and freeway can be obtained using GIS systems. Some data were rejected due to the following reasons:

(i) The error in the data (e.g., the error in time or speed makes it impossible to accurately determine whether the vehicle is driving.)

(ii) Failure of the GPS sensor for a long time, so that the location data were not available

(iii) Failure of the yawn detecting device for a long time, so that the yawn data were not available

(iv) Too much location data were interrupted due to signal blocking (e.g., too much data of the latitude and longitude are interrupted, so that excessive mileage cannot be accurately calculated)

These cases finally led to a reduction of the location data by 5 vehicles to 161. The outliers were not further eliminated because we believed that their impact on the prediction results was insignificant due to the large data size.

2.2. Predictor Variables. The predictor variables are derived from raw location data. Traffic safety researchers have inferred some risk factors related to fatigue from observed-accident statistics, such as travel time, average velocity, mileage, road type, and so on [48, 49]. In addition, studies have shown that the risk factors such as travel time, average velocity, road type, and driving environment have significant impact on truck safety [50–52]. According to these risk factors, we designed six different categories of the predictor set including travel time, day of week, road type, continuous driving time, average velocity, and overall mileage. By accumulating mileage between every adjacent two data points of CDT, overall mileage M of each CDT can be calculated using latitude and longitude. Continuous driving time is an important index for predicting driver fatigue, so we take the average continuous driving time and the longest continuous driving time (C1-2) to measure driving time. Except for overall mileage and continuous driving time, we discretize the four other categories of the predictor set into a fixed number of intervals, where each interval corresponds to a predictor variable. Travel time is divided into five variables T1-5 that catch vehicle traveling at different times. Two other predictor variables catch vehicle traveling on weekdays and weekends (W1-2), while another variable triplet differentiates the three road types (R1-3). We separate average velocity into four variables V1-4, where the fourth interval includes mileage accumulated at velocities larger than 80 km/h (i.e., 80 km/h is the maximum speed limit for the CDT in China). The overview of predictor variables is shown in Table 1.

We assume that cumulative fatigue driving on the target day is strongly related to the task of the previous week. The predictor variables for specific target day were calculated using data from the previous week. We define the accumulated mileage of the previous week as the mileage accumulated from day $t-1$ to day $t-7$ on day t. The accumulated mileage of day t is described as

$$\text{PAM}_t = \sum_{i=t-7}^{t-1} P_i, \tag{1}$$

where PAM_t represents the accumulated mileage of the day t, P_i represents the mileage of the ith day, and t is an integer greater than 7.

We use the location data of the 161 vehicles to calculate predictor variables of each day which was described in Table 1. Except for continuous driving time, values of the 15 predictor variables are the mileage accumulated in a week before target day (i.e., the predictor variables of day t are accumulated for day $t-1$ to day $t-7$ on the dependent variable). The predictor variables, which have different dimensions and change intervals, may result in some indicators to be ignored and affect the results of data analysis. Therefore, we normalize all predictor variables, where the normalized equation for all predictor variables except for overall mileage and continuous driving time is

$$X^* = \frac{X}{M}, \tag{2}$$

where X is the values of all predictor variables except for overall mileage and continuous driving time and M is the values of overall mileage.

The normalized equation for continuous driving time C is

$$C^* = \frac{C - C_{\min}}{C_{\max} - C_{\min}}, \tag{3}$$

where $C_{\max}$ is the values of the maximum continuous driving time, $C_{\min}$ is the values of the minimum continuous driving time, and their values are obtained across 161 vehicles in 11 months. And we furthermore normalize overall mileage M by taking the logarithm of M and dividing it by the logarithm of the M maximum:

$$M^* = \frac{\log_{10}(M)}{\log_{10}(M_{\max})}, \tag{4}$$

where the maximum of M is also obtained across 161 vehicles in 11 months. The descriptive statistics of all predictor variables are shown in Table 2.

Except for continuous driving time, values of the 15 predictor variables are the mileage accumulated in a week. It may cause the collinearity problem of the generated predictor variables at the same time. This problem is an unwanted property for most classifiers and is especially troublesome for logistic regression [53, 54].

Therefore, we select the method of factor analysis to solve the collinearity problem of logistic regression. Factor analysis is a multivariate analysis method that converts multiple variables into several integrated variables (or latent variables), which are mainly used to reduce the number of variables and classify variables with high correlation, using common factors instead. In this study, principal component analysis is used to extract factors with eigenvalues greater than 1 as common factor. Table 3 presents the eigenvalues, the percentage of variance, the

Table 1: Predictor variables.

Group	Variable description
Travel time (T)	$T1$: 0 am and 5 am
	$T2$: 5 am and 9 am
	$T3$: 9 am and 5 pm
	$T4$: 5 pm and 10 pm
	$T5$: 10 pm and 12 pm
Day of week (W)	$W1$: weekdays except Friday
	$W2$: weekends with Friday
Road type (R)	$R1$: urban roads
	$R2$: highway except the freeway
	$R3$: freeway
Continuous driving time (C)	$C1$: average continuous driving time
	$C2$: longest continuous driving time
Average velocity (V)	$V1$: 0 km/h–40 km/h
	$V2$: 40 km/h–60 km/h
	$V3$: 60 km/h–80 km/h
	$V4$: over 80 km/h (has only a lower bound)
Overall mileage (M)	M: mileage traveled per day

Table 2: Descriptive statistics of all predictor variables.

Variable	Mean	1st Q	2nd Q (median)	3rd Q
$T1$	0.076	0.000	0.034	0.123
$T2$	0.168	0.078	0.159	0.243
$T3$	0.496	0.395	0.484	0.604
$T4$	0.208	0.124	0.221	0.285
$T5$	0.052	0.000	0.040	0.092
$W1$	0.592	0.497	0.577	0.678
$W2$	0.408	0.321	0.423	0.503
$R1$	0.263	0.130	0.238	0.345
$R2$	0.227	0.072	0.136	0.355
$R3$	0.510	0.265	0.566	0.774
$C1$	0.501	0.333	0.487	0.669
$C2$	0.314	0.178	0.273	0.413
$V1$	0.213	0.084	0.153	0.260
$V2$	0.285	0.151	0.245	0.375
$V3$	0.497	0.292	0.538	0.699
$V4$	0.005	0.000	0.000	0.000
M	0.717	0.695	0.742	0.775

cumulative eigenvalue, and the cumulative percentage of variance associated with each factor. It reveals that the first four factors explain approximately 76.9% of total variance. Finally, the number of common factors is determined to be 4.

