Detection of dangerous driving behaviour using machine learning techniques and big data

Thodoris Garefalakis, Eva Michelaraki, George Yannis

National Technical University of Athens, Department of Transportation Planning and Engineering, Athens, Greece

Introduction

Road crashes account for over 1.19 million fatalities and countless injuries globally each year, with human error being a significant contributor. Efforts to improve road safety have focused on intelligent transportation systems (ITS) and advanced driver-assistance technologies.

The current study investigates the role of **Safety Tolerance Zones (STZ)** in assessing and predicting driving risks, aiming to enhance real-time intervention and reduce crashes. By leveraging **machine learning techniques** and **naturalistic driving data**, this study contributes to the development of a safer and more efficient road environment.

Objective

This study aims to:

1. Evaluate **key factors** influencing Safety Tolerance Zone (STZ) prediction using XGBoost for feature selection.
2. Develop a real-time prediction model for **identifying risky driving** behaviour.

Data Collection

A **naturalistic driving experiment** conducted, involving 50 car drivers from Belgium over a 15-month timeframe (from April 2021 to July 2022), resulting in the development of a substantial database comprising 7,160 trips. Key performance indicators were collected as shown in Table 1.

Table 1: *Description of driving performance indicators*

Variable	Description
GPS_distances_sum	Distance travelled (km)
ME_AWS_hw_measurement_mean	Headway measurement (seconds)
ME_AWS_fcw_mean	Forward collision warning
ME_AWS_pcw_mean	Pedestrian collision warning
DEM_evt_hb_lvl_M_mean	Medium level harsh braking events
DEM_evt_ha_lvl_M_mean	Medium level harsh acceleration events
ME_Car_wipers_median	Indicates weather conditions (wipers on/off)
ME_AWS_time_indicator_median	Indicates lighting conditions (day, dusk, night)

Figure 1 reveals that "DEM_evt_ha_lvl_M_mean" and "DEM_evt_hb_lvl_M_mean" were highly correlated, leading to the exclusion of "DEM_evt_hb_lvl_M_mean" from the analysis.

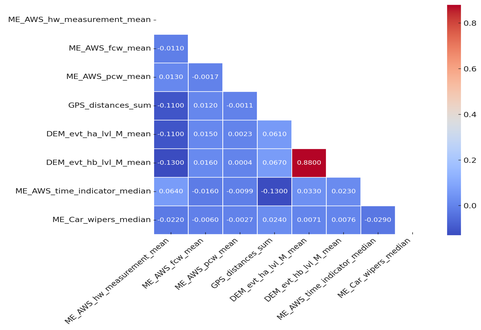

Figure 1: *Correlation heatmap*

Methodology

- **eXtreme Gradient Boosting (XGBoost)** algorithm was utilized to identify critical features influencing STZ levels. XGBoost, known for its efficiency and accuracy, evaluated feature importance using metrics such as:
  - Gain: Contribution of a feature to improving model accuracy.
  - Frequency: Occurrence of a feature in decision trees.
  - Cover: Magnitude of a feature's impact on predictions.
- **Long Short-Term Memory (LSTM)** a variant of Recurrent Neural Networks (RNNs), was employed for real-time STZ prediction. Its architecture is well-suited for handling sequential and temporal data. The model:
  - Included three neurons in the input layer for critical features (headway, collision warnings, and distance).
  - Used a single neuron in the output layer to predict STZ levels.

The proposed approach followed using LSTMs is given in Figure 2:

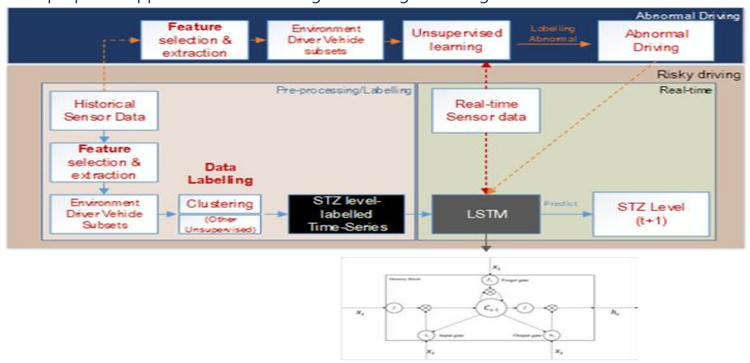

Figure 2: *STZ modelling using LSTMs*

Feature Importance Results

The feature importance analysis was conducted using the **Extreme Gradient Boosting (XGBoost)** algorithm to determine the relevance of variables in predicting Safety Tolerance Zone (STZ) levels.

