

Object-focused Risk Evaluation of AI-driven Perception Systems in Autonomous Vehicles

Subhadip Ghosh, *Senior Member, IEEE*, Aydin Zaboli, *Graduate Student Member, IEEE*,
Junho Hong, *Senior Member, IEEE*, Jaerock Kwon, *Senior Member, IEEE*

Abstract— One of the primary motivations for autonomous vehicle (AV) technology is to reduce road accidents compared to human-driven cars. This necessitates having robust perception systems to detect and classify objects correctly in real-time environments. Various factors, including the complexity of the scene, the type of object, the capability of the perception sensors, and the performance of AI-based algorithms, can affect its robustness. Furthermore, vulnerabilities in these factors can be exploited as cyber-physical attacks. Hence, this paper presents a novel mathematical model for system-level risk evaluation of AV perception systems that incorporates the relevant objects for AV applications and the machine learning (ML) algorithms used to detect and classify them. This model is adapted from the ISO/SAE 21434 threat analysis and risk assessment (TARA) model with an enhancement in impact rating and attack feasibility assessment. Additionally, a case study for impact rating is demonstrated with real data from traffic crashes where the most important objects are impacted. Also, the effect of the robustness of the detection algorithm on attack feasibility assessment is illustrated with some AI/ML-based state-of-the-art detection algorithms used in AVs.

Index Terms—Attack, Autonomous Vehicles, Objects, Perception System, Risk Assessment, Robustness factor, TARA.

I. INTRODUCTION

The high-level working principle of AVs requires perception of the surrounding environment so that motion planning and control of driving actions can be performed accordingly. For this purpose, AVs collect data with perception sensors and process this data with an AI/ML-based perception algorithm to extract meaningful information about the scene. However, a compromised perception system can expose AVs to driving hazards. Few researchers have demonstrated some cyber-physical attacks (e.g., sensor jamming and spoofing attacks) on sensors. Some researchers have focused on adversarial methods (e.g., perturbation, inference, and data poisoning) to exploit vulnerabilities in perception algorithms that result in incorrect classification of traffic signs and objects on the road [1]–[5]. Thus, a robust TARA method is crucial to assess the safety and performance risk of AVs when the perception system is under attack. Traditional threat modeling techniques for automotive applications primarily focus on electrical and electronic (E/E) systems with attacker, asset, or software-centric approaches. However, these methods are not adequate to capture system-level threats from a compromised cyber-physical interaction of the perception system [6]. An integrated TARA framework is proposed after conducting a comparative

analysis of the AV perception system with the ISO/SAE21434 TARA guidelines and a systems theoretic process approach known as STPA-Sec [7]. Furthermore, ISO/SAE 21434 is refined to incorporate the object-centric approach and AI robustness factor into the mathematical modeling for risk calculations. The principal contributions of this paper are articulated as follows:

- An integrated TARA framework introduced customized for the AV perception systems. Designed to enable a rigorous end-to-end assessment of mission risks arising from security weaknesses in AVs, the framework encompasses a theoretical basis, a formalized mathematical model, and preliminary demonstrations of its applicability to the object (i.e., humans, bicyclists & motorcyclists, animals, and vehicles) detection process within AV systems.
- Also, a modified methodology proposed for the risk assessment analysis using real traffic crash data within AV systems, guided by insights derived from the ISO/SAE 21434 standards. The refinement centers upon enhancing calculations of impact ratings and attack feasibility for vulnerable interactions or elements. This is accomplished through a comprehensive analysis of traditional AV perception system architectures and functions, alongside the subsequent integration of AI algorithm vulnerability (i.e., robustness factor) considerations and the significance of detected objects within the AV risk assessment framework. Consequently, this refinement yields more precise and applicable risk evaluations specifically tailored to the AV domain.