Fatigue may be determined according to the physical activities and human behavior [55]. The driver's yawn data are used as a proxy for ground truth for the classification algorithm. If the driver yawns in target day, the driver is considered to be fatigued in this day. In this paper, fatigue, indicated by yawning, is predicted by our approach using location data of CDT. According to Kiang's suggestions on classifier selection [56], our approach considers three types of classifiers, namely, logistic regression, neural networks, and random forest.

In a supervised classification problem, a training set is usually used to construct classification models and the independent testing set is used to testify the predictive performance of these models [57]. Therefore, for logistic regression, we randomly divided the dataset with 4 common factors into two subsets, in which 70% of the whole dataset were included in the training set and the remaining 30% were included in the testing set. For neural network, we randomly divided the normalized dataset with 17 predictor variables in Table 1 into two subsets, in which 70% of the whole dataset were used for training and 30% were used for testing. Cutler et al. suggested that the random forest algorithm included the interactions among the variables, so there was no collinearity problem faced by other models [58]. Therefore, for random forest, we randomly divided the unnormalized dataset with 17 predictor variables in Table 1 into two subsets, in which 70% of the whole dataset were used for training and 30% were used for testing.

3. Research Approach

3.1. Logistic Regression. This paper judges whether the driver is fatigued by whether the driver is yawning. Since the dependent variable is binary, we establish a binary logistic model:

TABLE 3: Eigenvalues, percent of variance, cumulative eigenvalue, and cumulative percent of variance for factor analysis.

Factor	Eigenvalue	Percentage of variance	Cumulative eigenvalue	Cumulative percentage of variance
1	6.343	37.313	6.343	37.313
2	2.860	16.826	9.203	54.139
3	2.413	14.192	11.616	68.331
4	1.458	8.574	13.074	76.905
5	0.931	5.479	14.005	82.384
6	0.658	4.028	14.663	86.412
7	0.519	3.052	15.182	89.465
8	0.445	2.616	15.627	92.080
9	0.392	2.307	16.019	94.387
10	0.345	2.029	16.364	96.416
11	0.259	1.523	16.623	97.939
12	0.245	1.443	16.868	99.382
13	0.104	0.610	16.972	99.992
14	0.001	0.006	16.973	99.998
15	0.000	0.002	17	100.000
16	$1.022E-13$	$6.010E-13$	17	100.000
17	$7.399E-14$	$4.317E-13$	17	100.000

$$\text{Logit}(P) = \ln\left(\frac{P}{1-P}\right) = \alpha + \sum \beta_j X_j + \varepsilon, \tag{5}$$

where P is the probability of the dependent variable $Y = 1$ (i.e., the probability of the driver yawning), the independent variable X_j is the various factors affecting the driver's fatigue (i.e., the common factor extracted by the factor analysis method), β_j is the regression coefficients of the independent variable X_j, α is a constant term, and ε is an error term.

3.2. Neural Network.

In this paper, the multilayer perceptual neural network algorithm is used to train data samples. The multilayer perceptual neural network is a forward-structured artificial neural network that uses a backpropagation algorithm for training. The network consists of an input layer, hidden layers, and an output layer. The input layer corresponds generally to features to classify and is used to receive input data. The hidden layer may have multiple layers for learning data and storing training results. The output layer corresponds to the defined classes and each class corresponds to a node in the output layer. It is used to output results. Each layer consists of multiple nodes, each of which can be passed to the next layer up to the output layer. Excluding the input node, all other nodes multiply the input by its own weighting factor ω, plus the offset b, and then combine its own nonlinear activation function to produce the output [59].

The optimization algorithm of multilayer perceptual neural network adopts the adjusted conjugate gradient algorithm and the activation function of each layer is different. The middle layer node uses the hyperbolic tangent function as the activation function:

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{6}$$

The output layer node uses the Softmax function as the activation function:

$$f(x_i) = \frac{e^{x_i}}{\sum_{n=1}^{N} e^{x_n}}, \tag{7}$$

where x_i represents the input from the previous layer and N represents the total number of nodes in the previous layer.

We use the 17 predictor variables normalized in Table 1 as the network input and choose to use a layer of hidden neurons based on the data characteristics. In order to determine the optimal number of nodes in the hidden layer, we first make the number of nodes in the hidden layer equal to the number of nodes in the input layer. Then, we gradually reduce the number of nodes and simultaneously calculate generalization errors, training errors, deviations, and variances. The number of nodes at this point is our choice when the generalization error has dropped and before it begins to increase again. We finally determined that the optimal number of nodes in the hidden layer is 13. Figure 1 shows the structure of the established neural network model. The comparison between neural network and logistic regression is common, and related studies have found that neural network is superior to logistic regression due to its complex model structure [60, 61].

3.3. Random Forest.

Random forest proposed by Breiman is an ensemble learning algorithm which constructs multiple decision trees through bootstrap aggregation [62]. Each tree is a standard Classification or Regression Tree (CART) that uses the so-called Decrease of Gini Impurity (DGI) as a splitting standard of the node [63]. Instead of using all input variables, random forest selects at random a subset of the input variables to split each node when growing a CART [64]. Each tree predicts a classification independently and "votes" for the corresponding class. The majority of the votes determine the optimal result of the random forest model [65]. The operating principle of random forest is summarized as follows and shown in Figure 2.

(i) k subsets of the sample $D1$, $D2$,..., Dk are drawn from the total sample set D using the bagging technique. The sample size of subsets Dk is the same as the total sample set D.

(ii) k decision trees are constructed according to the k subsets and obtain k classification results.
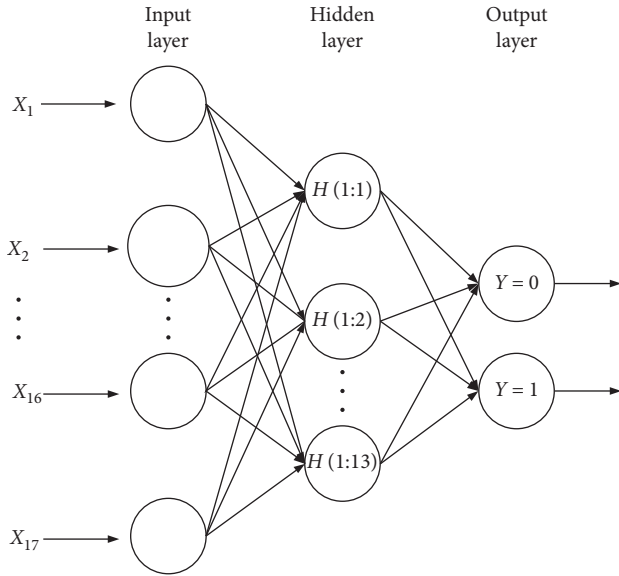
FIGURE 1: Schematic diagram of multilayer perceptual neural network structure.

