Towards the Responsible/Trustworthy AI in Multi-Domain Operations for Cyber Social Security: A Black-Box AML Case Study in the CAN Bus Frame Detection Task

Vita Santa Barletta[1,*,†], Danilo Caivano[1,†], Christian Catalano[1,†], Samuele del Vescovo[2,*,†] and Michele Scalera[1,†]

[1] *Università degli studi di Bari Aldo Moro, Piazza Umberto I, 70121 Bari, Apulia, Italy*

[2] *Scuola IMT Alti Studi Lucca, Piazza S.Francesco, 19, 55100 Lucca, Apulia, Italy*

Abstract

The automotive sector has witnessed significant advancements driven by the enhanced connectivity of vehicles even with Smart City infrastructures. This evolution expands the vehicles' attack surface in unforeseen ways, exposing continental-wide critical networks important for governmental and military operations. Considering today's threat of Multi-Domain Operations (MDOs), it isn't difficult to imagine vehicles as perfect victims of attacks related to future complex MDOs with the ultimate goal of affecting people-related effect dimension's. Special attention should be paid to the security of Machine Learning (ML) based Intrusion Detection Systems (IDSs), useful to detect intrusions in Controller Area Network (CAN) protocol-based In-Vehicles networks. Adversarial Machine Learning (AML) attacks in the Black-Box scenario pose a concrete threat to such IDSs making the task of defense really challenging. Therefore, the main goal of this work is to understand the possible importance of some hyperparameters related to Decision Tree (DT)-based Ensemble models (on which the Supervised ML IDS is based) in the CAN Bus Frame Detection Task, recognized as an inherent defense (or deterence) tool from Black-Box AML attacks (seen as the "Cyber" part of a MDO). The victim core models are Technology Transfer state-of-the-art Random Forest bagging-based (RF), Gradient Boosting (GB) and Extreme Gradient Boosting (XGB). The attack considered is Zeroth Order Optimization. The experimental results show the hyperparameters related to the bagging trees number's per RF and to the boosting rounds number's per GB influence the attack time. This cannot be seen for the one related to the boosting rounds number's per XGB. The correct choice of these values can be a perfect Responsible/Trustworthy AI best practices' example for the Robustness/Security of automotive ML systems. The secondary goal is to study the impact (qualitative) of such evidence on the organizational units (Detection, Response and Prevention) in the Cyber Social Security (CSS) in MDOs. Generally, a Very High impact is estimated considering the importance of such evidence in threat detection and response.

Keywords

Trustworthy AI, Automotive Cybersecurity, Black-Box Adversarial Machine Learning, Multi-Domain Operations

1. Introduction

Recently, the automotive industry is advancing at a rapid pace, recognizing Connected and Autonomous Vehicles (CAVs) and Internet of Vehicles (IoV) technologies as essential assets for achieving long-term sustainability within Smart Cities [1, 2].

The evolution of shared mobility as well as the evolution of electric mobility are strategies for optimising any travel with a view to greater sustainability [2, 3]. Like any innovation, this development climate brings new and demanding challenges (especially) related to automotive cybersecurity and ultimately to individuals' safety [4, 5].

This concept of technological and social development can be linked to the concept of Multi-Domain Operations [6] (MDO) which is purely a military one. Considering the strong interconnection that binds

Joint National Conference on Cybersecurity (ITASEC & SERICS 2025), February 03-8, 2025, Bologna, IT

[*] Corresponding author.

[†] These authors contributed equally.

vita.barletta@uniba.it (V. S. Barletta); danilo.caivano@uniba.it (D. Caivano); christian.catalano@uniba.it (C. Catalano); samuele.delvescovo@imtlucca.it (S. d. Vescovo); michele.scalera@uniba.it (M. Scalera)

0000-0002-0163-6786 (V. S. Barletta); 0000-0001-5719-7447 (D. Caivano); 0000-0003-4038-2317 (C. Catalano); 0009-0001-8220-5135 (S. d. Vescovo); 0000-0002-2455-2032 (M. Scalera)

the five operational warfare domains (i.e. “Land”, “Sea”, “Air”, “Space” and “Cyber” [6]), it would be very easy for any attacker (more or less organized) to attack Controller Area Network (CAN) protocol-based In-Vehicles networks (IVNs) related to civilian-use vehicles by exploiting a set of activities (including military ones) conducted through different domains to perceive, understand, and orchestrate “effects” aimed at generating events at a rate beyond the adversary’s decision-making capability [7, 8].

In such a complex scenario, the integration of Artificial Intelligence (AI) and Machine Learning (ML) into the automotive sector should be regarded not only as innovative enhancements to vehicle security but also as essential tools for preventing situations that may compromise the psychological and physical well-being of passengers [9]. One of the most worrisome threats is malicious intrusions into In-Vehicle Controller Area Network (CAN) based networks [10]. So, advanced security measures incorporating not only traditional defense systems but also AI/ML based approaches are needed like the ML-based Intrusion Detection Systems (IDSs) [11, 12].

