Leveraging Trustworthy AI for Automotive Security in Multi-Domain Operations: Towards a Responsive Human-AI Multi-Domain Task Force for Cyber Social Security

Vita Santa Barletta[1,†], Danilo Caivano[1,†], Gabriel Cellammare[1,†], Samuele del Vescovo[2,*,†], Massimiliano Morga[3,†] and Annita Larissa Sciacovelli[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, Italy*
[3]*SER&P, Spin-off of University of Bari Aldo Moro*

Abstract

Multi-Domain Operations (MDOs) emphasize cross-domain defense against complex and synergistic threats, with civilian infrastructures like smart cities and Connected Autonomous Vehicles (CAVs) emerging as primary targets. As dual-use assets, CAVs are vulnerable to Multi-Surface Threats (MSTs), particularly from Adversarial Machine Learning (AML) which can simultaneously compromise multiple in-vehicle ML systems (e.g., Intrusion Detection Systems, Traffic Sign Recognition Systems). Therefore, this study investigates how key hyperparameters in Decision Tree-based ensemble models—Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB)—affect the time required for a Black-Box AML attack i.e. Zeroth Order Optimization (ZOO). Findings show that parameters like the number of trees or boosting rounds significantly influence attack execution time, with RF and GB being more sensitive than XGB. Adversarial Training (AT) time is also analyzed to assess the attacker's window of opportunity. By optimizing hyperparameters, this research supports Defensive Trustworthy AI (D-TAI) practices within MST scenarios and contributes to the development of resilient ML systems for civilian and military domains, aligned with Cyber Social Security framework in MDOs and Human-AI Multi-Domain Task Forces.

Keywords

Multi-Domain Operations, Trustworthy AI, Cyber Social Threat Intelligence, Automotive Security, Human-AI Responsive Collaboration

1. Introduction

Multi-Domain Operations (MDOs) constitute the central paradigm of modern military strategy, emphasizing the integrated coordination of capabilities across *Space*, *Air*, *Land*, *Sea*, and *Cyber* domains to produce synergistic malicious effects [1]. These operations generate complex, multi-surfaced challenges aimed at complicate the adversary's Military Decision-Making Process (MDMP) [2, 3]. Within this framework, Multi-Domain Task Forces (MDTFs) play a key role by aligning defence resources across both physical and informational environments [4]. A core challenge in MDOs is the lack of defensive synchronization between kinetic and cyber capabilities across domains [5]. To address this, the Cyber Social Security (CSS) framework has been introduced to integrate cyber defence within traditional domain-based strategies, forming the CSS-MDO framework [6, 7].

COL-SAI 2025: Workshop on COllaboration and Learning through Symbiotic Artificial Intelligence, in conjunction with the 16th Biannual Conference of the Italian SIGCHI Chapter (CHItaly 2025), October 6-10 2025, Salerno, Italy (2025)

*Corresponding author.
†These authors contributed equally.
vita.barletta@uniba.it (V. S. Barletta); danilo.caivano@uniba.it (D. Caivano); g.cellammare1@studenti.uniba.it (G. Cellammare); samuele.delvescovo@imtlucca.it (S. d. Vescovo); m.morga@serandp.com (M. Morga); annitalarissa.sciacovelli@uniba.it (A. L. Sciacovelli)
0000-0002-0163-6786 (V. S. Barletta); 0000-0001-5719-7447 (D. Caivano); 0009-0001-8220-5135 (S. d. Vescovo); 0000-0003-4795-9238 (A. L. Sciacovelli)

The CSS-MDO framework adopts a multidimensional approach, where the five active warfare domains are represented along the horizontal axis, each equipped with specialized tools, methods, and procedures [7]. These are aligned with the vertical axis representing the Detection-Response-Prevention processes. In this framework, unlike other military domains, the "cyber" domain is not only a "link" between the various "traditional" domains but it has its own offensive/defensive identity [3]. This bi-dimensional structure identify operational tiers (i.e. horizontally and vertically) in which MDTFs may operate [8].

Smart cities constitute vulnerable assets within future MDOs, especially involving the *Cyber* and *Land* domains. Among the critical components of a smart city's Internet of Things (IoT) infrastructure, Connected and Autonomous Vehicles (CAVs) represent a prominent point of vulnerability [9]. As foundational enablers of shared and electric mobility, CAVs are central to optimizing urban transportation and improving sustainable travel models [10, 11]. However, in this rapidly evolving and innovation context, this high-value assets' attack surfaces are expanding and complicating in ways challenging the capacity of any "cyber-social" blue team to effectively detect and mitigate threats. Specifically, attackers (more or less skilled) may exploit known vulnerabilities in In-Vehicle Networks (IVNs) based on the Controller Area Network (CAN) protocol [12]. These attacks may be part of broader, coordinated military/civilian activities conducted across multiple domains to create complex and orchestrated dilemmas for defenders [13].

