Evaluation of Hybrid Machine Learning Models for Risky Driving Behavior Classification: A Comparative Study Using RNN-AdaBoost, GANs, and XGBoost

Eleni Maria Theodoraki*, Thodoris Garefalakis, Paraskevi Koliou, George Giannis

1. National Technical University of Athens, Greece
Corresponding author: e_theodoraki@mail.ntua.gr

Abstract

Driver behavior is a critical determinant of road safety, influencing the likelihood and severity of traffic crashes. This study evaluates the performance of three advanced machine learning models: (i) RNN-AdaBoost, (ii) Generative Adversarial Networks (GANs), and (iii) XGBoost-RF for classifying hazardous driving behaviors across multiple safety levels. A key focus is on assessing the impact of Conditional GANs (cGANs) for generating synthetic data to address class imbalances and enhance classification accuracy.

The dataset used in this research originates from a naturalistic driving study conducted in Belgium and the UK, capturing a diverse range of real-world driving behaviors. Initially, driving behaviors were categorized into three safety levels: Normal, Dangerous, and Avoidable Accident. However, following data augmentation with cGANs, a revised classification schema was introduced, consolidating risk levels into a binary system: Normal and Avoidable Accident. This adjustment aimed to optimize the training process while maintaining interpretability.

Each model employed in this study offers distinct advantages. RNN-AdaBoost leverages temporal dependencies in driving data and integrates boosting to refine classification accuracy. GANs, both before and after data augmentation, are evaluated for their effectiveness in improving model generalization. XGBoost, a powerful ensemble learning algorithm, provides robust and scalable risk classification. The study further employs SHAP (Shapley Additive Explanations) analysis to interpret model predictions, identifying key factors such as harsh acceleration and braking as dominant risk indicators.

The results reveal that while the cGAN-enhanced dataset significantly improves GAN model performance boosting accuracy from 76% to 90% in Belgium and 79% to 91% in the UK it introduces overfitting risks in the hybrid models. The XGBoost-RF and RNN-AdaBoost models achieve near-perfect accuracy on augmented data but struggle to generalize effectively to real-world scenarios. This study underscores both the potential and the challenges of cGAN-generated synthetic data in driving behavior classification, highlighting the need for careful validation and adaptive augmentation strategies.

The insights gained from this research provide a foundation for policymakers and industry stakeholders to refine risk classification models, implement proactive safety interventions, and develop data-driven approaches to enhance road safety on a global scale.

Keywords: *road safety; risky driving behavior; machine learning; RNN-AdaBoost; Generative Adversarial Networks (GANs); XGBoost; SHAP; Variance Threshold; Mutual Information*.

1. Introduction

Road transport is a fundamental component of modern society, facilitating economic growth and personal mobility. However, the increasing number of privately owned vehicles has also introduced significant challenges, with road safety being a major concern. According to the World Health Organization (WHO), road traffic injuries rank among the leading global causes of death, accounting for approximately 1.19 million fatalities annually (Global Status Report on Road Safety 2023). Beyond the tragic loss of life, road accidents have severe economic and social consequences.

Human error remains the primary cause of most traffic incidents, with research indicating that 90-95% of crashes are linked to driver behavior. Risk factors such as speeding, traffic violations, distracted driving, fatigue, and impaired driving play a significant role in accident causation. Addressing these risk factors is a crucial aspect of road safety interventions. While emerging autonomous vehicle technologies offer promising solutions, mitigating human error through advanced driver monitoring and behavior prediction remains essential. Studies suggest that autonomous vehicles could reduce traffic accidents by up to 93% by eliminating common driver mistakes (Khashayarfard & Nassiri, 2021). Furthermore, enhancing decision-making and attention monitoring systems in human-driven vehicles could contribute to substantial crash reductions (Mueller et al., 2020).

Recent advancements in intelligent transportation systems (ITS) and machine learning (ML) have introduced powerful tools for real-time driver behavior analysis and accident risk prediction. These technologies enable the detection of hazardous driving patterns, providing opportunities for proactive interventions. Deep learning (DL) models have demonstrated remarkable success in identifying unsafe driving behaviors based on naturalistic driving data, making them highly relevant for predictive safety applications.

Several studies have explored ML techniques for assessing accident risks and evaluating driver behavior. For instance, Shangguan et al. (2021) examined how behavioral and environmental factors contribute to driving risks, while Peppes et al. (2021) demonstrated the role of ITS in enhancing autonomous vehicle safety by predicting driver errors. Similarly, Shi et al. (2019) developed an unsupervised learning framework for assessing driving risks based on real-world datasets. In addition, models such as Random Forests, Long Short-Term Memory (LSTM) networks, and Deep Neural Networks (DNN) have shown strong accuracy in categorizing driver behavior based on speed, acceleration, braking intensity, and proximity to other vehicles. The integration of ensemble models, such as DNNs combined with Random Forests, has been particularly effective in classifying driver behavior into distinct risk categories (J. Wang et al., 2021, i-DREAMS project).

Despite the predictive power of these models, a major challenge remains interpretability. Many ML and DL models operate as "black boxes," offering little transparency regarding the factors that influence their predictions. This lack of interpretability is particularly problematic in safety-critical applications such as autonomous driving and road risk assessment, where understanding the reasoning behind predictions is crucial for trust, accountability, and decision-making. To address this, interpretability techniques, such as SHAP (Shapley Additive Explanations), have been introduced to highlight the most influential driving features in classification models, ensuring that predictions align with real-world driving risks. However, despite their importance, interpretability techniques remain underutilized in ML-based road safety studies, leading to a gap in ensuring model transparency.