Compromising that system allows for the execution of various attacks (targeting assets within the "Transport" sector [13]) through corrupted Electronic Control Units (ECUs), considered as nodes within In-Vehicle Networks, ultimately causing abnormal behavior within the vehicle itself [14]. Within this context, the cognitive well-being of passengers may also be compromised, with serious repercussions stemming from the covert actions conducted by malicious actors against ML-based IDS. This concept is the link between automotive cyber security and Cyber Social Security (CSS) [15].

The most worrisome tool for conducting attacks on ML-based systems is Adversarial Machine Learning (AML) [16, 17, 18]. In the case of evasive paradigm, it’s possible to manipulate the input data (at the testing/monitoring time [19]) of the victim model, for example by altering any type of image (or any CAN processed frame) in imperceptible ways, to fool the victim ML model which will misclassify that example [20, 21]. In general, AML attacks can be executed in three distinct scenarios, each varying based on the attacker’s knowledge of the target system’s architecture and parameters. The Black-Box scenario is both the most probable and accessible for attackers, as it requires no prior insight into the victim system’s structure [16, 22]. Current literature on the application or conceptualization of Black-Box attacks within the CAN bus frame detection task remains limited and in an early stage of development.

Therefore, this paper shows an empirical case study on the importance of some Decision Tree (DT)-based Ensemble models hyperparameters used as the core of Supervised ML-based IDS in the CAN Bus Frame Detection Task in automotive cyber security scenarios. It’s supposed the IDS is installed in the vehicle itself and attacked via a Black-Box AML attack i.e. Zeroth Order Optimization (ZOO) in a pure Black-Box Evasive Scenario. This attack is labelled as the "Cyber" part of an MDO. Several algorithms underlying this analysis are considered: Random Forest bagging based (RF), Gradient Boosting (GB) and Extreme Gradient Boosting (XGB). Basically, the time needed to generate adversarial examples (for each victim ML model) is empirically evaluated as the values associated with these parameters change. This provides a qualitative estimate of the evolution of that time with the goal of providing a concrete demonstration to any defense team (in any Vehicle-SOC) regarding the most appropriate values to associate with such hyperparameters. The experimental results show the hyperparameters related to the bagging trees number’s per RF and to the boosting rounds number’s per GB influence the attack time needed. This cannot be seen for the one related to the boosting rounds number’s per XGB. These can be seen as an inherent defense (or deterrence) tool from Black-Box AML attacks capable of controlling the attack time (for the RF and GB case). Choosing these values correctly can be a perfect example of Responsible/Trustworthy AI (by Design) best practice [23, 24, 19]. In particular, we allude to the Robustness [25] and Security [26] of the chosen ML models under Black-Box AML attacks [27].

Moreover, the secondary goal of this work is to qualitatively identify the (negative) impact resulting from worst practices programming (i.e. inappropriate values for the previously mentioned hyperparameters) related to the application of the ZOO attack (in a pure Black-Box scenario) on the "Detection", "Response" and "Prevention" axes of the framework for Cyber Social Security in MDOs. Figure 1 presents two dimensions, Horizontal and Vertical, which combine Technical and Organisational requirements for integrating various operations. The Horizontal dimension is defined by five operational warfare domain which allow to identify for each layer methodologies, tools and techniques necessary to define in each dimension the Detection-Response-Prevention life cycle. It enables the definition of different tactical and strategic levels, aiming to vertically organize security operations through the integration of technical aspects of each domain facilitating the inclusion of cyber aspects in MDO. The definition of the three operational units along the Vertical dimension allow to manage the impact on the civil context and, specifically, on Cyber Social Security [15, 28].

There is no shadow of a doubt that setting the right values can negatively affect the time required for the attack by favouring (hopefully as much as possible) the defense team by giving away valuable time to unearth the previous Vulnerability Assessment and Penetration Testing (VAPT) attempt suffered by the In-Vehicle network.

This entire analysis can be recognized as a critical component of any Cyber Threat Intelligence (CTI), designed to comprehensively assess the risks posed by such threats. This intelligence could prove indispensable in countering adversaries during Multi-Domain Operations (MDOs) [7]. In summary, the following research questions (RQs) can be highlighted:

- RQ1: Can the hyperparameters related to the bagging trees number's per RF, to the boosting rounds number's per GB and to the boosting rounds number's per XGB influence the ZOO attack related adversarial examples' generation time applied to Supervised ML-based IDS in the CAN Bus Frame Detection Task in a Black-Box attack scenario?
- RQ2: Is it possible to qualitatively quantify the (negative) impact of these values on the "Detection", "Response" and "Prevention" axes of the framework for CSS in MDOs?

So, the main contributions are:

- Empirically detect the possible influence of the hyperparameters related to the bagging trees number's per RF, to the boosting rounds number's per GB and to the boosting rounds number's per XGB under on the ZOO adversarial examples generation time applied to Supervised ML-based IDS in the CAN Bus Frame Detection Task in a Black-Box attack scenario;
- Qualitatively quantify the (positive) impact of these values on the "Detection", "Response" and "Prevention" axes of the CSS framework in MDOs.

The idea underlying the research work is to accelerate the process of innovation and awareness associated with exploitable disruptive technologies in one or more domains during an MDO and to contribute to developing a multidimensional national deterrence approach [7].

The rest of the paper is organized as follows: section 2 provides some related works, section 3 shows the experimental setup; section 4 illustrates the results and the discussions; section 5 concludes the paper.

