Adversarial Training to Improve Deep-Learning Intrusion Detection in the Internet of Vehicles

Research proposal — MSc Artificial Intelligence

Yulan Galagoda
PROJ518: MSc Dissertation & Research Skills
Supervisor: Dr. Shaymaa Al-Juboori
University of Plymouth · 2026

Research proposal — work in progress

Contents

1. Introduction to the research area

Can the very techniques used to attack deep-learning intrusion detectors be repurposed to fix their greatest weakness on realistic data? This proposal sets out to find out.

The rapid evolution of the Internet of Vehicles (IoV) has turned modern transport into a highly connected ecosystem. Vehicles rely heavily on internal protocols such as the Controller Area Network (CAN) bus to manage critical functions from braking to engine control. This connectivity improves efficiency and safety, but it also introduces significant security risk: the CAN bus was designed without native encryption or authentication, leaving it inherently vulnerable. As Arafah et al. (2025) note, the expanding attack surface of connected vehicles demands robust defences to prevent unauthorised access that could compromise passenger safety.

Artificial intelligence and machine learning have emerged as the leading approach to securing these networks, detecting malicious activity by learning normal vehicle behaviour and flagging anomalies. Alkadi et al. (2024) observe that while traditional machine-learning models are often favoured for their simplicity in resource-constrained environments, they struggle to adapt to complex, evolving threats — leading researchers towards deep-learning models that can detect more sophisticated attacks with greater precision.

Deep learning in this domain faces two critical obstacles. The first is vulnerability to adversarial attacks: Mamun et al. (2025) demonstrate that attackers can introduce imperceptible noise into network traffic to make deep models misclassify malicious activity as benign. The second is data dependence: Le and Alsmadi (2026) argue that many high-performing models rely on massive data redundancy, and that when datasets are strictly cleaned to remove duplicates and reflect real-world conditions, these complex models fail to generalise for lack of unique training examples.

This research focuses on the intersection of those two problems. It investigates whether the techniques used to generate adversarial attacks can instead be repurposed to solve data scarcity. By applying adversarial training to strictly de-duplicated data, the project aims to develop a deep-learning detector that is both robust against manipulation and high-performing in realistic, data-constrained conditions.

2. Aim and objectives

The primary aim is to develop a robust deep-learning intrusion-detection system for IoV networks that overcomes the limitations of data scarcity — demonstrating that adversarial training can be used as a data-augmentation strategy to enable deep-learning models to outperform lightweight machine-learning algorithms on strictly de-duplicated, realistic forensic datasets. To achieve this, five objectives are defined:

1. **Strict data hygiene.** Process the CICIoV2024 benchmark with strict de-duplication, isolating unique attack signatures to replicate the realistic, resource-constrained environment described by Le and Alsmadi (2026).

2. **Performance baselines.** Evaluate and compare a 1D Convolutional Neural Network against a Random Forest on the de-duplicated dataset, quantifying how far deep models degrade when denied redundant training data.

3. **Adversarial stress tests.** Synthesise adversarial examples using the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) to assess the fragility of both models against evasion attacks.

4. **Adversarial augmentation.** Implement a multi-strategy adversarial-training framework that uses the generated samples to retrain the network, enriching the training distribution of the clean dataset.

5. **Robustness and superiority.** Rigorously test the adversarially-trained model against the Random Forest baseline to determine whether the defence restores the superiority of deep learning against both standard and adversarial threats.

3. The research context

Recent work on the CICIoV2024 benchmark presents a contradiction about the optimal approach to IoV security. Arafah et al. (2025) established the dataset's validity, reporting near-perfect detection (F1 ≈ 1.0) with ensemble methods such as Extra Trees, suggesting standard machine learning was sufficient for vehicular networks.

The accuracy trap and adversarial fragility

Mamun et al. (2025) challenged that optimism, showing deep-learning architectures suffer performance drops of up to 40% under attacks such as FGSM and arguing that adversarial training is essential for robustness. A more fundamental issue was identified by Le and Alsmadi (2026), who revealed that the high accuracy reported by Arafah et al. relied on massive data redundancy (99.75% duplicate records). When strict de-duplication isolated unique attack signatures, complex 1D-CNN models failed to generalise (F1 < 0.55) through data scarcity, whereas lightweight Random Forests remained stable.

The proposed synthesis

This creates a dilemma: deep learning is required for complex threats yet fails on realistic forensic data. Alkadi et al. (2024) confirm that lightweight models, though efficient for IoT, remain catastrophically vulnerable to black-box evasion. To resolve this, the project integrates findings from Zhu et al. (2026) — who showed multi-strategy adversarial training could restore deep-learning accuracy from 24% to 97% with minimal overhead — and Yang et al. (2020), who validated adversarial mechanisms as data-augmentation tools in data-starved environments. The research therefore hypothesises that applying adversarial training to the strictly de-duplicated dataset of Le and Alsmadi will resolve data scarcity, enabling deep-learning models to finally outperform Random Forests in resource-constrained environments.

