Enhancing Risky Driving Behavior Classification Using Conditional GANs (cGANs): A Data Augmentation Approach

Eleni Maria Theodoraki[1*[0009-0001-6033-0728]], Paraskevi Koliou[1[0000-0003-3926-3345]], George Yannis[1[0000-0002-7017-0756]]

[1] National and Technical University of Athens, Iroon Polytechniou 5-9, 15772 Athens, Greece
*e_theodoraki@mail.ntua.gr

Abstract. Accurately classifying risky driving behaviors is essential for road safety, but is hindered by class imbalance, with dangerous behaviors underrepresented in most datasets. This study investigates the use of Conditional Generative Adversarial Networks (cGANs) to generate synthetic data for rare risk categories and improve classification performance. Driving behaviors were categorized as Normal, Dangerous, or Avoidable Accident. Following augmentation, a binary classification scheme (Normal vs. Avoidable Accident) was adopted. A cGAN was trained to generate synthetic high-risk scenarios, which were used to balance the dataset. Classifiers—GAN-based, XGBoost-RF, and RNN-AdaBoost—were evaluated on both original and augmented datasets. Results showed that cGAN-generated data improved model accuracy (Belgium: 76%→90%; UK: 79%→91%). However, hybrid models achieved near-perfect accuracy on augmented data, indicating overfitting. This study highlights both the promise and risks of cGAN-based augmentation, emphasizing the need for careful validation to ensure real-world applicability.

Keywords: Risky Driving Behavior, Classification Models, Data Augmentation, Generative Adversarial Networks (GANs), Conditional GANs (cGANs).

1 Introduction

Road traffic accidents remain a leading cause of injury and mortality worldwide. Risky driving behaviors, such as speeding, harsh braking, and aggressive maneuvers, are primary contributors to crashes [1]. Accurate classification of such behaviors enables proactive interventions, but imbalanced datasets hinder model performance: normal driving dominates, while hazardous cases are underrepresented. This imbalance causes machine learning models to bias toward the majority class, leading to poor detection of risky events.

Traditional machine learning models have been used to classify driver behaviors, but are limited in addressing imbalance [2]. Deep learning methods, including Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks,

better capture temporal dependencies but still overfit to the majority classes [3]. Generative Adversarial Networks (GANs) address imbalance by synthesizing minority-class samples, while Conditional GANs (cGANs) enable label-controlled generation suitable for targeted augmentation. Recent studies have applied GAN-based augmentation in transportation safety, including crash risk prediction [4], autonomous driving validation [2], urban mobility modeling [3], and intelligent agent training [5]. Methodological advances, such as diversity-sensitive GANs [6], generative oversampling frameworks [7], and Conditional Tabular GANs [8], have demonstrated improved representation of minority classes. Additionally, cGANs have been explored for anomaly detection and dangerous scenario generation [9,10]. While these studies report performance gains, limited attention has been given to the trade-off between accuracy improvement and generalization risks in safety-critical classification tasks. This gap motivates the present study. Unlike prior studies focusing primarily on performance gains, this work systematically evaluates the trade-off between accuracy improvement and generalization risk, providing practical guidance for applying synthetic data in safety-critical transportation analytics.

2 Methodology

2.1 Data Collection

This study draws on the i-DREAMS naturalistic driving project, where drivers in Belgium and the UK were monitored across four phases: baseline, ADAS intervention, post-drive feedback, and gamification. A total of 43 Belgian drivers (7,163 trips, 147,337 minutes) and 26 UK drivers (8,226 trips, 118,175 minutes) contributed data.

The experiment was conducted over four months and was divided into four distinct phases: the Baseline Phase, ADAS Intervention Phase, Performance Feedback Phase, and Gamification Phase. Throughout all phases, real-time driving behavior was continuously monitored. The study assessed the effectiveness of ADAS warnings and post-drive feedback mechanisms in influencing driver behavior.

