Driver Drowsiness Detection based on Non-intrusive Metrics Considering Individual Specifics

Xuesong Wang, PhD
Professor
School of Transportation Engineering,
Tongji University
4800 Cao’an Road, Jiading District
Shanghai, 201804, China
Phone: +86-21-69583946
Fax: +86-21-69582897
Email: wangxs@tongji.edu.cn

Chuan Xu*
PhD student
School of Transportation Engineering
Tongji University
Room 338, Building of School of Transportation Engineering
4800 Cao’an Road, Shanghai, 201804, P. R. of China
Tel: +86 15216706076
Email: xuchuan7@gmail.com
(Corresponding Author)

Xiaohong Chen, PhD
Professor
School of Transportation Engineering
Tongji University
4800 Cao’an Road, Shanghai, 201804, P. R. of China
Tel) +86-21-65989270 Fax) +86-21-65982897
Email: chenxh@tongji.edu.cn

Word Count: 4732 (Text) +6 Figures+3 Tables=6982words

July 2014

Submitted for Presentation at the TRB Annual Meeting, and Publication in the Journal of the Transportation Research Board
ABSTRACT

Drowsy driving is a serious highway safety problem. If drivers could be warned before they became too drowsy to drive safely, some of these crashes could be prevented. Being able to reliably detect drowsiness depends on the presentation of timely warnings of drowsiness. To date, the effectiveness of drowsiness detection methods has been limited by their failure to consider individual differences. The present study sought to develop a drowsiness detection model that accommodates the varying effects of drowsiness on individual driving performance. Nineteen driving behavior variables and four eye feature variables were measured as participants drove a fixed road course in a high fidelity motion-based driving simulator after having worked an 8-hour night shift. During the test, participants were asked to report their drowsiness level using the Karolinska Sleepiness Scale (KSS) at the midpoint of each of the six rounds through the road course. A multilevel Ordered Logit model (MOL), an Ordered Logit model (OL), and an Artificial Neural Network model (ANN) were used to determine drowsiness. The MOL had the highest drowsiness detection accuracy, and this finding shows that consideration of individual differences improves the models’ ability to detect drowsiness. According to the results of the models, percentage of eyelid closure, average pupil diameter, standard deviation of lateral position and steering wheel reversals were most important in the 23 variables.

Keywords: Drowsiness Detection, Multilevel Ordered Logit Model, Non-Intrusive, Driving Behavior, Eye Feature, Driving Simulator
1 INTRODUCTION

Drowsy driving is a serious threat to road safety. In 2009, over 800 fatalities and 30,000 injuries from car crashes were attributed to drowsy driving in the United States (1). In Europe, an estimated 20% of traffic crashes are caused by drowsy driving (2). The problem seems more serious in China where, in 2007, 1768 fatalities were attributed to drowsy driving (3). Considering that as of 2012, China had over 85,000 kilometers of expressways (4), and the annual growth in kilometers exceeded 14% during 2002-2011, the problem requires immediate attention. A recent study using a naturalistic driving method, estimated the increase in crash risk associated with drowsy driving to be four times greater than when driving while alert (5). Besides increasing crash risk, drowsy driving crashes are often more severe than other crashes because they frequently occur on high speed expressways, and are frequently run-off-the-road crashes with no braking prior to impact.

Unlike drinking and driving, drowsy driving does not provide an objective measure of its occurrence, and therefore enforcement cannot be used to counter this problem (6). An alternative would be to notify the driver if he becomes too drowsy to drive safely. This requires the reliable detection of drowsy driving—a problem that has been extensively researched. Based on the type of data used, drowsiness detection can be conveniently separated into the two categories of intrusive and non-intrusive methods. Intrusive methods, such as electroencephalograms (EEGs) (7) or electrocardiograms (EKGs) (8), show good detection accuracy, however, are limited to the research laboratory. In contrast, methods based on non-intrusive measures detect drowsiness by measuring driving behavior and sometimes eye features, and so are useful for real world driving situations.

