Identifying Crucial Indicators of Task Complexity and Coping Capacity Associated with Crash Risk through Machine Learning Techniques: A Comparative Study using On-Road and Simulator Data

Eva Michelaraki
PhD, Research Associate
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece, GR15773
Email: evamich@mail.ntua.gr

Thodoris Garefalakis
Research Associate
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece, GR15773
Email: tgarefalakis@mail.ntua.gr

Muhammad Wisal Khattak
Ph.D. Candidate, Research Associate
UHasselt, School for Transportation Sciences
Transportation Research Institute (IMOB), Agoralaan, 3590 - Diepenbeek, Belgium
Email: muhammadwisal.khattak@uhasselt.be

Muhammad Adnan
Associate Professor
UHasselt, School for Transportation Sciences
Transportation Research Institute (IMOB), Agoralaan, 3590 - Diepenbeek, Belgium
Email: muhammad.adnan@uhasselt.be

Evita Papazikou
PhD, Research Associate
Transport Safety Research Centre
School of Design and Creative Arts, LDS 1.20, Loughborough University, LE11 3TU, UK
Email: A.Papazikou@lboro.ac.uk

Rachel Talbot
PhD, Research Associate
Transport Safety Research Centre
School of Design and Creative Arts, LDS 1.20, Loughborough University, LE11 3TU, UK
Email: r.k.talbot@lboro.ac.uk

Christelle Al Haddad
PhD, Research Associate
Chair of Transportation Systems Engineering
Technical University of Munich, Arcisstrasse 21, 80333, Munich, Germany
Email: christelle.haddad@tum.de

Constantinos Antoniou
Professor
Chair of Transportation Systems Engineering
Technical University of Munich, Arcisstrasse 21, 80333, Munich, Germany

Email: c.antoniou@tum.de

Tom Brijs
Professor
School for Transportation Sciences
Transportation Research Institute (IMOB), Agoralaan, 3590 Diepenbeek, Belgium
Email: tom.brijs@uhasselt.be

George Yannis
Professor
Department of Transportation Planning and Engineering
National Technical University of Athens, Athens, Greece, GR15773
Email: geyannis@central.ntua.gr

Word Count: 6,489 words + 4 tables (250 words per table) = 7,489 words

Submitted: July 31, 2024

ABSTRACT

Task demand is the objective complexity of the task and arises out of a combination of features of the environment, the behavior of other road users, control and performance characteristics of the vehicle. On the other hand, coping capacity refers to the ability of drivers and road systems to manage and respond effectively to various challenges and stressful situations encountered while driving. The aim of this study was to identify crucial indicators of task complexity and coping capacity associated with crash risk through machine learning techniques. Towards that end, data from an on-road driving experiment (involving 135 drivers) along with data from a simulator experiment (involving 55 drivers) were collected and analysed. In order to fulfill these objectives, a feature importance algorithm extracted from Extreme Gradient Boosting (XGBoost) was used to evaluate the significance of variables on forecasting STZ. Additionally, a Neural Network model was implemented for real-time data prediction, taking into account the most important and significant risk indicators. Furthermore, a comprehensive assessment of the performance of three machine learning classifiers (i.e. Decision Trees, Random Forests and k-Nearest Neighbors) across two distinct datasets (i.e. on-road and simulator experiment dataset) was performed to predict STZ levels for headway. Results indicated that RF model outperformed the DT and kNN models across all metrics, making it the most effective for predicting headway with accuracy up to 90%. It was also revealed that Neural Networks demonstrated that the level of STZ can be predicted with an exceptional accuracy of up to 89.8%.

Keywords: Task Complexity, Coping Capacity, On-Road Experiment, Simulator Experiment, Machine Learning.

INTRODUCTION

The driving task can be characterised as the 'dynamic control task in which the driver has to select relevant information from a vast array of mainly visual inputs to make decisions and execute appropriate control responses' and 'drivers execute planned actions which are shaped by their expectations of the unfolding road, pedestrian and traffic scenario in front of them and the reality that they actually observe' *(1)*. Thus, it is partly determined by exogenous factors of the driving environment and partly by the driver's perception, planning and execution abilities.

Task complexity is related to the current status of the real-world context in which a vehicle is being operated. Since this context is consistent of various individual elements which, together, determine the complexity of the task imposed on the vehicle operator, a multi-dimensional approach in further operationalizing this concept is adopted.

Learning to drive demands a lot of practice before expert levels are reached. To begin with, task complexity is determined by goals that have to be reached by performance *(2, 3)*. The driving task is partly determined by the demands of the road environment, traffic restrictions, weather conditions and time of the day or location *(4)*. However, the complexity of the driving task is also associated with driver performance, such as harsh events, driving speeds and following distances or exposure indicators, such as distance travelled and total duration.

What the driver brings to the problem of managing task complexity is determined by the driver's upper limit of competence and their momentary capability. Firstly, competence refers to the driver's attainment in the range of skills broadly described as roadcraft, a concept which includes control skills, the ability to read the road (i.e. hazard detection and recognition), and anticipatory and defensive driving skills. Secondly, capability refers to the momentary ability of the driver to deliver their level of competence. It refers to what the driver actually is able to do at any given moment. This distinction is roughly equivalent to one made in different terms by previous studies *(5)*.

Coping capacity refers to the ability of drivers and road systems to manage and respond effectively to various challenges and stressful situations encountered while driving. This includes the resources, skills and strategies that drivers, vehicles, and infrastructure employ to ensure safe and efficient travel. This concept is dependent upon two underlying factors and it consists of several aspects of both vehicle and operator state. These are also multi-dimensional in nature.

The i-DREAMS project, funded by the European Commission Horizon 2020 initiative, aims to address these challenges by establishing, developing, testing, and validating a 'Safety Tolerance Zone' (STZ) to ensure safe driving behavior. By continuously monitoring risk factors associated with task complexity (e.g., traffic conditions and weather) and coping capacity (e.g., driver's mental state, driving behavior, and vehicle status), i-DREAMS aims to determine the appropriate level within the STZ and implement interventions to maintain drivers' operations within acceptable safety limits. The STZ comprises three levels: 'Normal', 'Dangerous', and 'Avoidable Accident'. The 'Normal' level indicates a low likelihood of a crash, while the 'Dangerous' level suggests an increased possibility of a crash without inevitability. The 'Avoidable Accident' level signifies a high probability of a crash, but it also allows sufficient time for drivers to take action and prevent it.

Taking all the aforementioned into account, the aim of this study was to identify crucial indicators of task complexity and coping capacity associated with crash risk through machine learning techniques. A comparative assessment between on-road and simulator data was provided. Towards that end, data collected from an on-road driving experiment (involving 135 drivers) along with data collected from a simulator experiment (involving 55 drivers) were analysed.

In order to fulfill these objectives, a feature importance analysis (e.g. XGBoost) was implemented in order to evaluate the significance of various variables in forecasting STZ levels in terms of headway. This approach allowed for the selection of the most appropriate independent variables, ensuring that the most influential factors were identified and prioritized in the analysis. Then, machine learning analysis (e.g. Neural Networks) was applied to make accurate and data-driven predictions by identifying complex patterns between task complexity and coping capacity on crash risk. Furthermore, a comprehensive assessment of the performance of three machine learning classifiers (i.e. Decision Trees, Random Forests and k-Nearest Neighbors) across the abovementioned distinct datasets (i.e. on-road and simulator experiment dataset) was performed to predict STZ levels for headway.