2. Related Work

2.1. Black-Box Evasive AML for CAN Bus Frame Detection

To the best of our knowledge, the scientific literature concerning Black-Box AML attacks on ML-based IDS within the context of CAN bus frame detection remains relatively limited.

For example, Aloraini et al. [20] have conducted an adversarial attack using a substitute victim IDS, trained on data extracted from the OBD-II interface. This dataset is different from the one used to train the real victim IDS [20]. This scenario is not a pure Black-Box one since the transferability of the adversarial examples is exploited [20]. The victim IDS models were: a baseline proprietary DNN-based IDS and one state-of-the-art model i.e. MTH-IDS. The surrogated models were a DNN and a DT. The dataset exploited for the surrogated model is the Car Hacking Dataset [29]. Several White-Box AML attacks were considered like Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), Projected Gradient Descent (PGD) and Jacobian-based Saliency Map Attack (JSMA) [20]. The experimental results have shown the decrease of the F1 scores from 95% to 38% and from 97% to 79% respectively for the real victim models [20].

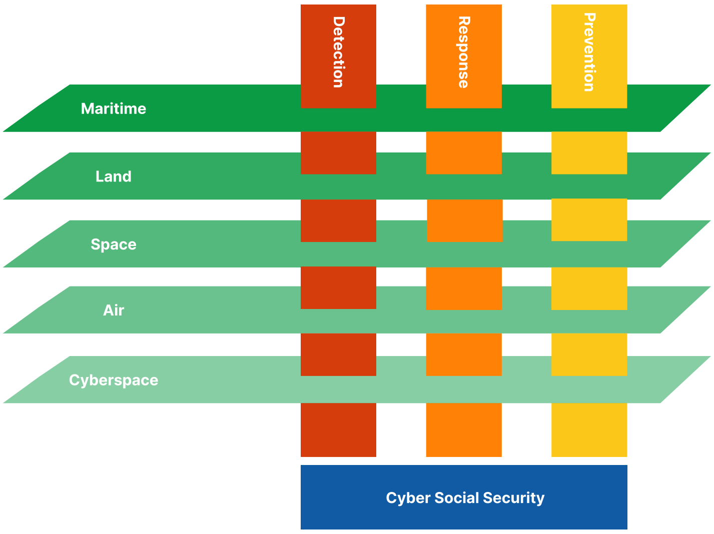

Figure 1: CSS-WAR Architecture

The Zenden et al. work [30] investigated the impact of FGSM attack on the performance of many ML and DL models in a (surrogated) Black-Box attack scenario. Furthermore, researchers have tested that the adversarial training is a countermeasure to this attack (with excellent results). The models tested consist of BL-DNN, BL-Ensemble, SOTA-CNN, and SOTA-LSTM [30]. The dataset exploited in this work is the Survival dataset [31] (only on the 2010 HYUNDAI YF Sonata part). The results showed the ML models are the most afflicted by this attack (loss of accuracy of about 40%) with respect to DL algorithms.

Longari et al. [21] have proposed a novel approach to deploy (pure) Black-Box evasion attacks to ML-based IDS in the examined task (in an online situation). Specifically, that methodology considers the transmission flow portion of CAN frames for the attack (thus not considering these frames in their own right). The approach taken into account analyzes a "sliding window" of Payload [21] unlike to consider a single (entire) example once. The dataset used in this experimentation is called "ReCAN", collected from real vehicles (the "ID C-1" part related to an Alfa Giulia Veloce) [21]. They have evaluated different ML algorithms as the core of IDS, i.e. Small-LSTM, Small-GRU, Large-GRU, CANnolo, Neural Network, Vector Auto-Regressive (VAR), Hamming based [21].

The aforementioned research studies do not account for attacks directly targeting the victim IDS within a true Black-Box scenario (in TT contexts). They do not address state-of-the-art Black-Box AML attacks that are not specifically tailored for the task at hand. The Barletta et al. [32] research work has investigated applying the ZOO attack in a pure Black-Box scenario in the same task (on the Supervised ML algorithms). This attack is often used in tasks related to image recognition (including automotive). Researchers exploited the OTIDS dataset [33] following the Bari et al. [34] workflow pipeline. The victim ML models exploited were: DT, RF, GB and XGB (in their default configurations). The experimental results showed that the attack's impact on the ML models' weighted accuracy was about 70%. Also in this work, the adversarial training proved to be a valuable countermeasure to the problem of adversarial examples. This work is essential for this paper since it's the experimental backbone.

2.2. Responsible AI & Black-Box Evasive AML Attacks in CAN Bus Frame Detection

This paper seeks to delve into the Robustness/Security of ML-based systems (in the automotive sector). Adversarial Training is the most exploited countermeasure to raise the security level of ML-based systems and avoid Black-Box AML attacks [34, 35, 32]. In the literature, there is an important lack of works dealing with the importance of ML-based system programming practices focused on Robustness/Security (by Design) concerning Black-Box AML attacks in the CAN bus frame detection task.

Consequently, it is necessary to consider the impact of best AI/ML-based systems programming practices (for Responsible/Trustworthy AI and especially for Robustness/Security of ML models) on Cyber Social Security (CSS) in MDOs. The CSS links the security of physical systems with the safety of the most valuable assets, i.e. physical persons.