To date, non-intrusive methods have been less reliable than intrusive methods, partly because individual differences in non-intrusive measures have prevented the identification of the point at which drowsiness impairs driving (9; 10). In the previous research, individual differences of non-intrusive methods were frequently mentioned, for both driving behavior and eye features. For driving behavior, Ingre et al. (9) report the individual difference of standard deviation of lateral position (SDLP), an important driving behavior measure to drowsiness. In this research, for the same drowsiness level, different drivers have different SDLPs. Another study finds the individual difference of driver’s lane departure behavior (11). Thiffault and Bergeron (12) reported the individual difference of the standard deviation of steering wheel movements. Regarding eye features, individual differences of blink duration (9; 13), percentage of eyelid closure (PERCLOS) (14) were also observed in many studies. However, most drowsiness detection methods such as decision trees, logistic regression, Bayesian networks (15), artificial neural networks (8), and support vector machines (16) have not properly handled the problem of differences in the manifestation of drowsiness among individuals. Ignoring such differences reduces the accuracy and reliability of these models, especially for non-intrusive measures.

To address and increase the accuracy for drowsiness detection model based on non-intrusive measures, a multilevel logit model was built based on both driving
behavior measures and eye features, which detect drowsiness by using individual-specific criterions. To compare the detection accuracy, two non-individual specific models (i.e., ordered logit model and artificial neural network) were established. All the data in this research is based on a high fidelity driving simulator experiment, in which driving behavior, eye features, and subjective drowsiness were scaled and collected.

2 METHOD

2.1 Participants

Sixteen male participants aged 24-40 (mean 32.8, SD, 5.0) with valid Chinese drivers licenses were recruited from students and staff at Tongji University. They were required to be in good health, have no sleep related disorders, and not to have taken any pharmaceuticals within one month prior to entering the study. Subjects who had a history of motion sickness were screened out. All subjects provided written consent and were paid about 200 RMB Yuan, depending on the total time in the laboratory. During the experiment, one subject’s eye movement data was lost because of a technical problem. One subject fell asleep before completing the task, however, his data up to that point was used.

2.2 Apparatus

The Tongji University driving simulator is shown in FIGURE 1. This simulator, currently the most advanced in China, incorporates a fully instrumented Renault Megane III vehicle cab in a dome mounted on an 8 degree-of-freedom motion system with an X-Y range of 20 × 5 meters. An immersive 5 projector system provides a front image view of 250° × 40° at 1000 × 1050 resolution refreshed at 60 Hz. LCD monitors provide rear views at the central and side mirror positions. SCANeR® studio software presented the simulated roadway and controlled a force feedback system that acquired data from the steering wheel, pedals and gear shift lever.

FIGURE 1 Tongji advanced driving simulator

Eye movement data were recorded using a Smarteye® eye tracking system. The
cameras and interface of this system are shown in FIGURE 2. The system uses four
cameras located in the front of the vehicle to record the driver’s eye movements at 60
Hz sampling rate.

![FIGURE 2 Smarteye® eye tracking system hardware and software interface](image)

2.3 Procedure
2.3.1 Experiment design
All drivers received were presented with the same conditions in the same order. The
driving course, diagrammed in FIGURE 3, simulated a 20km ring shaped 6-lane rural
highway:, 3.75m in width composed of the straight segments numbered 1, 3, 5, 9, 11,
13; two circle curves each with a 700 meter radius, (segments numbered 7, 15); and
several transition curves (numbered 2,4,6,9,10,12,14,16). Only the straight line road
segments were used for the analysis. The length of both straight line and curve segments
was a total of 2 km. Grass and trees were placed beside the highway, as well as a few
small villages along the straight segments.

![FIGURE 3 Ring shaped highway and segment ID numbers](image)

To assess driving performance at high levels of drowsiness, night shift workers
were selected and tested just after shift completion around 8:00 a.m. Upon arrival at the
simulator facility, participants were asked to complete questionnaires on their basic
driver information and current levels of drowsiness, after which they spent five minutes
familiarizing themselves with driving the simulator. At about 8:30 a.m., they received
the main test, in which each subject was asked to drive and respect road rules for one
hour.
In order to induce drowsiness,
• The driving task was reduced to one lane, eliminating the need to change lanes
• Drivers were required to drive at a constant speed (120 km/h), eliminating the need
  for manual gear changes
• No radio or music was played
• No environmental disturbances (e.g. crosswinds) were introduced
All driving was during daytime periods and no tunnel or weather changes occurred eliminating the need to adjust headlights.

Only occasional and uneventful traffic was present.

After the main test, participants were asked to complete post experiment questionnaires on their levels of drowsiness.