FIGURE 2: The operating principle of random forest.

FIGURE 3: Average error rate of hyperparameter tuning.

(iii) Optimal results are obtained by voting.

To execute the random forest algorithm, the open source software, Python, which provides a language and environment for statistical calculation, was used. Before training the random forest model, tuning its hyperparameter is necessary to obtain random forest model with the best predictive performance. Two important hyperparameters, namely, the number of classification trees (ntree) and the number of variables tried at each split (mtry), have a significant effect on the performance of the model. Regarding the hyperparameters mtry, many studies use the value recommended by Breiman mtry = sqrt (M), where M is the number of predictor variables [66]. In this study, mtry = 4. Therefore, we only tuned the hyperparameters ntree and its tuning range was 10–4000. We compared the random forest models with different hyperparameters *n*tree using the average error rate from 5-fold cross-validation. As shown in Figure 3, the average error rate decreased sharply when ntree increased from 10 to 60. When ntree increased from 60 to 2200, the average error rate had slightly different trends; however, generally, the average error rate decreased slightly. When increased from 2200 to 4000, the average error rate almost remained stable. Therefore, *n*tree = 2200 was determined as the optimal value. Finally, the optimum hyperparameters were determined to be 2200 trees with the number of variables tried at each split being 4.

3.4. Model Evaluation. For the training results, three indexes including accuracy, F1-measure, and area under the ROC curve (AUROC) are used to evaluate the predictive performance of the classifiers. Although more indexes can be used to evaluate the predictive performance of the classifiers, we believe that these three indexes can complete the comparison between logistic regression and neural network classifiers. The numbers of true negatives (TN), true positives (TP), false positives (FP), and false negatives (FN) are used as a
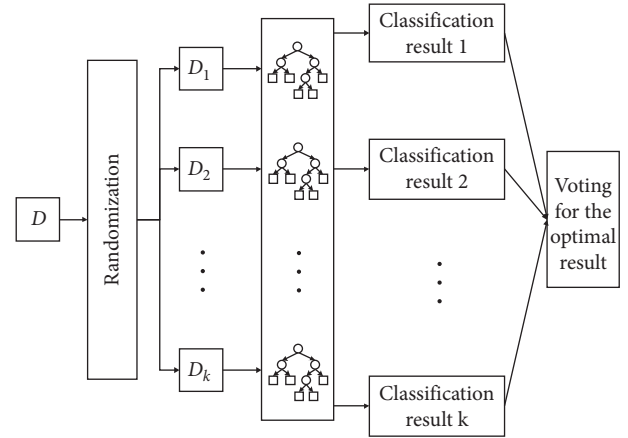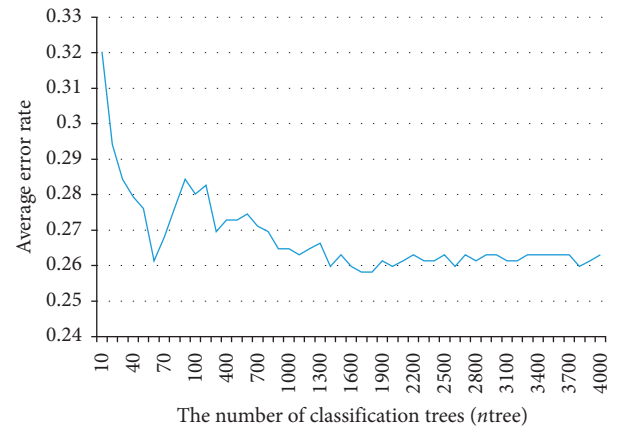
measurement to assess the performance of classifiers. Different terms are used in different domains. Accuracy is the most basic index for assessing the performance of classifiers. It is used as an overall measure and calculated as

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \qquad (8)$$

where TP and TN indicate correctly classified cases and FP and FN indicate the incorrectly classified cases. However, the skewed class distribution of samples, in reality, makes traditional metrics such as accuracy unable to properly reflect the performance of the classifiers [67]. Therefore, another index, F1-measure, is proposed to evaluate performance and calculated as

$$F1 - measure = \frac{2TP}{2TP + FP + FN}. \qquad (9)$$

Accuracy and F1-measure evaluate the performance of the classifiers by comparing predicted class labels. In this sense, they can actually be thought to measure different aspects of the same coin, and show recognized disadvantages [68]. Therefore, the receiver operating characteristic (ROC) curve is used to measure the performance of classifiers. The curve is generated by plotting true positives as the percentage of all positives and negative ones in the sample [69]. We hope to reduce ROC performance to a single scalar value

representing expected performance to evaluate the performance of classifiers, so the AUROC is considered as an additional index. AUROC gives a single measure of overall accuracy that is independent of any particular threshold [70, 71]. Larger AUROC value indicates that better predictive model is a generally accepted rule for determining a model's performance when comparing various models.

4. Results

4.1. Model Comparison and Selection. We trained logistic regression, neural network, and random forest using the training set and calculated accuracy, F1-measure, and AUROC value of every model based on the testing set.

Figure 4 depicts the results for classification performance of logistic regression, neural network, and random forest models. Obviously, the accuracy, F1-measure, and AUROC of the random forest are higher than the logistic regression and neural network (i.e., the predictive performance of the random forest is better). Therefore, the random forest is more suitable to use location data to predict fatigue driving than logistic regression and neural network. It can be seen from Figure 4 that the accuracy of the random forest is 74.18%. Although this accuracy is not too high, it can be accepted compared to other fatigue driving detection methods based on vehicle information. In addition, more than 60% of F1-measure reveals its ability to detect real yawn, which means the number of a missed yawn is reduced using the random forest classification. The random forest was selected to predict fatigue driving using predictor variables.

4.2. Variable Importance Analysis. After determining the random forest model as the optimal prediction model, we analyzed the relationship between fatigue driving and driving task of last week according to variable importance described by random forest.