4. The proposed research design

The study uses a quantitative experimental design to evaluate adversarial training as a data-augmentation strategy, isolating data scarcity by using strictly de-duplicated data to replicate the conditions under which deep learning typically fails.

Methodology

The CICIoV2024 dataset is processed with Python and Pandas. Following Le and Alsmadi (2026), a strict de-duplication algorithm reduces millions of records to unique attack signatures. Two comparative models are developed: a Random Forest baseline representing standard IoT stability (Alkadi et al., 2024), and a 1D-CNN selected for temporal feature extraction but known to underperform on de-duplicated data.

Experimental phases

1. **Baseline evaluation.** Both models are trained on clean data; the Random Forest is expected to outperform the 1D-CNN given the network's sensitivity to data scarcity.

2. **Adversarial stress testing.** The Adversarial Robustness Toolbox (ART) generates FGSM and PGD attacks against the models, empirically proving the vulnerability of lightweight models to evasion (Mamun et al., 2025).

3. **Adversarial augmentation.** Following Zhu et al. (2026), the generated adversarial samples are integrated into the 1D-CNN training set, testing the core hypothesis: that synthetic "hard" examples can compensate for lack of volume, enabling the 1D-CNN to surpass the Random Forest.

Risk mitigation

Stratified K-fold cross-validation is used throughout to ensure rare attack classes are evenly distributed across training and testing iterations.

5. Project risk assessment

Successful completion relies on complex computational experiments. The principal risks and their mitigations are summarised below.

Risks and mitigations

Risk	Description	Mitigation
Technical implementation	Generating valid CAN-bus frames with the Adversarial Robustness Toolbox (ART) that preserve structural integrity may be complex.	Use established open-source implementations and run a prototyping phase on a small subset to validate the attack pipeline before scaling.
Data sufficiency	Strict de-duplication (removing ~99.75% of records) may leave a dataset too small for the deep model to converge, even with augmentation.	If convergence fails, apply stratified relaxed de-duplication — retaining a controlled share of duplicates for minority classes only — keeping the challenge hard but learnable.

Risk	Description	Mitigation
Result inconclusiveness	The adversarially-trained 1D-CNN may still not outperform the Random Forest.	A negative result is a valid contribution: the analysis would shift to the feature limitations preventing generalisation, critiquing the limits of adversarial defence in low-data regimes.

6. Impact

This research addresses critical safety requirements for the Internet of Vehicles. Scientifically, it validates adversarial training as a data-augmentation strategy: by showing adversarial examples can resolve data scarcity in forensic datasets, it offers a methodological blueprint for applying deep learning in data-sensitive domains — such as medical and financial — where privacy law necessitates strict de-duplication.

UN Sustainable Development Goals

The project aligns with Goal 9 (Target 9.5) by upgrading automotive cybersecurity capability without massive computational resources, and with Goal 11 (Target 11.2) by helping prevent malicious vehicle hijacking, protecting passengers and pedestrians in future smart cities.

Commercial and economic impact

Economically, the findings help manufacturers meet stringent standards such as UN Regulation No. 155. A lightweight, adversarially-robust model offers a commercially viable solution deployable on existing hardware, lowering the barrier to secure autonomous technologies.

Maximising impact

To support real-world adoption, source code and datasets will be open-sourced for community validation, findings will be submitted to peer-reviewed venues (e.g. IEEE Transactions on Intelligent Transportation Systems), and a technical white paper will be shared with industry bodies such as SAE International to advocate adoption of robust deep-learning models over legacy systems.

References

Alkadi, S., Al-Ahmadi, S. and Ismail, M.M.B. (2024) 'RobEns: robust ensemble adversarial machine learning framework for securing IoT traffic', *Sensors*, 24(8), p. 2626.

Arafah, M. et al. (2025) 'Advancing IoV security through machine learning: design and evaluation of the CICIoV2024 benchmark dataset'. Amman: IEEE.

Le, H. and Alsmadi, I. (2026) 'Intrusion detection for Internet of Vehicles CAN bus communications using machine learning: an empirical study on the CICIoV2024 dataset', *Future Internet*, 18(1), p. 14.

Mamun, Q. et al. (2025) 'Adversarial attacks on machine learning-based IDS for V2X networks: a CICIoV2024 study'. Oslo: IEEE.

Yang, Y. et al. (2020) ‘Network intrusion detection based on supervised adversarial variational auto-encoder with regularization’, *IEEE Access*, 8, pp. 42169–42184.

Zhu, J. et al. (2026) ‘STS-AT: a structured TensorFlow adversarial training framework for robust intrusion detection’, *Sensors*, 26(2), p. 536.