Within the context of MDOs and the realted Multi-Domain Threats (MDTs), one of the attackers goals may be to compromise Electronic Control Units (ECUs), which function as critical communicational nodes within IVNs [14]. This threat can also affect the autonomous vehicle's driving system by exploiting various IVNs levels. Such intrusions can lead to anomalous or unsafe behavior in the affected vehicle, thereby posing direct risks to its functionality and safety [15]. Furthermore, an attacker can compromise physically road signs to compromise autonomous driving's features of the vehicle. In response to these threats, the integration of Artificial Intelligence (AI) and Machine Learning (ML) technologies has emerged as a promising avenue for reinforcing vehicular cybersecurity. Among these approaches, there are ML-based Intrusion Detection Systems (IDSs) as well as Traffic Sign Recognition Systems (TSRS), which are specifically designed to detect abnormal communication patterns and unauthorized access attempts within IVNs [16].

So, an adversary can manipulate input data, such as images or CAN bus frames, at the testing or deployment phase [17]. These perturbations are imperceptible to human observers yet are sufficient to deceive the targeted ML model into producing incorrect classifications [18, 19]. AML attacks are generally categorized into three scenarios. These are discriminated by the level of knowledge the attacker possesses about the internal architecture and parameters of the victim model. Among these, the Black-Box setting is considered the most plausible and accessible from an adversarial perspective, as it assumes no prior access to or knowledge of the model's internal workings [20, 21]. Moreover, aligning with the conceptual framework of MDTs, in which adversaries operate across multiple domains simultaneously, it is plausible to hypothesize scenarios in which attacks are exploited on different surfaces of a single asset. This concept can be referred to as Multi-Surface Threat (MST). Current literature on the application of Black-Box attacks within the CAN bus frame detection task remains limited and in an early stage of development (even in MDTs and MSTs scenarios).

Therefore, the primary goal of this paper is to investigate about the role of specific hyperparameters associated with Decision Tree (DT)-based ensemble Technology Transfer (TT) models. These are Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGB). These are the core of the supervised ML-based victim IDS for the CAN bus frame detection task. It is assumed that the IDS (installed onboard the vehicle) is subjected to a Black-Box AML attack i.e. the Zeroth Order Optimization (ZOO) in a pure evasive Black-Box setting. This type of attack is conceptualized as a part of a MST, so a Single-Surface Threat (SST), falling into the *Cyber* component of a complex MDT. This threat is framed into a MDO. So, this work tries to address how variations in selected hyperparameters affect the time required to generate adversarial examples for each targeted ML model. Results indicate that the number of bagging trees in RF and the number of boosting rounds in GB have a significant impact on the attack time. Thus, the same does not hold for the boosting rounds in XGB. These hyperparameters, in the cases of RF and GB, can be interpreted as intrinsic (or deterrence) defense against the ZOO attack.

Appropriately values for these hyperparameters may lead to a trustworthy AI-by-design approach for In-Vehicles (I-V) ML systems' robustness [22], [17]. All of that can contribute to the concept of Defensive Trustworthy AI (D-TAI) i.e. AI for defence purposes. In particular, the work underscores the relevance of robustness [23] and security [24] properties in the design and deployment of ML models within adversarial environments [25]. Additionally, Adversarial Training (AT) time is analyzed to better understand the attacker's "window of opportunity", proving the combo between hyperparameters' correct tuning and AT can slow the attacker's steps, fostering the activities of the blue team. Finally, this work aims to contribute to the proper education regarding responsible development of ML systems in both civilian and military (public/private) contexts for industries and academies, promoting their integration into the CSS-MDO [7] framework as part of a defence strategy coordinated by Human-AI Multi-Domain Task Forces (H-AI MDTF). The proof of concept is the qualitative identification of the positive impact of this trustworthy AI-by-design practice on the vertical axes of the CSS-MDO framework.

In summary, the research questions (RQs) are:

- RQ1: Can the hyperparameters related to the number of bagging trees in RF, the number of boosting rounds in GB, and the number of boosting rounds in XGB affect the time needed to generate ZOO adversarial examples, when applied to supervised ML-based IDS for the CAN bus frame detection task (in a Black-Box attack scenario)?
- RQ2: Can the combination of the timing for the AT and the correct setting of the previous parameters slow down the attacker's actions by reducing the attacker's "window of opportunity"?
- RQ3: Is it possible to qualitatively quantify the (positive) impact of these values on the Detection-Response-Prevention axes of the CSS-MDO framework (for H-AI MDTFs educational purpose)?

So, the main contributions are:

- Empirically evaluate the influence of hyperparameters as the number of bagging trees in RF, the number of boosting rounds in GB, and the number of boosting rounds in XGB on the time needed to generate adversarial examples exploiting the ZOO attack, when applied to supervised ML-based IDS in the CAN bus frame detection task (considering a Black-Box threat model).
- Understand whether the combination of the correct setting of the previously mentioned parameters and the time needed to execute the AT can slow down or reduce the attacker's "window of opportunity".
- Qualitatively quantify the (positive) impact of these values on the "Detection", "Response" and "Prevention" axes of the CSS-MDO framework (for H-AI MDTFs educational purpose).