Variable importance (called "variable importance score" in this study) reflects every predictor variable's contribution to the total risk. The random forest model computes variable importance scores by assessing the importance of every predictor variable using the Gini decrease index [72]. The computation was implemented based on the "feature_importances_" in the random forest package of open source software, Python. Figure 5 provides the normalized variable importance scores (i.e., the sum of the importance scores for all variables is one).

Fatigue driving has a close relationship with the driving task of last week. By comparing the variable importance in Figure 5, the paper draws the following conclusions:

(1) It is not difficult to see from Figure 5 that the importance scores of average continuous driving time ($C1$) and longest continuous driving time ($C2$) are the highest among all the predictor variables. This shows that continuous driving time is closely related to fatigue driving. Fatigue driving refers to the phenomenon that the driver produces dysfunction of physiology and mental function after driving for a long time so that driving skills decline objectively. Prolonged driving will
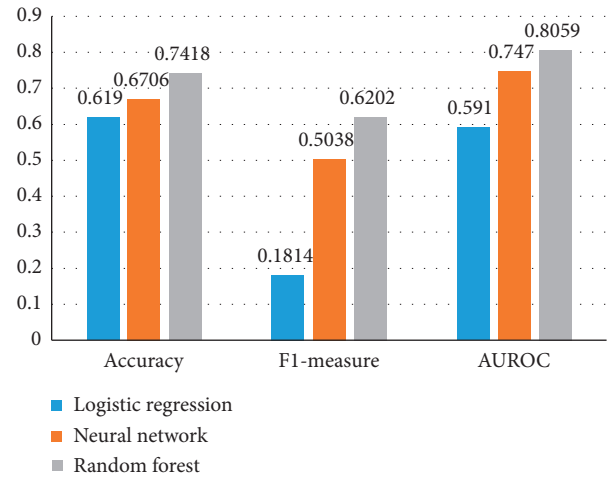

FIGURE 4: Prediction accuracy, F-measure, and AUROC for every model.
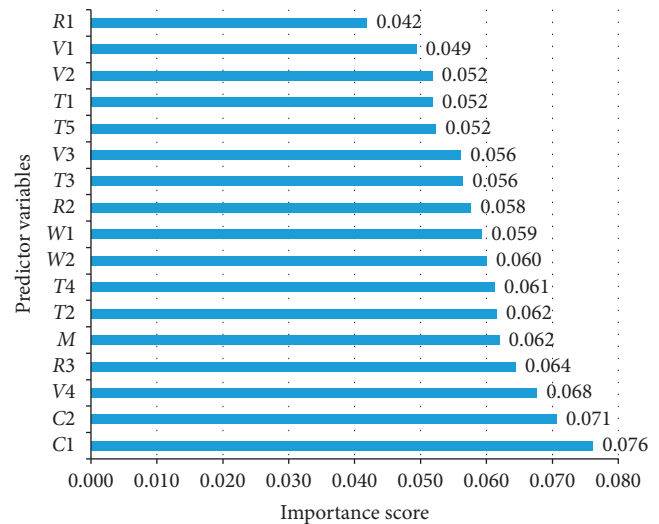

FIGURE 5: Relative importance of predictor variables.

make driver mental overload and cause task-related fatigue [45]. Therefore, the continuous driving time of the driver must be strictly controlled to avoid accidents caused by prolonged driving.

(2) During the predictor variables of travel time group, the importance scores of travel time between 5 am and 9 am ($T2$) and travel time between 5 pm and 10 pm ($T4$) are the highest. This shows that the driver is more likely to be fatigued when driving in these two time periods for a long time, which is basically consistent with the previous research results [73, 74]. In addition, it has been extensively proven that the number of accidents related to fatigue driving increase in the early morning and late evening [45, 75]. The time period indicated by $T4$ is extremely fragile. This is because, after a day of hard work, the driver will have a series of tired symptoms such as dry eyes, dry throat, and yawning. The time period indicated

by $T2$ is early morning, which is the time period when fatigue driving and traffic accidents are most likely to occur. During this time, the human circadian rhythm is in a state of slow brain reaction, lower blood pressure, and stiff and paralyzed blood vessels in the hands and feet. Therefore, in order to avoid fatigue driving, the driver's driving time should be reasonably arranged and the driver should try to avoid driving in these two periods. The importance scores of travel time between 10 pm and 12 pm ($T5$) and travel time between 0 am and 5 am ($T1$), in contrast, were found to be lower than all other variables of travel time (T). Apparently it seems surprising. But we can see from Table 2 that the mean of these two time periods is relatively small, which indicates that the driver rarely travels during these two time periods. One reason for this may be due to the relatively higher accident rate compared to other time periods; the company deliberately controls the driver not to drive during this time period.

(3) It is not surprising to observe that the importance scores of the average velocity (V) consistently decrease from average velocity over 80 km/h ($V4$) to average velocity between 0 and 40 km/h ($V1$). Therefore, the importance scores for $V4$ (0.104) > $V3$ (0.040) > $V2$ (0.036) > $V1$ (0.035). This shows that the driver is more likely to be fatigued when driving at a high speed for a long time. It has been extensively proven that the higher the driving speed is, the easier the driver is to be fatigued [76]. The higher the driving speed, the greater the degree of tension or concentration of the driver's central nervous system and the greater the mental and physical energy consumed. At the same time, the driver's field of vision narrows as the speed of the vehicle increases, and the information that is missed increases, making the driver more nervous. It is worth mentioning that the importance scores of average velocity between 40 and 60 km/h ($V2$) and average velocity between 0 and 40 km/h ($V1$) are relatively lower than other predictor variables and the importance of average velocity over 80 km/h ($V4$) is relatively higher. This shows that when the driver drives in the environment in which the speed is lower than 60 km/h for a long time, the driver is not prone to fatigue, but once the driver drives in the environment where the speed exceeds 80 km/h for a long time, the driver is more likely to be fatigued. Therefore, the driving speed of the vehicle should be reasonably controlled to avoid the driver driving at a high speed for a long time.

(4) It can be seen from Figure 5 that the importance scores of the road type (R) consistently increase from urban roads ($R1$) to freeway ($R3$). This shows that the driver is more likely to be fatigued when driving on the freeway for a long time, which is consistent with previous research [74, 76, 77]. In addition, it has been reported that 40 percent of the accidents caused by fatigue driving occur on freeways [78]. The freeway has neither traffic signal control nor pedestrians, nonmotor vehicles, and other low-speed motor vehicles. Driving on this road for a long time is easy to cause the driver to sleep. In addition, when driving on the freeway, the driver's energy is always in a state of high tension, and the physical exertion is increased, and the speed of the vehicle will be unconsciously increased, and even the brake deceleration consciousness will be lost. Driving in such an environment for a long time can also make the driver feel tired. Therefore, it is necessary to adopt the fatigue warning device when the driver is driving on the freeway.