The paper is structured in the following manner. Firstly, a general overview of motivation and objectives is highlighted. Then, the collection process (i.e. on-road and simulator experiments) and the processing of the dataset are described. The research methodology is outlined, including the explananation of collecting the data and the theoretical foundations of the underlying models employed. Finally, the results of the study are presented, followed by significant conclusions regarding the relationship between key factors of task complexity and coping capacity on risk.

THE EXPERIMENTS

To begin with, the participant selection criteria for both on-road and simulator experiments required drivers to have at least 10,000 km annual mileage, be at least 18 years old, hold a valid driving license and vehicle insurance, and be mentally fit to drive. Gender representation aimed for at least 40% per gender, with a preference for equal division. On-road trials additionally required participants to have a mixed driving pattern with at least 20% exposure to urban, rural and highway environments.

On-road experiment

For the purpose of this analysis, an on-road driving experiment was carried out involving 135 car drivers (with total duration of 4 months) and a large database of 31,954 trips was collected and analysed. The most prominent driving behavior indicators, including speeding, headway, duration, distance and harsh events were assessed.

With regards to the on-road experiment, the field trials were structured into four phases, as depicted in **Figure 1**. In particular, Phase 1 served as a reference period where driving behavior was monitored without any interventions. Phase 2 involved a period of monitoring where only real-time warnings from Advanced Driver Assistance Systems (ADAS) were provided inside the vehicle. In phase 3, these in-vehicle warnings were supplemented with feedback delivered via a smartphone app, while phase 4 introduced gamification features in the app, supported by a web dashboard.

Figure 1 Four phases of the on-road experiment

Simulator experiment

Supplementary to the on-road experiment, a simulator driving experiment was carried out involving 55 drivers (with total duration of 2 months) and a database consisting of 165 trips (55 drivers x 3 driving scenarios) was created. The simulator trials consisted of three phases, as shown in **Figure 2**. Two practice scenarios were developed for all participants in order to become familiarised with the simulator, each one with a total duration of 5-10 minutes. The first practice scenario did not contain traffic situations. In this way, participants became acquainted with driving through a scenario (e.g. visual environment, use of the mock-up). The second practice scenario contained traffic situations (e.g. intersection with a stop sign) requiring the execution of simple manoeuvres in order to become more acquainted with the driving simulator (e.g. use of pedals, steering wheel to manage safety margins while driving).

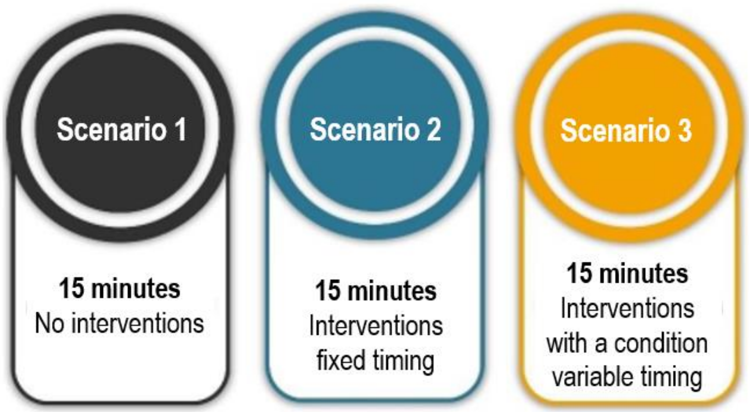

Figure 2 Three scenarios of the simulator experiment

The intervention drives then followed using a series of high-risk scenarios. The experimental scenarios focused on speeding, headway and fatigue as a modifying condition. Risk factors were investigated through a series of risky events tested during the drive-1, drive-2, and drive-3 scenarios. In total, the trials consisted of three 15-minute drives, including a baseline monitoring scenario followed by two intervention scenarios, one with fixed timing warnings and one with variable timing warnings and the inclusion of a condition (i.e. fatigue). There was a break between the two intervention scenarios. If, for any logistical reason, the break was extended, for example, if session 1 and session 2 were conducted on different days, or if a break was required between scenarios, a practice drive was completed if needed to ensure participants were re-familiarized. In addition, the order of sessions, scenarios, and events within trials was randomized. Simulation sickness is an important consideration when conducting simulator trials. Therefore, participants were screened for sickness throughout the simulator trials, and the trials were stopped if symptoms of simulation sickness were apparent, or the participant reported feeling unwell.

A **custom car simulator** developed by DriveSimSolutions was designed, as shown in **Figure 3**. The simulator is based on a Peugeot 206 and uses many Original Equipment Manufacturer (OEM) parts, such as the complete dashboard, a working instrument cluster and driving seat to recreate the cockpit of the actual vehicle. The simulator uses fully customizable STISIM Drive 3 software, allowing for creation of custom scenarios and data collection at every simulation update frame. It is also visualized on a triple monitor setup consisting of three 49 inch 4K monitors, providing an 135° field of view.

Figure 3 Car simulator developed by DriveSimSolutions, using OEM Peugeot 206 parts

Figure 4 depicts an overview of an intersection in STISIM Drive 3 for an example of a road environment.

Figure 4 Example of an intersection in STISIM Drive 3

Participants were also requested to fill in a driving behavior questionnaire in order to gather comprehensive data on various aspects of driving, socio-demographic background, safety attitudes and psychology. More specifically, the questionnaire included personal and vehicle details, current use of and opinions on different ADAS, driving style and confidence, opinions on driving and safety, self-assessment of driver's risk-taking behaviors, history of crashes and traffic violations, fatigue and sleepiness during driving, and health and medical conditions.

Overview of the variables used

A vast library of data from on-road and simulator experiments was created in order to investigate the most prominent driving behavior indicators available. From task complexity, the variables used were time indicator and wipers, for coping capacity – vehicle state, the most appropriate indicators found to be fuel type, vehicle age and gearbox, while for coping capacity – operator state, the performance measures included speeding, headway, overtaking, duration, distance and harsh events (i.e. harsh acceleration and harsh braking), gender and age.

METHODS

Neural Networks (NNs)

Neural Networks (NNs) represent a powerful computational model capable of capturing complex non-linear patterns within datasets *(6)*. These networks emulate the parallel processing of human neurons and are commonly employed in classification tasks. The architecture used, known as the multi-layer perceptron NN, is composed of three essential layers: an input layer, one or more hidden layers, and an output layer. In the context of analysing risky driving behavior, the input layer functions as the initial data receiver, encompassing various driving attributes, like vehicle speed, acceleration and headway.

The hidden layer, featuring a variable number of neurons, conducts computations by combining weighted inputs from these attributes. Each neuron in the hidden layer is equipped with an activation function, introducing the necessary non-linearity to the model. This non-linearity is crucial for capturing intricate patterns and relationships between these attributes and the target variable, which, in this work, pertains to different levels of risky driving behavior. The determination of the number of neurons in the hidden layer often involves experimentation, as it significantly impacts the net-work's capacity to learn and generalise. Simple problems may require just one hidden layer, whereas more complex tasks might demand multiple hidden layers. Moving to the output layer, it serves as the central hub for consolidating information from the hidden neurons to generate the final output of the network. In the context of this classification task regarding risky driving behavior, the output layer includes multiple neurons, each corresponding to distinct classes or levels of risk.

Decision Trees (DTs)

Decision Tree (DT) is a supervised learning technique used for both classification and regression problems, although they are most commonly preferred for solving classification problems. The algorithm is structured as a tree, where internal nodes represent the features of a dataset, branches represent decision rules, and each leaf node represents an outcome *(7)*. A DT consists of two types of nodes: decision nodes and leaf nodes. Decision nodes are used to make decisions and have multiple branches, whereas leaf nodes represent the output of those decisions and do not contain any further branches *(8)*.