Accordingly, this paper seeks to underscore the critical need for future research.

3. Methodology

In this section (useful for answering all RQs), details about the Black-Box attack scenario, the ZOO attack pipeline, the empirical estimation of attack time (and its possible evolution) and the qualitative analysis of the parameters' impact on the CSS framework for MDOs are discussed. This work is based on Python (version 3.9). The implementation of the ML models is carried out using the Scikit-learn library, with the exception of the XGBoost model which utilized the xgboost library. Data manipulation and processing are facilitated by the Pandas framework. The ZOO attack implementation is provided by the Adversarial Robustness Toolkit (ART) [36]. The working machine is equipped with an AMD Ryzen 5 2600 Six-Core Processor and 16 GB of RAM.

3.1. Attack Scenario

The attack scenario examined aligns with that presented in [32]. The attack begins by conducting a Vulnerability Assessment and Penetration Test (VAPT) on the target In-Vehicle Network (IVN) to compromise a single ECU. This phase facilitates the exfiltration and injection of CAN frames, enabling the attacker to gather extensive insights into the behavior of the target IDS. The attacker's ultimate goal is to infiltrate the IDS module directly [37] and, obtain the correct label for each preprocessed frame to generate the corresponding adversarial example. Additionally, the attacker may monitor the IDS-generated predictions by gaining control of any module interfacing with the IDS system. The attacker knows nothing about the victim system (i.e. the pure Black-Box scenario) [32].

3.2. Attack Pipeline

The work presented in this paper is based on the Barletta et al. [32] attack pipeline, useful for training the victim ML models (to be attacked later). Specifically, the OTIDS dataset [33] is prelaborated following the Bari et al. [34] pipeline. The final dataset version is splitted into three parts: A (i.e. the 60% part), B (i.e. the first 20% part) and C (i.e. the second 20% part).

ML models useful for empirical estimation (discussed later) are: RF bagging-based, GB and XGB (in their default configurations). The attack pipeline (for each victim ML model) follows these steps:

1. Training of the IDS on A dataset (obtaining $Model_A$);
2. Generating the adversarial examples sets B' and C' (related to the B and C) on $Model_A$;

Before performing the ZOO attack on the A subdataset, a K-Fold Stratified Cross Validation (K-FSCV) (with $k = 5$) is performed. The ZOO configuration follows the default configuration except for:

- the *learning_rate* is set to 0.1 (default is 0.01) since the attacker's probably would probably want to converge very fast (during the gradient descent step);
- the *max_iter* is set to 50 (default is 10) since the attacker would probably want to get examples very close to the normal ones, by increasing the number of trials;
- the *variable_h* is set to 0.2 (default is 0.0001) since the attacker probably wants the adversarial examples very quickly (enlarging the extremes of the search range). However, the global minimum (i.e. the minimum adversarial perturbation) of the gradient descent is not guarantee.

3.3. Empirical Estimation of the Hyperparameters' Influence

This phase is useful for answering RQ1. Ideally, every defense team (Vehicle-SOC) would want to exploit an ML-based IDS system that involves as much time as possible to generate adversarial examples. Considering this idea, some hyperparameters related to Ensemble-based ML models (i.e. RF, GB, XGB) could be labelled as an intrinsic defense tool for the IDS system useful to control the needed generation time for the adversarial examples. This approach aims to strengthen the organization's defense mechanisms by maximizing the effort required of the attacker. The analysis' modus operandi is the following: for each ML model (i.e. RF, GB and XGB) and for each hyperparameters' value (incremental), the value (seconds) related to the generation time of 92270 examples is detected after about five minutes of computation. The examined hyperparameters are the bagging trees number's per RF, the boosting rounds number's per GB and the boosting rounds number's per XGB. So, the empirical estimation is performed on $Model\_A$ by exploiting the second step of the attack pipeline (mentioned above). The empirical analysis is performed only on B' since the examples follow the same distribution. The core goal of this analysis is to assess whether a direct proportionality exists between the previously discussed parameters and the time required to generate the considered adversarial examples. Accordingly, this study aims to offer practical recommendations regarding optimal values for these hyperparameters and the most effective ensemble model.

3.4. Impact Qualitative Analysis on CSS Framework for MDOs

To answer RQ2, the qualitative analysis is carried out considering the "Land" and "Cyber" domains. In general, assessing the impact of cyber threats on assets (across all domains within MDOs scenarios) requires an estimation of their inherent risk level [38, 39]. Even in this situation where the goal is to estimate the (negative) impact resulting from programming practices that do not also consider the resilience of ML models to Black-Box attacks, this is critical.

Considering the lack (previously mentioned), an high-level qualitative risk assessment is adopted taking into account not only the severity of social consequences resulting from such attacks (i.e. the instigation of terrorism's climate linked to the abnormal vehicle behavior as well as the disorientation of civilian/military operators And the reputational damage in the "Country System" [32]) but also the "risk-averse" approach that informs our perspective.