(5) We can also find that the importance score of weekends with Friday ($W2$) is higher than the importance score of weekdays except Friday ($W1$) from Figure 5. This indicates that the driver is more likely to be fatigued when driving on weekends for a long time. This is because the driver will continue to accumulate fatigue as he drives on weekdays, which makes the driver's fatigue index relatively higher on weekends. This provides a basis for a reasonable arrangement of driver travel time.

5. Discussion

This paper has offered a brand new approach to predict driving fatigue using location data of CDT. The existing approach to predict driving fatigue mainly uses physiological and behavioral indicators. However, physiological and behavioral measurements may interfere with the driver's normal driving, and the corresponding detection devices are relatively expensive. Our approach predicts fatigue using the location data of CDT which are collected without interfering with the driver's normal driving. Location data acquisition equipment is relatively inexpensive and is generally installed in commercial trucks. These are beneficial to the future popularization and application of real driving conditions. In addition, most studies on the prediction of fatigue driving focus on short-time forecasting. Few studies research on long-term prediction methods, specifically on commercial trucks. Our approach addresses the long-term prediction of fatigue driving in commercial trucks.

5.1. Model Application Illustration. The proposed approach can be used not only for driving fatigue prediction of commercial dangerous goods transport vehicles but also for other transport vehicles. In addition, our approach can directly use the location data of the vehicle to predict fatigue, which not only solves the problem that most domestic commercial transport vehicles do not have the image acquisition device installed, but also has no disadvantages of other detection approaches that interfere with the driver. Our approach can also aid decision making and is a useful complement to real-time monitoring. Even if the transport vehicles are equipped with the image acquisition device, our approach is also necessary to help prevent fatigue. What is also worth noting about our approach is that it can not only be used to predict fatigue, but also provide a basis for transportation companies to arrange transportation mission reasonably. Figure 6 depicts the application of the prediction approach.
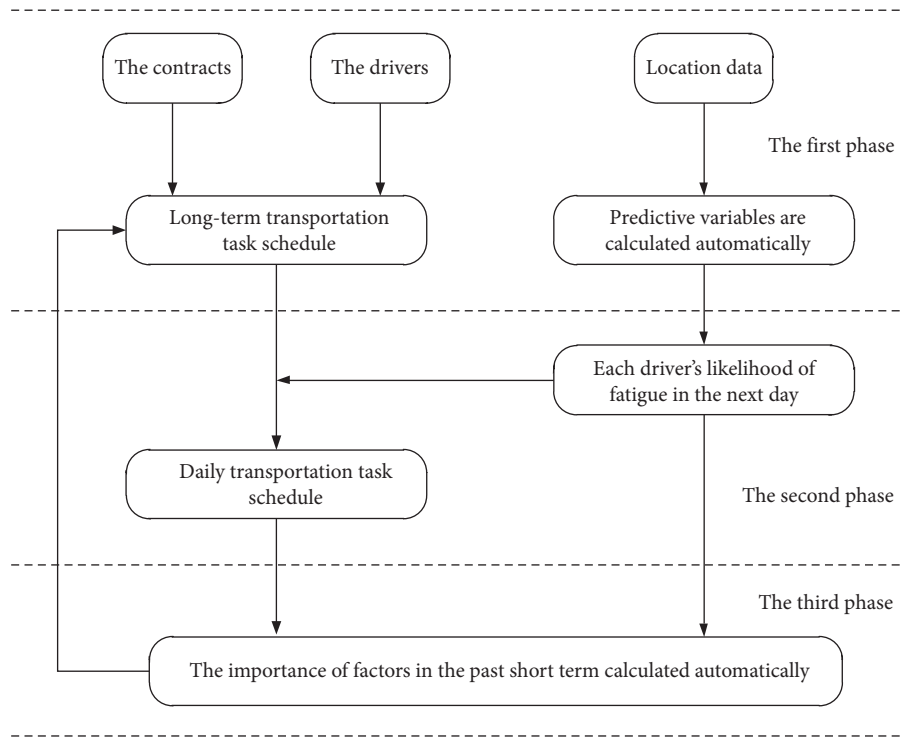
FIGURE 6: Application of the prediction approach.

The proposed approach can be used to optimize the daily task arrangement of the transportation company. In the first phase, the transportation company will complete a long-term transportation task schedule, which may be used for a week or a month, based on the contracts signed and the number of drivers. At the same time, the location data of the driver during the transportation task are automatically collected.

In the second phase, 17 indicators can be calculated automatically and each driver's likelihood of fatigue in the next day was dynamically predicted. The prediction results can help optimize the daily schedule of transportation tasks.

In the third phase, the importance of each factor in the past short term can help the transportation company managers to formulate the long-term task schedule. Do your best to avoid long hours in which drivers are prone to fatigue.

5.2. Research Limitations and Future Research Needs. Some methodological and conceptual limitations should be considered in the interpretation of our results. These limitations make us consider using other models to further improve prediction accuracy in future research. This would involve the combination of different classifiers [79]. We should also use the data from other dangerous goods transportation companies to verify our results. Due to the limited data acquisition properties, our approach only analyzes the influence of six predictor sets. In the future, other available related variables should also be considered to extend the set of predictor variables and yield further

improvements to predictive performance and the guidance of analysis results. Driver's physique, lifestyle, stress, and other factors have a certain impact on the predictive performance of our model. Therefore, if the driver's relevant information can be obtained and used as predictor variables, the accuracy of the model will be further improved. In addition, seasonal changes have a significant impact on our approach; future efforts should be made to eliminate the effects of the seasons, thus making our approach more complete.