Each vehicle was equipped with an OBD-II device and a mobile data acquisition system, enabling continuous collection of speed, acceleration, braking, and headway information. Every 30-second segment was categorized into one of three safety levels: *Normal*, *Dangerous*, or *Avoidable Accident*. The dataset provides over 265,000 minutes of driving experience across two countries collected under a variety of behavioural intervention conditions, allowing the performance of classification models to be assessed across a representative range of driving and behavioral states.

2.2 Definition of ‘Safety Tolerance Zone’

The target risk modulation, the Safety Tolerance Zone, is a classifier for the driving performance in the a priori defined discrete time slots. In the current study, the entire driving trajectory is divided into 30-second time intervals and classified into risk

categories corresponding to the highest level of the safety tolerance zone. This method provides a consistent temporal resolution that allows for comparison between drivers, trips, and experimental conditions. The classification method using defined intervention thresholds is divided into three categories. The first category is Normal, which includes low-risk driving behavior. The second is Dangerous, medium-risk driving behavior. The third is Avoidable Accident, a high-risk driving behavior requiring immediate intervention. This classification was based on two primary performance indicators:

- Headway distance (distance between the driver's vehicle and the preceding vehicle).
- Speeding behavior (excessive deviation from speed limits).

Given that safety-critical driving events are less frequent, it was expected that Normal driving instances would dominate the dataset, while Dangerous and Avoidable Accident cases would be underrepresented. The dataset contained intervention variables labeled as iDreams_Headway_Map_level_i and iDreams_Speeding_Map_level_i, where *i* represented intervention levels from -1 to 3. When 0: The intervention level does not correspond to i, and 1: The intervention level corresponds to *i*. The final classification framework was:

- Normal: Highest intervention level recorded as -1, 0, or 1.
- Dangerous: Highest intervention level recorded as 2.
- Avoidable Accident: The highest intervention level recorded was

This hierarchical classification ensured an accurate representation of real-world driving risks, providing a foundation for subsequent machine learning models aimed at predicting and mitigating unsafe driving behaviors.

2.3 Data Collection and Preprocessing

Risk Level Categorization translates the conceptual Safety Tolerance Zone into measurable thresholds. The dataset utilized in this study consists of real-world driving data collected from naturalistic driving experiments. To ensure consistency and enhance model training, several preprocessing steps were performed:

1. Data Cleaning: Missing values (-9999) were identified and removed to prevent inconsistencies.

2. Feature Normalization: All numerical features, including *GPS_distances_sum* and *GPS_spd_mean*, were normalized using Min-Max scaling to standardize feature distributions.

3. Risk Level Categorization: Each driving instance was classified into one of three risk categories based on excessive speeding and harsh braking behaviors (see Table 1).

4. Class Balancing: Since the dataset exhibited significant class imbalance, with "Normal" behaviors dominating, data augmentation techniques were applied to ensure a more equitable class distribution before training the models.

Table 1: Risk Level Classification Criteria

Risk Level	Condition
Normal	iDreams_Speeding_Map_level_3_mean ≤ 0.2 AND DEM_evt_hb_lvl_H_mean = 0
Dangerous	0.2 < iDreams_Speeding_Map_level_3_mean ≤ 0.6 OR 0 < DEM_evt_hb_lvl_H_mean ≤ 0.3
Avoidable Accident	iDreams_Speeding_Map_level_3_mean > 0.6 OR DEM_evt_hb_lvl_H_mean > 0.3

2.4 Conditional GAN (cGAN) Architecture

To create fictitious examples of dangerous driving behavior, a Conditional GAN (cGAN) was used. The model was trained mainly on "Normal" driving trajectories, which were the majority class. Its goal was to generate balanced representations of the underrepresented "Dangerous" and "Avoidable Accident" categories.

The architecture consists of a generator using fully connected layers with batch normalization and Leaky ReLU activation. The generator creates realistic, label-specific driving data after receiving a noise vector and risk label. Also, a Discriminator that **uses** convolutional layers trained with gradient penalty and Wasserstein loss for stable convergence and reduced mode collapse to assess sample authenticity and class consistency.

This configuration improved classifier robustness and increased data diversity by enabling the controlled synthesis of high-risk driving patterns. The model was trained using alternating generator–discriminator updates until convergence.