4. Result & Discussion

4.1. Empirical Evalution of Attack Time

Figure 2 provides an estimation of the time required to generate 92270 adversarial examples (corresponding to the B' part) as a function of the hyperparameter representing the number of bagging trees in the RF model. The results clearly illustrate a direct proportional relationship between these two variables (also through the regression line). For example, for 91 estimators about 33 hours are expected. This evidence shows that this parameter contributes not only to the accuracy of the model but also to the robustness of this attack. So, the answer to RQ1 (related to RF) is affirmative.

Figure 3 shows the same estimation of the RF case (but for the GB model). The independent variable is the number of boosting rounds. The results unequivocally demonstrate a directly proportional relationship between these two variables, as further corroborated by the regression line. For example, for 90 estimators about 2 hours are expected. The same consideration can be highlithed for the GB model. So, the answer to RQ1 (related to GB) is affirmative.

Figure 4 shows the same estimation of the RF and GB case (but for the XGB model). In this situation, the independent variable is the number of boosting rounds. The results don't demonstrate a real directly proportional relationship between these two variables, as further corroborated by the regression line. For example, for 90 estimators about 9 hours are expected. So, the answer to RQ1 (related to XGB) is negative.

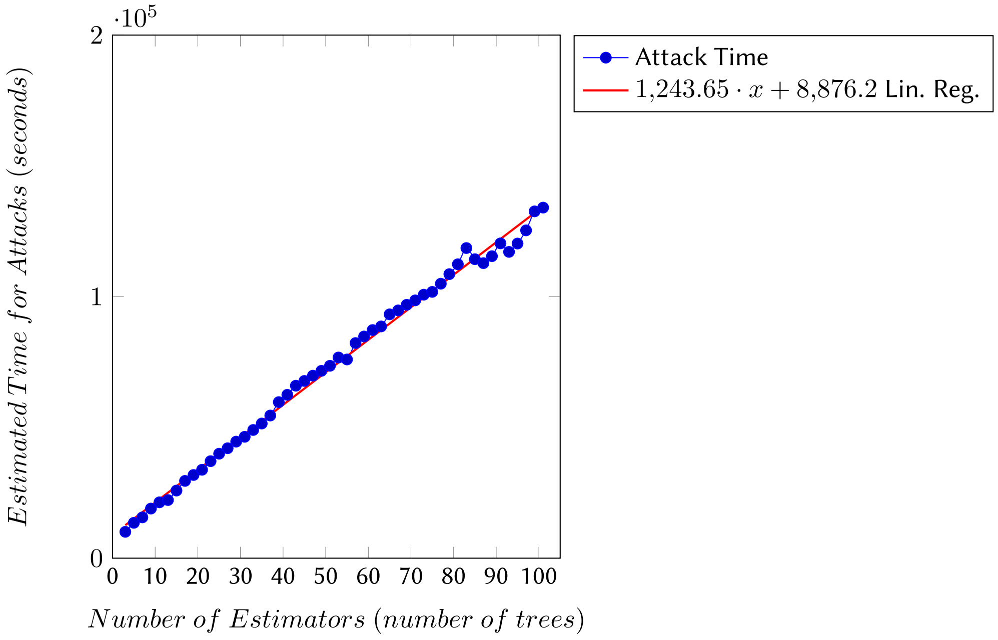

Figure 2: Time needed to generate B' as the value of the bagging trees number's for the RF model is increased.

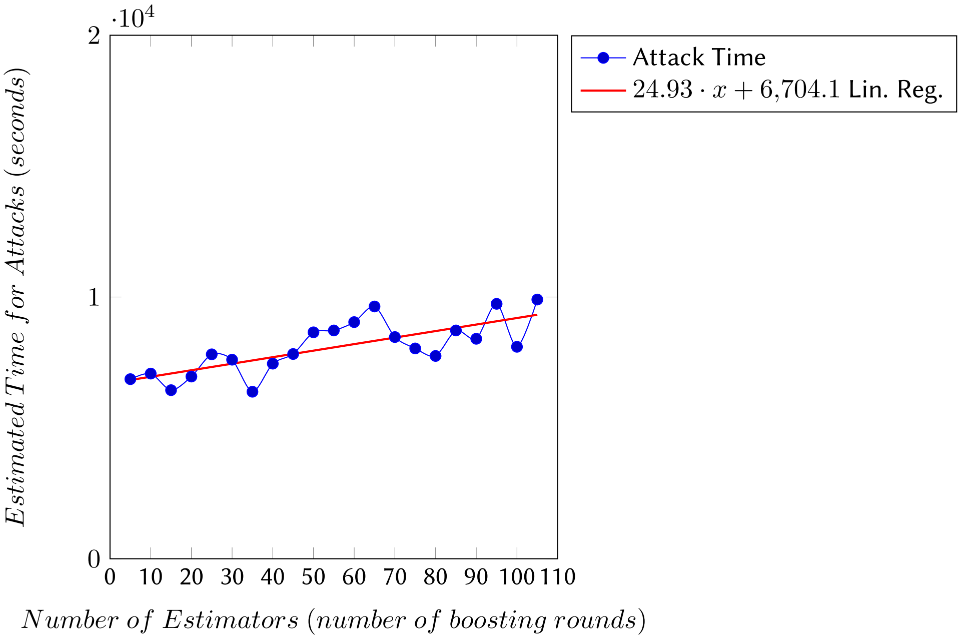

Figure 3: Time needed to generate B' as the value of the boosting rounds number's for the GB model is increased.