The paper is organized as follows: section 2 describes the related works; section 3 describes the experimentation setup; section 4 shows the results; section 5 conclude the work and explain the future developments.

2. Related Work & Research Gap

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

To the best of our knowledge, the scientific literature addressing Black-Box AML attacks against ML-based IDS in the context of CAN bus frame detection remains relatively underdeveloped actually.

Zenden et al. [26] examined the Fast Gradient Sign Method (FGSM) attack's impact on various ML and DL models' performance within a surrogate Black-Box attack scenario. Their study further demonstrated that adversarial training serves as an effective mitigation strategy, yielding notably positive outcomes. The evaluated models included BL-DNN, BL-Ensemble, SOTA-CNN, and SOTA-LSTM architectures [26]. The experiments were conducted using a subset of the Survival dataset [27]. The results indicated that ML models were particularly susceptible to the FGSM attack, exhibiting an accuracy degradation of approximately 40%.

Longari et al. [19] proposed a novel methodology for executing Black-Box evasion attacks (in a pure scenario) against ML-based IDS within the context of online CAN bus frame detection. Their method targets the entire transmission flow by analyzing segments of CAN payloads. The attack strategy employs a sliding window technique over the payload data, rather than processing entire examples in isolation [19]. The experimentation utilizes the "ReCAN" dataset, specifically the "ID C-1" subset collected from a real Alfa Romeo Giulia Veloce vehicle [19]. A range of ML algorithms were evaluated as the IDS core, including Small-LSTM, Small-GRU, Large-GRU, CANnolo, Neural Network, Vector Auto-Regressive (VAR), and Hamming distance-based models [19].

Instead, Aloraini et al. [18] 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 [18]. This scenario constitutes a non-pure Black-Box due to the transferability of the adversarial examples exploited [18]. The victim IDS models were: a baseline proprietary IDS based on Deep Neural Network (DNN) 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 [28]. Several White-Box AML attacks were considered like FGSM, Basic Iterative Method (BIM), Projected Gradient Descent (PGD) and Jacobian-based Saliency Map Attack (JSMA) [18]. 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 [18].

These works do not consider attacks conducted directly against the target IDS in a pure Black-Box scenario in a TT context. Moreover, they do not explore the application of state-of-the-art Black-Box AML techniques that are not explicitly tailored to this specific task. In contrast, the study by Barletta et al. [29] investigates the application of the ZOO attack within a pure Black-Box setting for the same task, focusing on supervised ML algorithms. Originally developed for image recognition tasks, the ZOO attack was exploited using the OTIDS dataset [30]. The victim models included DT, RF, GB, and XGB (contestualized in TT scenarios). Experimental findings revealed a reduction in weighted accuracy of approximately 70%. Furthermore, adversarial training was once again validated as an effective countermeasure against such attacks.

2.2. Evasion Black-Box AML Attacks & Defensive Trustworthy AI in CAN Bus Frame Detection for MDOs

This paper aims to explore the concept of trustworthy AI for defensive purposes (specifically the robustness dimension) as a proactive defense mechanism (i.e. D-TAI) for ML-based systems in the automotive domain. Among existing countermeasures, AT is the most widely adopted technique to enhance the security posture of ML-based systems and mitigate the risks posed by Black-Box AML attacks [31, 32, 29]. However, the current literature reveals a significant gap concerning the role of robustness-by-design oriented programming practices in the development of ML-based systems for CAN bus frame detection under Black-Box AML attack scenarios.

2.3. Defensive Trustworthy AI Impact Evaluation on CSS-MDO Framework for H-AI MDTFs

It is necessary to consider the impact of best ML-based systems programming practices (for D-TAI and especially for robustness of ML models) on Single Surface threats (considered as a part os MSTs targetting various I-V Systems) mapped within the CSS-MDO framework's policies. In other words, it is necessary to consider the positive impact of properly educating MDTFs about these practices emphasizing the virtuous collaboration between human and (well-setted) artificial agents in H-AI MDTFs. All of that can be a piece of the MDTs big puzzle in MDOs scenario. Accordingly, this paper seeks to underscore this critical need for future research.