This research aims to bridge these gaps by (1) incorporating interpretability techniques into ML models to enhance transparency in predicting hazardous driving behaviors and (2) evaluating the generalizability of these models across datasets from multiple countries (Belgium and the UK). By integrating synthetic data augmentation through Conditional GANs (cGANs) and assessing its impact on classification accuracy, this study provides a comprehensive evaluation of how data-driven techniques can contribute to road safety improvements.

The remainder of this paper is structured as follows: Section 2 includes a review of previous research work related to hybrid machine learning models, outlines the objective of the study, and highlights its practical applications. Section 3 outlines the methodology, including data collection, preprocessing, and model development. Section 4 presents the experimental results, and Section 5 discusses key findings, limitations, and future research directions.

2. Literature Review

The study of driving behavior and risk assessment is a pivotal area of research, particularly in the fields of transportation safety and human behavior optimization. Numerous methodologies have been proposed in the literature, focusing on analyzing driving data to improve predictive models for hazardous situations. Techniques based on machine and deep learning, such as eXtreme gradient boosting (XGBoost), OneClassSVM, Recurrent Neural Networks (RNN) and Generative Adversarial Networks (GANs), offer innovative approaches for analyzing and modeling driving behavior.

This literature review explores key contributions in this domain, including methods for predicting and assessing driving risks, with an emphasis on their challenges and achievements. By comparing methodologies and findings, this review aims to highlight trends and directions for future research, as well as their practical applications in areas such as vehicle safety and usage-based insurance pricing.

Driving Simulator Studies (DSS) and Naturalistic Driving Studies (NDS) are the two primary methodologies commonly utilized in the examination of driving behavior analysis research (Osman et al., 2019). These study approaches have yielded significant insights into the complex nature of unsafe driving behaviors and have become essential instruments for comprehending the aspects that lead to road safety issues. The study of (Wijayaratna et al., 2019) investigated the use of both methodologies to assess the impact of mobile phone conversations on driving performance. The findings indicated that DSS often highlight an increased crash risk associated with mobile phone use, whereas NDS suggested a reduction in crash risk. Each approach offers distinct advantages, making their comparison beneficial for drawing well-rounded conclusions. For example, DSS enables the efficient collection of diverse driving scenario data in controlled environments. Conversely, NDS provides greater realism, offering a more accurate reflection of natural driving conditions (Wang et al., 2022).

Machine learning-based models are extensively employed in the field of road safety due to their high accuracy and effectiveness in predicting risky driving behavior. In this context, recent studies have utilized various models, including random forests (RFs), multilayer perceptrons (MLP), support vector machines (SVMs), eXtreme gradient boosting (XGBoost), decision trees (DT), gradient boosting (GB), and logistic regression (LR).

The study of (Wang et al., 2022) focuses on detecting driver distraction caused by phone use during driving, employing vehicle dynamics data from the Shanghai Naturalistic Driving Study (SH-NDS). Researchers developed a Bidirectional Long Short-Term Memory (Bi-LSTM) model with an attention mechanism to analyze variables like speed, acceleration, lane offset, and steering wheel rate. The model achieved 91.2% accuracy in identifying distraction, surpassing traditional machine learning methods.

This approach emphasizes time-series data processing, offering a realistic and efficient solution for integrating into advanced driver assistance systems, thus enhancing road safety.
The study of (Masello et al., 2023) explores the use of contextual data from telematics to predict risky driving events with XGBoost and Random Forest models. Analyzing 77,859 km of naturalistic driving, they identify factors like speed limits, weather, road quality, and traffic patterns that influence speeding, distractions, and near-misses. SHAP values are used for feature importance, providing insights for insurers and road safety stakeholders to assess and mitigate risks.

(Ziakopoulos et al., 2023) examines the role of smartphone-based applications in monitoring and influencing driving behavior, focusing on mobile phone use as a source of driver distraction. Using data from a naturalistic experiment with 230 drivers, the researchers combined smartphone sensor data and self-reported questionnaire responses. Machine learning algorithms, including XGBoost, identified total trip distance as the most significant factor, with other variables such as driving experience, feedback received, and family size playing secondary roles. The results suggest that tailored feedback and smartphone-based interventions could reduce distracted driving and improve road safety.

(Shangguan et al., 2021) presents a real-time driving risk prediction method using naturalistic data, employing the Rear-end Crash Risk Index (RCRI) and Fuzzy C-means clustering to classify risk into four levels. The methodology extracts features like speed difference, headway distance, and acceleration, achieving up to 89.2% accuracy in predicting medium and high-risk statuses with MLP models. Based on 1,440 car-following events in the Shanghai study, the findings aim to enhance safety in connected and autonomous vehicles (CAVs).

Addressing the challenges of traffic crash analysis, (Shao et al., 2024) utilize machine learning and natural language processing (NLP) to enhance predictive accuracy and interpretability. Through XGBoost and SHapley Additive exPlanations (SHAP), their approach achieves a prediction accuracy of 0.79 for crash injury severity while identifying critical behavior-cause relationships, such as distracted driving and excessive speeding. The study categorizes crash causes into five primary groups linked to 141 specific behaviors, suggesting targeted interventions like improved road signage and stricter enforcement of traffic laws. A behavior-cause matrix and a proposed traffic safety knowledge graph further advance the understanding and management of traffic safety.

(Osman et al., 2019) investigated the role of secondary tasks in distracted driving and proposed a bi-level classification methodology to detect and classify such tasks using machine learning. The study used driving behavior parameters like speed, acceleration, and yaw rate to identify engagement in tasks such as handheld phone calls, texting, and passenger interaction. Among the tested classifiers, Decision Trees achieved a remarkable 99.8% accuracy in detecting task engagement, while Random Forest excelled in classifying task types with an accuracy of 82.2%. These results highlight distinct driving patterns associated with specific secondary tasks, offering potential for enhanced in-vehicle safety systems.