2.3.2 Measurement of drowsiness—participants’ assessments

To track drivers’ drowsiness changes during the one hour driving task, the participants were asked to report their Karolinska Sleepiness Scale (KSS) level at the middle point of pass through the driving course. KSS is on a 9 point ordinal scale, however, it’s not necessary to distinguish all nine levels when using the KSS. It is, however necessary to identify the drowsiness level with a high crash risk. Several studies (15; 17) suggested that serious behavioral and physiological changes do not occur until KSS>=7. In addition, the description of KSS 7 is “sleep, but no effort to keep alert,” but KSS 8, 9 are described as “need some or great effort to keep alert.” Therefore, drowsiness in this study is categorized into three levels as follows:

- Level 1 (DL=1): KSS range from 1 to 6, no drowsiness or low-level of drowsiness;
- Level 2 (DL=2): KSS is 7, moderate-level of drowsiness;
- Level 3 (DL=3): KSS range from 8 to 9, high-level of drowsiness.

2.3.3 Measurement of the effects of drowsiness

Vehicle-based signals measuring driving behavior, with 10Hz sample frequency, were obtained from the system, including vehicle speed, lateral position and steering wheel angle. Based on these signals, several drowsiness related behavior indicators were extracted. Eye movement signals including eyelid opening and pupil diameter were also recorded, with 60Hz sample frequency. Using the Smarteye Pro software, eye blink was identified. Eye activity indicators including percentage of eye closure (PERCLOS), average pupil diameter, blink frequency, and average blink duration were calculated. The measures are summarized in TABLE 1 below.
### TABLE 1 Driving Behavior and Eye Feature Metrics

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Description of the Variables</th>
<th>Mean</th>
<th>S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Driving Behavior</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LP_stdev</td>
<td>Standard deviation of lateral position (m)</td>
<td>0.306</td>
<td>0.134</td>
</tr>
<tr>
<td>LP_avg</td>
<td>Average of lateral position (m)</td>
<td>0.214</td>
<td>0.269</td>
</tr>
<tr>
<td>LD_Area</td>
<td>Sum of lane departure time-space area (m•s)</td>
<td>1.627</td>
<td>4.207</td>
</tr>
<tr>
<td>LD_TArea</td>
<td>Sum of lane departure time-space area weighted by lane crossing time (m•s)</td>
<td>6.129</td>
<td>35.383</td>
</tr>
<tr>
<td>LD_Frequency</td>
<td>Lane departure frequency</td>
<td>0.660</td>
<td>1.118</td>
</tr>
<tr>
<td>LD_Speed</td>
<td>Lane departure lateral speed (m/s)</td>
<td>0.046</td>
<td>0.099</td>
</tr>
<tr>
<td>LD_Tc</td>
<td>Time percentage of lane crossing of the vehicle center</td>
<td>0.002</td>
<td>0.017</td>
</tr>
<tr>
<td>LD_Te</td>
<td>Time percentage of lane crossing of the vehicle edge</td>
<td>0.021</td>
<td>0.045</td>
</tr>
<tr>
<td>SW_Speed_stdev</td>
<td>Standard deviation of steering angular speed (degree/s)</td>
<td>0.012</td>
<td>0.008</td>
</tr>
<tr>
<td>SW_Area_MA</td>
<td>Area surrounded by steering angle and its moving average</td>
<td>0.440</td>
<td>0.257</td>
</tr>
<tr>
<td>SWM_Re</td>
<td>Steering wheel reversals</td>
<td>190.485</td>
<td>29.522</td>
</tr>
<tr>
<td>SW_Range_1</td>
<td>Percentage of steering speed in 0-2.5 degree/s</td>
<td>0.876</td>
<td>0.085</td>
</tr>
<tr>
<td>SW_Range_2</td>
<td>Percentage of steering speed in 2.5-5 degree/s</td>
<td>0.077</td>
<td>0.037</td>
</tr>
<tr>
<td>SW_Range_3</td>
<td>Percentage of steering speed in 5-7.5 degree/s</td>
<td>0.024</td>
<td>0.020</td>
</tr>
<tr>
<td>SW_Range_4</td>
<td>Percentage of steering speed in 7.5-10 degree/s</td>
<td>0.010</td>
<td>0.012</td>
</tr>
<tr>
<td>SW_Range_5</td>
<td>Percentage of steering speed exceeding 10 degree/s</td>
<td>0.013</td>
<td>0.024</td>
</tr>
<tr>
<td>Speed</td>
<td>Average speed (km/h)</td>
<td>117.424</td>
<td>6.650</td>
</tr>
<tr>
<td>Speed_stdev</td>
<td>Standard deviation of speed (km/h)</td>
<td>2.866</td>
<td>2.860</td>
</tr>
<tr>
<td>Speeding_T</td>
<td>Time percentage of speed exceeding the limit speed 120 km/h</td>
<td>0.311</td>
<td>0.372</td>
</tr>
<tr>
<td><strong>Eye Features</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blink_Frequency</td>
<td>Average blink frequency per second</td>
<td>0.504</td>
<td>0.318</td>
</tr>
<tr>
<td>Blink_duration</td>
<td>Average blink duration (second)</td>
<td>0.402</td>
<td>0.054</td>
</tr>
<tr>
<td>PERCLOS</td>
<td>Percentage of eyelid closure</td>
<td>0.132</td>
<td>0.099</td>
</tr>
<tr>
<td>Pupil</td>
<td>Average pupil diameter (mm)</td>
<td>3.807</td>
<td>0.894</td>
</tr>
</tbody>
</table>