The DT provides a graphical representation for getting all the possible solutions to a problem or decision based on given conditions. It is called a decision tree because it starts with the root node, which expands into further branches, constructing a tree-like structure. A decision tree simply asks a question, and based on the answer (Yes/No), it further splits the tree into subtrees. The tree splits by asking questions at each node, where the answer (typically yes or no) determines the next branch to follow. This splitting continues until a stopping criterion is met, such as all data points in a node belonging to the same class or reaching a pre-defined depth of the tree.

In particular, the root node is where the decision tree begins. It represents the entire dataset, which is then divided into two or more homogeneous sets. The leaf nodes are the final output nodes, and the tree cannot be further divided once it reaches a leaf node. Splitting is the process of dividing the decision node or root node into sub-nodes based on given conditions. A branch or sub-tree is formed by splitting the tree into smaller sections.

Random Forests (RFs)

Random Forest (RF) is a powerful tree learning technique in machine learning. It combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have led to widespread adoption, as it effectively handles both classification and regression problems *(9)*. Ensemble learning methods, like RFs, are composed of a set of classifiers, such as DTs, whose predictions are aggregated to identify the most popular result. The most well-known ensemble methods include bagging, also known as bootstrap aggregation and boosting.

In RF, each tree is constructed using a random subset of the data set to measure a random subset of features in each partition. This randomness introduces variability among individual trees, reducing the risk of overfitting and improving overall prediction performance. In prediction, the algorithm aggregates the results of all trees, either by voting (for classification tasks) or by averaging (for regression tasks) This collaborative decision-making process, supported by multiple trees with their insights, provides an example stable and precise results. RFs are widely used for classification and regression functions, which are known for their ability to handle complex data, reduce overfitting, and provide reliable forecasts in different environments.

Breiman *(10)* introduced the bagging method, where a random sample of data in a training set is selected with replacement, allowing individual data points to be chosen more than once. After generating several data samples, these models are trained independently. Depending on the task, regression or classification, the average or majority of these predictions yields a more accurate estimate. This approach is commonly used to reduce variance within a noisy dataset.

K-Nearest Neighbors (kNNs)

The k-Nearest Neighbors (kNN) algorithm is a popular machine learning technique also used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values. During the training phase, the kNN algorithm stores the entire training dataset as a reference. When making predictions, it calculates the distance between the two points of the input data and all the training examples, using a chosen distance metric.

In the case of classification, the algorithm assigns the most common class label among the K Neighbors as the predicted label for the input data point. As per regression, it calculates the average or weighted average of the target values of the K Neighbors to predict the value for the input data point. It should be noted that the kNN algorithm is straightforward and easy to understand, making it a popular choice in various domains. However, its performance can be affected by the choice of K and the distance metric, so careful parameter tuning is necessary for optimal results.

The accuracy of the kNN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not consistent with their importance. Much research effort has been put into selecting or scaling features to improve classification. A particularly popular approach is the use of evolutionary algorithms to optimize feature scaling *(11)*.

Model evaluation metrics

In order to compare the classification performance of the several configurations, well-established model evaluation metrics were calculated. The following metrics were utilized, based on the confusion matrix, which provides True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) metrics. The classification algorithms were evaluated using the accuracy, precision, recall, f1-score, and false alarm rate as defined below.

Accuracy, which represents the proportion of correctly classified observations, is defined as:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

Precision, which quantifies the number of positive class predictions that actually belong to the positive class, is defined as follows:

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall, also known as True Positive Rate, is defined as follows:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

F1-score, which combines precision and recall into a single measure, is defined as follows:

$$f1 - score = \frac{2x\,(Precision)x\,(Recall)}{(Precision)+(Recall)} \quad (4)$$

False alarm rate is defined as follows:

$$False\ Alarm\ Rate = \frac{FP}{FP+TN} \quad (5)$$

RESULTS

The structure methodology along with the proposed characteristics to estimate the STZ headway is depicted in **Figure 5**.

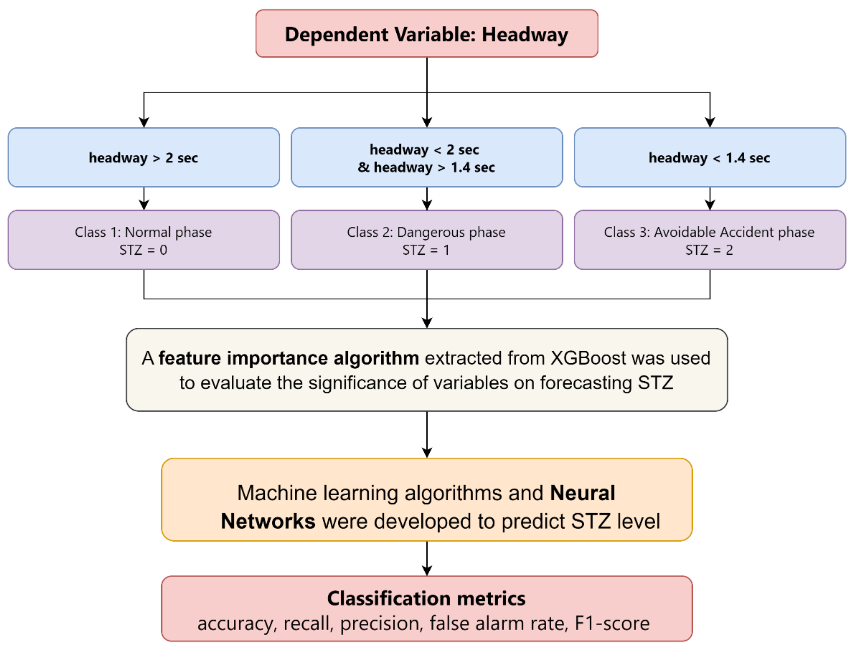

Figure 5 Proposed methodology for the definition of the STZ headway

On-road driving analyses

A feature importance algorithm derived from Extreme Gradient Boosting (XGBoost) was implemented in order to evaluate the significance of various variables in forecasting STZ. This approach allowed for the selection of the most appropriate independent variables, ensuring that the most influential factors were identified and prioritized in the analysis.

It was revealed that duration, average speed, vehicle age, time indicator, overtaking, gearbox, forward collision warning and car wipers found to be the most influential factors among all examined indicators. Conversely, parameters such as pedestrian collision warning, harsh events (i.e. harsh acceleration and harsh braking) and gender were less significant. Lastly, variables related to distance travelled and fuel type had a negligible impact on STZ headway. **Figure 6** provides an overview of the feature importance of independent variables for headway based on XGBoost algorithm.