4.2. Impact on CSS Framework for MDOs

Table 1 shows the impact of the previous analysis on the CSS framework for MDOs (i.e. the answer to RQ2). A "High" impact is observed on the "Prevention" axis (and not "Very High") since deterrence does not categorically prevent the execution of the attack.

5. Conclusion & Future Work

At a time in history when CAVs are becoming increasingly central assets in Smart Cities and MDOs are a concrete threat, it's critical to hardening vehicle defenses to prevent damage to passengers' physical

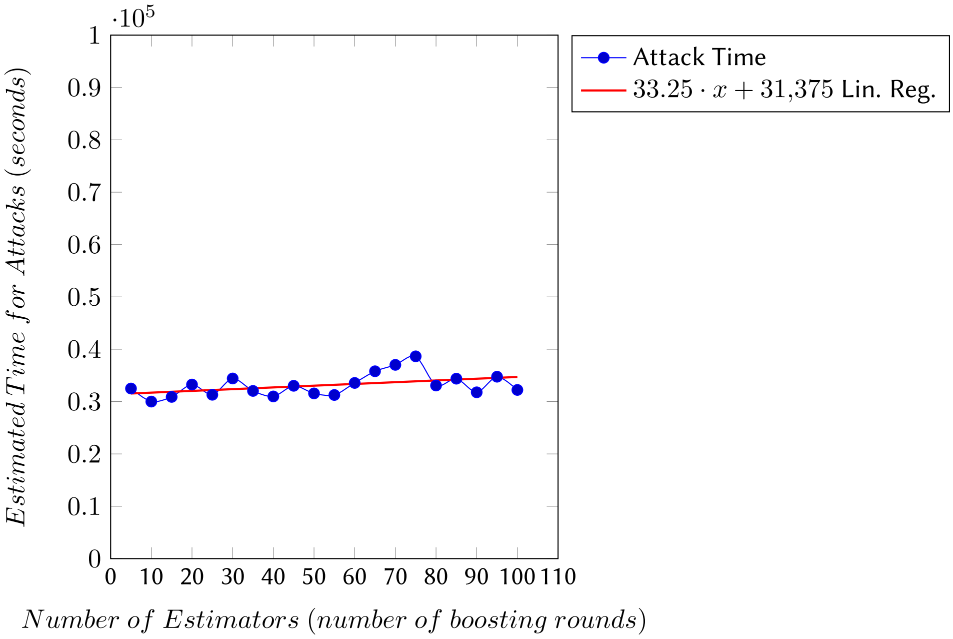

Figure 4: Time needed to generate B' as the value of the boosting rounds number's for the XGB model is increased.

Table 1
Inappropriate values' impact assigned to the hyperameters under consideration for RF and GB on the axes of the CSS framework for MDOs.

Axis of CSS Framework	Impact	Motivation
Detection	VH	- Gain of time to detect/
Response	VH	respond to VAPT attempt
Prevention	H	- Deterrence

and cognitive dimensions. AI/ML represent a tool that can fulfil this task but there is a need to develop IDSs for CAN bus IVNs that are robust (as much as possible) to Black-Box AML attacks as well.

Therefore, in this paper the possible influence of some hyperparameters related to DT-based TT state-of-the-art Ensemble models (i.e. RF, GB, XGB) underlying an IDS, victim of the ZOO attack (in a purely Black-Box scenario), on the time needed to the generation of adversarial examples is evaluated (RQ1). In addition, the impact of such on the CSS framework for MDOs is qualitatively evaluated (RQ2).

The experimental results reveal a direct proportional relationship between the bagging trees' number for the RF and the estimated time required to execute the attack. This trend is also confirmed for the boosting rounds' number for the GB model but does not hold for the corresponding parameter for the XGB model. Consequently, only in the first two cases can these hyperparameters be considered intrinsic defense mechanisms (or deterrents) against the attack under investigation. The RF model is recommended given its greater robustness. Generally, the impact of such evidence is rated very high considering the important possibility of controlling the timing of the attack (to the point of diverting attention away from the attacker).

Some future directions of this work, it would be interesting to perform the empirical analysis by considering different values related to the attack hyperparameters (even the default ones), to base the analysis on additional Black-Box (and White-Box) attacks as well as additional state-of-the-art datasets in the automotive context. In addition, it is also consedered to extend the analysis on datasets and IDSs (based on ML and Deep Learning algorithms) exploited in additional systems of national interest (i.e. IoT networks, Aircraft, Submarines). Regarding impact analysis, a clear development is to quantify the analysis itself.

6. Acknowledgments

This work was partially supported by the following projects: SERICS - “Security and Rights In the CyberSpace - SERICS” (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU; Casa delle Tecnologie Emergenti del Comune di Bari – “Bari Open Innovation Hub” – CUP J99J19000300003.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] M. A. Richter, M. Hagenmaier, O. Bandte, V. Parida, J. Wincent, Smart cities, urban mobility and autonomous vehicles: How different cities needs different sustainable investment strategies, Technological Forecasting and Social Change 184 (2022) 121857. URL: https://www.sciencedirect.com/science/article/pii/S004016252200381X. doi:https://doi.org/10.1016/j.techfore.2022.121857.