The study (Zhu et al., 2022) focuses on developing an improved algorithm for identifying dangerous driving behaviors using highway vehicle trajectory data obtained via video monitoring. The research defines five key risk indicators: dangerous car following, lateral deviation, frequent acceleration and deceleration, frequent lane changes, and forced insertion. By employing unsupervised anomaly detection methods like OneClassSVM and improving classification through an imbalanced XGBoost algorithm, the study demonstrates significant advancements in accurately identifying dangerous drivers. These methods enhance the precision of detection in highly imbalanced datasets, contributing to road safety improvements.

The study (Vanhoeyveld & Martens, 2018) investigates techniques to address the challenges of learning from imbalanced and sparse behavioral data. The research evaluates various oversampling, undersampling, and cost-sensitive learning methods, particularly on high-dimensional and sparse data like user behavior in online platforms. The EasyEnsemble method, which combines balanced subsets with boosting, outperformed others, demonstrating efficiency and robustness in handling imbalanced data. This work provides valuable insights into adapting machine learning algorithms for behavior data and improving predictive performance in domains with inherent data imbalances.

Advancements in driving behavior analysis, including machine learning models, DSS, and NDS, provide valuable insights into road safety and driver risk assessment. By integrating controlled experiments with real-world data, these methods address critical challenges like distraction and aggressive driving. Future research should focus on refining predictive models, enhancing real-time applications, and fostering interdisciplinary approaches to improve transportation safety and efficiency.

3. Materials and Methods

3.1 Data Collection

As part of the i-DREAMS research initiative, a naturalistic driving study was conducted with participants from Belgium and the UK. The Belgian cohort consisted of 43 drivers, generating a substantial dataset of 7.163 trips and 147.337 minutes of driving data. In the UK, 26 drivers participated, contributing 8.226 trips and 118.175 minutes of recorded driving time. The study aimed to gather extensive data on driving behavior and road conditions to enable a detailed analysis of risky driving behaviors.

The experiment was conducted over a four-month period and was divided into four distinct phases. The first phase, lasting four weeks, served as a baseline without any intervention. In the second phase, adaptive Advanced Driver Assistance Systems (ADAS) provided real-time vehicle warnings for another four weeks. During the third phase, participants received performance feedback through a mobile app, while the fourth phase expanded on this by incorporating gamification to promote safer driving habits. Across all phases, the study continuously monitored real-time driving behavior and assessed the effectiveness of both real-time ADAS warnings and post-drive feedback mechanisms. Figure 1 illustrates the different phases of the i-DREAMS on-road study's experimental design.

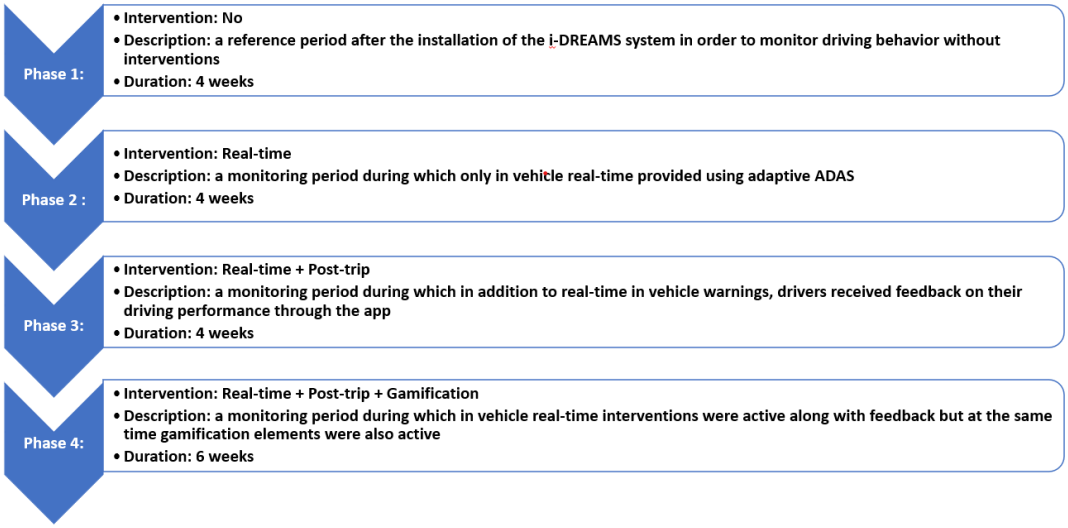

Figure1: *Overview of the different phases of the experimental design*

Data collection utilized advanced technologies, including an OBD-II device installed in each vehicle to capture hundreds of diving parameters. Additionally, the Mobile system, connected via mobile networks,

enabled seamless data acquisition without requiring user input. To assess driving behavior, each 30-second segment of a trip was classified into one of three safety categories: Normal, Dangerous, or Avoidable Accident. These classifications were based on intervention thresholds established in existing research and key variables such as speed and headway distances.

3.2 Definition of 'Safety Tolerance Zone'

Before implementing classification algorithms, it was essential to categorize the driving data into distinct levels of the Safety Tolerance Zone. The Safety Tolerance Zone was divided into three categories - Normal, Dangerous and Avoidable Accident- based on intervention thresholds from literature and the i-DREAMS project. This structured categorization enabled targeted analysis of driving behaviors and their associated safety risks.
To ensure consistency with established research methodologies, the classification framework was based on two key driving performance indicators: headway distance (the distance between the driver's vehicle and the preceding vehicle) and speeding behavior. These indicators were consistently applied across datasets from both Belgium and the UK. Given that safety-critical driving events are generally less frequent, it was expected that Normal driving instances would dominate the dataset, while Dangerous behaviors and Avoidable Accidents would constitute minority classes.

