International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Adversarial Attacks and Defense Mechanisms in Autonomous Vehicles

Antony Adwin Luiz [1], Sudha D[2]
[1] Student, Department of Computer Applications, SCMS School of Technology & Management
[2] Assistant Professor, Department of Computer Applications, SCMS School of Technology & Management

https://doi.org/10.55041/ijsmt.v2i5.125
Cite this Article: Luiz, A. A. (2026). Adversarial Attacks and Defense Mechanisms in Autonomous Vehicles. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.125

Abstract
The integration of deep learning (DL) and artificial intelligence (AI) has significantly advanced the capabilities of autonomous vehicles (AVs), enabling intelligent perception, planning, and decision-making. However, these enhancements have also introduced new cybersecurity vulnerabilities, particularly in the form of adversarial attacks. This review focuses on four key adversarial attack types—Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), physical adversarial patch attacks, and sensor spoofing—that pose serious threats to AV safety. It further explores advanced and practical defense mechanisms such as hybrid deep learning models, patch-based occlusion-aware detection, autoencoders with memory modules, and sensor fusion-based spoofing detection frameworks. The review is supported by recent research papers and highlights both theoretical and applied perspectives, covering classification and perception tasks, sensor vulnerabilities, and real-world validation. Through detailed analysis of attack mechanisms, impacts, and mitigation strategies, the urgent need for scalable, real-time, and integrated defense systems threats is emphasized to secure AI-driven transportation from evolving adversarial.

Keywords: Autonomous Vehicles, Adversarial Attacks, Deep Learning Security, Fast Gradient Sign Method, Projected Gradient Descent, Physical Adversarial Patch Attacks, Sensor Spoofing, Defense Mechanisms, Anomaly Detection, Sensor Fusion

1 Introduction
The integration of deep learning and AI has significantly enhanced autonomous vehicle (AV) capabilities, but this introduces cybersecurity risks, particularly from adversarial attacks. These attacks manipulate AI perception systems, causing AVs to misinterpret their surroundings and make incorrect decisions, threatening safety and reliability. Examples include altering traffic signs or spoofing LiDAR sensors to create false objects. Adversarial attack techniques are systematically reviewed, and defense mechanisms are explored like adversarial training, robust deep learning models, and anomaly detection. The goal is to highlight vulnerabilities and propose solutions to enhance AV cybersecurity, ensuring safe and reliable operation in the face of evolving threats.

2 Literature Review

2.1 Ibrahim et al. (2025) – Safety-Oriented Review of Adversarial Attacks
This systematic review by Ibrahim et al. [1] analyzes 117 papers published between 2020 and 2024, focusing on adversarial attacks and defenses in autonomous vehicles (AVs) from a safety perspective. The authors propose a safety scenarios taxonomy matrix inspired by ISO 21448, also known as Safety Of The Intended Functionality (SOTIF), categorizing attacks into environment-based (such as LiDAR spoofing and camera patches), agent-based (like data poisoning), and camouflage-based scenarios. Defenses are grouped into robustness-based methods (e.g., adversarial training) and detection-based approaches (e.g., anomaly detection). The review emphasizes the feasibility of real-world attacks, such as adversarial patches on traffic signs, while also discussing the limitations of current defense mechanisms, including issues like computational overhead. While the paper offers comprehensive coverage of multi-sensor attack

vectors—including LiDAR, camera, and radar—and introduces a valuable safety-centric taxonomy, it does not include experimental validation of the proposed defenses and leans more toward classification than introducing new mitigation techniques. The study effectively highlights deep neural network vulnerabilities and emphasizes defense strategies such as adversarial retraining.

2.2 Girdhar et al. (2023) – Cybersecurity of Autonomous Vehicles

This paper by Girdhar et al. [2] presents a detailed and structured literature review that explores the cybersecurity landscape of autonomous vehicles (AVs) with a focus on adversarial machine learning attacks and corresponding defense models. It outlines how deep learning (DL) techniques used in AVs—such as for perception, prediction, and planning—are highly susceptible to attacks like evasion and data poisoning. The paper discusses real-world adversarial examples, including spoofing traffic signs, LiDAR sensor manipulation, and misclassification due to adversarial perturbations. It also categorizes defenses into proactive (e.g., adversarial training, defensive distillation, feature masking) and reactive strategies (e.g., adversarial detection, input reconstruction, and network verification). While the study provides an exhaustive overview of AI vulnerabilities, various attack methods (like FGSM, PGD, C&W), and defense techniques in AV ecosystems, it lacks novel empirical contributions and focuses largely on summarizing existing frameworks. The paper's strength lies in its wide scope, covering both theoretical concepts and industrial case studies, and its value in identifying gaps—such as the limited evaluation of adversarial defenses in regression-based AV models—thus encouraging future research.