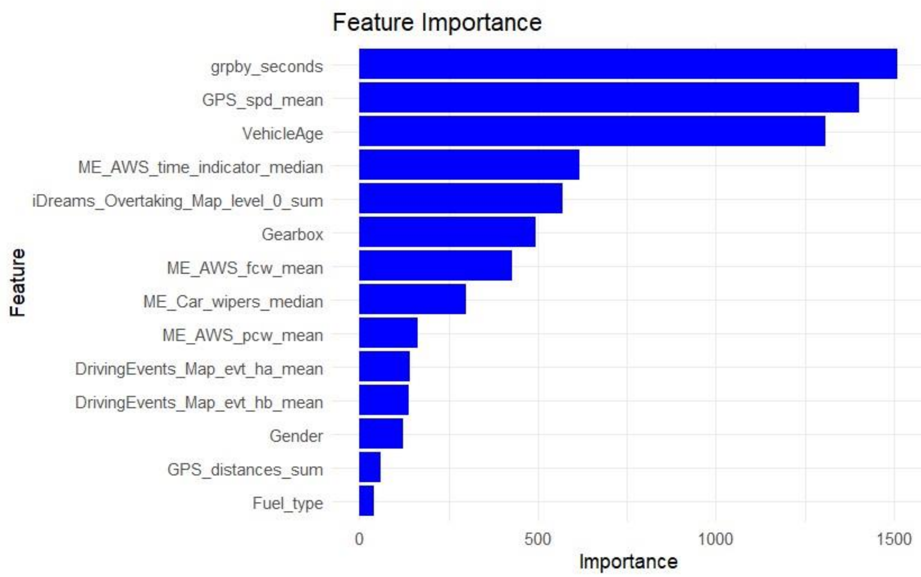

Figure 6 XGBoost feature importance of independent variables for headway

Based on the feature importance and the significance of the relevant indicators, a dataset of 998,358 rows from the on-road experiment was used and a feed-forward multilayer perceptron NN model was implemented. There were ten neurons in the input layer (i.e. distance travelled, duration, headway, harsh acceleration, harsh braking, time indicator, gearbox, fuel type, gender and wipers) and three neurons in the output layer (i.e. STZ1, STZ2, STZ3), as shown in **Figure 7**. It should be noted that STZ1 headway refers to normal phase, STZ2 headway refers to danger phase, while STZ3 headway refers to avoidable accident phase.

The model was run with deep neural networks, making use of two hidden layers (represented by circles in the middle of the diagram) where the computations take place. Each hidden layer node receives inputs from the previous layer, processes them, and passes the output to the next layer. The connections between nodes have weights (shown as numbers on the connecting lines), which are adjusted during training to minimize prediction error. In addition, weights represent the strength of the connection between nodes. Positive weights (black lines) indicate a positive influence, while negative weights (blue lines) indicate a negative influence on the connected nodes. **Figure 7** illustrates the NN model used to predict STZ headway based on various input features.

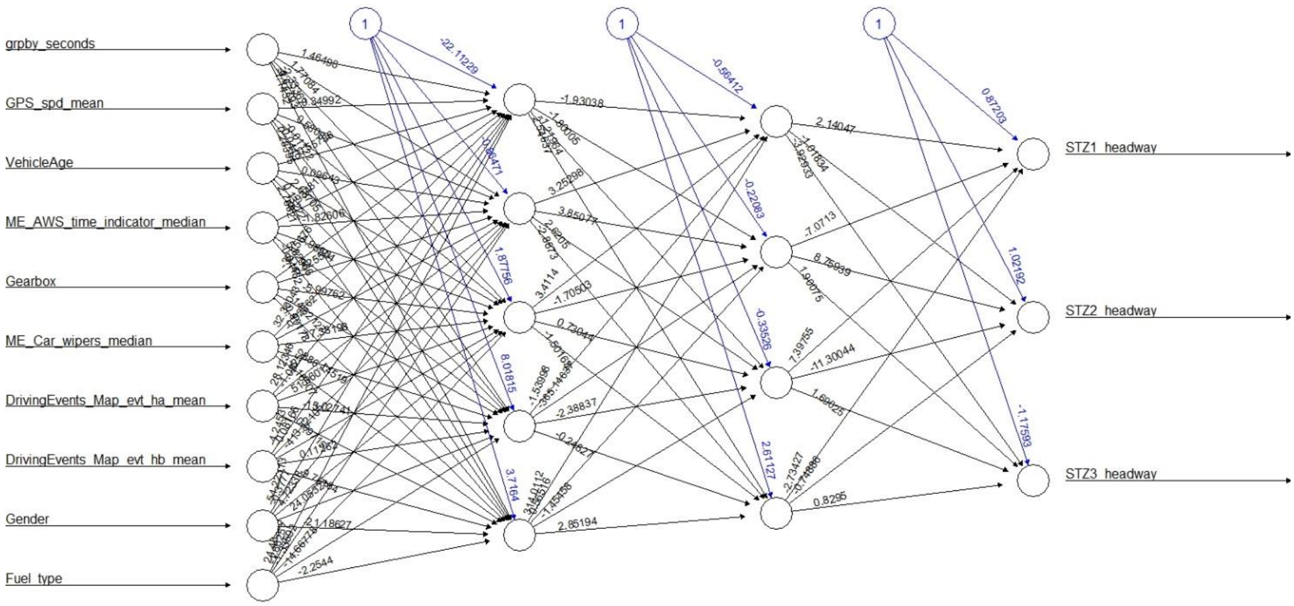

Figure 7 The multi-layer Neural Network model layout for STZ headway – on-road experiment

Based on the confusion matrix calculated, the variable STZ headway holds the result of dividing the sum of True Positives and True Negatives over the sum of all values in the matrix. The data were split into 80% train and 20% test in order to evaluate the models. The confusion matrix illustrates the classification performance for three classes: class 0 (normal level), class 1 (dangerous level), and class 2 (avoidable accident level), as shown in **Figure 8**. The model correctly classified 8,343 instances (31.75%) in the normal phase, 6,129 instances (23.33%) in the dangerous phase, and 6,986 instances (26.59%) in the avoidable accident phase. Misclassifications include 1,397 instances (5.32%) in the normal phase misclassified as dangerous, and 642 instances (2.44%) as avoidable accidents. For the dangerous phase, 1,284 instances (4.89%) were misclassified as normal, and 214 (0.81%) as avoidable accidents. In the avoidable accident phase, 570 instances (2.17%) were misclassified as normal, and 711 instances (2.71%) as dangerous. Overall, the model shows a reasonable performance, with the highest accuracy in the normal and avoidable accident phases.

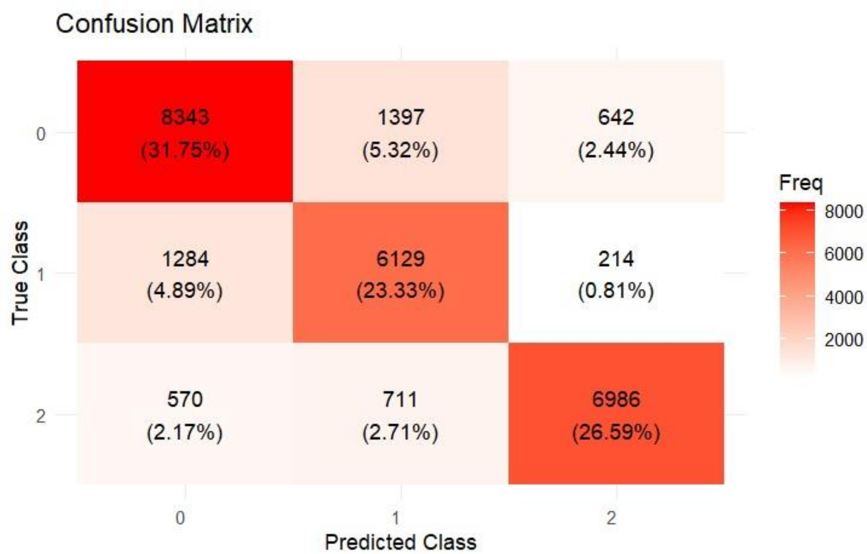

Figure 8 Confusion matrix for the test dataset for Neural Networks – headway

Table 1 provides the assessment of classification model for headway. Focusing on the results of all classes combined, the classifiers achieve 81.7% accuracy, 80.8% precision, 83.4% recall and an F1-score of 81.9%. The overall accuracy indicates that the model is 81.7% accurate in making correct predictions. The precision of 80.8% shows that the model is highly accurate regarding positive samples. The recall of 83.4% demonstrates the model's ability to detect safety-critical classes (i.e., "dangerous" and "avoidable accident") effectively. It should be noted that normal driving comprised the majority of the data, predicted with 86.3% accuracy, 84.5% precision and 89.1% recall. The dangerous driving classification showed 80.5% precision and 81.8% recall, while the avoidable accident phase presented the lowest rates, with 80.3% precision and 74.4% recall. Overall, these findings indicate that the NN model can adequately predict the STZ for headway.