[2] T. Campisi, A. Severino, M. A. Al-Rashid, G. Pau, The development of the smart cities in the connected and autonomous vehicles (cavs) era: From mobility patterns to scaling in cities, Infrastructures 6 (2021). URL: https://www.mdpi.com/2412-3811/6/7/100. doi:10.3390/infrastructures6070100.

[3] H. Olufowobi, G. Bloom, Chapter 16 - connected cars: Automotive cybersecurity and privacy for smart cities, in: D. B. Rawat, K. Z. Ghafoor (Eds.), Smart Cities Cybersecurity and Privacy, Elsevier, 2019, pp. 227–240. doi:https://doi.org/10.1016/B978-0-12-815032-0.00016-0.

[4] V. S. Barletta, D. Caivano, C. Catalano, M. De Vincentiis, Quantum-based automotive threat intelligence and countermeasures, in: Proceedings of the 28th International Conference on Evaluation and Assessment in Software Engineering, EASE ’24, ACM, New York, NY, USA, 2024, p. 548–554. doi:10.1145/3661167.3661278.

[5] E. Commission, Ai act, https://digital-strategy.ec.europa.eu/it/policies/regulatory-framework-ai, 2024.

[6] N. S. W. D. Command, Multi-domain operations in nato – explained, https://www.act.nato.int/article/mdo-in-nato-explained/, 2023.

[7] S. G. della Difesa Italiana, The italian defence approach to multi-domain operations (approccio della difesa alle operazioni multidominio), https://www.difesa.it/assets/allegati/31787/2.1defence_approach_to_mdos.pdf, 2022.

[8] F. Tommasi, C. Catalano, M. Fornaro, I. Taurino, Mobile session fixation attack in micropayment systems, IEEE Access 7 (2019) 41576–41583. doi:10.1109/ACCESS.2019.2905219.

[9] E. U. A. for Cybersecurity, G. Dede, R. Naydenov, A. Malatras, C. E. C. C. de Investigación, R. Hamon, H. Junklewitz, I. Sanchez, E. C. J. R. Centre, Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving, EUR (Luxembourg. Online), Publications Office of the European Union, 2021. URL: https://books.google.it/books?id=9oZbzgEACAAJ.

[10] K. N, V. Ravi, V. Sowmya, Unsupervised intrusion detection system for in-vehicle communication networks, Journal of Safety Science and Resilience 5 (2024) 119–129. URL: https://www.sciencedirect.com/science/article/pii/S2666449624000070. doi:https://doi.org/10.1016/j.jnlssr.2023.12.004.

[11] A. Alfardus, D. B. Rawat, Intrusion detection system for can bus in-vehicle network based on machine learning algorithms, in: 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mo-

bile Communication Conference (UEMCON), 2021, pp. 0944–0949. doi:10.1109/UEMCON53757.2021.9666745.

[12] K. A, H. R, F. L. D, J. H, S. M. JI, G. G. E, Toward explainable, robust and fair ai in automated and autonomous vehicles (2023). doi:10.2760/95650(online).

[13] D. Morris, G. Madzudzo, A. Garcia-Perez, Cybersecurity threats in the auto industry: Tensions in the knowledge environment, Technological Forecasting and Social Change 157 (2020) 120102. URL: https://www.sciencedirect.com/science/article/pii/S0040162520309288. doi:https://doi.org/10.1016/j.techfore.2020.120102.

[14] F. Sommer, J. Dürrwang, R. Kriesten, Survey and classification of automotive security attacks, Information 10 (2019). URL: https://www.mdpi.com/2078-2489/10/4/148. doi:10.3390/info10040148.

[15] V. S. Barletta, D. Caivano, C. Catalano, M. de Gemmis, D. Impedovo, Cyber social security education, in: Extended Reality: International Conference, XR Salento 2024, Lecce, Italy, September 4–7, 2024, Proceedings, Part IV, Springer-Verlag, Berlin, Heidelberg, 2024, p. 240–248. URL: https://doi.org/10.1007/978-3-031-71713-0_16. doi:10.1007/978-3-031-71713-0_16.

[16] B. Wu, Z. Zhu, L. Liu, Q. Liu, Z. He, S. Lyu, Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective, 2024. arXiv:2302.09457.

[17] C. Xie, Z. Cao, Y. Long, D. Yang, D. Zhao, B. Li, Privacy of autonomous vehicles: Risks, protection methods, and future directions, 2022. arXiv:2209.04022.

[18] H. Cao, W. Zou, Y. Wang, T. Song, M. Liu, Emerging threats in deep learning-based autonomous driving: A comprehensive survey, 2022. arXiv:2210.11237.

[19] and European Union Agency for Cybersecurity, A. Malatras, I. Agrafiotis, M. Adamczyk, Securing machine learning algorithms, 2021. URL: https://op.europa.eu/publication-detail/-/publication/c7c844fd-7f1e-11ec-8c40-01aa75ed71a1. doi:doi/10.2824/874249.