### 2.3.4 Data analysis

Using DL as the dependent variable, and the driving behavior and eye feature metrics described in TABLE 1 as the independent variables, three models, individual-specific model multilevel ordered logit model, and two non-individual-specific models, an ordered logit model and neural network model.

The data set was divided into a training set and a validation set. The data in the training set was used to build the models, and the data in the validation set was used to test the models. To ensure each of the two data sets contained the data from every subject at every drowsiness level, the training set was generated by randomly selecting 70% of the data for each subject at each drowsiness level, and the rest of the data was assigned to the validation set. The individual specific multilevel ordered logit model was established first, and then the non-individual specific ordered logit model and neural network model were constructed using the same variables as the multilevel ordered logit model.

- **Multilevel Ordered Logit Model**

  Multilevel ordered logit models are often presented as cumulative logit models. Suppose an ordered DL$_{ij}$ is the drowsiness level for $i$th subject on the $j$th road segment. A latent continuous variable DL*$_{ij}$ is established as the unobserved measure of DL$_{ij}$. DL*$_{ij}$ is related to DL$_{ij}$ by a series of latent thresholds. Differing from the ordered logit...
model, the multilevel model accounts for each subject’s individual performance by using a set of variable thresholds specific to each subject: \( \gamma_{ki} \) (k=1, 2), see formula (2).

\[
DL_{ij} = \begin{cases} 
1 & \text{if } DL^*_{ij} < \gamma_{1i} \\
2 & \text{if } \gamma_{1i} < DL^*_{ij} < \gamma_{2i} \\
3 & \text{if } \gamma_{2i} < DL^*_{ij}
\end{cases}
\] (1)

The \( DL^*_{ij} \) can be written in the same form as the regular linear regression model.

\[
DL^*_{ij} = \theta_{ij} + \epsilon_{ij} \quad \text{and} \quad \theta_{ij} = \sum_{p=1}^P \beta_p x_{pij}
\] (2)

Where \( x_{pij} \) is the explanatory variable for \( i \)-th subject on \( j \)-th segment. \( \epsilon_{ij} \) is the disturbance term, which is assumed as a logistic distribution as the cumulative density function. Thus, the cumulative response probabilities of the ordinal DL may be denoted as:

\[
P_{ij(k)} = \Pr(DL^*_{ij} \leq k) = \Phi(\gamma_{ki} - \theta_{ij}) = \frac{\exp(\gamma_{ki} - \theta_{ij})}{1 + \exp(\gamma_{ki} - \theta_{ij})}, k = 1, 2
\] (3)

\[
\text{Logit}(P_{ij(k)}) = \log \left[ \frac{P_{ij(k)}}{1 - P_{ij(k)}} \right] = \log \left[ \frac{\Pr(DL^*_{ij} \leq k)}{\Pr(DL^*_{ij} \geq k)} \right] = \gamma_{ki} - \theta_{ij}, k = 1, 2
\] (4)

In order to accommodate differences among subjects, the thresholds \( \gamma_{ki} \) were specified as random effects.

\[
\gamma_{ki} = \gamma_k + b_i, k = 1, 2
\] (5)

Where the intercept \( \gamma_k \) represents a constant component for thresholds for all subjects. A random effect component \( b_i \) is formulated to accommodate the between-subject heterogeneities.