2.3 Deng et al. (2021) – Survey of Attacks and Defenses in Autonomous Driving Systems

Deng et al. [3] offer a detailed survey on the various adversarial threats faced by deep learning-based autonomous driving systems (ADSs), highlighting how attacks can target different layers—from sensing and perception to decision-making and cloud services. It categorizes attacks into physical (like LiDAR spoofing and camera blinding), cyber (such as GPS spoofing and message falsification), and learning-based adversarial attacks (including evasion and poisoning), demonstrating their potential to mislead deep neural networks into dangerous decisions. While the paper thoroughly reviews both white-box and black-box adversarial attack strategies, it also discusses a wide range of defense mechanisms including adversarial training, defensive distillation, redundancy in sensors, network regularization, and real-time anomaly detection systems. However, many of these defenses are either resource-intensive or validated only on simplified models, making them less practical for real-time AV deployment. Nonetheless, the paper emphasizes the need for future research into full-system robustness, semantic adversarial attacks, and lightweight, scalable defense strategies suitable for real-world autonomous driving environments.

2.4 Mahima et al. (2021) – Adversarial Attack and Defense Technologies

Mahima et al. [4] present a detailed review of adversarial attacks and defense technologies in autonomous vehicles (AVs), focusing on how machine learning models in AVs—used for tasks such as traffic sign recognition, semantic segmentation, object detection, and steering control—are susceptible to various adversarial threats. It categorizes attacks based on their phase (training vs. testing), nature (black-box, white-box, grey-box), and purpose (targeted, untargeted, exploratory), and elaborates on notable examples like lenticular printing, GAN-based perturbations, adversarial billboards, and steering deviation using physical-world setups. The paper also surveys multiple defense strategies including adversarial training, defensive distillation, GAN-based denoisers, and encoder-decoder restorers, showing varying degrees of robustness across tasks. While the paper is commendable for covering both theoretical and applied aspects of adversarial ML in AVs, including simulation and real-world experiments, many defenses remain only partially effective, especially under physical transformations and complex driving environments. Furthermore, the study emphasizes the pressing need for generalizable and lightweight defense frameworks capable of countering both man-made and natural adversarial scenarios in real-time conditions, which remains an open challenge in the field.

2.5 Almutairi et al. (2023) – Security of Deep Neural Networks in Smart Vehicles

Almutairi et al. [5] provide a comprehensive survey on the security challenges faced by deep neural networks (DNNs) in smart vehicles, highlighting how adversarial attacks—both digital and physical—can compromise object detection and

decision-making in autonomous systems. It categorizes attacks based on the adversary's knowledge (white-box or black-box), intent (targeted or untargeted), and medium (digital or physical), and presents a thorough taxonomy of generation techniques such as FGSM, PGD, and semantic attacks like sticker manipulation or lighting changes. The paper also explores an extensive range of defense strategies, including adversarial training, denoisers like DIP, certified robustness, preprocessing filters like DCT-based transforms, and detection models based on temporal video consistency or frame-level anomaly scores. Moreover, it discusses DNN evaluation frameworks like DeepTest, which test model behavior in extreme scenarios. While the paper excels in organizing existing techniques and pinpointing critical gaps in real-time defense performance and evaluation, it also notes that many defenses compromise accuracy on clean data and struggle with scalability or physical-world robustness. The authors call for future research into certified defenses, 3D adversarial robustness, sensor-level redundancies, and smarter test data generation for more resilient autonomous systems.

2.6 Khan et al. (2022) – Hybrid Defense Method for Traffic Sign Classification

Khan et al. [6] propose a deep learning-based hybrid defense method that combines random filtering, ensembling, and local feature mapping to improve the robustness of traffic sign classifiers in autonomous vehicles under adversarial attacks. Using transfer learning, the authors retrain Inception-V3 and ResNet-152 models on traffic sign images and integrate them with optical character recognition (OCR) to detect text-based local features. The hybrid method is tested against common white-box adversarial attacks such as FGSM, MIM, PGD, and C&W, achieving up to 89% classification accuracy under attack, significantly outperforming traditional defenses like JPEG filtering and feature squeezing. The ensemble-based voting mechanism and random image transformations make it harder for adversaries to exploit the classifiers, while the inclusion of OCR enhances resilience for text-based signs. Although highly effective, the approach is computationally intensive and currently focused only on traffic sign recognition, suggesting the need for broader validation across other AV perception tasks and attack types.