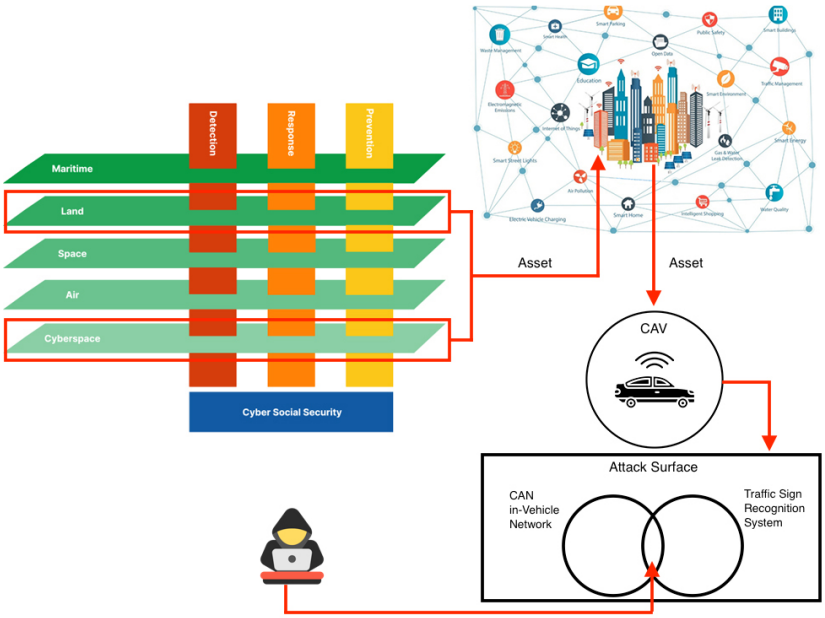

Figure 1: MST considered

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 and AT time and the qualitative analysis of the parameters' impact on the CSS-MDO framework's axes are discussed. All this is about the CAN-based SST. This work is based on Python 3.9. The implementation of the ML models is provided by the Scikit-learn library. The XGBoost model implementation is based on the xgboost library. The Pandas framework is used to handle the dataset. The ZOO attack implementation is provided by the Adversarial Robustness Toolkit (ART) [33]. The working machine is supported by an AMD Ryzen 5 2600 Six-Core Processor and 16 GB of RAM.

3.1. CAN-based SST Attack Scenario

This phase addresses the RQ1. Considering the possibility of MSTs (in the examined case impacting several I-V systems simultaneously), an example of that could be a simultaneous threat on the CAN-based IVN protected by an IDS for the CAN bus frame detection task and on the network enabling the Traffic Sign Recognition System (TSRS). Figure 1 describe the attack scenario.

The CAN-based SST attack begins with a Vulnerability Assessment and Penetration Testing (VAPT) phase targeting the CAN based IVN, aiming to compromise a single ECU. This initial step facilitates both the exfiltration and injection of CAN frames, enabling the attacker to infer behavioral patterns of the target IDS. The ultimate goal is to penetrate the IDS module itself [34], thereby obtaining the true label assigned to each preprocessed frame for the generation of corresponding adversarial examples. Additionally, the attacker may observe the IDS's output by compromising a module that interfaces with the IDS. Importantly, the attacker operates under a pure Black-Box scenario, with no prior knowledge of the victim system's internal architecture or parameters [29].

3.2. Attack Pipeline for the CAN-based SST

This work is based on the Barletta et al. [29] attack pipeline, useful for training the victim ML models. The OTIDS dataset [30] is prelaborated following the Bari et al. [31] pipeline. The final dataset version

is splitted into three parts: A (i.e. the 60% part), B and C (i.e. the other 20% parts). ML models useful for empirical estimation are: RF bagging-based, GB and XGB (in their default configurations). The attack pipeline extracted is composed by these steps (for each victim ML model):

1. IDS training on the A dataset (obtaining $Model_A$);
2. Adversarial examples sets' generation i.e. B' and C' (starting from the B and C sets) on $Model_A$;
3. Training on $A + B + B'$ dataset, obtaining $Model_{A+B+B'}$ (Adversarial Training);

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

- the *learning_rate* setted to 0.1 (default is 0.01). The attacker probably want to converge very fast (during the gradient descent);
- the *max_iter* setted to 50 (default is 10). The attacker probably want to get examples very close to the normal ones (by increasing the number of trials);
- the *variable_h* setted to 0.2 (default is 0.0001). The attacker probably wants the adversarial examples quickly (enlarging the extremes of the search range);

This research work is not interested in quantifying the impact of the attack on the victim ML performances' since Barletta et al. [29] have already explored that. For this reason, phase three of the attack pipeline directly considers the execution of the adversarial training.

3.3. Empirical Estimation of the Hyperparameters' Influence & AT time on the CAN-based SST

This subsection is useful for answering RQ1 and RQ2. Ideally, a Vehicle-Security Operations Center (Vehicle-SOC) involved in a MDTF would prefer to exploit a ML-based IDS that maximizes the time required for generating adversarial examples while minimizes the coutermeasure's time, thereby increasing the attack's operational cost. Considering that, certain hyperparameters of ensemble-based ML models specifically RF, GB, and XGB can be conceptualized as intrinsic defensive mechanisms, potentially influencing the computational effort needed to generate input adversarial examples. The approach adopted in this study involves measuring the time (seconds) required to generate 92270 adversarial examples for each model (i.e. RF, GB, XGB), as a function of incrementally varied hyperparameter values. Time is detected after about five minutes of computation. These include the number of bagging trees in RF and the number of boosting rounds in both GB and XGB. The empirical assessment is conducted on $Model_A$, focusing on the second phase of the previously described attack pipeline, and it is limited to the B' dataset, assuming its consistent with C'. The goal is to determine whether a direct proportional relationship exists between these hyperparameters and the adversarial example generation time. This analysis seeks to provide actionable insights into optimal hyperparameter configurations and the selection of the most defensive ensemble model. Finally, only for models demonstrating a "defensive" tendency of the mentioned hyperparameters, the time required to perform an AT (considering the training set $A + B + B'$ i.e. 461350 examples) is evaluated as the values of the hyperparameters vary. This is useful to better understand the extent of the attacker's "window of opportunity." It is assumed that once introduced into the CAN network, the attacker proceeds with the compromise of the IDS system.