The remainder of the paper is organized as follows: Section II states a representation of the integrated threat model framework. Different level of object-centric evaluation according to the risk analysis along with the modified mathematical modeling are mentioned in Section III. Section IV presents the results and discussion of based on the impact rating according to the traffic crash real data and the robustness factor, integrated in the risk formulations. Finally, this paper is concluded in Section V.

II. AN INTEGRATED THREAT MODEL FRAMEWORK

This threat model integrates system-centric and asset-centric approaches to analyze the threats of perception from system interaction with the environment and from attacks on the hardware and software components. As shown in Fig. 1, steps of this TARA model are adapted from the STPA-Sec method and guidance in the ISO/SAE 21434 standard. In

S. Ghosh, A. Zaboli, J. Hong and J. Kwon are with the Department of Electrical and Computer Engineering, University of Michigan – Dearborn, Dearborn, MI 48128, USA. emails: {subhagh, azaboli, jhwr, jrkwon}@umich.edu

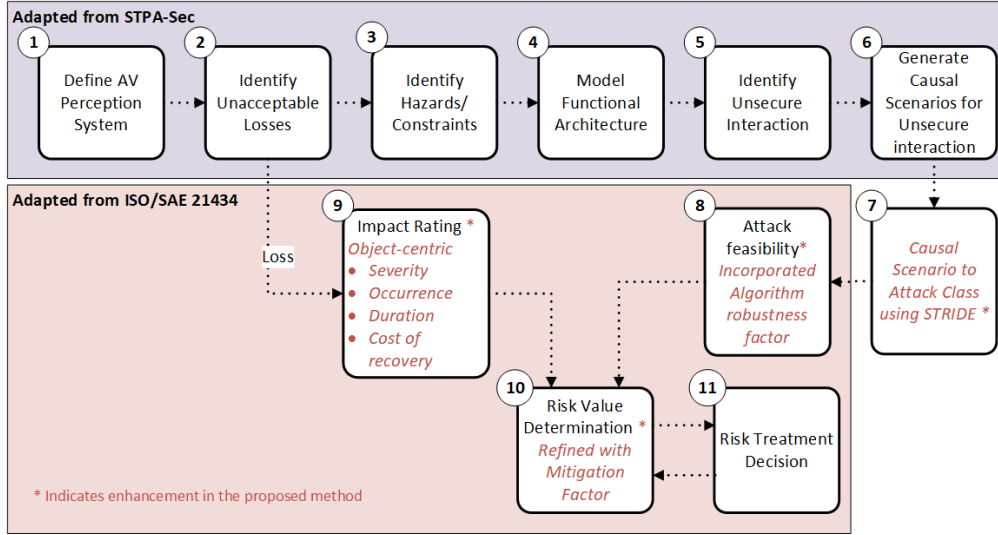


Fig. 1. An integrated TARA framework.

this framework, analysis starts by defining the AV perception system and then identifying the losses that are unacceptable (e.g., loss of life or injury). In the 3rd step, the hazards are identified that can cause the loss. Then, a model is created to identify the interaction that can cause the hazard when compromised. In the next step, corresponding possible causal scenarios for this unsecure interaction are generated. These tasks are performed in 4th, 5th and 6th steps. In the 7th step, the causal scenario is transferred from the system domain to the attack domain by using the STRIDE method. In the 8th step, the AI robustness factor is incorporated as an attack potential factor for attack feasibility assessment, along with other factors from ISO/IEC 18045. In the 9th step, impacts are rated for unacceptable losses depending on the types of objects on the road. In the last steps, risk is managed after considering the attack feasibility assessment, impact rating, and mitigation factor to detect and control the risk. In the next section, the mathematical models for attack feasibility, impact rating, and risk calculations are discussed in detail.