An Intra-class Correlation Coefficient (ICC) is normally defined to examine the proportion of specific subject-level variance:

\[
\text{ICC} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}
\] (6)

Where \( \sigma_b^2 \) is within group variance and \( \sigma_w^2 \) is between group variance. A value of ICC close to zero indicates there is a very small variation between the different subjects, and a model without multilevel structure is adequate for the data. Otherwise, a multilevel model would be preferred.

(2) Artificial Neural Network Model

Artificial Neural Networks (ANN), a popular class of computational intelligence models, have been widely applied to drowsiness detection, partly because of their ability to work with massive amounts of multi-dimensional data, their modeling flexibility, and their generally good predictive ability.

In this study, we built a feed-forward neural network with one hidden layer
consisting of the interconnection of neurons only between two adjacent layers. A back
propagation training method was used. Before modeling, the following standardization
procedure was carried out for each metric:

\[ x_i = \frac{x_i - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \]  

(7)

A basic computational element is called a node. Each node receives input from
an external source or from other nodes. Each input has an associated weight \( w_{ij} \),
which can be modified to model synaptic learning by the process of training. The input
of the \( j \)th node in the hidden layer is calculated as follows:

\[ Z_j = \sum_i w_{ij} (x_i + b_j) \quad i = 1,2, \ldots, N \quad j = 1,2, \ldots, M \]  

(8)

Where \( Z_j \) is the input to \( j \)th node in the hidden layer, \( w_{ij} \) is the weight of \( i \)th node in
the input layer to \( j \)th node in the hidden layer, \( x_j \) is the value of \( i \)th node in the input
layer, \( b_j \) is bias value for the \( j \)th node in the hidden layer, \( N \) is the number of nodes in
the input layer, and \( M \) is the number of nodes in the hidden layer.

The output of a node is decided by its input as well as the activation function.
Different activation functions such as sigmoid functions, hyperbolic tangent functions,
and logistic functions can be used. A hyperbolic tangent function was used as the
activation function of the hidden layer in our study. It was calculated as follows:

\[ H_j = f(Z_j) = \tanh(Z_j) = \frac{e^{2z_j} - 1}{e^{2z_j} + 1} \]  

(9)

Where \( H_j \) is the output of \( j \)th node in hidden layer.

In our study, we want the outputs of ANN to be interpretable as probabilities
for a categorical target variable (DL), for those outputs to lie between 0 and 1, and to
have a sum of 1. Therefore, a Softmax activation function is used for the output layer,
which is written as follows:

\[ O_k = \frac{\exp(Z_k)}{\sum_{m=1}^{c} \exp(Z_m)} \]  

(10)

Where \( O_k \) is the output of \( k \)th node in the output layer, \( c \) is the number of categories
for the target variable.

In the training process, the network output, in general, may not be equal to the
desired output. Therefore, the output error is calculated as the difference between the
network output and the desired output. If the output error does not satisfy the tolerance
level, the network modifies the connection weights \( w_{ij} \) according to the value of the
output error; then, training data is inputted again to the network and the network output
is calculated. The training cycle is continued until the network achieves the desired
tolerance level.
3 RESULTS

In the MOL model development, beginning with the 23 variables, each variable was tested for the statistical significance and the insignificant ones were eliminated. Among the variables studied, five variables were identified as significant, as judged by 95% credible interval (CI): PERCLOS, Pupil, Blink duration, SWM_Re, LP_stdev. Then, among the significant variables, the Pearson correlation coefficients were examined. Blink durations were highly correlated with PERCLOS, but PERCLOS are more significantly related to DL. Therefore, in the final model, Blink duration was also eliminated.

The results of the MOL model are shown in TABLE 2. Two eye feature metrics, one steering variability metric and one lane variability metric are used as the explanatory variables in the final model. Among these explanatory variables, the fixed effects of PERCLOS, LP_stdev and SWM_Re are positive, while Pupil was negative. The threshold $\gamma_1$ is significant ($t = -2.255$, sig.=0.025) and $\gamma_2$ is insignificant ($t = -0.427$, sig.=0.669). For the random effects, ICC of the data set (training set) is 0.402, which shows a large between-group heterogeneity and within-group homogeneity. Therefore, it can be inferred that if an ordered logit model was implemented without considering the random effects between subjects, the results may be biased and inaccurate. It is also implied that the drowsiness detection algorithm shouldn’t stay the same for different subjects.