3.4. Hyperparameters' Impact Qualitative Analysis on CSS-MDO Framework

This subsection is useful for answering RQ3. Considering the critical nature of the scenario that surrounds this research work, an impact (positive) analysis realted to the MDTFs education about the best programming practices (improving the resilience of ML models to Black-Box attacks) is an important milestone to underline the right importance of these. All this is intended to emphasize the fruitful collaboration between human and artificial agents in future Human-AI MDTFs. The actual qualitative analysis is based on the *Land* and *Cyber* domains. This analysis comes from an high-level

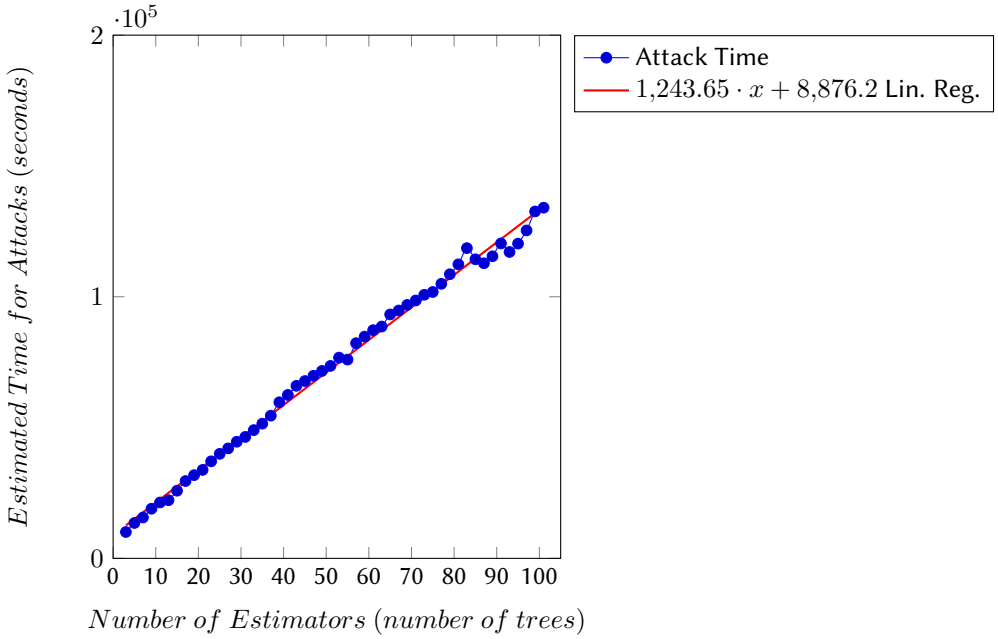

Figure 2: Time required to generate B' as a function of the increasing number of bagging trees in the RF model.

qualitative risk assessment related to this SST. This second analysis is adopted considering multiple hypothetical negative consequences: the potential to incite a climate of terror through anomalous vehicle behavior and the cognitive disruption of civilian and military operators, the reputational damage to the national infrastructure and institutions [29], and the inherently risk-averse perspective guiding the human evaluator point-of-view [35].

4. Result & Discussion

4.1. Empirical Evalution of Attack & AT Time

Figure 2 presents an estimation of the needed time to generate 92270 adversarial examples (corresponding to subset B') as a function of the bagging trees number (i.e. the examinated hyperparameter) in the RF model. The results demonstrate a linear relationship between the number of estimators and the generation time, as evidenced by the regression line. For instance, with 81 estimators, the expected generation time approaches approximately 31 hours. This finding indicates that the selected hyperparameter influences not only the model's predictive performance but also its robustness against ZOO attack. Therefore, the response to RQ1, concerning the RF model, is affirmative.

Figure 3 presents the same type of the previous estimation but applied to the GB model. In this case, the examined independent variable is the number of boosting rounds. The results clearly reveal a directly proportional relationship between the number of boosting rounds and the time required to generate the adversarial examples, as indicated by the regression line. For instance, approximately 2 hours are needed for 80 estimators. These findings support the same conclusions drawn for the RF model, thereby affirmatively answering RQ1 in the context of GB.

Figure 4 depicts the estimation for the XGB model, again using the number of boosting rounds as the independent variable. Unlike the RF and GB cases, the results do not exhibit a consistent linear relationship between the number of estimators and the adversarial generation time, as confirmed by the regression analysis. For example, around 9 hours are required when 80 estimators are used. Consequently, the answer to RQ1 in the case of XGB is negative.