III. OBJECT-CENTRIC RISK EVALUATION MODEL

The assessment of risk levels involves a comparative analysis between the severity of possible damage scenarios and the exploitable attack vectors. The likelihood of an attack path is quantified by attributing numerical values to each contributing factor, such as specialist expertise (*SE*), knowledge of the time or component (*KoIC*), equipment (*Eq*), elapsed time (*ET*), window of opportunity (*WoO*), and the proposed robustness factor (*RF*). Attack potentials are classified into four distinct categories, each assigned a numerical range from 0 to 3, with 0 representing the most relaxed scenario and 3 indicating the highest level of concern. These parameters are systematically delineated and expounded in alignment with the ISO/IEC 18045 standard, specifically for the automotive industry. According to Fig. 2, *SE* categorizes attackers based on their automotive knowledge and technical skills, ranging from laymen with limited expertise to multiple experts with

SE	KoIC	Eq	RF
Layman 0	Public 0	Standard 0	Performance: Normal scenarios = Poor Abnormal scenarios = Poor Adversarial attacks = Poor
Proficient 1	Restricted 1	Specialized 1	Performance: Normal scenarios = Good Abnormal scenarios = Poor Adversarial attacks = Poor
Expert 2	Sensitive 2	Custom-built 2	Performance: Normal scenarios = Good Abnormal scenarios = Good Adversarial attacks = Poor
Multiple Experts 3	Critical 3	Multiple Custom-built 3	Performance: Normal scenarios = Good Abnormal scenarios = Good Adversarial attacks = Good

ET	WoO
< a day 0	Unlimited 0
< a week 1	Easy 1
< a month 2	Moderate 2
> a month 3	Difficult 3

Parameter (ϕ) Attack potential description and value (V)

Fig. 2. Evaluation of attack potential in compliance with ISO/IEC 18045 parameters.

multidisciplinary knowledge. *KoIC* describes the accessibility of target information, from public to critically restricted. *Eq* assesses the resources needed for an attack, from standard to multiple custom-built tools. *ET* measures the duration required to execute an attack, which varies from less than a day to over a month. *WoO* refers to the time frame in which a target must be accessible for a successful attack, ranging from unlimited to difficult access [8]. Lastly, the *RF* parameter evaluates the susceptibility of perception subsystems in AVs, considering both software elements and AI algorithms under normal, abnormal, and adversarial attack scenarios. This framework provides a comprehensive view of the factors influencing the potential for attacks on AV systems.

A. Proposed Integrated Mathematical Modeling for Risk Assessment

A novel mathematical model is presented in Eq. (1) for assessing the risk value in the context of AI vulnerability analysis, specifically for USecX (R_{USecX}). This model, distinct from previous approaches, incorporates a mitigation factor (M_{USecX}) that accounts for the controllability (C_{USecX}) and detectability (D_{USecX}) of attacks, assigning values from 0 to 2 to various levels of these parameters, as illustrated in Eq. (1) [9], [10]. For example, in case of AV perception,

introducing multi-modal hi-fidelity sensors and robust AI/ML algorithm validated with exhaustive normal, abnormal and adversarial scenarios can improve the detectability. Further, having a real-time response system to take the AV to a safe state can improve the controllability [11]. The model's innovation lies in its consideration of the mitigation factor, a critical aspect often overlooked in prior research. This factor plays a pivotal role in gauging the extent to which attacks can be detected and managed, thereby linking the concepts of attack types, their controllability and detectability, and overall risk assessment in a comprehensive framework.

$$R_{USecX} = \frac{1 + I_{USecX} \times F_{USecX}}{1 + M_{USecX}}, \quad (1)$$

$$M_{USecX} = C_{USecX} + D_{USecX}.$$

This approach also prioritizes various road elements (e.g., humans, animals, and vehicles) based on factors such as their damage severity ($S_{L,USecX}$), presence rating ($O_{H,USecX}$), average recovery time ($T_{L,USecX}$), and financial loss ($\Gamma_{L,USecX}$). Additionally, the impact rating (I_{USecX}) ranges from 0 to 3 (low to severe levels, respectively), encompassing the overall impact of cyberattacks, which can vary across environmental, financial, and operational safety aspects, defined as Eq. (2):

$$I_{USecX} = S_{L,USecX} \times O_{H,USecX} + T_{L,USecX} + \Gamma_{L,USecX} \quad (2)$$