[20] F. Aloraini, A. Javed, O. Rana, Adversarial attacks on intrusion detection systems in in-vehicle networks of connected and autonomous vehicles, Sensors 24 (2024). URL: https://www.mdpi.com/1424-8220/24/12/3848. doi:10.3390/s24123848.

[21] S. Longari, F. Noseda, M. Carminati, S. Zanero, Evaluating the robustness of automotive intrusion detection systems against evasion attacks, in: Cyber Security, Cryptology, and Machine Learning: 7th International Symposium, CSCML 2023, Be'er Sheva, Israel, June 29–30, 2023, Proceedings, Springer-Verlag, 2023, p. 337–352. URL: https://doi.org/10.1007/978-3-031-34671-2_24. doi:10.1007/978-3-031-34671-2_24.

[22] S. Kotyan, A reading survey on adversarial machine learning: Adversarial attacks and their understanding, 2023. arXiv:2308.03363.

[23] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, A. Jacquet, Responsible ai pattern catalogue: A collection of best practices for ai governance and engineering, ACM Comput. Surv. 56 (2024). URL: https://doi.org/10.1145/3626234. doi:10.1145/3626234.

[24] E. U. A. for Cybersecurity (ENISA), Artificial intelligence and cybersecurity research, 2023. URL: https://www.enisa.europa.eu/publications/artificial-intelligence-and-cybersecurity-research. doi:10.2824/808362.

[25] H.-L. E. G. on AI European Commission, Ethics guidelines for trustworthy ai, 2024. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[26] N. I. of Standards, Technolgy, Ai fundamental research - security, 2023. URL: https://www.nist.gov/artificial-intelligence/ai-fundamental-research-security.

[27] S. Goellner, M. Tropmann-Frick, B. Brumen, Responsible artificial intelligence: A structured literature review, 2024. URL: https://arxiv.org/abs/2403.06910. arXiv:2403.06910.

[28] V. S. Barletta, M. Calvano, A. Sciacovelli, Cyber social security in multi-domain operations, in: 2024 IEEE International Workshop on Technologies for Defense and Security (TechDefense), 2024, pp. 41–46. doi:10.1109/TechDefense63521.2024.10863352.

[29] H. M. Song, J. Woo, H. K. Kim, In-vehicle network intrusion detection using deep convolutional neural network, Vehicular Communications 21 (2020) 100198.

[30] I. Zenden, H. Wang, A. Iacovazzi, A. Vahidi, R. Blom, S. Raza, On the resilience of machine learning-based ids for automotive networks, in: 2023 IEEE Vehicular Networking Conference

(VNC), IEEE, 2023. doi:10.1109/vnc57357.2023.10136285.

[31] M. L. Han, B. I. Kwak, H. K. Kim, Anomaly intrusion detection method for vehicular networks based on survival analysis, Vehicular Communications 14 (2018) 52–63. URL: https://www.sciencedirect.com/science/article/pii/S2214209618301189. doi:https://doi.org/10.1016/j.vehcom.2018.09.004.

[32] V. S. Barletta, D. Caivano, C. Catalano, S. D. Vescovo, Black-box adversarial ml attacks on ids and multi-domain impact analysis for threat intelligence in automotive scenarios, in: 2024 IEEE International Workshop on Technologies for Defense and Security (TechDefense), 2024, pp. 132–137. doi:10.1109/TechDefense63521.2024.10863442.

[33] H. Lee, S. H. Jeong, H. K. Kim, Otids: A novel intrusion detection system for in-vehicle network by using remote frame, in: 2017 15th Annual Conference on Privacy, Security and Trust (PST), 2017, pp. 57–5709. doi:10.1109/PST.2017.00017.

[34] B. S. Bari, K. Yelamarthi, S. Ghafoor, Intrusion detection in vehicle controller area network (can) bus using machine learning: A comparative performance study, Sensors 23 (2023). doi:10.3390/s23073610.

[35] B. Badjie, J. Cecílio, A. Casimiro, Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review, ACM Computing Surveys 57 (2024) 1–52.

[36] M. Nicolae, M. Sinn, T. N. Minh, A. Rawat, M. Wistuba, V. Zantedeschi, I. M. Molloy, B. Edwards, Adversarial robustness toolbox v0.2.2, CoRR abs/1807.01069 (2018). arXiv:1807.01069.

[37] V. S. Barletta, D. Caivano, C. Catalano, M. De Vincentiis, A. Pal, Machine learning for automotive security in technology transfer, in: A. Rocha, H. Adeli, G. Dzemyda, F. Moreira, V. Colla (Eds.), Information Systems and Technologies, Springer Nature Switzerland, Cham, 2024, pp. 341–350.

[38] M. Angelelli, S. Arima, C. Catalano, E. Ciavolino, A robust statistical framework for cyber-vulnerability prioritisation under partial information in threat intelligence, Expert Systems with Applications 255 (2024) 124572. URL: https://www.sciencedirect.com/science/article/pii/S0957417424014398. doi:https://doi.org/10.1016/j.eswa.2024.124572.

[39] M. T. Baldassarre, V. S. Barletta, D. Caivano, D. Raguseo, M. Scalera, Teaching cyber security: The hack-space integrated model, in: Italian Conference on Cybersecurity, volume 2315, 2019. URL: https://ceur-ws.org/Vol-2315/paper06.pdf.