The dataset included intervention variables labeled as iDreams_Headway_Map_level_i and iDreams_Speeding_Map_level_i, where i represented the intervention level ranging from -1 to 3. These variables captured real-time headway and speed deviations relative to predefined safety thresholds. Each intervention level was encoded using a binary system: 0: The intervention level does not correspond to i and 1: The intervention level corresponds to i.

To determine the Safety Tolerance Zone for each recorded timeframe, the intervention variables were evaluated to identify the most severe safety level encountered during that period. The classification process followed a hierarchical approach, prioritizing the highest risk level present in the data. The final categorization was defined as follows:

- Normal: When the highest recorded intervention level was -1, 0, or 1.
- Dangerous: When the highest recorded intervention level was 2.
- Avoidable Accident: When the highest recorded intervention level was 3.

This classification approach ensured that every instance reflected the most critical safety condition within the analyzed timeframe, allowing for an accurate representation of real-world driving risks. By structuring the dataset in this manner, the classification process laid the groundwork for subsequent machine learning applications aimed at predicting and mitigating unsafe driving behaviors.

3.3 Machine Learning Models

The main projective of this study is to accurately classify and predict risk levels based on real-world driving data. The classification problem was structured into three distinct risk levels: Normal, Dangerous, and Avoidable Accident. Three hybrid models combining deep learning, machine learning, and ensemble methods were developed to improve prediction accuracy and robustness. These models effectively handle sequential data and enhance classification through decision tree-based techniques. The developed models include:

1. Recurrent Neural Network (RNN) - AdaBoost
2. Extreme Gradient Boosting (XGBoost) – Random Forest (RF)
3. Generative Adversarial Networks (GANs)

The application of these hybrid models in road safety research is still in its early stages, making this study a key step toward advanced risk prediction. Each model was chosen for its strengths: **RNN-AdaBoost** combines sequential learning with boosting for better accuracy; **XGBoost-RF** merges gradient boosting and Random Forests to improve generalization and reduce overfitting; and **GANs** offer a generative approach for modeling rare, high-risk scenarios and enhancing data diversity.

To enhance model interpretability, the **SHapley Additive Explanations (SHAP)** algorithm was employed, offering deeper insights into how predictions were made. SHAP provides a game-theoretic approach to explain individual predictions by assigning contribution values to each feature, making it particularly effective for complex machine learning models. This step is crucial for addressing the "black-box" nature of many predictive models, increasing transparency, and ensuring the reliability of risk assessment in driving behavior classification.

3.3.1 Recurrent Neural Network (RNN)-AdaBoost

The integration of **RNN-LSTM** and **AdaBoost** offers a novel hybrid method for road safety assessment by combining temporal pattern recognition with ensemble boosting. This approach is rare in the field but effective for handling imbalanced data and detecting complex driving behaviors.

- **RNN-LSTM:** LSTM networks capture long-term dependencies in driving sequences, allowing the model to recognize patterns that may indicate risky behavior, such as gradual acceleration followed by sudden braking.
- **AdaBoost**: AdaBoost improves classification by focusing on hard-to-classify cases, making it especially valuable for identifying rare but critical safety risks. Applying it to LSTM outputs boosts accuracy, particularly for minority classes. Together, this hybrid model effectively analyzes sensor-based driving data to enhance risk detection.

3.3.2 Extreme Gradient Boosting (XGBOOST)-Random Forest

The hybrid use of **XGBoost and Random Forest** offers a novel approach to road safety assessment, combining the strengths of decision tree-based methods to capture complex patterns and enhance performance, particularly with imbalanced datasets.

- **Random Forest:** As an ensemble of decision trees, Random Forest captures nonlinear relationships in driving data and improves generalization by reducing overfitting, making it effective for identifying risk-related patterns.

- **XGBoost**: This gradient boosting method refines model performance by correcting prior errors, making it well-suited for detecting rare hazardous behaviors. When applied to Random Forest outputs, XGBoost improves accuracy and reduces misclassification in minority classes.

3.3.3 Generative Adversarial Networks (GANS)

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014), are deep learning models that generate realistic synthetic data through adversarial training between two neural networks: a Generator and a Discriminator.

- **Generator**: Transforms random noise into synthetic samples resembling real data, aiming to deceive the Discriminator.

- **Discriminator**: Distinguishes real from synthetic data, guiding the Generator through feedback to improve realism.

This adversarial process continues until the Generator produces data indistinguishable from real samples, following a min-max optimization framework. GANs have demonstrated success in image synthesis, data augmentation, and anomaly detection, though they face challenges such as mode collapse and training instability. Variants like **cGANs**, **WGANs**, and **StyleGANs** address these issues, enhancing GAN applicability across domains.

3.4 Multi-class Classification and Model Evaluation

To address the class imbalance commonly found in real-world driving data—where hazardous events are much rarer than normal behavior—SMOTE was used to generate synthetic instances for minority classes (Dangerous and Avoidable Accident). This improved the model's ability to detect and classify critical safety-related behaviors.

Prior to model training, a rigorous feature selection was carried out to identify the most influential variables for assessing driving risk levels. First, Variance Threshold was applied to remove features with low variability, ensuring that only informative variables were retained. Next, Mutual Information was calculated to evaluate the relevance of each feature to the target variable, and only those meeting both criteria were selected. The resulting features were normalized using Min-Max Scaling to standardize the data range. A Random Forest Classifier was then trained on the processed dataset, and model performance was assessed using standard classification metrics. Additionally, feature importance analysis was performed to identify which variables most significantly influenced predictions. This approach helped improve both the model's interpretability and classification accuracy.