This method highlights the heightened risk and necessity for effective risk management strategies. The approach categorizes risk levels into four intuitive groups, aiding in easier interpretation and application in risk assessments, in which low, moderate, high, and severe impact ratings can be assigned. The framework, based on the sum of attack potential values, establishes a link between these values and the attack feasibility rating (F_{USecX}), as outlined in Eq. (3) and depicted in Fig. 2. The attack potential values, determined by parameters in Eq. (4), range from 0 to 18 and are categorized into three groups (i.e., 0-6, 7-12, and 13-18) for better comprehension of attack feasibility.

$$F_{USecX} = \max(F_{USecX}^{\eta_i}) \quad (3)$$

$$\sum V_{\eta_i} = V_{\eta_i}^{SE} + V_{\eta_i}^{K o I C} + V_{\eta_i}^{Eq} + V_{\eta_i}^{ET} + V_{\eta_i}^{W o O} + V_{\eta_i}^{RF} \quad (4)$$

for $i = 1, 2, 3, \dots, n.$

Eq. (4) delineates the correlation among attack potential parameters, highlighting that basic attack paths (utilizing standard tools, unskilled attackers, and public information) have higher possibility of success due to their simplicity. The complexity and feasibility of an attack path inversely correlate with the required attack potential degree. The novel index, $V_{\eta_i}^{RF}$ measures the resilience of ML algorithms (i.e., object classification) against typical, abnormal, and adversarial scenarios. This index effectively assesses the performance of AVs under attacks, regarding different objects, reflecting the AVs' environmental dependence.

IV. RESULTS AND DISCUSSION

A. Impact Rating for Objects from Traffic Crash Data

This section presents the real traffic crash data along with the calculations for impact ratings for different objects. Table I demonstrates the different parameters in impact rating (I_{USecX}) considering humans, bicyclists & motorcyclists, animals, and vehicles involved in car crashes. The real data was extracted from the Michigan Traffic Crash Facts (MTCF) and National Safety Council – Injury Facts [12], [13] for crashes in December 2022 in Michigan State. This data is categorized based on four groups, including fatal injury, suspected serious injury, suspected minor & possible injury, and no injury, to calculate the parameters involved in Eq. (2). The severity levels are considered severe (3), high (2), moderate (1), and low (0), respectively, for different objects involved in crashes. The first parameter, severity, can be calculated for the pedestrians involved in crashes based on different levels of injuries as follows:

$$\begin{aligned} & \text{Severe} \times \frac{\text{Fatal Injury}}{\text{Total Crash Count}} + \text{High} \times \frac{\text{Suspected Serious Injury}}{\text{Total Crash Count}} \\ & + \text{Moderate} \times \frac{\text{Suspected Minor \& Possible Injury}}{\text{Total Crash Count}} \\ & + \text{Low} \times \frac{\text{No Injury}}{\text{Total Crash Count}} \\ & = 3 \times \frac{23}{198} + 2 \times \frac{37}{198} + 1 \times \frac{114}{198} + 0 \times \frac{24}{198} = 1.298 \end{aligned}$$

Similarly, other severity values can be found for different objects on the road. Occurrence can be found according to the total crash count (i.e., 198) for pedestrians divided by the total crash count (i.e., $198 + 52 + 5907 + 1683 = 7840$) for all objects (e.g., $\frac{198}{7840} = 0.0253$) to show the presence of objects according to the traffic crashes on the road. This is the same procedure to find other *Occurrence* values for different objects. Assume a range of recovery times within each category including fatal (9 – 12 months), serious (6 – 9 months), minor/possible (1 – 3 months), and no injury (0). The average recovery time for each category using the midpoint can be found. A sample calculation of the average recovery time (months) for pedestrians can be $\frac{23 \times 10.5 + 37 \times 7.5 + 114 \times 2 + 24 \times 0}{198} = 4.65$ months, and a similar process can be carried out for bicyclists and motorcyclists, animals, and vehicle items. Regarding the average costs per injury category, it is necessary to define the average costs for each injury category. These can vary significantly depending on factors (e.g., location, healthcare costs, legal settlements, and lost wages). Some rough estimates can be mentioned as follows:

- Fatal Injury: \$10,000,000 (assuming high cost of life, medical expenses, and lost wages)
- Suspected Serious Injury: \$500,000 (assuming moderate cost of medical treatment and lost wages)
- Suspected Minor Injury & Possible Injury: \$100,000 (assuming lower cost of medical treatment)
- No Injury: \$0 (assuming no immediate financial loss)

An example of the financial loss calculations based on the pedestrian object can be represented as follows:

TABLE I
IMPACT RATING PARAMETERS FOR DIFFERENT OBJECTS ON THE ROAD BASED ON REAL TRAFFIC CRASH DATA.

Objects	Severity ($S_{L,U SecX}$)	Occurrence ($O_{H,U SecX}$)	Recovery time (months) ($T_{L,U SecX}$)	Financial loss (M\$) ($\Gamma_{L,U SecX}$)
Human (i.e., Pedestrian)	1.298	0.0253	4.65	259.9
Bicyclist & Motorcyclist	0.788	0.00663	2.15	25.3
Animal	0.0208	0.753	0.044	22.9
Vehicle	0.191	0.2147	0.48	204.8

- Fatal Injury: 23 pedestrians \times \$10,000,000/pedestrian = \$230,000,000
- Suspected Serious Injury: 37 pedestrians \times \$500,000/pedestrian = \$18,500,000
- Suspected Minor Injury & Possible Injury: 114 pedestrians \times \$100,000/pedestrian = \$11,400,000
- No Injury: 24 pedestrians \times \$0/pedestrian = \$0

Total Estimated Cost: \$230,000,000 + \$18,500,000 + \$11,400,000 + \$0 = \$259,900,000

According to Table I, human object category demonstrates the highest impact rating among objects, according to real traffic crash data. $I_{U SecX}$ index can be advantageous in terms of different parameters including damage severity, presence of different objects on the road, average recovery time after damage occurrence, and financial loss. According to real crash data in Michigan State, a thorough analysis of different objects considering indexes can be carried out. For instance, even though the occurrence index ($O_{H,U SecX}$) for animals is significantly higher than other objects, this object category has lower average values for severity, recovery time, and the loss. Also, different level of injuries considered which the most vulnerabilities have been obtained to humans and vehicles, respectively.

B. Attack Feasibility Assessment with AI Robustness Factor

The section details three case studies that utilize a modified formula incorporating an AI robustness factor to assess attack feasibility. Currently, in the absence of a standardized robustness metric, the prevailing practice is to gauge AI's resilience based on conventional performance metrics amidst defensive scenarios. ROC curve, accuracy, precision, recall, and F1-score are some of the common metrics that are used for evaluating AI performance. In this paper, the performance metrics provided by the authors are translated to RF factor as per Figure 2 along with other parameters. For first two cases, it is assumed that performance of the perception algorithm is good in normal and abnormal scenarios to highlight the effect on RF and attack feasibility due to adversarial attacks. First example is for speed limit sign detection with same defense model for multiple attacks and the second one is for object detection under the same attack with multiple defense methods [14], [15]. These case studies are shown with attack type, defense method, attack potential and attack feasibility values in Table II. The third case is presented to show the comparison of attack feasibility ratings when the defense method is completely missing for AI perception algorithm under adversarial attack and performance is poor for abnormal

scenarios. This example is crafted based on the performance of traffic cone detection in abnormal scenarios [16] and then assumed perception will be poor under adversarial attack. In Fig. 3, the a comprehensive risk assessment approach with an AI robustness factor from our proposed integrated TARA method is presented. In this approach, potential AI/ML models