6. Conclusion

In order to solve the fatigue driving problem of dangerous goods transportation, this paper proposed an approach that used location data obtained from a transportation company to predict fatigue driving and further analyzed the relationship between fatigue driving and driving environment. The proposed approach can be used to predict fatigue driving using the location data of CDT which were collected without interfering with the driver's normal driving and provide a basis for transportation companies to arrange transportation mission reasonably. The main findings were concluded as follows:

(1) We used logistic regression, neural network, and random forest techniques to predict fatigue driving from the location data. To choose a more suitable classifier as a predictive model, we obtained a set of 17 predictor variables from the six different categories of the predictor set related to fatigue to train and compare logistic regression, neural network, and random forest

classifiers. By analyzing and comparing the classification performance results of logistic regression, neural networks, and random forest models, we found that accuracy (74.18%), F1-measure (62.02%), and AUROC (0.8059) of random forest were separately best, so random forest was more suitable for predicting fatigue driving using location data

(2) To provide a basis for the transportation company to arrange transportation reasonably, after determining the random forest model as the optimal prediction model, we analyzed the relationship between fatigue driving and driving environment according to variable importance described by random forest. We found that fatigue driving was closely related to driving conditions such as travel time, continuous driving time, driving speed, road type, and so on. The period extremely prone to fatigue driving is early morning and the evening, and the driver is more prone to fatigue on weekdays than on weekends. The higher the driving speed is, the easier the driver is to be fatigued. The probability of fatigue driving on the freeway is higher than that of highway and urban road. These conditions can provide a basis for the company to avoid driver fatigue driving, thereby reducing traffic accidents.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

References

[1] Q. Yang, K.-S. Chin, and Y.-L. Li, "A quality function deployment-based framework for the risk management of hazardous material transportation process," *Journal of Loss Prevention in the Process Industries*, vol. 52, no. 52, pp. 81–92, 2018.

[2] Y. Leung, R. Li, and N. Ji, "Application of extended Dempster-Shafer theory of evidence in accident probability estimation for dangerous goods transportation," *Journal of Geographical Systems*, vol. 19, no. 3, pp. 249–271, 2017.

[3] E. Erkut, S. A. Tjandra, and V. Verter, "Hazardous materials transportation," in *Handbooks in Operations Research and Management Science*, G. Laporte and C. Barnhart, Eds., vol. 14, pp. 539–621, Elsevier, Haarlem, Netherlands, 2007.

[4] M. H. Laarabi, A. Boulmakoul, R. Sacile, and E. Garbolino, "A scalable communication middleware for real-time data collection of dangerous goods vehicle activities," *Transportation Research Part C: Emerging Technologies*, vol. 48, pp. 404–417, 2014.

[5] P. Leonelli, S. Bonvicini, and G. Spadoni, "New detailed numerical procedures for calculating risk measures in hazardous materials transportation," *Journal of Loss Prevention in the Process Industries*, vol. 12, no. 6, pp. 507–515, 1999.

[6] A. M. Tomasoni, *Models and Methods of Risk Assessment and Control in Dangerous Goods Transportation (DGT) Systems*, École Nationale Supérieure des Mines de Paris, Paris, France, 2010.

[7] G. Landucci, G. Antonioni, A. Tugnoli, S. Bonvicini, M. Molag, and V. Cozzani, "HazMat transportation risk assessment: a revisitation in the perspective of the Viareggio LPG accident," *Journal of Loss Prevention in the Process Industries*, vol. 49, no. SI, pp. 36–46, 2017.

[8] R. Bubbico, G. Maschio, B. Mazzarotta, M. F. Milazzo, and E. Parisi, "Risk management of road and rail transport of hazardous materials in Sicily," *Journal of Loss Prevention in the Process Industries*, vol. 19, no. 1, pp. 32–38, 2006.

[9] R. Fu, H. Wang, and W. Zhao, "Dynamic driver fatigue detection using hidden Markov model in real driving condition," *Expert Systems with Applications*, vol. 63, pp. 397–411, 2016.

[10] A. Williamson and R. Friswell, "The effect of external non-driving factors, payment type and waiting and queuing on fatigue in long distance trucking," *Accident Analysis & Prevention*, vol. 58, no. 3, pp. 26–34, 2013.

[11] I. D. Brown, "Prospects for technological countermeasures against driver fatigue," *Accident Analysis & Prevention*, vol. 29, no. 4, pp. 525–531, 1997.

[12] G. Maycock, "Driver sleepiness as a factor in car and HGV accidents," Trl Report, The Royal Society for the Prevention of Accidents, Birmingham, UK, 1995.

[13] G. Zhang, K. K. W. Yau, X. Zhang, and Y. Li, "Traffic accidents involving fatigue driving and their extent of casualties," *Accident Analysis & Prevention*, vol. 87, pp. 34–42, 2016.

[14] America USO, *Safety Study: Fatigue, Alcohol, Other Drugs, and Medical Factors in Fatal-to-the-Driver Heavy Truck Crashes*, Vol. 1, Bureau of Justice Statistics, Washington, DC, USA, 1990.

[15] P. C. Morrow and M. R. Crum, "Antecedents of fatigue, close calls, and crashes among commercial motor-vehicle drivers," *Journal of Safety Research*, vol. 35, no. 1, pp. 59–69, 2004.

[16] Federal Motor Carrier Safety Administration, *Report to Congress on the Large Truck Crash Causation Study*, U.S. Department of Transportation, Washington, DC, USA, 2006.

[17] M. J. M. Sullman, M. L. Meadows, and K. B. Pajo, "Aberrant driving behaviours amongst New Zealand truck drivers," *Transportation Research. Part F: Traffic Psychology and Behaviour*, vol. 5, no. 3, pp. 1–232, 2002.

[18] D. Blower, P. E. Green, and A. Matteson, "Condition of trucks and truck crash involvement," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2194, no. 1, pp. 21–28, 2010.

[19] F. Sagberg, "Road accidents caused by drivers falling asleep," *Accident Analysis & Prevention*, vol. 31, no. 6, pp. 639–649, 1999.

[20] B. C. Tefft, "Prevalence of motor vehicle crashes involving drowsy drivers, United States, 1999–2008," *Accident Analysis & Prevention*, vol. 45, no. 45, pp. 180–186, 2012.

[21] G. Tzamalouka, M. Papadakaki, and J. E. Chliaoutakis, "Freight transport and non-driving work duties as predictors of falling asleep at the wheel in urban areas of Crete," *Journal of Safety Research*, vol. 36, no. 1, pp. 75–84, 2005.

[22] H. M. Kanazawa, T. Onoda, and N. Yokozawa, "Excess workload and sleep-related symptoms among commercial long-haul truck drivers," *Sleep and Biological Rhythms*, vol. 4, no. 2, pp. 121–128, 2006.

[23] M. Suzuki, T. Kontogiannis, G. Tzamalouka, C. Darviri, and J. Chliaoutakis, "Exploring the effects of lifestyle, sleep factors and driving behaviors on sleep-related road risk: a study of Greek drivers," *Accident Analysis & Prevention*, vol. 40, no. 6, pp. 2029–2036, 2008.