TABLE 1 Evaluation metrics for NN for headway

Model Fit measures	0	1	2	Total
Accuracy	0.863	0.852	0.819	**0.817**
Precision	0.845	0.805	0.803	**0.808**
Recall	0.891	0.818	0.744	**0.834**
F1 Score	0.867	0.811	0.773	**0.819**
False alarm rate	0.317	0.413	0.348	**0.392**

**0 refers to normal phase, 1 refers to dangerous phase, 2 refers to avoidable accident phase*

Table 2 presents the evaluation metrics for each level of headway (0: normal, 1: dangerous, 2: avoidable accident) along with the total values. The comparison of classification models (DT, RF, kNN) for predicting headway reveals key differences in performance metrics across three phases (0: normal, 1: dangerous, and 2: avoidable accident). The RF model outperforms the others, achieving the highest overall accuracy (86.9%), precision (88.7%), recall (90.7%), and F1 score (85.5%). The DT model shows moderate performance, with an overall accuracy of 84.4%, precision of 84.6%, recall of 85.6%, and F1 score of 83.8%. The kNN model has the lowest performance, with an overall accuracy of 78.4%, precision of 72.0%, recall of 74.9%, and F1 score of 71.7%. These results highlight that RF is the most effective classifier for predicting headway, followed by DT, with kNN being the least effective.

TABLE 2 Evaluation metrics for classification models for headway

Model Fit measures	0	1	2	Total
Accuracy				
DT	0.858	0.847	0.826	**0.844**
RF	0.890	0.867	0.849	**0.869**
kNN	0.817	0.784	0.751	**0.784**
Precision				
DT	0.842	0.871	0.841	**0.846**
RF	0.903	0.888	0.852	**0.887**
kNN	0.787	0.737	0.716	**0.720**
Recall				
DT	0.897	0.857	0.801	**0.856**
RF	0.936	0.875	0.861	**0.907**
kNN	0.772	0.746	0.697	**0.749**
F1 Score				
DT	0.863	0.846	0.817	**0.838**
RF	0.879	0.857	0.828	**0.855**
kNN	0.752	0.734	0.693	**0.717**

**0 refers to normal phase, 1 refers to dangerous phase, 2 refers to avoidable accident phase*

Figure 9 presents the comparison of classifier metrics of the three machine learning techniques for headway. In summary, RF exhibits the best performance, leading in accuracy, precision, and F1 score, while showing competitive recall scores. DT and kNN show similar performance, though kNN tends to lag slightly behind in precision.

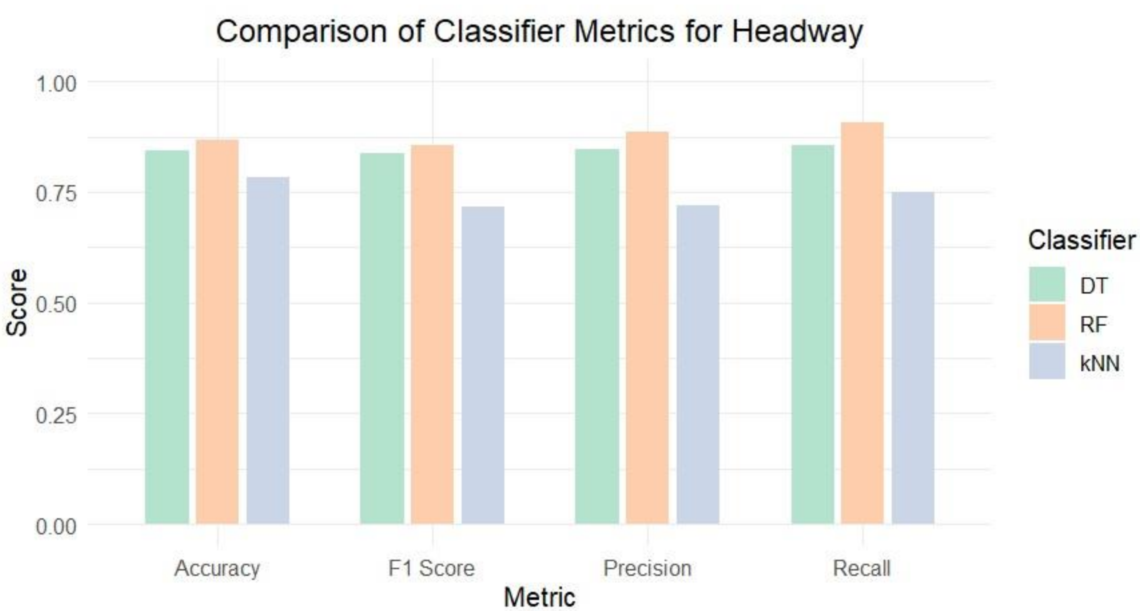

Figure 9 Comparison of classifier metrics of machine learning techniques for headway

Simulator analyses

With regards to headway, it was revealed that TTC, average speed, duration, hands-on event and fatigue found to be the most influential factors among all examined indicators. Conversely, parameters such as LDW was less significant, while FCW had a negligible impact on STZ headway. **Figure 10** provides an overview of the feature importance of independent variables for headway based on XGBoost algorithm.