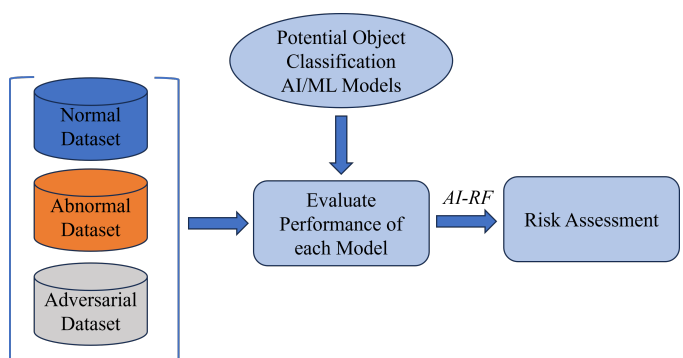


Fig. 3. A comprehensive risk assessment based on the AI robustness factor for AV perception datasets.

for AV perceptions systems can be evaluated against normal, abnormal and adversarial dataset. For each model AI RF can be assigned based on Fig. 2 and provided to risk assessment formula according to Eq. 1. As per our analysis, a decrease in AI RF enables lower values of SE , ET , and Eq for a successful attack. As a result, the attack feasibility is high when AI RF is low. When this is combined with abnormal scenarios, as shown in Case 3, the cumulative attack feasibility also increases. It can be interpreted as a higher risk according to Eq. 1. In Fig. 4, an AI/ML model performance for normal, abnormal and adversarial scenarios based on different case studies from Table II is shown as a bar chart for a comparative view. A

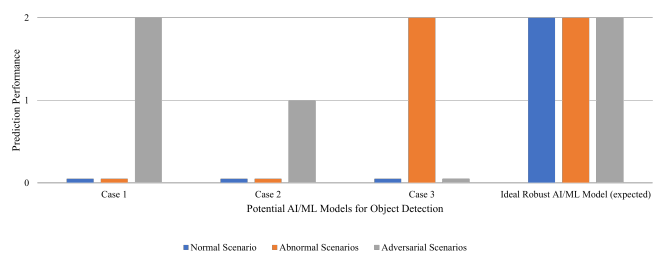


Fig. 4. A comparative example of an AI robustness factor evaluation considering case studies from Table II for normal, abnormal, and adversarial scenarios.

prediction value near to 0 represents model performance is

TABLE II
ATTACK FEASIBILITY CALCULATIONS FOR DIFFERENT AI ALGORITHMS' ROBUSTNESS FOR PERCEPTION SYSTEMS.

Case #	Attack feasibility assessment							Attack feasibility
	∇^{ET}	∇^{SE}	$\nabla^{K\sigma IC}$	$\nabla^{W\sigma O}$	∇^{Eq}	∇^{RF}	$\sum V$	
1.	4	3	3	0	4	3	17	Low
2.	1	3	3	0	0	2	9	Moderate
3.	1	3	3	0	0	1	8	High
<p>Case 1 - Attack: poster-printing attack I-FGSM, C&W, Deepfool, JSMA. Defense method: SVD-based optimal approximation with 5G. Performance: accuracy score (80%–90%). hence determined as good performance.</p> <p>Case 2 - Attack: a patch on the back of a truck placed in front of the camera. Defense method: FPDA, Z-mask, HyperNeuron. Performance: 0.5–0.6 AUROC. hence determined as poor performance.</p> <p>Case 3 - Performance for the abnormal scenario is 65.8%.</p> <p>Assumption 1: For Cases 1 and 2, performance is good in normal and abnormal scenarios. Assumption 2: For Case 3, performance is good in normal but there is no defense against adversarial scenarios.</p>								

not evaluated or shown poor performance and 1 and 2 values represent moderate and good performance, respectively. It can be noted that these models have not shown the comprehensive good results against normal, abnormal and adversarial scenarios as shown in an hypothetical example of ideal robust AI/ML model on the right side of the chart.