As a result, the following four variables were selected for use in the classification models:

Table 1: Selected variables for Belgium and UK

Belgium	UK
GPS_distances_sum– Total distance traveled by the vehicle	**GPS_distances_sum**– Total distance traveled by the vehicle
GPS_spd_mean– Average speed of the vehicle during the trip	**GPS_spd_mean**– Average speed of the vehicle during the trip
ME_Car_speed_mean– Average speed of the vehicle	**ME_Car_speed_mean**– Average speed of the vehicle
DEM_evt_hb_lvl_H_mean– Mean level of harsh braking events recorded during the trip	**IBI_value_mean**– Mean time interval between successive heart beats

Each model was evaluated using metrics such as accuracy, precision, recall, false alarm rate, and f1-score, providing a comprehensive assessment of performance, defined by Equation (1) to Equation (5):

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad \textbf{(1)}$$

$$Precision = \frac{TP}{TP+FP} \quad \textbf{(2)}$$

$$Recall = \frac{TP}{TP+TN} \quad \textbf{(3)}$$

$$False\ Alarm\ Rate = \frac{FP}{FP+TN} \quad \textbf{(4)}$$

$$f1-score = \frac{2x(\text{Presicion})x(\text{Recall})}{(\text{Precision})+(\text{Recall})} \quad \textbf{(5)}$$

where: True Positive (TP) represents the instances which belong to class i and were correctly classified in it; True Negative (TN) represents the instances which do not belong to class i and were not classified in it; False Positive (FP) represents the instances which do not belong to class i but were incorrectly classified in it; False Negative (FN) represents the instances which belong to class i but were not classified in it.

SHAP (SHapley Additive Explanations) is a technique used to interpret complex machine learning models, based on Shapley values from game theory. It quantifies each feature's contribution —such as speed, headway, and acceleration—to the model's predictions, showing how features impact decisions. In this study, SHAP enhances model transparency, crucial for road safety applications, by providing clear explanations of feature importance, fostering trust and aiding policymaking.

4. Results

The presentation of the results includes the performance of the three machine and deep learning models that were developed to classify the driving behavior into three risk categories: Normal, Dangerous, and Avoidable Accident. The models – Recurrent Neural Network (RNN) combined with AdaBoost, Extreme Gradient Boosting (XGBOOST) combined with Random Forest (RF), and Generative Adversarial Networks – were applied to naturalistic driving data collected in Belgium and the UK. Key metrics, including accuracy, precision, recall, false positive rate (FPR), and F1-score were utilized for the evaluation of these models.

4.1 Identification of the Safety Tolerance Zone Levels

Table 2 provides a summary of the performance of the three machine learning models, comparing their accuracy, precision, recall, false positive rate (FPR), and F1-score across both datasets (Belgium and the UK). Among them, the Random Forest (RF) combined with Extreme Gradient Boosting (XGBOOST) consistently delivered the best results in both datasets. Specifically, in Belgium, the RF-XGBOOST model achieved 93% accuracy, 93% precision, and 93% recall. Similarly, in the UK dataset, it reached 92% accuracy, with 92% precision and 92% recall. The model's effectiveness in minimizing misclassifications is further reflected in its low FPR, recorded at 7.4% for Belgium and 9.8% for the UK. In contrast, the RNN-AdaBoost model exhibited lower performance, particularly in the Belgium dataset, where it achieved an accuracy of 83%, recall of 83%, and a significantly higher FPR of 14.7%. These results suggest that while the RNN-AdaBoost model can identify risky behaviors to some extent, it is more prone to misclassification, making its predictions less reliable than those of the RF-DNN model. Performance in the UK dataset was relatively better, with an accuracy of 85% and a recall of 85%, though the FPR remained elevated at 10.3%. The GANS model demonstrated suboptimal performance, especially in terms of precision, which reached 64% in the Belgium dataset. However, its overall accuracy was lower than that of the RF-XGBOOST model, standing at 76% in Belgium and 79% in the UK. The FPR for this model was 23.2% in Belgium and 21% in the UK, indicating difficulties in correctly distinguishing certain risk categories.

Table 2: Comparison of classification model evaluation metrics for Belgium and UK

Dataset	*Model*	*Accuracy*	*Precision*	*Recall*	*FPR*	*F1-score*
Belgium	XGBOOST & RF	93%	93%	93%	7.4%	93%
	RNN & AdaBoost	83%	82%	83%	14.7%	82%
	GANS	76%	64%	76%	23.2%	66%
UK	XGBOOST & RF	92%	92%	92%	9.8%	91%
	RNN & AdaBoost	85%	84%	85%	10.3%	84%
	GANS	79%	80%	79%	21%	74%

Figure 2 illustrates model performance across Belgium and UK datasets. Both XGBoost & RF and RNN & AdaBoost achieve high and consistent scores in accuracy, precision, recall, and F1-score, whereas the GAN model underperforms significantly. A deeper examination of recall highlights its significance in road safety, as it represents the model's effectiveness in accurately recognizing risky driving behaviors, including hazardous or preventable accidents. A higher recall indicates an improved ability to detect such behaviors, which is essential for reducing the likelihood of crashes.
Based on the presented results, several key observations can be made regarding the performance of each model across the two datasets (Belgium and UK).