2.5 Classification Model Training and Evaluation

Three models were used to evaluate the efficacy of the cGAN-augmented dataset. The first model was Extreme Gradient Boosting and Random Forests combined in XGBoost-RF, a tree-based ensemble that was chosen for its interpretability and good performance on structured data. The second model, RNN-AdaBoost, addresses class imbalance by combining adaptive boosting with recurrent neural networks to capture temporal patterns. The last model, the cGAN Classifier, prior to and following augmentation, the cGAN's capacity to distinguish between Normal and Avoidable Accident cases was assessed through testing as a classifier.

The classification performance was assessed by using standard evaluation metrics: 1) Accuracy, 2) Precision, 3) Recall, and 4) F1-score. To ensure robustness, k-fold cross-validation was applied. Additionally, SHAP analysis was used to interpret feature importance and explain model decision-making. Stratified 5-fold cross-validation was

applied to preserve class distribution across folds. The dataset was split into training and validation subsets at an 80/20 ratio before augmentation.

2.6 Assessing the Impact of Data Augmentation

To evaluate the effectiveness of cGAN-generated synthetic data, model performance was compared across two datasets. First, the original dataset was used to train and test models solely on real-world driving data. Second, the augmented dataset incorporates the cGAN-generated synthetic data into training to balance class distributions.

3 Results

This section presents the results of the classification models trained on both the original and augmented datasets.

3.1 Performance of Classification Models

Table 2 provides a comparative analysis of classification performance across the XGBoost-RF, RNN-AdaBoost, and GAN-based models for both the original and augmented datasets.

Table 2: Performance Comparison of Classification Models

Dataset	***Model***	***Accuracy***	***Precision***	***Recall***	***F1-score***
Belgium (original dataset)	XGBOOST & RF	93%	93%	93%	93%
	RNN & AdaBoost	83%	82%	83%	82%
	GANS	76%	64%	76%	66%
UK (original dataset)	XGBOOST & RF	92%	92%	92%	91%
	RNN & AdaBoost	85%	84%	85%	84%
	GANS	79%	80%	79%	74%
Belgium (Augmented dataset)	XGBOOST & RF	100%	100%	100%	100%
	RNN & AdaBoost	100%	100%	100%	100%
	cGANS	90%	90%	90%	90%
UK (Augmented dataset)	XGBOOST & RF	100%	100%	100%	100%
	RNN & AdaBoost	100%	100%	100%	100%
	cGANS	91%	90%	91%	91%

Beyond absolute performance, differences in the relative performance of models also give an understanding. For example, while the relative accuracy of the XGBoost-RF model was highest on the original dataset, the cGAN classifier showed the greatest relative improvement. This suggests that the use of synthetic data is mostly helpful for improving the representation of minority classes.

3.2 Overfitting Concerns and Generalization

It is believable that the higher performance of XGBoost-RF and RNN-AdaBoost obtained on the augmented data set is owing to overfitting the training data and, hence, learning spurious correlations only present in synthetic data rather than actual behavioural correlations. The smaller variance of samples produced in this case may have led to memorization of training data by the models rather than generalizing the distribution. Models may have become overconfident in detecting risky behaviors, leading to a higher false positive rate, as evidenced by K-fold cross-validation, which demonstrated a significant recall improvement but only modest precision gains. While the recall scores improved considerably after augmentation, the precision scores only improved marginally, meaning that the models learned to identify risky behaviour but did not get better at distinguishing true positives from false positives.

3.3 Feature Importance and Model Interpretability

Key feature contributions to risk classification were identified by SHAP analysis:

- The most significant variables across datasets were those related to speed (GPS_spd_mean and iDreams_Speeding_Map_level_3_mean).
- To differentiate between "Dangerous" and "Avoidable Accident" situations, harsh braking events (DEM_evt_hb_lvl_H_mean) were essential.
- Concerns about overfitting were raised by the fact that models trained on the augmented dataset depended more on synthetic features.