The results of the OL model using the same explanatory variables with the MOL model are also shown in TABLE 2. All the explanatory variables are significant in 95% CI, and the coefficient for each variable shows different values but the same sign as that in MOL model. The threshold $\gamma_2$ is significant ($t = 2.742$, sig. = 0.006) and $\gamma_1$ is insignificant ($t = -0.786$, sig. = 0.432).
For ANN, PERCLOS, and Pupil, SWM_Re were standardized before input into the model. Three neurons are used in the hidden layer. Based on the ANN model, the importance of each variable were also calculated. The normalized importance of each variable is PERCLOS (100%), Pupil (74.5%), LP_stdev (65.2%), and SWM_Re (41.1%). This implies that eye feature metrics performed better than driving behavior metrics in drowsiness detection. The receiver operating characteristic (ROC) curve of ANN is also formed. The area under ROC curve for DL=3 (0.887) is larger than DL=1 (0.779) and DL=2 (0.647).

The summary of drowsiness detection accuracy is shown in TABLE 3. For each data set (train set and test set), the overall accuracy of the MOL model is the highest among the three models. However, the detection accuracy difference between the train set and test set for MOL is also the largest among the three models, which implies the MOL also has the largest generalization error. For the other two models, ANN performed better than the OL model in both the train set and test set. For the test set, the detection accuracy of MOL varies in a very small range across drowsiness levels (63.33% ~ 65.32%), while the other two models vary greatly (OL: 50.00%~55.95%; ANN: 47.92%~65.83%). The top three possible detection errors for the three models were the same, which are mistaking 1 for 2, mistaking 3 for 2 and mistaking 2 for 1. Mistaking 1 for 3 and mistaking 3 for 1 have the smallest possibility.