1. **XGBOOST & RF:** This model consistently outperforms the alternatives in both datasets, with the highest overall accuracy and lowest false positive rate, demonstrating strong generalization and robustness.
2. **RNN & AdaBoost:** While performing worse than XGBOOST & RF, this model still achieves relatively high accuracy, recall and F1-score.
3. **GANs:** The GAN-based model exhibits the lowest performance across all key metrics. It has the lowest accuracy and F1-score, as well as the highest FPR, indicating a weaker ability to correctly classify instances. Particularly, in the Belgium dataset, the precision drops significantly (64%), highlighting its struggle in making confident predictions.

Despite its strong performance, XGBoost-RF has notable limitations. One of its main drawbacks is its high computational cost, particularly when handling large datasets with numerous features. The iterative nature of boosting and the ensemble approach of Random Forests require significant memory and processing power, making real-time inference challenging in embedded systems or mobile applications. Additionally, while XGBoost is more interpretable than deep learning models, its reliance on decision trees can sometimes lead to overfitting, especially when applied to noisy data.

On the other hand, the RNN-AdaBoost model effectively captures temporal dependencies, making it particularly suited for sequential driving behavior analysis. but it is computationally expensive and prone to the vanishing gradient problem. Although AdaBoost enhances weak learners, it is also sensitive to noisy data, meaning that misclassified instances can significantly affect the final model's predictions. This sensitivity could explain the relatively higher False Positive Rate (FPR) observed in this model compared to XGBoost-RF. Future research should focus on mitigating these limitations by optimizing the hyperparameters of XGBoost-RF to reduce overfitting and computational cost while exploring alternative architectures for RNNs, such as Transformer-based models, which can handle long-range dependencies more efficiently. Additionally, integrating feature selection techniques before training could help streamline model complexity while maintaining classification accuracy.

The GAN-based model underperformed due to several factors. GANs are difficult to train because of their adversarial nature, leading to unstable convergence and poor generalization. Unlike RNN-AdaBoost, which captures temporal dependencies, and XGBoost-RF, which enhances generalization, the GAN lacks mechanisms for effective classification of sequential driving behaviors. Additionally, the GAN exhibited high false positive rates (23.2% in Belgium and 21% in the UK), indicating significant misclassification, especially in distinguishing risk categories. These issues suggest that GANs are better suited for data augmentation than direct classification tasks. Additionally, GANs are primarily designed for generating synthetic data rather than direct classification, which inherently limits their ability to minimize classification errors. Unlike supervised learning models such as XGBoost and RNN-AdaBoost, which optimize specifically for classification accuracy, GANs focus on generating realistic data distributions. This design limitation affects their ability to differentiate between various driving risk levels effectively.

Another notable challenge is the issue of mode collapse, a common limitation where the generator produces highly similar synthetic samples without capturing the full range of real-world driving behaviors. This reduces the diversity of the training data and affects the model's generalization, ultimately limiting its classification capability. Furthermore, GANs lack interpretability compared to tree-based models like XGBoost, which leverage SHAP (Shapley Additive Explanations) to provide insights into feature importance. This black-box nature of GANs makes it difficult to determine which specific driving behaviors contribute most to risk classification, limiting their applicability in safety-critical environments.

Moreover, while the XGBoost-RF hybrid demonstrated the highest classification accuracy, its computational complexity could be a limitation in real-time applications. Similarly, the RNN-AdaBoost model efficiently captures temporal dependencies, making it well-suited for time-series analysis but at the cost if the increased training time. Future research could explore optimizing these models for deployment in edge computing environments, where computational efficiency is critical.

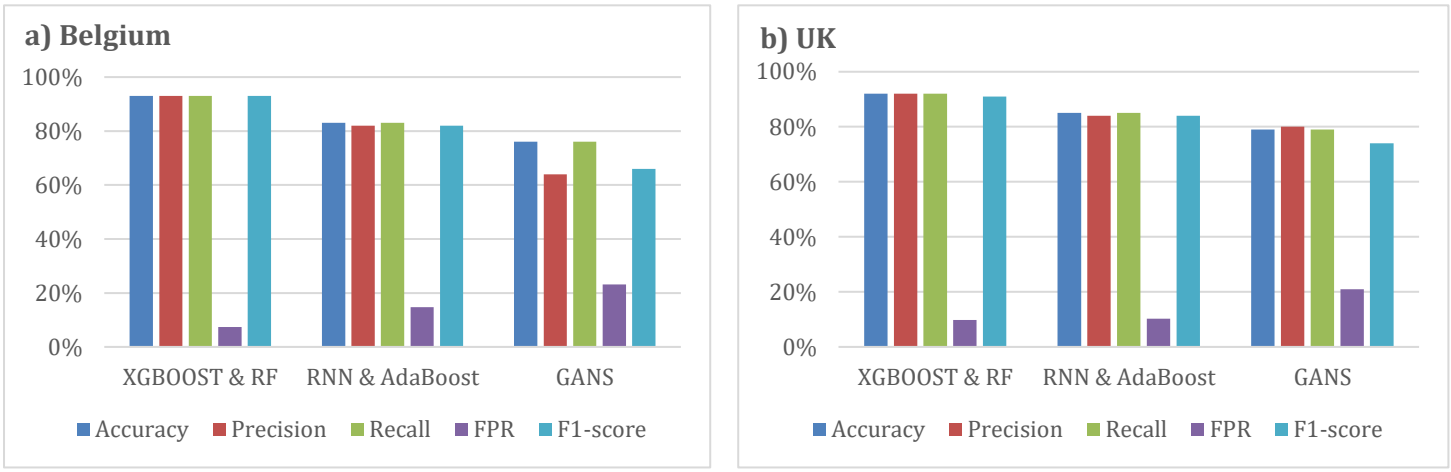

Figure 2: *Performance of the classification models according to the evaluation metrics for a) Belgium and b) UK*

4.2 Interpretability of Classification Models

SHAP (SHapley Additive exPlanations) values quantify the impact of each variable on predictions, aiding the interpretation of machine learning models. In travel-related analysis, SHAP values assess the influence of factors like **travel speed, trip distance, harsh braking, and heart rate intervals** on model outcomes. **Higher SHAP** values for speed and distance highlight their strong impact on driving behavior, while harsh braking indicates aggressive driving. Heartbeat intervals reflect physiological stress, linking driver state to safety.
The results, as shown in Fig. 3 and Fig. 4, identify the most influential features in predicting risky driving behaviors for Belgium and UK respectively.