V. CONCLUSION AND FUTURE WORK

The TARA model presented here incorporates the effect of objects on the road and the robustness of AV perception algorithms. A case study with humans, cyclists, animals, and vehicles from traffic crash data demonstrates that the proposed model embeds object-level granularity while evaluating the risk of every perception system under an attack. This case study shows that an assessment of attack feasibility augmented with an AI robust factor captures the holistic performance of the perception algorithm under normal, abnormal, and adversarial scenarios with attack and defense methods. We believe that both of these enhancements are useful tools to assess the risk of the AV perception system and evaluate potential solutions to mitigate the threat. In the future, the plan is to analyze this model with traffic data for various scenarios and other combinations of attack and defense methods. Also, this is our goal to develop the AI/ML models for AV scenarios which performs close to the robust AI/ML model.

REFERENCES

- [1] N. Worzyk, H. Kahlen, and O. Kramer, "Physical adversarial attacks by projecting perturbations," in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 649–659.
- [2] R. Duan, X. Mao, A. K. Qin, Y. Chen, S. Ye, Y. He, and Y. Yang, "Adversarial laser beam: Effective physical-world attack to dnns in a blink," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 062–16 071.
- [3] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1625–1634.
- [4] H. Zhang, W. Zhou, and H. Li, "Contextual adversarial attacks for object detection," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, 2020, pp. 1–6.
- [5] D. Li, J. Zhang, and K. Huang, "Universal adversarial perturbations against object detection," *Pattern Recognition*, vol. 110, p. 107584, 2021.
- [6] M. Girdhar, Y. You, T.-J. Song, S. Ghosh, and J. Hong, "Post-accident cyberattack event analysis for connected and automated vehicles," *IEEE Access*, vol. 10, pp. 83 176–83 194, 2022.
- [7] S. Ghosh, A. Zaboli, J. Hong, and J. Kwon, "An integrated approach of threat analysis for autonomous vehicles perception system," *IEEE Access*, vol. 11, pp. 14 752–14 777, 2023.
- [8] S. R. Esmaili and A. S. Esterabadi. (2019) Attack analysis methodologies. Master Thesis, Chalmers University of Technology, Gothenburg, Sweden.
- [9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [10] F. Wu, L. Xiao, W. Yang, and J. Zhu, "Defense against adversarial attacks in traffic sign images identification based on 5g," *EURASIP Journal on Wireless Communications and Networking*, vol. 2020, no. 1, pp. 1–15, 2020.
- [11] A. Zaboli, J. Hong, J. Kwon, and J. Moore, "A survey on cyber-physical security of autonomous vehicles using a context awareness method," *IEEE Access*, vol. 11, pp. 136 706–136 725, 2023.
- [12] "Michigan traffic crash facts (MTCF)," <https://www.michigantrafficcrashfacts.org/>, accessed on Nov. 2023.
- [13] "National Safety Council - Injury Facts," <https://injuryfacts.nsc.org/>, accessed on Nov. 2023.
- [14] L. Xiao, L. Xiao, W. Yang, and J. Zhu, "Defense against adversarial attacks in traffic sign images identification based on 5g," *EURASIP Journal on Wireless Communications and Networking*, pp. 1687–1499, 2020.
- [15] F. Nesti, G. Rossolini, G. D'Amico, A. Biondi, and G. Buttazzo, "Carla-gear: a dataset generator for a systematic evaluation of adversarial robustness of vision models," 2022.
- [16] H. Zhu, T. M. Tam Tran, A. Benjumea, and A. Bradley, "A scenario-based functional testing approach to improving dnn performance," in *2023 IEEE International Conference on Service-Oriented System Engineering (SOSE)*, 2023, pp. 199–207.