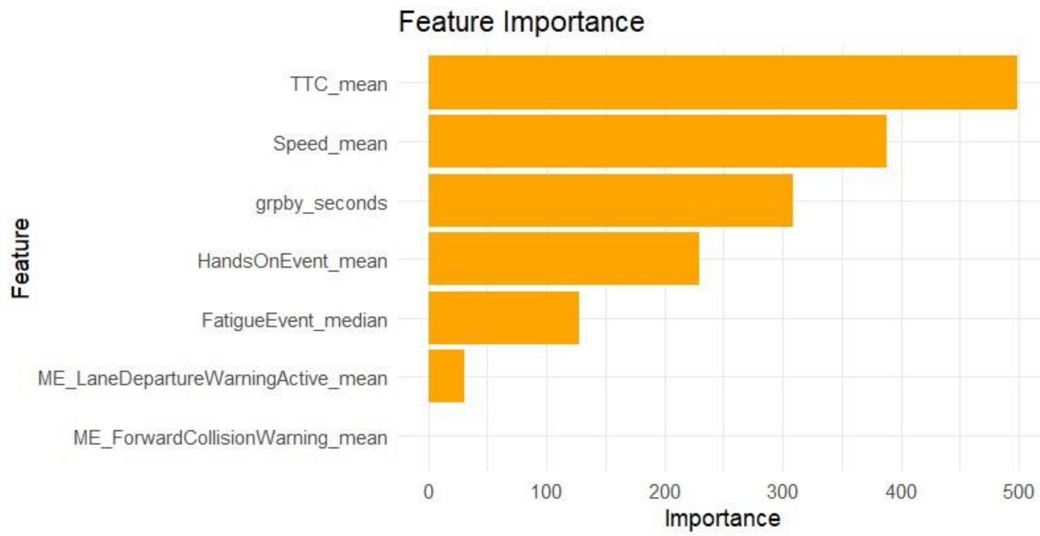

Figure 10 XGBoost feature importance of independent variables for headway

Based on the feature importance and the significance of the relevant indicators, a dataset of 745,251 rows from the simulator experiment was used and a feed-forward multilayer perceptron NN model was implemented. The multi-layer NN model applied consisted of five neurons in the input layer (i.e. TTC, average speed, duration, hands-on event and LDW) and three neurons in the output layer (i.e. STZ1, STZ2, STZ3). It should be noted that STZ1 headway refers to normal phase, STZ2 headway refers to danger phase, while STZ3 headway refers to avoidable accident phase. The model was run with deep neural networks, making use of two hidden layers (represented by circles in the middle of the diagram) where the computations take place. Positive weights (black lines) indicate a positive influence, while negative weights (blue lines) indicate a negative influence on the connected nodes. **Figure 11** illustrates the NN model used to predict STZ headway based on various input features.

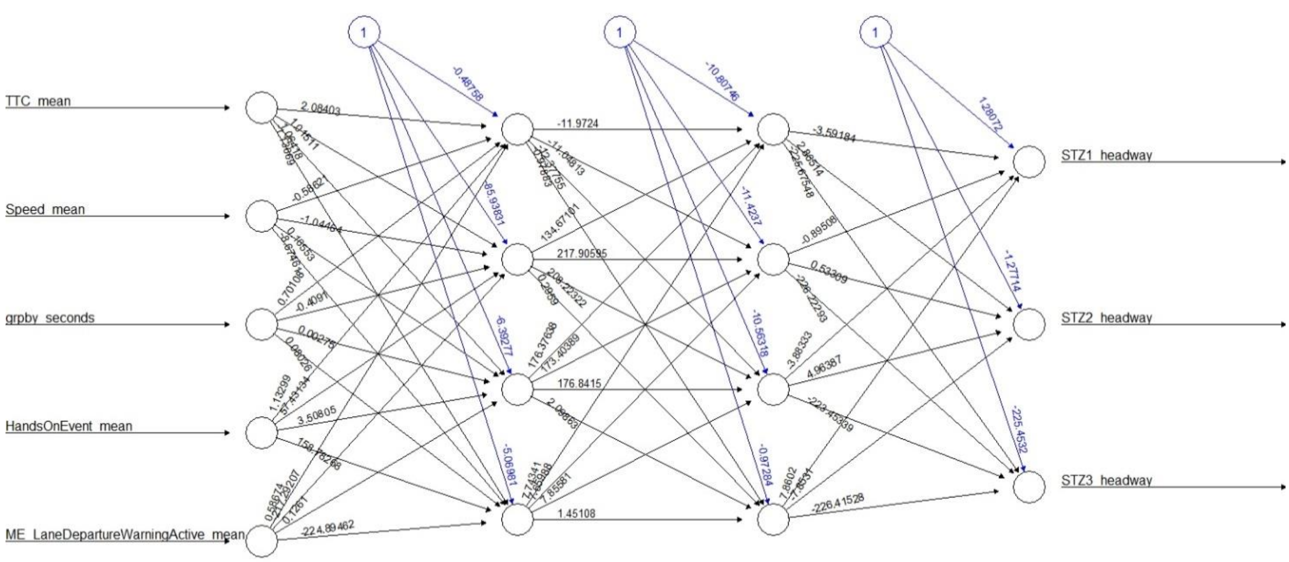

Figure 11 The multi-layer Neural Network model layout for STZ headway – simulator experiment

Based on the confusion matrix calculated, the variable STZ headway holds the result of dividing the sum of True Positives and True Negatives over the sum of all values in the matrix. The confusion matrix illustrates the classification performance for three classes: class 0 (normal level), class 1 (dangerous level), and class 2 (avoidable accident level), as shown in **Figure 12.**

For the normal class, 408 instances (35.02%) were correctly classified, while 7 instances (0.6%) were misclassified as dangerous, and 69 instances (5.92%) as avoidable accidents. In the dangerous class, 334 instances (28.67%) were correctly identified, with 12 instances (1.03%) misclassified as normal and 7 instances (0.6%) as avoidable accidents. For the avoidable accident class, 266 instances (22.83%) were correctly classified, while 58 instances (4.98%) were misclassified as normal, and 4 instances (0.34%) as dangerous. Overall, the model shows reasonable performance with the highest accuracy in classifying the normal phase, despite some misclassifications, particularly between the normal and avoidable accident phases. The model demonstrates a balanced performance but could improve in distinguishing between similar classes.

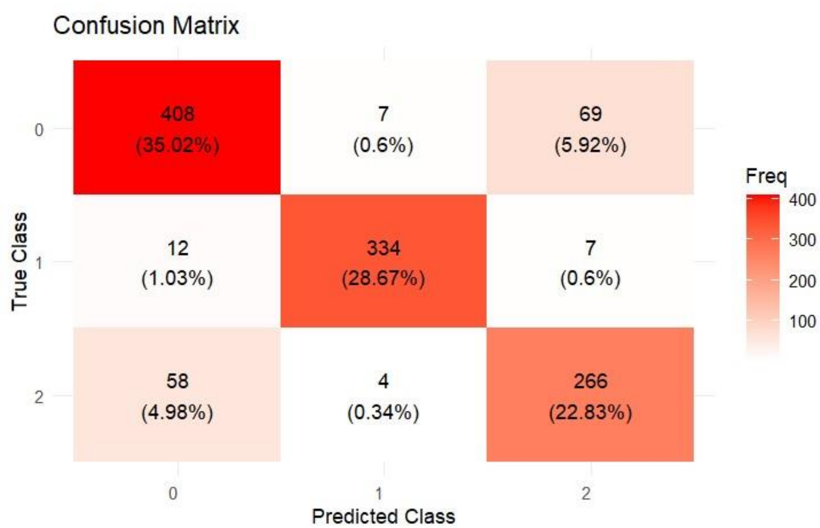

Figure 12 Confusion matrix for the test dataset for Neural Networks – headway

TABLE 3 Evaluation metrics for NN for headway

Model Fit measures	0	1	2	Total
Accuracy	0.907	0.973	0.915	**0.898**
Precision	0.876	0.968	0.853	**0.912**
Recall	0.899	0.946	0.842	**0.906**
F1 Score	0.887	0.957	0.847	**0.899**
False alarm rate	0.287	0.114	0.257	**0.153**

**0 refers to normal phase, 1 refers to dangerous phase, 2 refers to avoidable accident phase*

Table 3 provides the assessment of classification model for headway. The overall model metrics are as follows: an accuracy of 89.8%, precision of 91.2%, recall of 90.6%, F1 score of 89.9%, and a false alarm rate of 5.3%. The overall accuracy indicates that the model is 89.8% accurate in making correct predictions. The precision of 91.2% shows that the model is highly accurate regarding positive samples. The recall of 90.6% demonstrates the model's ability to detect safety-critical classes (i.e., "dangerous" and "avoidable accident") effectively. It should be noted that normal driving comprised the majority of the data, predicted with 90.7% accuracy, 87.6% precision, and 89.9% recall. The dangerous driving classification showed 96.8% precision and 94.6% recall, while the avoidable accident phase presented the lowest rates, with 85.3% precision and 84.2% recall.