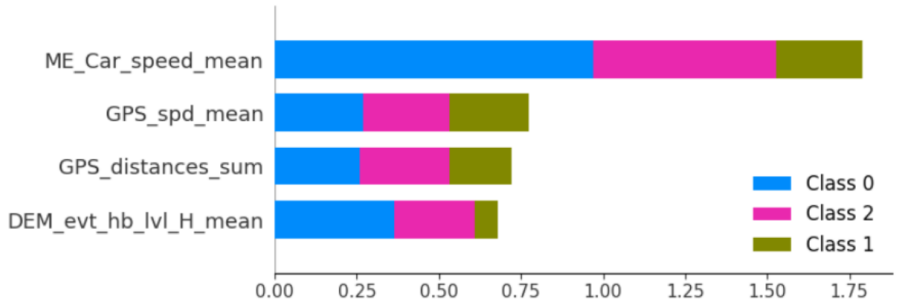

Figure 3: *SHAP results for Belgium*

The SHAP summary bar plot for Belgium reveals that speed-related features, particularly ME_Car_speed_mean, have the highest overall impact on the model's predictions across all three classes, with GPS_spd_mean also contributing significantly but more evenly. GPS_distances_sum plays a moderate role, mostly affecting Class 1 and Class 2, while DEM_evt_hb_lvl_H_mean (likely related to heart rate variability or stress) has the least influence, though still relevant for Class 0 and Class 2. In conclusion, speed metrics are the most influential, with distance and physiological factors supporting but less impactful roles in the classification process.

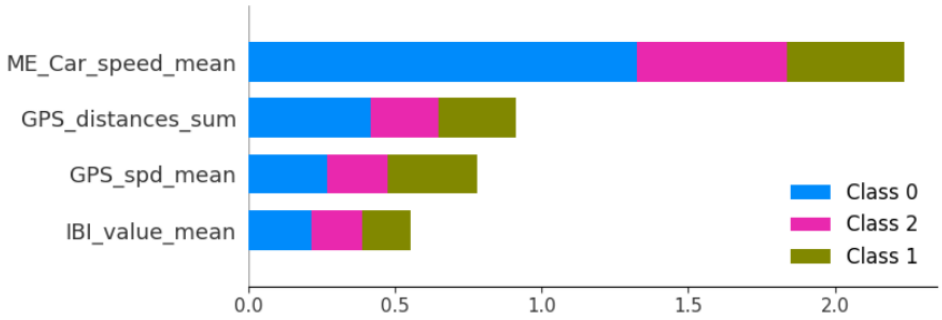

Figure 4: *SHAP results for UK*

The SHAP analysis for UK highlights **ME_Car_speed_mean** as the most influential feature in model predictions across all classes, indicating that average car speed is the primary determinant in classification. **GPS_distances_sum** and **GPS_spd_mean** have moderate contributions, with the former mainly affecting Class 0 and Class 2, while the latter impacts all classes more evenly. **IBI_value_mean**, representing heart rate variability, has the least impact, suggesting that physiological responses play a minor role compared to speed-related factors. Overall, speed metrics dominate model decision-making, while distance and physiological variables provide supporting insights.

Interestingly, SHAP results indicate that while average vehicle speed is the dominant feature in both datasets, the influence of physiological variables (e.g., heart rate) is more pronounced in the UK dataset. This suggests potential cultural or environmental differences in driver stress responses, which warrants further investigation.

5. Discussion

This study developed highly effective models for classifying dangerous driving behaviors using real-world data from drivers in Belgium and the UK. By integrating advanced machine learning and deep learning techniques, the models demonstrated high accuracy in detecting risky driving patterns. A key finding was that specific behaviors, such as travel speed and, total trip distance, serve as critical indicators of driver safety.

The comparative analysis between the two countries revealed distinct driving patterns, with speed and travel distance playing a more significant role in the UK, while speed also has a significant role in Belgium. The hybrid approach, combining Extreme Gradient Boosting (XGBOOST) and Random Forests (RF), proved particularly effective, consistency achieving strong performance across both datasets. Additionally, the study found that higher speeds and longer travel distances we re more critical in both countries. These findings emphasize the need for region-specific safety interventions tailored to local driving behaviors.

A key innovation of this research is the application of the SHAP (Shapley Additive exPlanations) values to interpret model decisions, enhancing transparency in understanding how specific driving factors influence safety assessments. This interpretability addresses concerns about the opaque nature of machine learning models, ensuring that insights are both accurate and actionable for improving road safety policies.

Future research could expand by increasing the dataset size to improve model reliability, particularly for rare events like severe crashes. Integrating additional variables, such as driver demographics and psychological traits, could enable more personalized risk assessments, while considering diverse driving conditions, such as urban versus rural environments and varying weather, would enhance model comprehensiveness.

While the XGBoost-RF hybrid demonstrated the highest classification accuracy, its computational complexity could be a limitation in real-time applications. Similarly, the RNN-AdaBoost model efficiently captures temporal dependencies, making it well-suited for time-series analysis but at the cost of increased training time. Future research could explore optimizing these models for deployment in edge computing environments, where computational efficiency is critical.