[24] J. Duke, M. Guest, and M. Boggess, "Age-related safety in professional heavy vehicle drivers: a literature review," *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 364–371, 2010.

[25] F. Meng, S. Li, L. Cao et al., "Driving fatigue in professional drivers: a survey of truck and taxi drivers," *Journal of Crash Prevention and Injury Control*, vol. 16, no. 5, p. 10, 2015.

[26] C. J. de Naurois, C. Bourdin, C. Bougard et al., "Adapting artificial neural networks to a specific driver enhances detection and prediction of drowsiness," *Accident Analysis & Prevention*, vol. 121, pp. 118–128, 2018.

[27] Y. Wang, Y. Zhang, D. Liu et al., "Driving fatigue detection based on EEG signal," in *Proceedings of the Fifth International Conference on Instrumentation & Measurement, Computer, Communication, and Control (IMCCC)*, IEEE, Qinhuangdao, China, pp. 714–717, September 2015.

[28] X. Q. Huo, W. L. Zheng, and B. L. Lu, "Driving fatigue detection with fusion of EEG and forehead EOG," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, IEEE, Vancouver, Canada, pp. 897–904, July 2016.

[29] Z. Li, J. Peng, L. Chen et al., "Fatigue driving detecting method based on time-space features in real driving conditions," in *Proceedings of the 10th IEEE Conference on Industrial Electronics and Applications Industrial Electronics and Applications*, IEEE, Auckland, New Zealand, pp. 183–187, June 2015.

[30] Y. Wang, X. Liu, Y. Zhang et al., "Driving fatigue detection based on EEG signal," in *Proceedings of the Fifth International Conference on Instrumentation & Measurement*, IEEE, Shenzhen, China, September 2016.

[31] Y. Dong, Z. Hu, K. Uchimura et al., "Driver inattention monitoring system for intelligent vehicles: a review," in *Proceedings of the IEEE Intelligent Vehicles Symposium*, IEEE, Xi'an, China, 2009.

[32] J. Chen and Q. Ji, "Drowsy driver posture, facial, and eye monitoring methods," in *Handbook of Intelligent Vehicles*, A. Eskandarian, Ed., pp. 913–940, Springer, London, UK, 2012.

[33] C. C. Liu, S. G. Hosking, and M. G. Lenné, "Predicting driver drowsiness using vehicle measures: recent insights and future challenges," *Journal of Safety Research*, vol. 40, no. 4, pp. 239–245, 2009.

[34] Y. J. Zhong, L. P. Du, K. Zhang et al., "Localized energy study for analyzing driver fatigue state based on wavelet analysis," in *Proceedings of the International Conference on Wavelet Analysis & Pattern Recognition*, IEEE, Hong Kong, China, 2008.

[35] Y. Takei and Y. Furukawa, "Estimate of driver's fatigue through steering motion," in *Proceedings of the IEEE International Conference on Systems*, IEEE, Los Angeles, CA, USA, 2006.

[36] L. Wang and Y. Pei, "The impact of continuous driving time and rest time on commercial drivers' driving performance and recovery," *Journal of Safety Research*, vol. 50, pp. 11–15, 2014.

[37] G. X. Chen, Y. Fang, F. Guo, and R. J. Hanowski, "The influence of daily sleep patterns of commercial truck drivers on driving performance," *Accident Analysis & Prevention*, vol. 91, pp. 55–63, 2016.

[38] Y. Wang, C. Ma, and Y. Li, "Effects of prolonged tasks and rest patterns on driver's visual behaviors, driving performance, and sleepiness awareness in tunnel environments: a simulator study," *Iranian Journal of Science and Technology, Transactions of Civil Engineering*, vol. 42, no. 2, pp. 143–151, 2018.

[39] K. Mahajan, N. R. Velaga, A. Kumar, A. Choudhary, and P. Choudhary, "Effects of driver work-rest patterns, lifestyle and payment incentives on long-haul truck driver sleepiness," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 60, no. 60, pp. 366–382, 2019.

[40] A. T. Mccartt, J. W. Rohrbaugh, M. C. Hammer, and S. Z. Fuller, "Factors associated with falling asleep at the wheel among long-distance truck drivers," *Accident Analysis & Prevention*, vol. 32, no. 4, pp. 493–504, 2000.

[41] M. M. Ohayon, "Demographic factors, fatigue, and driving accidents: an examination of the published literature," *Accident Analysis and Prevention*, vol. 43, no. 2, pp. 516–532, 2011.

[42] K. Sadeghniiat-Haghighi, Z. Yazdi, and A. M. Kazemifar, "Sleep quality in long haul truck drivers: a study on Iranian national data," *Chinese Journal of Traumatology*, vol. 4, no. 19, pp. 225–228, 2016.

[43] P. K. Arnold, L. R. Hartley, A. Corry, D. Hochstadt, F. Penna, and A. M. Feyer, "Hours of work, and perceptions of fatigue among truck drivers," *Accident Analysis & Prevention*, vol. 29, no. 4, pp. 471–477, 1997.

[44] T. Åkerstedt, A. Knutsson, P. Westerholm, T. Theorell, L. Alfredsson, and G. Kecklund, "Sleep disturbances, work stress and work hours," *Journal of Psychosomatic Research*, vol. 53, no. 3, pp. 741–748, 2002.

[45] J. F. May and C. L. Baldwin, "Driver fatigue: the importance of identifying causal factors of fatigue when considering detection and countermeasure technologies," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 12, no. 3, pp. 1–224, 2009.

[46] A. Watson and G. Zhou, "Microsleep prediction using an EKG capable heart rate monitor," in *Proceedings of the 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, 2016.

[47] C. Jacobé de Naurois, C. Bourdin, A. Stratulat, E. Diaz, and J.-L. Vercher, "Detection and prediction of driver drowsiness using artificial neural network models," *Accident Analysis & Prevention*, vol. 126, pp. 95–104, 2019.

[48] J. Paefgen, T. Staake, and F. Thiesse, "Evaluation and aggregation of pay-as-you-drive insurance rate factors: a classification analysis approach," *Decision Support Systems*, vol. 56, pp. 192–201, 2013.

[49] D. Lord and F. Mannering, "The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives," *Transportation Research Part A*, vol. 44, no. 5, pp. 0–305, 2010.