TABLE 2 MOL & OL Model Estimated Results

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Effect Estimate Mean</th>
<th>Effect Estimate S.D.</th>
<th>t</th>
<th>sig</th>
<th>95% Confidence Level Lower</th>
<th>95% Confidence Level Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOL Threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₁</td>
<td>-3.778</td>
<td>1.675</td>
<td>-2.255</td>
<td>0.025</td>
<td>-7.070</td>
<td>-0.486</td>
</tr>
<tr>
<td>γ₂</td>
<td>-0.712</td>
<td>1.666</td>
<td>-0.427</td>
<td>0.669</td>
<td>-3.986</td>
<td>2.562</td>
</tr>
<tr>
<td>Fixed effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERCLOS</td>
<td>5.226</td>
<td>2.129</td>
<td>2.455</td>
<td>0.014</td>
<td>1.043</td>
<td>9.410</td>
</tr>
<tr>
<td>Pupil</td>
<td>-1.780</td>
<td>0.318</td>
<td>-5.60</td>
<td>0.000</td>
<td>-2.403</td>
<td>-1.156</td>
</tr>
<tr>
<td>SWM_Re</td>
<td>0.010</td>
<td>0.005</td>
<td>2.182</td>
<td>0.030</td>
<td>0.001</td>
<td>0.020</td>
</tr>
<tr>
<td>LP_stdev</td>
<td>5.227</td>
<td>1.007</td>
<td>5.193</td>
<td>0.000</td>
<td>3.249</td>
<td>7.206</td>
</tr>
<tr>
<td>Random effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Between-subject variance</td>
<td></td>
<td>3.496</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-subject variance</td>
<td></td>
<td>5.195</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ICC</td>
<td>0.402</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OL Threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>γ₁</td>
<td>-0.545</td>
<td>0.694</td>
<td>-0.786</td>
<td>0.432</td>
<td>-1.907</td>
<td>0.816</td>
</tr>
<tr>
<td>γ₂</td>
<td>1.914</td>
<td>0.698</td>
<td>2.742</td>
<td>0.006</td>
<td>0.544</td>
<td>3.283</td>
</tr>
<tr>
<td>Fixed effects</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PERCLOS</td>
<td>9.891</td>
<td>1.441</td>
<td>6.865</td>
<td>0.000</td>
<td>7.063</td>
<td>12.719</td>
</tr>
<tr>
<td>Pupil</td>
<td>-1.065</td>
<td>0.121</td>
<td>-8.779</td>
<td>0.000</td>
<td>-1.303</td>
<td>-0.827</td>
</tr>
<tr>
<td>SWM_Re</td>
<td>0.011</td>
<td>0.003</td>
<td>3.212</td>
<td>0.001</td>
<td>0.004</td>
<td>0.017</td>
</tr>
<tr>
<td>LP_stdev</td>
<td>3.606</td>
<td>0.833</td>
<td>4.329</td>
<td>0.000</td>
<td>1.971</td>
<td>5.240</td>
</tr>
<tr>
<td>Model &amp; Dataset</td>
<td>Observed</td>
<td>Predicted</td>
<td>Correct</td>
<td>Overall Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td>----------</td>
<td>-----------</td>
<td>---------</td>
<td>-----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>67.90%</td>
<td>31.00%</td>
<td>1.20%</td>
<td>67.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>22.80%</td>
<td>68.90%</td>
<td>8.30%</td>
<td>68.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.90%</td>
<td>29.60%</td>
<td>68.50%</td>
<td>68.50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOL (Train set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>64.17%</td>
<td>30.28%</td>
<td>5.56%</td>
<td>64.17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>27.13%</td>
<td>63.33%</td>
<td>9.54%</td>
<td>63.33%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5.00%</td>
<td>29.68%</td>
<td>65.32%</td>
<td>65.32%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOL (Test set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>59.52%</td>
<td>39.29%</td>
<td>1.19%</td>
<td>59.52%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>40.00%</td>
<td>51.67%</td>
<td>8.33%</td>
<td>51.67%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.70%</td>
<td>43.52%</td>
<td>52.78%</td>
<td>52.78%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OL (Train set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>50.00%</td>
<td>50.00%</td>
<td>0.00%</td>
<td>50.00%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30.83%</td>
<td>52.50%</td>
<td>16.67%</td>
<td>52.50%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.19%</td>
<td>42.86%</td>
<td>55.95%</td>
<td>55.95%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OL (Test set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>54.17%</td>
<td>40.28%</td>
<td>5.56%</td>
<td>54.17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>26.67%</td>
<td>63.33%</td>
<td>10.00%</td>
<td>63.33%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9.09%</td>
<td>36.36%</td>
<td>54.55%</td>
<td>54.55%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANN (Train set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>47.92%</td>
<td>52.08%</td>
<td>0.00%</td>
<td>47.92%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>30.00%</td>
<td>65.83%</td>
<td>4.17%</td>
<td>65.83%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>12.20%</td>
<td>36.59%</td>
<td>51.22%</td>
<td>51.22%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANN (Test set)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### 4 DISCUSSION

In order to find out a group of suitable indicators, twenty-three non-intrusive indicators were developed in this study. Eight metrics are based on lane lateral position, eight metrics are based on steering wheel angel, three metrics are based on vehicle speed and four metrics are based on eye features. The MOL results indicate that the indicators group formed by PERCLOS, Pupil, SWM_Re, and LP_stdev are appropriate to detect drowsiness. To test whether the variables selected in MOL are acceptable, all twenty-three variables were input into an ANN model with 10 neurons in hidden layers. The importance of each variable is shown in FIGURE 5. In order of importance, PERCLOS, Pupil, Blink_duration, SWM_Re, and LP_Stdev were the most important variables to predict the likelihood of DL. With the exception of Blink_duration, which was eliminated because of the correlation with PERCLOS, these variables are also significant in the MOL model.
According to FIGURE 5, the top three important variables are all eye feature metrics. These results can be interpreted that eye feature metrics perform better than driving behavior metrics in drowsiness detection. The following test also verified this inference. After removing eye feature metrics, the detection accuracy of the ANN model for the test set was reduced to 45.8%. Meanwhile, keeping only eye feature metrics, the detection accuracy of the ANN model was reduced to 49.3%. Because of this, some drowsiness detection studies have used only eye feature metrics to detect drowsiness (10; 16). However, the detection accuracy of the ANN model using both metrics has the highest detection accuracy. Use of driving behavior metrics is also needed. Moreover, the eye features are often measured by camera and image processing, which may not be reliable. The driving behavior metrics can be a supplement to increase the reliability of the detection system.