Considering the suboptimal performance of the GAN model, uture research should focus on using GANs for data augmentation to address class imbalances, enhancing the robustness of other models like XGBoost-RF and RNN-AdaBoost. A promising direction is integrating GANs with structured frameworks, such as XGBoost, to improve data synthesis and classification accuracy. Additionally, exploring stable variants like Conditional GANs (cGANs) or Wasserstein GANs (WGANs) could address training instability and enhance performance in driving behavior classification. By adopting these approaches, future studies can further enhance the effectiveness of machine learning models in identifying and mitigating risky driving behaviors.

Moreover, addressing the interpretability challenges of GANs is crucial. Unlike tree-based models, which provide clear explanations for their predictions, GANs function as black-box models, making it difficult to derive actionable insights for road safety applications. Future work should explore techniques to enhance GAN interpretability, such as integrating SHAP-based explanations or alternative model architectures that provide greater transparency in decision-making. Additionally, overcoming mode collapse by employing improved GAN training strategies could help ensure a broader and more representative range of synthetic driving behaviors, enhancing their overall contribution to road safety research.

Interestingly, SHAP results indicate that while average vehicle speed is the dominant feature in both datasets, the influence of physiological variables (e.g., heart rate) is more pronounced in the UK dataset. This suggests potential cultural or environmental differences in driver stress responses, which warrants further investigation.

A practical implementation of these models in Advanced Driver Assistance Systems (ADAS) could provide real-time risk assessments and generate personalized driver feedback. By integrating real-time feature extraction with optimized model inference, such systems could enhance road safety by offering proactive warnings to drivers before hazardous behaviors escalate. Expanding the geographic scope of studies would further validate the models across different road systems and driving cultures, providing deeper insights into regional safety dynamics.

Overall, this study represents a major step forward in using machine learning to enhance road safety. By prioritizing interpretability, regional variations, and real-world applications, it lays a strong foundation for future advancements in driver risk assessment and crash prevention.

Data Availability Statement: The data can be provided upon request.

Acknowledgments: This project is carried out within the framework of the National Recovery and Resilience Plan Greece 2.0,funded by the European Union – NextGenerationEU (Implementation body: HFRI), research project “OptiMo - Optimising driver behaviour for safe, green and energy efficient mobility”.

Conflicts of Interest: The authors declare no conflict of interest.

6. References

2017 IEEE Intelligent Vehicles Symposium (IV) : June 11-14, 2017, Redondo Beach, California, USA. (2017). IEEE.

Chen, Y., Wang, K., & Lu, J. J. (2023). Feature selection for driving style and skill clustering using naturalistic driving data and driving behavior questionnaire. *Accident Analysis and Prevention*, *185*. https://doi.org/10.1016/j.aap.2023.107022

Masello, L., Castignani, G., Sheehan, B., Guillen, M., & Murphy, F. (2023). Using contextual data to predict risky driving events: A novel methodology from explainable artificial intelligence. *Accident Analysis and Prevention*, *184*. https://doi.org/10.1016/j.aap.2023.106997

Osman, O. A., Hajij, M., Karbalaieali, S., & Ishak, S. (2019). A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. *Accident Analysis and Prevention*, *123*, 274–281. https://doi.org/10.1016/j.aap.2018.12.005

Seo, H., Shin, J., Kim, K. H., Lim, C., & Bae, J. (2022). Driving Risk Assessment Using Non-Negative Matrix Factorization with Driving Behavior Records. *IEEE Transactions on Intelligent Transportation Systems*, *23*(11), 20398–20412. https://doi.org/10.1109/TITS.2022.3193125

Shangguan, Q., Fu, T., Wang, J., Luo, T., & Fang, S. (2021). An integrated methodology for real-time driving risk status prediction using naturalistic driving data. *Accident Analysis and Prevention*, *156*. https://doi.org/10.1016/j.aap.2021.106122

Shao, Y., Shi, X., Zhang, Y., Shiwakoti, N., Xu, Y., & Ye, Z. (2024). Injury severity prediction and exploration of behavior-cause relationships in automotive crashes using natural language processing and extreme gradient boosting. *Engineering Applications of Artificial Intelligence*, *133*. https://doi.org/10.1016/j.engappai.2024.108542

Vanhoeyveld, J., & Martens, D. (2018). Imbalanced classification in sparse and large behaviour datasets. *Data Mining and Knowledge Discovery*, *32*(1), 25–82. https://doi.org/10.1007/s10618-017-0517-y

Wang, X., Xu, R., Zhang, S., Zhuang, Y., & Wang, Y. (2022). Driver distraction detection based on vehicle dynamics using naturalistic driving data. *Transportation Research Part C: Emerging Technologies*, *136*. https://doi.org/10.1016/j.trc.2022.103561

Wijayaratna, K. P., Cunningham, M. L., Regan, M. A., Jian, S., Chand, S., & Dixit, V. V. (2019). Mobile phone conversation distraction: Understanding differences in impact between simulator and naturalistic driving studies. *Accident Analysis and Prevention*, *129*, 108–118. https://doi.org/10.1016/j.aap.2019.04.017

Zhu, S., Li, C., Fang, K., Peng, Y., Jiang, Y., & Zou, Y. (2022). An Optimized Algorithm for Dangerous Driving Behavior Identification Based on Unbalanced Data. *Electronics (Switzerland)*, *11*(10). https://doi.org/10.3390/electronics11101557

Ziakopoulos, A., Kontaxi, A., & Yannis, G. (2023). Analysis of mobile phone use engagement during naturalistic driving through explainable imbalanced machine learning. *Accident Analysis and Prevention*, *181*. https://doi.org/10.1016/j.aap.2022.106936