As shown in Figure 3, the analysis identified headway measurement, forward collision warning indicator, and distance traveled as **the most critical features**. On the other hand, weather indicators (car wipers) and lighting conditions were found to be less significant. Variables such as pedestrian collision warning and medium-level harsh acceleration events also showed minimal impact on predictions.

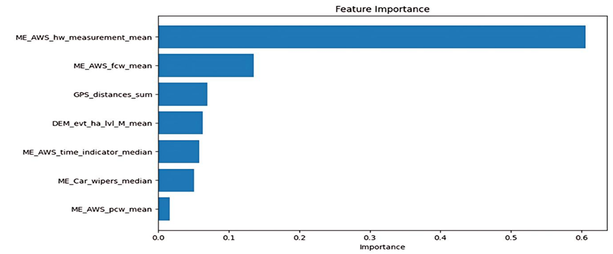

Figure 3: *XGBoost feature importance of independent variables*

Classification Results

A dataset consisting of approximately 275,000 data points was employed to train an **LSTM Neural Network model**. In accordance with the feature importance and the significance of relevant indicators, the input layer was structured with three neurons, representing headway measurement, forward collision warning indicator, and distance travelled. Additionally, the output layer was configured with a single neuron denoting the STZ. The model aimed to classify Safety Tolerance Zone (STZ) levels into **'Normal'**, **'Dangerous'**, and **'Avoidable Accident'** categories.

The performance metrics, including accuracy, precision, recall, and F1-score, are presented in Table 2, highlighting the model's strengths and limitations. While the overall accuracy of 71% indicates reliable predictions, the recall score of 67% reflects the model's moderate ability to identify actual instances of risky driving behaviors. **Recall is especially critical in this context**, as it indicates the model's capacity to detect safety-critical situations, even at the cost of occasional false positives.

Table 2: *Assessment of the classification LSTM on STZ level*

Model	Accuracy	Precision	Recall	f1-score
Long Short-Term Memory (LSTM)	71%	55%	67%	55%

The **Confusion Matrix** (Figure 4) offers a detailed breakdown of prediction outcomes, showing the true positives, true negatives, false positives, and false negatives for each STZ level. This visualization demonstrates that the model performed better at identifying 'Normal' and 'Dangerous' levels but faced challenges in accurately classifying the 'Avoidable Accident' category, which requires immediate attention for road safety interventions.

Further insights into the model's overall performance and its classification ability are visualized in Figure 5. While the LSTM model shows **potential for real-time prediction** of risky driving behaviors, it emphasizes the need for optimization, particularly in refining its ability to detect high-risk situations that require immediate intervention.

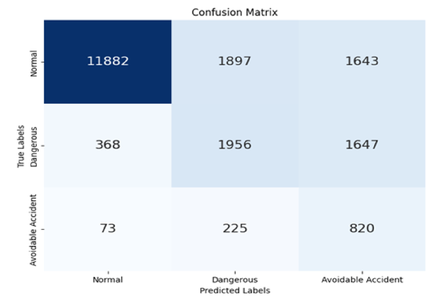

Figure 4: *Confusion Matrix*

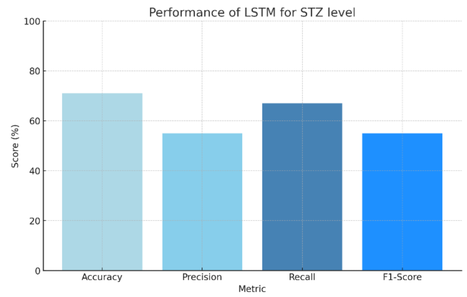

Figure 5: *Performance of LSTM for STZ level*

Conclusions

- The study highlighted headway measurement, forward collision warning indicator, and distance traveled as **the most critical features** influencing driving safety.
- The **Long Short-Term Memory (LSTM)** model achieved a **71% accuracy** and **67% recall**, demonstrating potential for detecting risky driving behaviors.
- The model struggled with accurately detecting **'Avoidable Accident' level**, indicating the need for enhanced data collection, additional contextual variables, or LSTM architecture refinement.
- The combination of **XGBoost and LSTM** offers a promising framework for real-time detection of dangerous driving behavior and data-driven road safety interventions.
- Future research could expand the sample size, incorporate **socio-demographic variables**, and test **alternative machine learning methods** for improved classification performance

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).

Contact Information:
Eva Michelaraki, PhD , Research Associate NTUA
Department of Transportation Planning and Engineering
Email: evamich@mail.ntua.gr
Website: https://www.nrso.ntua.gr/p/evamich/

104th Annual Meeting of the Transportation Research Board, January 5–9, 2025
TRBAM-25-01848