Figure 5 illustrates the evolution of AT time as a function of the number of bagging trees in the RF model. A clear direct proportionality is observed between the two variables, with a maximum of approximately 150 seconds for 105 trees. Notably, some configurations show local minima in training

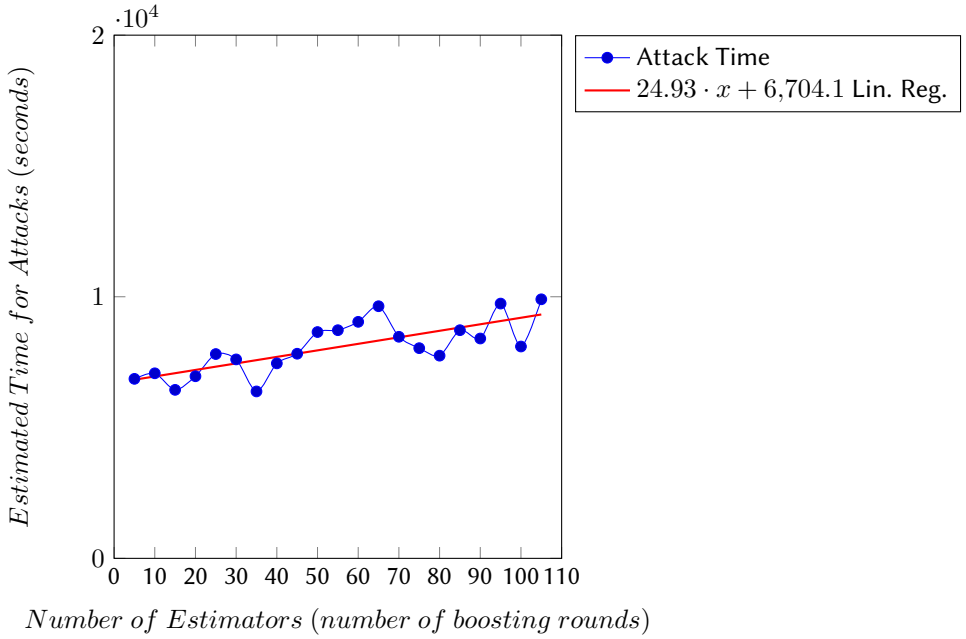

Figure 3: Time required to generate B' as a function of the increasing number of boosting rounds in the GB model.

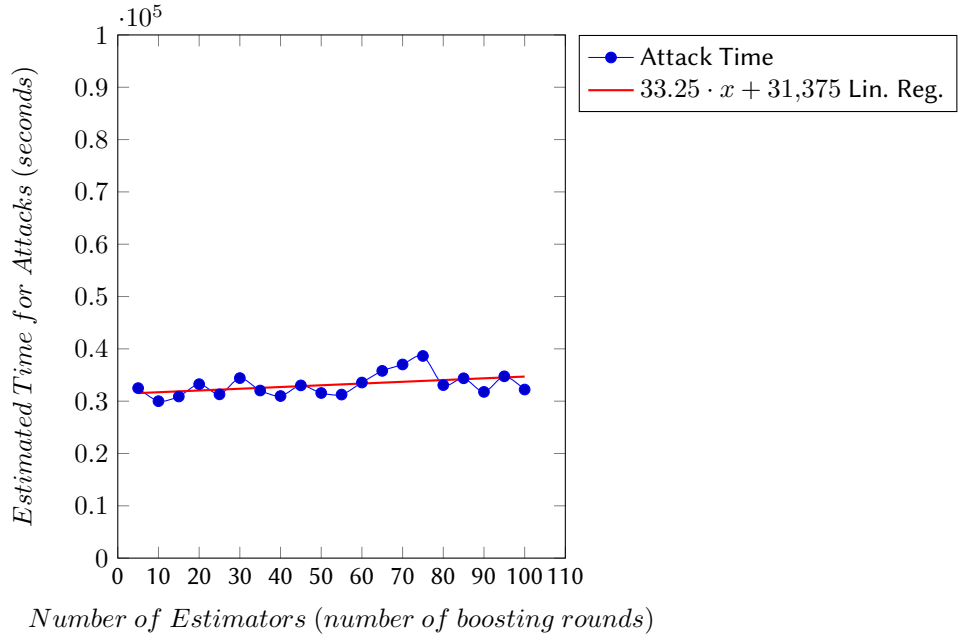

Figure 4: Time required to generate B' as a function of the increasing number of boosting rounds in the XGB model.

time, which can be attributed to tree configurations that limit the depth of the learned patterns. A similar phenomenon is observable in Figure 6, which depicts the impact of the number of boosting rounds on training time for the GB model. In this case, the maximum time reaches approximately 670 seconds for 105 boosting rounds. When comparing the estimated attack and AT times for both models, the RF model demonstrates a more favorable trade-off.