Similar to the on-road experiment dataset, the same machine learning classifiers, namely DT, RF and kNN were developed and a comprehensive evaluation of their performance was implemented. Recognizing the limitations posed by the "accuracy paradox", the assessment utilized multiple metrics, including accuracy, precision, recall, and F1-score, to provide a more reliable evaluation. Given that risky driving occurs less frequently than normal driving, and because classification algorithms typically assume an equal distribution of samples, the ADASYN technique was employed to address the issue of data imbalance.

Table 4TABLE 4 presents the evaluation metrics for each level of headway (0: normal, 1: dangerous, 2: avoidable accident) along with the total values. RF consistently outperforms the other classifiers, achieving the highest overall accuracy at 90.1%, precision at 87.2%, and F1 score at 84.7%, with a solid recall of 84.1%. DT shows moderate performance with an accuracy of 87.1%, precision of 83.0%, recall of 82.6%, and an F1 score of 80.4%, while kNN demonstrates the lowest performance, with an accuracy of 85.0%, precision of 76.3%, recall of 78.6%, and an F1 score of 77.9%. These results suggest that RF is the most effective classifier among the three, followed by DT, with kNN lagging behind.

TABLE 4 Evaluation metrics for classification models for headway

Model Fit measures	0	1	2	Total
Accuracy				
DT	0.959	0.846	0.807	**0.871**
RF	0.961	0.884	0.858	**0.901**
kNN	0.922	0.833	0.795	**0.850**
Precision				
DT	0.865	0.832	0.826	**0.830**
RF	0.902	0.887	0.834	**0.872**
kNN	0.790	0.781	0.707	**0.763**
Recall				
DT	0.835	0.771	0.766	**0.826**
RF	0.865	0.735	0.704	**0.841**
kNN	0.795	0.725	0.679	**0.786**
F1 Score				
DT	0.810	0.793	0.780	**0.804**
RF	0.830	0.849	0.811	**0.847**
kNN	0.793	0.771	0.752	**0.779**

**0 refers to normal phase, 1 refers to dangerous phase, 2 refers to avoidable accident phase*

Figure 13 presents the comparison of classifier metrics of the three machine learning techniques for headway. In summary, the RF model outperforms the DT and kNN models across all metrics, making it the most effective for predicting headway. The DT model shows moderate performance, while the kNN model consistently has the lowest scores, indicating that it is the least effective for this task.

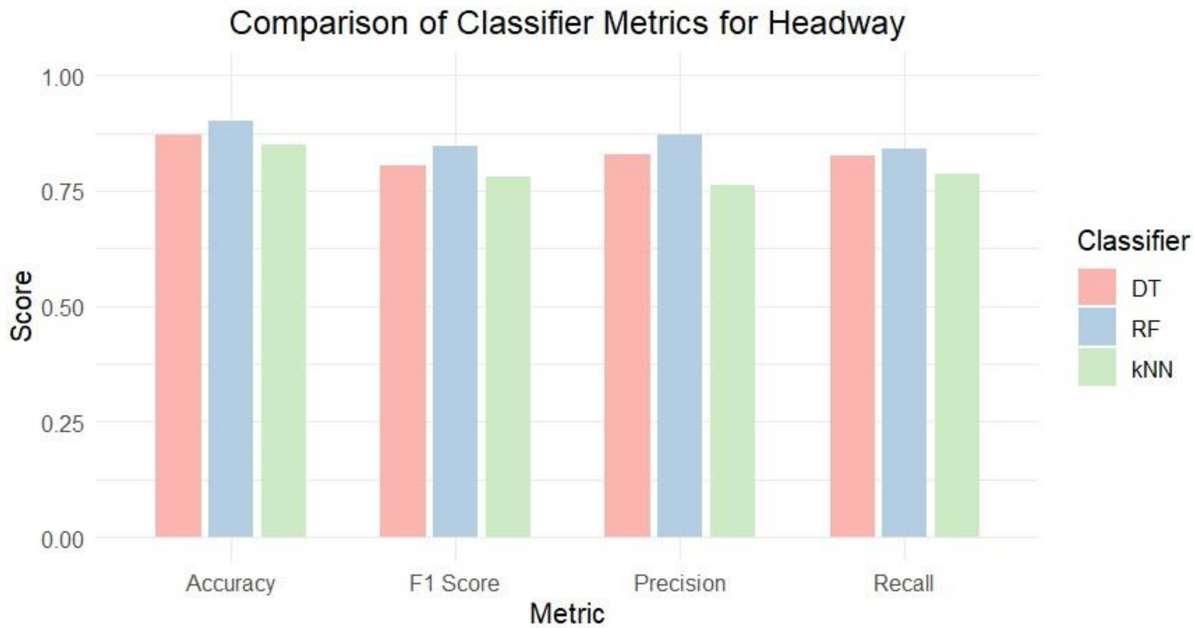

Figure 13 Comparison of classifier metrics of machine learning techniques for headway

DISCUSSION

Within the framework of the machine learning analysis, two NNs were conducted for the STZ headway prediction. It should be noted that among several machine learning algorithms, NNs proved to be the best approach for capturing complex relationships among various driving parameters and predicting the likelihood of potential risks or crashes.

The results of predictive analyses demonstrated that the level of STZ can be predicted with an exceptional accuracy of up to 89.8%. Additionally, the models exhibited a low false alarm rate, maxing out at 4%, showcasing their ability to minimise incorrect predictions and un-necessary alerts. In the on-road experiment with regards to STZ headway, the NN exhibited an overall accuracy of 81.7%. The precision of 80.8% showed that the model was highly accurate regarding positive samples, while the recall of 83.4% demonstrated the model's ability to detect safety-critical classes. In the simulator experiment, the overall model metrics were impressive, with an accuracy of 89.8%, precision of 91.2% and recall of 90.6%. These metrics indicated that the model was highly accurate in making correct predictions and excels in identifying positive samples, as evidenced by its high precision. The model's ability to detect safety-critical classes effectively was also demonstrated by its high recall. This performance suggested a well-rounded and effective predictive capability for headway in the simulator environment.

Overall, all models showed a reasonable performance with the highest accuracy in the normal phase, probably due to several key factors. Firstly, normal driving conditions likely constituted the majority of the training data, providing the model with more examples to learn from and thus improving its accuracy for this phase. Additionally, normal driving behavior is generally more consistent and predictable compared to dangerous or avoidable accident scenarios, making it easier for the model to identify and classify correctly. The lower complexity and less varied nature of normal driving conditions, compared to the erratic changes often seen in hazardous conditions, further contributed to the model's accuracy. Lastly, the features and indicators of normal driving are likely more distinct and less ambiguous than those of dangerous or avoidable accidents, reducing the chances of misclassification.

The effectiveness of the NN models in predicting headway levels was encouraging. The high accuracy, precision, and recall rates observed demonstrated the potential of these models for real-world applications. NN models from the on-road experiment, while strong, presented difficulties in achieving high precision. As for simulator models, the headway level metrics showcased similar findings, with the headway incidents having slightly higher results. In addition, in the simulator experiment, the model for STZ headway showed a balanced performance, with the highest accuracy in classifying the normal phase. NN models, with their feed-forward structure, offer distinct advantages in capturing patterns within data and excel at discerning intricate relationships, making it effective when temporal dependencies are not prominent.