The reliance on synthetic-feature structures suggests that, despite its helpful effect on class distribution, augmentation may alter feature importance distributions. Thus, interpretability analyses should always be performed on any augmentation experiment to avoid misinterpreting reliance on behavioural signals as model reliance on artefacts of the generated data.

4 Discussion

While model performance improved across key evaluation metrics, several critical challenges and limitations must also be considered.

4.1 Effectiveness of Data Augmentation

The results show that adding synthetic data notably improved recall and overall accuracy across all models, indicating that cGAN-generated samples effectively supplemented high-risk instances and helped address class imbalance.

However, the gain in precision was limited. Although models detected more risky events, they also generated more false positives, likely due to overfitting on synthetic patterns that may not generalize well to real-world data.

4.2 Overfitting and Generalization Challenges

Although accuracy improved, analysis of precision and false positive rates reveals concerns about generalization. The growing gap between precision and recall in the augmented dataset suggests models may have overfit to synthetic patterns instead of learning true high-risk behaviors. This was especially clear in the XGBoost-RF and RNN-AdaBoost models, which achieved perfect accuracy on augmented data but showed inconsistent performance on real-world samples.

The SHAP analysis further confirmed overfitting tendencies, revealing an over-reliance on synthetic features. While the synthetic dataset improved class balance, the uniformity of generated samples may have introduced biases that the models exploited. This highlights the need for future research to explore more diverse data augmentation and regularization strategies to improve robustness.

4.3 Practical Implications for Road Safety

From a real-world perspective, enhancing the identification of high-risk driving behaviors has substantial implications for road safety applications. Improved classification models can be leveraged in real-time risk assessment systems, including driver assistance warnings, autonomous vehicle safety measures, and insurance telematics applications.

However, ensuring that machine learning models generalize effectively across diverse driving conditions and driver behaviors is essential for practical deployment. Future research should evaluate robustness using external validation datasets and controlled simulations.

5 References

1. World Health Organization: Global status report on road safety 2023. World Health Organization (2023). Available at: https://www.who.int/publications/i/item/9789240068888
2. Zhang, M., Zhang, Y., Zhang, L., Liu, C., Khurshid, S.: Deeproad: GaN-based metamorphic testing and input validation framework for autonomous driving systems. In: Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE 2018), pp. 132–142. Association for Computing Machinery (2018).

3. Rajagopal, B.G. et al.: A hybrid Cycle GAN-based lightweight road perception pipeline for road dataset generation for urban mobility. PLoS One 18(11) (2023).
4. Park, N., Park, J., Lee, C.: Conditional generative adversarial network-based roadway crash risk prediction considering heterogeneity with dynamic data. Journal of Safety Research 92, 217–229 (2025).
5. Giannoulopoulos, A.: Utilising GANs for training intelligent agents for autonomous vehicles in a simulator with deep learning (2024). Available at: https://ikee.lib.auth.gr/record/359279/files/Giannoulopoulos.pdf
6. Yang, D., Hong, S., Jang, Y., Zhao, T., Lee, H.: Diversity-sensitive conditional generative adversarial networks (2019). Available at: http://arxiv.org/abs/1901.09024
7. Li, L., Zhang, X.: Addressing data imbalance in collision risk prediction with active generative oversampling. Scientific Reports 15(1) (2025).
8. Chen, J., Pu, Z., Zheng, N., Wen, X., Ding, H., Guo, X.: A generative deep learning approach for crash severity modeling with imbalanced data (2024). Available at: https://doi.org/10.48550/arXiv.2404.02187
9. Qiu, Y., Misu, T., Busso, C.: Driving anomaly detection with conditional generative adversarial network using physiological and CAN-bus data. In: Proceedings of the 2019 International Conference on Multimodal Interaction (ICMI 2019), pp. 164–173. Association for Computing Machinery (2019).
10. Xu, S.: A framework for generating dangerous scenes: Towards explaining realistic driving trajectories. PhD thesis, Santa Cruz (2023). Available at: https://www.proquest.com/openview/2eefcff7aeef3030d4304fafa47e2b00