For driving behavior metrics, both lane related and steering related metrics are important in the models. Among lane related metrics, the lane variability measure (LP_stdev) performed better than other lane departure measures in drowsiness detection models. The possible reason is the lane departure metrics only measure the features of lane departure events and the lane variability information is missed for the non-departure parts. SWM_Re is the most important variable among steering related metrics, which measures steering variability. Rapid steering wheel movement is suggested to be a drowsiness measurement (18). In this study, it is measured by SWM_Rang_5 and is not significant in the MOL model. Moreover, some studies (19) concluded the lane variability is highly correlated with steering variability. We calculated the Pearson correlation of SWM_Re and LP_stdev at 0.089 (sig. = 0.059), which indicates the
correlation between these variables is small.

In addition, previous research found there might be a curvilinear relationship between KSS and drowsiness metrics (9), with a stronger change at high KSS levels when compared with low KSS levels. In this study, among the four significant variables in the MOL model, the mean change between DL 3 and DL 2 are larger than that between DL 2 and DL 1 (see FIGURE 6). PERCLOS is a typical example and it is inferred to be sensitive to a high drowsiness level. This also might explain why it is the variable with the highest importance in the ANN model. Due to the larger change of metrics on higher DL, it can be inferred that higher drowsiness is easier to be detected, which is also verified by the detection accuracy of MOL and OL models. The correct detection rate for DL=3 is the highest.

In this study, the OL model can be considered as the basic model, while MOL and ANN models can be viewed as two improvements on the OL model. The MOL model’s improvement is considering individual difference, while the ANN’s improvement is insuring higher adaptability for the data. The results show MOL has the highest detection accuracy among the three models. This was attributed to using a series of individual specific thresholds, which were achieved by adding a random intercept in the subject level of MOL. That means the improvement by using an individual specific logit model is larger than by using a more complex algorithm with higher flexibility. However, in MOL, this random intercept for each driver is hard to predict. It can be decided only by a procedure like “training.” Driving experience,
gender, age, and other characteristic variables of the driver were attempted as explanations of individual differences, but no strong results were found. Yet the individual difference is still the main obstacle to increasing the drowsiness detection accuracy.

Some research that has classified drowsiness in only two levels (alert or drowsy) has achieved high detection accuracy. But two levels of drowsiness aren’t enough to support warning the driver before the crash risk becomes critical. Therefore, three levels of drowsiness were used in this study. However, the detection accuracy is highly related with lower numbers of drowsiness levels. If we used two levels of drowsiness (Level 1: KSS 1~7; Level 2: 8~9), the detection accuracy increases from 64.15% to 88.6% for the MOL model with the same variables, and the detection accuracy increases from 56.04 to 83.3% for ANN. It can be inferred that the more drowsiness levels to be classified, the lower detection accuracy we would get in the models. Therefore, when comparing the detection accuracy among detection models, the way to classify drowsiness levels should also be considered.

5 CONCLUSION AND RECOMMENDATIONS

Twenty-three non-intrusive metrics including driving behavior and eye feature metrics were evaluated in a simulated shift work study with a motion based high-fidelity driving simulator in a controlled laboratory environment. Based on these metrics, MOL was developed to detect three-level drowsiness. For comparison, two non-individual specific models, OL and ANN, were also established.

MOL has the highest detection accuracy. This may be attributed to using a series of individual specific criteria, which was achieved by adding a random intercept at the subject level. Among the twenty-three variables, PERCLOS, Pupil, LP_stdev and SWM_Re were significant in MOL and OL, and were also confirmed in ANN. Metrics of eye features performed better (showed higher importance) in the drowsiness detection models than other metrics, which was also verified using ANN by comparing the detection accuracy between eye features only and driving behaviors only. We also found higher DL is more easily detected because of higher heterogeneity between adjacent DLs.

Based on the analysis, using a user-specific method to detect driver drowsiness is recommended in order to address the inaccuracies caused by individual difference. In this study, the group characteristics (like age, gender) of participants were controlled. However, if group characteristics exist, we can build group-specific models that would simplify the model training process by group determinants. Therefore group characteristics are recommended for study. Also, because establishment of drowsiness level in this research was subjective, more accurate measures should be applied, for example, using EEG to determine drowsiness level.

ACKNOWLEDGEMENTS

Supported by Road and Traffic Engineering Key Laboratory of Tongji University, Ministry of Education.

REFERENCES

1. NHTSA’s National Center for Statistics and Analysis. Traffic Safety Facts: A Brief