The performance of three machine learning classifiers (i.e. DT, RF, kNN) across two distinct datasets (i.e. on-road and simulator experiment dataset) was thoroughly assessed in order to provide insights into the complex relationship between risk and the interdependence of task complexity and coping capacity. It is worth noting that these classification models were selected due to their strong performance and widespread use in the literature for identifying unsafe driving patterns and real-time risk prediction.

The evaluation of the three machine learning classifiers (DT, RF, kNN) revealed varying performance across the two datasets. In the on-road experiment for STZ headway, RF exhibited higher performance, leading in satisfactory accuracy (86.9%) and precision (88.7%), while showing competitive recall scores (90.7%). DT and kNN showed similar performance, though kNN tended to lag slightly behind in precision. The results from the simulator were similar to those observed in the on-road experiment. In particular, in the simulator experiment for STZ headway, RF emerged as the top-performing model with an accuracy of 90.1%, demonstrating its ability to accurately classify driving behavior in a controlled environment. Following the DT model which also performed well scoring a notable 87.1% accuracy. Regarding kNN model, they underperformed compared to the other two, displaying a lower weighted accuracy (85%) and recall (82.6%). Among the different algorithms, RF stranded out with the highest accuracy of 90% in STZ headway, indicating its ability to accurately classify driving behaviors in a controlled environment. RF also achieved a well-balanced precision (87.2%) and recall (84.1%), demonstrating its robustness and versatility.

CONCLUSIONS

The present research endeavored to identify crucial indicators of task complexity and coping capacity associated with crash risk through machine learning techniques. A comparative assessment between on-road and simulator data was provided. For that purpose, data from an on-road driving experiment (involving 135 drivers) along with data from a simulator experiment (involving 55 drivers) were collected and analysed.

In order to fulfill these objectives, a feature importance analysis (e.g. XGBoost) was implemented in order to evaluate the significance of various variables in forecasting STZ levels in terms of headway. This approach allowed for the selection of the most appropriate independent variables, ensuring that the most influential factors were identified and prioritized in the analysis. Then, machine learning analysis (e.g. Neural Networks) was applied to make accurate and data-driven predictions by identifying complex patterns between task complexity and coping capacity on crash risk. Furthermore, a comprehensive assessment of the performance of three machine learning classifiers (i.e. DT, RF, kNN) across the abovementioned distinct datasets (i.e. on-road and simulator experiment dataset) was performed to predict STZ levels for headway.

It was revealed that duration, average speed, vehicle age, time indicator, overtaking, gearbox, forward collision warning and car wipers found to be the most influential factors among all examined indicators. NNs demonstrated that the level of STZ can be predicted with an exceptional accuracy of up to

89.8%. It was revealed that the model was highly accurate in making correct predictions and excels in identifying positive samples. The model's ability to detect safety-critical classes effectively was also demonstrated by its high recall. This performance suggested a well-rounded and effective predictive capability for headway in the simulator environment.

Results indicated that RF models outperformed the DT and kNN models across all metrics, making them the most effective for predicting headway with accuracy up to 90%. The DT model showed satisfactory performance, while the kNN model consistently had the lowest but moderate scores, indicating that it is the least effective for this task. Overall, the three machine learning classifiers (DT, RF, kNN) had insightful results in terms of accuracy, precision and recall. The performance variations observed underscored the importance of selecting the right model based on data characteristics and precision-recall trade-offs, essential for real-world applications. Evaluating the results of both approaches (i.e. on-road and simulator experiment), the RF model emerged as the most efficient one. These findings are essential for advancing the understanding of driving behavior across various contexts, ultimately contributing to the development of safer and more efficient transportation systems.

Despite the robust analytical methodologies employed, it is crucial to acknowledge certain limitations of this study. Firstly, one primary limitation of this study is the simulator experimental sample size of drivers which may impact the generalizability of the findings. While the data collected provided valuable insights, a larger sample would strengthen the reliability and applicability of the results. Secondly, although the developed models showcased strength, the integration of deep learning approaches, such as Recurrent or Convolutional Neural Networks, could have potentially enhanced predictive capabilities. Moreover, incorporating interpretative machine learning techniques like LIME *(12)* alongside SHAP might have provided additional insights. Moreover, the performance of the kNN model was comparatively lower than that of the RF model, suggesting the need for additional optimisation and tuning to enhance outcomes.

As per future research directions, the examination of additional methods of analysis could be applied. In particular, imbalanced learning, factor analysis and models taking into account unobserved heterogeneity could be explored for the understanding of the relationship between task complexity, coping capacity and crash risk. Additional methodologies, such as econometric techniques could be also implemented. Future studies may consider delving into advanced deep learning models, such as Long Short-Term Memory (LSTM), which, based on relevant studies, have demonstrated superior performance. Lastly, future research endeavours should focus on integrating contextual information such as road infrastructure or traffic patterns to enhance the accuracy and applicability of the models.

ACKNOWLEDGMENTS

The research was funded by the EU H2020 i-DREAMS project (Project Number: 814761) funded by European Commission under the MG-2-1-2018 Research and Innovation Action (RIA).

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Constantinos Antoniou, Tom Brijs, George Yannis; data collection: Eva Michelaraki, Thodoris Garefalakis, Muhammad Adnan, Muhammad Wisal Khattak, Evita Papazikou, Rachel Talbot, Christelle Al Haddad; draft manuscript preparation: Eva Michelaraki, Thodoris Garefalakis, Evita Papazikou, Rachel Talbot, Christelle Al Haddad; analysis and interpretation of results: Eva Michelaraki, Thodoris Garefalakis, Muhammad Adnan, Muhammad Wisal Khattak, Christelle Al Haddad, Constantinos Antoniou, Tom Brijs, George Yannis. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Shinar, D., & Oppenheim, I. (2011). Review of models of driver behaviour and development of a unified driver behaviour model for driving in safety critical situations. In Human Modelling in Assisted Transportation: Models, Tools and Risk Methods (pp. 215-223). Springer Milan.

2. Fairclough, S. H., Venables, L., & Tattersall, A. (2005). The influence of task demand and learning on the psychophysiological response. International Journal of Psychophysiology, 56(2), 171-184.

3. Paxion, J., Galy, E., & Berthelon, C. (2014). Mental workload and driving. Frontiers in psychology, 5, 88843.

4. European Commission (2019). Lack of driving experience. Last accessed on 2/6/2024. Retrieved from: https://ec.europa.eu/transport/road_safety/specialist/knowledge/young/contributing_factors_to_crash_risk/lack_of_driving_experience_en.

5. Evans, L. (1991). Traffic safety and the driver. Science Serving Society.

6. Ahad, N., Qadir, J., & Ahsan, N. (2016). Neural networks in wireless networks: Techniques, applications and guidelines. Journal of network and computer applications, 68, 1-27.

7. Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), 130.

8. Suthaharan, S., & Suthaharan, S. (2016). Decision tree learning. Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning, 237-269.

9. Breiman, L., & Cutler, A. (2016). Random forests for scientific discovery.

10. Breiman, L. (1996). Bagging predictors. Machine learning, 24, 123-140.

11. Nigsch, F., Bender, A., van Buuren, B., Tissen, J., Nigsch, E., & Mitchell, J. B. (2006). Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. Journal of chemical information and modeling, 46(6), 2412-2422.

12. Man, X., & Chan, E. (2021). The best way to select features? comparing mda, lime, and shap. The Journal of Financial Data Science Winter, 3(1), 127-139.