Furthermore, in both scenarios, this combination of training and attack times effectively reduces the attacker's "window of opportunity" (thus strengthening the defence) during an attack on the IVN, particularly when considering the potential additional attack delay introduced by the execution of AT. The needed AT time is significantly less than the time provided to the H-AI MDTF to detect the

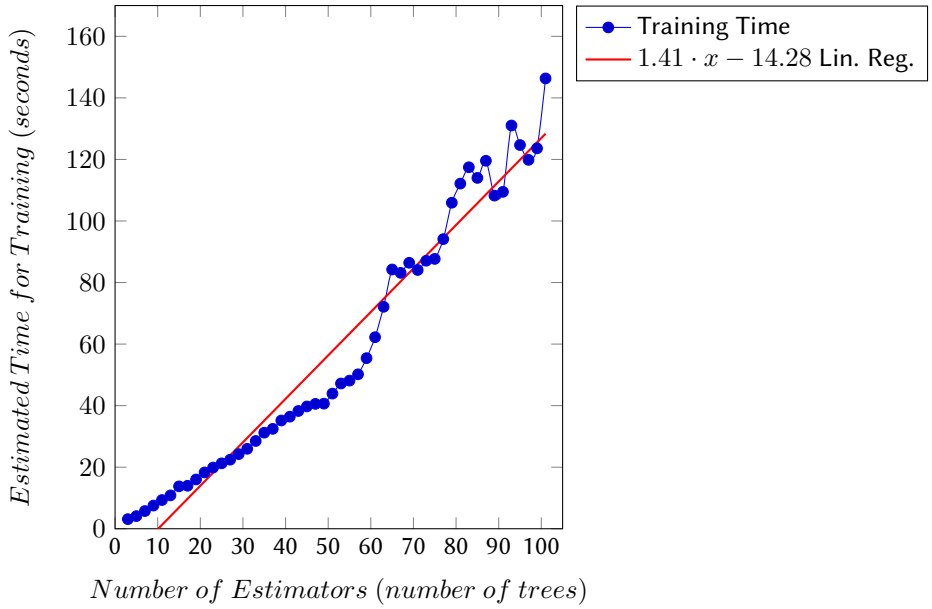

Figure 5: Time required to run the AT as a function of the increasing number of bagging trees in the RF model

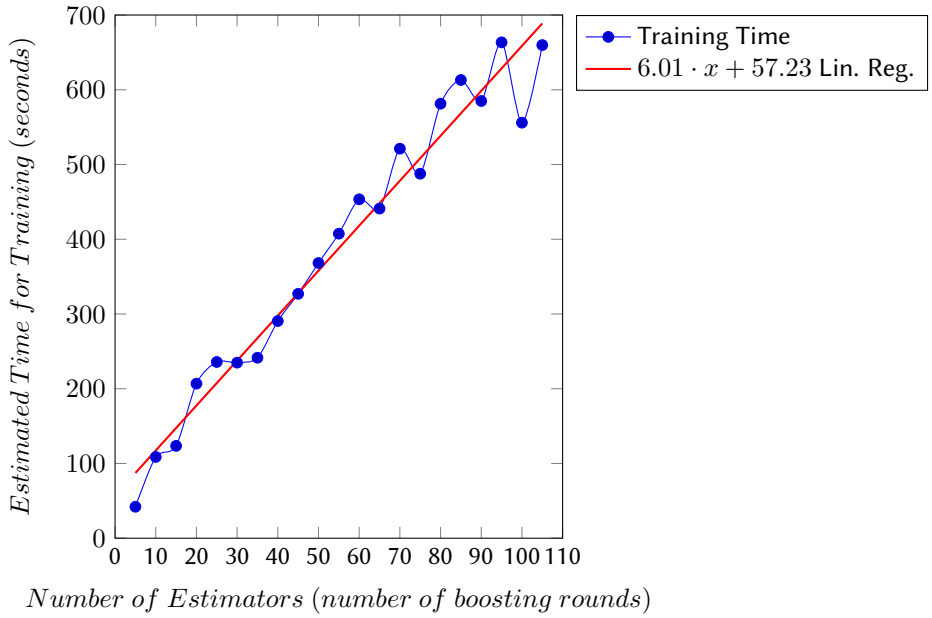

Figure 6: Time required to run the AT as a function of the increasing number of boosting rounds in the GB model

intrusion in the CAN network, making this countermeasure almost necessary in online systems. So, the answer to RQ2 is affirmative.

4.2. Hyperparameters' Impact on Educating MDTF along CSS-MDO Framework Axes for MDOs

Table 1 shows the impact of the previous analysis on the CSS-MDO framework axes for MDOs. This is the answer to RQ3. Generally, the education of MDTFs (even H-AI) about this analysis has a "Very High" impact due to the complementary motivations [29]. An "High" impact is observed on the "Prevention" axis since deterrence does not categorically prevent the execution of the attack. By properly educating MDTFs on this, it is also possible to gain valuable time in critical defensive operations. Indeed, exploiting

Table 1
Appropriate values' impact assigned to the hyperameters under consideration for RF and GB on the CSS-MDO framework's axes for MDOs.