2.7 Badjie et al. (2024) – Adversarial Attacks on Image Classification Models

Badjie et al. [7] provide a comprehensive and focused review on adversarial attacks specifically targeting image classification-based deep learning models in autonomous driving systems (ADSs). It introduces novel taxonomies for classifying attacks based on perturbation scope, visibility, metrics, intent, and adversary knowledge, and analyzes over 50 attack methods including FAB, C&W, ATTA, and DeepBillboard. It also evaluates defensive techniques such as adversarial training, gradient masking, and ensemble methods, offering insights into both proactive and reactive countermeasures. The paper emphasizes challenges like real-time constraints, dynamic driving environments, multi-sensor fusion, and the difficulty of achieving physical-world attack robustness. Although extensive in coverage, the study is centered primarily on classification tasks and does not delve into other perception components like detection or segmentation. Still, it stands out by offering recommendations and research directions aimed at enhancing model robustness and advancing secure image-based perception in ADSs.

2.8 Shibly et al. (2023) – Memory-Based Defense Model for Autonomous Driving

Shibly et al. [8] propose a defense framework for autonomous driving models using an autoencoder with a memory module to counter adversarial attacks such as Hijacking, Vanishing, Fabrication, and Mislabeling. Tested on the NVIDIA DAVE-2 model with the Udacity dataset, the system utilizes a memory-based reconstruction technique that transforms adversarial inputs to closely resemble clean data. The model demonstrates superior performance compared to traditional defenses like adversarial training and defensive distillation, achieving up to 93.8% success in White-box and 74.1% in Black-box settings. It proves especially effective against FGSM and AdvGAN attacks. Despite its effectiveness, the approach incurs computational overhead due to its complex memory and GAN-based architecture. The study suggests future improvements in optimizing memory usage and expanding real-time applicability.

2.9 Gupta et al. (2023) – Cyber-Attacks and Security Mechanisms in CAV Systems

Gupta et al. [9] present a structured examination of cyber threats and security solutions across the Edge–Fog–Cloud architecture of connected and autonomous vehicles (CAVs). The paper categorizes attacks into software (e.g., ML poisoning, Sybil), network (e.g., DoS, MiTM), and hardware-level threats (e.g., side-channel, node capture). For each threat domain, corresponding defense mechanisms are reviewed, including encryption, secure protocol design, anomaly detection, and formal verification. With over 200 references, the review offers comprehensive insight into current challenges and countermeasures across CAV systems. While the framework is conceptually sound, many proposed defenses remain theoretical and face implementation barriers in resource-constrained environments. Nevertheless, the work contributes a scalable architectural perspective for developing practical and layered cybersecurity solutions in CAV ecosystems.

2.10 Choi and Tian (2022) – Adversarial Attacks on YOLO Detectors

Choi and Tian [10] analyze the vulnerability of YOLO object detectors to adversarial perturbations that specifically target objectness scores, rather than conventional loss functions like classification or localization. Their proposed objectness-targeted attack significantly reduces detection performance—by 45.2% and 43.5% on KITTI and COCO datasets, respectively. To mitigate this, they introduce an objectness-aware adversarial training technique that increases mean Average Precision (mAP) by 21% and 12% on the same datasets. This novel focus on objectness brings a new dimension to understanding detection fragility. The model retraining strategy demonstrates high efficacy in countering adversarial samples without degrading performance on clean data. However, the approach is limited to YOLO detectors and does not evaluate generalization across other perception architectures or in real-world deployment scenarios.

2.11 Sobh et al. (2021) – Multi-Task Perception Vulnerability Analysis

Sobh et al. [11] examine how adversarial perturbations affect multi-task visual perception networks in AVs, which simultaneously handle distance estimation, motion detection, semantic segmentation, and object recognition. Using both white-box and black-box attacks—targeted and untargeted—the study evaluates performance degradation and inter-task vulnerability propagation. Results show that an attack targeting a single task can substantially impair the outputs of other tasks, revealing the inherent coupling in shared encoder-decoder structures. The authors demonstrate that even simple defenses like input preprocessing or normalization are often insufficient in complex multitask setups. The paper offers deep insights into the cascading effects of adversarial inputs across tasks, promoting the need for holistic and task-aware defense methods. Despite its strengths in highlighting cross-task vulnerabilities, the study relies mostly on simulation and lacks evaluation under real-world conditions or with more sophisticated mitigation techniques.