[50] F. Chen, H. Peng, X. Ma, J. Liang, W. Hao, and X. Pan, "Examining the safety of trucks under crosswind at bridge-tunnel section: a driving simulator study," *Tunnelling and Underground Space Technology*, vol. 92, Article ID 103034, 2019.

[51] S. Niu and S. V. Ukkusuri, "Risk Assessment of Commercial dangerous -goods truck drivers using geo-location data: a case study in China," *Accident Analysis & Prevention*, vol. 137, Article ID 105427, 2020.

[52] F. Chen and S. Chen, "Injury severities of truck drivers in single- and multi-vehicle accidents on rural highways," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1677–1688, 2011.

[53] F. Chen, M. Song, and X. Ma, "Investigation on the injury severity of drivers in rear-end collisions between cars using a random parameters bivariate ordered probit model," *International Journal of Environmental Research and Public Health*, vol. 16, no. 14, p. 2632, 2019.

[54] H. Midi, S. K. Sarkar, and S. Rana, "Collinearity diagnostics of binary logistic regression model," *Journal of Interdisciplinary Mathematics*, vol. 13, no. 3, pp. 253–267, 2010.

[55] M. Littner, C. A. Kushida, W. M. Anderson et al., "Practice parameters for the role of actigraphy in the study of sleep and circadian rhythms: an update for 2002," *Sleep*, vol. 26, no. 3, pp. 337–341, 2003.

[56] M. Y. Kiang, "A comparative assessment of classification methods," *Decision Support Systems*, vol. 35, no. 4, pp. 441–454, 2003.

[57] C. Qi, A. Fourie, X. Du, and X. Tang, "Prediction of open stope hangingwall stability using random forests," *Natural Hazards*, vol. 92, no. 2, pp. 1179–1197, 2018.

[58] D. R. Cutler Jr., T. C. Edwards, K. H. Beard et al., "Random forests for classification in ecology," *Ecology*, vol. 88, no. 11, pp. 2783–2792, 2007.

[59] F. K.-H. Chun, P. I. Karakiewicz, A. Briganti et al., "A critical appraisal of logistic regression-based nomograms, artificial neural networks, classification and regression-tree models, look-up tables and risk-group stratification models for prostate cancer," *BJU International*, vol. 99, no. 4, pp. 794–800, 2007.

[60] B. Swiderski, J. Kurek, and S. Osowski, "Multistage classification by using logistic regression and neural networks for assessment of financial condition of company," *Decision Support Systems*, vol. 52, no. 2, pp. 539–547, 2012.

[61] O. E. Askin and F. Gokalp, "Comparing the predictive and classification performances of logistic regression and neural networks: a case study on timss 2011," *Procedia-Social and Behavioral Sciences*, vol. 106, pp. 667–676, 2013.

[62] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*, vol. 52, pp. 456–462, 2019.

[63] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012.

[64] J.-H. Huang, H.-L. Xie, J. Yan, H.-M. Lu, Q.-S. Xu, and Y.-Z. Liang, "Using random forest to classify T-cell epitopes based on amino acid properties and molecular features," *Analytica Chimica Acta*, vol. 804, pp. 70–75, 2013.

[65] R. J. Marshall, "The use of classification and regression trees in clinical epidemiology," *Journal of Clinical Epidemiology*, vol. 54, no. 6, pp. 603–609, 2001.

[66] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery," *Remote Sensing of Environment*, vol. 118, pp. 259–272, 2012.

[67] J. Shreve, H. Schneider, and O. Soysal, "A methodology for comparing classification methods through the assessment of model stability and validity in variable selection," *Decision Support Systems*, vol. 52, no. 1, pp. 247–257, 2012.

[68] P. A. Flach, J. Hernández-Orallo, and C. F. Ramirez, "A coherent interpretation of AUC as a measure of aggregated classification performance," in *Proceedings of the 28th International Conference on Machine Learning*, pp. 657–664, Washington, DC, USA, July 2011.

[69] E. Gokgoz and A. Subasi, "Comparison of decision tree algorithms for EMG signal classification using DWT," *Biomedical Signal Processing and Control*, vol. 18, pp. 138–144, 2015.

[70] F. Chen, S. Chen, and X. Ma, "Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data," *Journal of Safety Research*, vol. 65, pp. 153–159, 2018.

[71] A. Subasi, "Classification of EMG signals using PSO optimized SVM for diagnosis of neuromuscular disorders," *Computers in Biology and Medicine*, vol. 43, no. 5, pp. 576–586, 2013.

[72] F.-f. Ai, J. Bin, Z.-m. Zhang et al., "Application of random forests to select premium quality vegetable oils by their fatty acid composition," *Food Chemistry*, vol. 143, pp. 472–478, 2014.

[73] J. A. Horne and L. A. Reyner, "Driver sleepiness and sleep disorders, with particular reference to Parkinson's disease," in *Proceedings of the Meeting on Causes and Management of Sleep Disturbances in Parkinsons Disease*, pp. 15–20, Cannes, France, January 2000.

[74] D. J. Beirness, H. M. Simpson, and K. Desmond, *The Road Safety Monitor 2004: Drowsy Driving*, Ottawa: Traffic Injury Research Foundation, Ottawa, Canada, 2005.

[75] A. I. Pack, A. M. Pack, E. Rodgman, A. Cucchiara, D. F. Dinges, and C. W. Schwab, "Characteristics of crashes attributed to the driver having fallen asleep," *Accident Analysis & Prevention*, vol. 27, no. 6, pp. 769–775, 1995.

[76] P.-H. Ting, J.-R. Hwang, J.-L. Doong, and M.-C. Jeng, "Driver fatigue and highway driving: a simulator study," *Physiology & Behavior*, vol. 94, no. 3, pp. 448–453, 2008.

[77] D. H. Li, Q. Liu, W. Yuan et al., "Relationship between fatigue driving and traffic accident," *Journal of Traffic and Transportation Engineering*, vol. 10, no. 2, pp. 104–109, 2010.

[78] A. T. Mccartt, S. A. Ribner, A. I. Pack, and M. C. Hammer, "The scope and nature of the drowsy driving problem in New York state," *Accident Analysis & Prevention*, vol. 28, no. 4, pp. 511–517, 1996.

[79] J.-Y. Yeh, T.-H. Wu, and C.-W. Tsao, "Using data mining techniques to predict hospitalization of hemodialysis patients," *Decision Support Systems*, vol. 50, no. 2, pp. 439–448, 2011.