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

an IDS that is highly robust to AML attacks helps ensure the sustained operational integrity of affected vehicles. Such robustness can be particularly critical during high-stakes MDOs, where maintaining functionality for as long as possible can be decisive.

5. Conclusion & Future Work

MDOs aim to integrate capabilities across all active domains of warfare to achieve coordinated, cross-domain effects against targeted systems. Within this context, Smart Cities and in particular CAVs emerge as critical and highly vulnerable assets, especially to cyberattacks leveraging Black-Box AML techniques. A perfect targets in such attacks can be ML-based IDSs employed for securing CAN-based IVNs. These threats are particularly concerning in the broader context of MDTs and, more specifically, within MSTs. Despite increasing attention, current research on Black-Box AML attacks targeting CAN frame detection remains sparse and at an early stage of development.

This paper evaluates (RQ1) the influence of hyperparameters related to DT-based state-of-the-art TT ensemble models (i.e. RF, GB, XGB), underlying an IDS targeted by the ZOO attack (seen as a SST) in a pure Black-Box scenario, on adversarial example generation time. Additionally, it understand (RQ2) whether the combo of the correct setting of the hyperparameters (under exam) and the AT needed time can slow down or reduce the attacker's "window of opportunity". Finally, it qualitatively assess (RQ3) the educational impact on MDTFs of this analysis mapped into the CSS-MDO framework contributing to the integration of artificial agents within MDTFs leading to H-AI MDTFs. The experimental results indicate a direct proportional relationship between the number of bagging trees in RF and boosting rounds in GB with the time required to exploit the ZOO attack, a trend not observed in XGB. Thus, only RF and GB exhibit hyperparameters that may serve as intrinsic defense mechanisms contributing to D-TAI. RF is recommended for its superior robustness considering the reduced AT time needed. Generally, the educational impact on MDTFs of such evidence is rated very high considering the important possibility of controlling the attack timing. All this, leads to labeling the collaboration between human and AI in H-AI MDTFs as functional to increase the resilience of social cyber attacks.

Some future directions of this work are 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 considered to extend the analysis on datasets and IDSs exploited in additional systems of national interest (i.e. IoT networks, aircraft, underwater vehicles), taking into account the possibility to integrate this intelligence to educate on risk management by exploiting a cyber social wargame also. Accompanying the analysis of AT times, it is useful to understand the attack times following the application of the countermeasure. Regarding the educational impact analysis, a clear development is to assess the attack impact also considering a victim TSRS in a MST.

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; Patto territoriale "Sistema universitario pugliese" – CUP F61B23000370006; CYBER-PREDICT: Cyber vulnerability ranking prediction by prescription, Avviso “Reti - Sostegno alla ricerca collaborativa”, - Codice Progetto DUQVKW0; SETH: Security Education and Training Hybrid-Wargame, Avviso “Reti - Sostegno alla ricerca collaborativa”

Declaration on Generative AI

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

References

[1] F. T. and, Nato’s approach to multi-domain operations: From the perspective of the economics of alliances, Defence and Peace Economics 35 (2024) 281–294. URL: https://doi.org/10.1080/10242694.2023.2235502. doi:10.1080/10242694.2023.2235502. arXiv:https://doi.org/10.1080/10242694.2023.2235502.

[2] A. Gilli, M. Gilli, G. G. and, Nato, multi-domain operations and the future of the atlantic alliance, Comparative Strategy 44 (2025) 73–91. URL: https://doi.org/10.1080/01495933.2024.2445491. doi:10.1080/01495933.2024.2445491. arXiv:https://doi.org/10.1080/01495933.2024.2445491.

[3] F. Thrope, E. Heinz, Improving cyberspace intelligence preparations for us army multi-domain operations (2019).

[4] T. Wójtowicz, D. Król, Multi-domain battle: new doctrine of the united states armed forces, Zeszyty Naukowe Akademii Sztuki Wojennej (2018) 64–78.

[5] F.-S. Gady, A. Stronell, Cyber capabilities and multi-domain operations in future high-intensity warfare in 2030, Cyber Threats and NATO 2030: Horizon Scanning and Analysis (2020) 151.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] H. Qin, M. Yan, H. Ji, Application of controller area network (can) bus anomaly detection based on time series prediction, Vehicular Communications 27 (2021) 100291. URL: https://www.sciencedirect.com/science/article/pii/S2214209620300620. doi:https://doi.org/10.1016/j.vehcom.2020.100291.

[13] 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.

[14] S. Rajapaksha, H. Kalutarage, M. O. Al-Kadri, A. Petrovski, G. Madzudzo, M. Cheah, Ai-based intrusion detection systems for in-vehicle networks: A survey, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3570954. doi:10.1145/3570954.

[15] 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.

[16] 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 & Mobile Communication Conference (UEMCON), 2021, pp. 0944–0949. doi:10.1109/UEMCON53757.2021.9666745.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.