2.12 Jiao et al. (2021) – Uncertainty-Based Mitigation Framework

Jiao et al. [12] propose a novel uncertainty-aware mitigation framework that integrates perception uncertainty into planning and control modules in lane-centering Advanced Driver Assistance Systems (ADAS). Implemented using OpenPilot and evaluated both in simulation and on real-world datasets, the system reduces lane deviation under adversarial conditions by 55–90%. This end-to-end approach is notable for going beyond perception to address downstream decision-making components. By modeling epistemic and aleatoric uncertainties, the system adjusts control outputs dynamically, improving robustness without explicit retraining of the perception model. The design marks a promising shift toward integrated AV defense strategies that bridge perception and control. Nonetheless, its practical applicability may be constrained by the added complexity of uncertainty modeling and computational resource requirements, especially for real-time, embedded AV systems.

2.13 Chen et al. (2025) – Evaluation of Perception Attacks and Defenses

Chen et al. [13] revisit the state of adversarial attacks on AV perception pipelines, including tasks like road sign detection and lead vehicle recognition, using production-grade L2 autonomous systems such as OpenPilot with YOLO-based models. The study evaluates multiple defense techniques, including adversarial training, input preprocessing, contrastive learning, and diffusion-based purification. Through extensive testing, the paper highlights which combinations of defense methods perform robustly across a range of physical-world perturbations and attack settings. The analysis offers valuable guidance for selecting context-aware defense solutions tailored to real-time AV applications. However, the scope remains

largely limited to perception tasks and does not account for cascading effects on planning or control modules, suggesting the need for more integrated, system-level studies in future work.

2.14 Strack et al. (2024) – Defense Against Physical Patch Attacks

Strack et al. [14] explore the vulnerability of infrared (IR)-based human detection systems—commonly used in autonomous vehicles and surveillance applications—to physical adversarial patch attacks. These attacks involve placing carefully designed patch patterns in the camera's field of view, which mislead object detectors into either ignoring humans or falsely detecting non-existent objects. The authors propose a Patch-based Occlusion-aware Detection (POD) method that augments training data with synthetic patch occlusions of varying size, shape, and location. This strategy helps the model learn to distinguish between actual human features and misleading patches. The method is lightweight, does not require architecture changes, and shows strong adaptability across different patch types. Experimental results demonstrate improved robustness on state-of-the-art IR detection benchmarks, with minimal performance loss on clean data. The paper emphasizes the practicality of POD in real-world autonomous systems where IR sensors are used in low-visibility environments. While highly effective, the study focuses mainly on human detection and does not extend the technique to other perception tasks, suggesting the need for broader applicability and real-time validation across full AV pipelines.

2.15 Dasgupta et al. (2024) – GNSS Spoofing Detection Using Sensor Fusion

Dasgupta et al. [15] propose a lightweight Global Navigation Satellite System (GNSS) spoofing detection method for autonomous vehicles using sensor fusion. Their system combines predictions from three models: (i) location shift estimation using LSTM, (ii) turn detection using k-Nearest Neighbors (k-NN) with Dynamic Time Warping (DTW), and (iii) motion state analysis based on speed data. The fused output is compared with GNSS readings to detect spoofing. Using the Honda Driving Dataset, the method accurately identifies false turns, stops, and position shifts with minimal delay. Turn detection achieved 100% accuracy, and location shift prediction had very low error. This approach enhances spoofing detection in real-world driving using only in-vehicle sensors and efficient computation. It achieves high accuracy with low computational cost using only onboard sensors, though it is limited to detecting only GPS spoofing attacks.

3 Types of Adversarial Attacks

3.1 Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) is a type of adversarial attack that can significantly impact the safety and reliability of autonomous vehicles (AVs) by causing deep neural network (DNN) models to misclassify critical information, such as traffic signs [2], [8]. These attacks introduce small, often imperceptible, perturbations to input data (e.g., images), leading the AV's AI system to make incorrect decisions [2].

FGSM attacks present significant challenges to autonomous vehicles by targeting their perception systems. One major issue is the misclassification of traffic signs, where even a slight perturbation can cause a stop sign to be misread, leading to hazardous outcomes [6]. These attacks compromise decision-making processes by feeding manipulated inputs to deep learning models, resulting in incorrect scene interpretations and potentially dangerous driving actions [7]. Furthermore, with the growing reliance on vehicle-to-everything (V2X) communication, the attack surface broadens, giving adversaries more opportunities to inject malicious data into sensor streams [2]. The subtle nature of FGSM perturbations, often invisible to the human eye, makes detection difficult without specialized defense mechanisms [2].

Autoencoder and Compressive Memory Module

The defense mechanism, highlighted in "Towards Autonomous Driving Model Resistant to Adversarial Attack" [8], involves the use of an autoencoder and a compressive memory module. This solution aims to prevent unexpected generalization on adversarial inputs by:

Storing Normal Image Features

The compressive memory module stores features of normal, unperturbed images.

Detecting Anomalies

When an adversarial input is presented, the system can compare it to the stored normal features, identifying and mitigating the adversarial perturbations.

This approach was evaluated against FGSM and AdvGAN attacks on the Nvidia Dave-2 driving model. It demonstrated significant effectiveness, achieving success rates of 93.8% (White-box setup) and 74.1% (Black-box setup) against FGSM attacks [8]. This represents an improvement of 24.7% in a White-box setup and 21.5% in a Black-box setup compared to previous results [8]. These detailed approaches demonstrate ongoing efforts to enhance the resilience of autonomous vehicles against sophisticated adversarial attacks like FGSM.

3.2 Projected Gradient Descent (PGD)

Projected Gradient Descent (PGD) is a white-box adversarial attack that enhances the Fast Gradient Sign Method (FGSM) by applying iterative gradient-based perturbations to input data. After each step, the modified input is projected back into a bounded norm space, ensuring the adversarial example remains visually similar to the original but can mislead the model into incorrect classifications [2], [5]. PGD is widely regarded as one of the strongest first-order adversarial attacks and serves as a benchmark for evaluating the robustness of deep learning (DL) models in adversarial settings [3], [4].

PGD attacks can seriously affect the safety of autonomous vehicles. They can fool the system into misreading traffic signs or objects, even if the model was trained to resist simpler attacks like FGSM [4], [6]. In vehicles that handle multiple tasks, such as detecting objects and understanding road layout, attacking one task can also weaken the others [11]. These attacks can confuse the vehicle's perception, which then affects its planning and control decisions, leading to unsafe driving behavior [2], [3], [11].

Hybrid Defense Method

An effective approach against PGD and other adversarial attacks in autonomous vehicles is a hybrid defense method that combines multiple strategies to improve the resilience of traffic sign classifiers [6]. This DNN-based hybrid approach utilizes a combination of three distinct strategies:

Random Filtering

This technique employs random cropping and resizing of input images to mitigate the effects of adversarial perturbations [6].

Ensembling

Plurality voting is used as an ensembling strategy, where multiple models are combined, and their outputs are aggregated to make a more robust decision [6].

Local Feature Mapping

An optical character recognition (OCR) model is used as a local feature mapper to extract and analyze specific features, such as text on traffic signs, making the system more resilient to subtle changes introduced by attacks [6].

This hybrid defense method has demonstrated significant improvements in accuracy against PGD attacks. For example, it improved traffic sign classification accuracy by 55% against PGD attacks compared to traditional defense methods, achieving an average accuracy of 87% in PGD attack scenarios [6]. The Inception-V3 and ResNet-152 models are retrained using transfer learning within this hybrid framework to function as traffic sign classifiers [6]. This approach aims to maintain classification performance even after an adversarial attack has impacted the input data [6].

3.3 Physical Adversarial Patch Attacks

Physical adversarial patch attacks involve manipulating real-world physical objects or signals to create adversarial examples. These attacks exploit vulnerabilities in a vehicle's interaction with its surroundings by manipulating inputs from various sensors and perception systems, significantly compromising AV safety and functionality [4].

Physical adversarial patch attacks can seriously mislead autonomous vehicles by tricking their visual systems. These attacks may cause deep learning models to misclassify traffic signs or objects, leading to wrong driving decisions and safety risks [7]. By projecting specific patterns, attackers can create false objects or change real ones, taking advantage of camera limitations like lens flare or auto-exposure [1]. They can also fool monocular depth estimation models into measuring distances incorrectly, even in real-world scenarios [1]. Traffic sign recognition systems are especially vulnerable, as these patches work across different conditions with high success rates [1]. Additionally, YOLO detectors, widely used for object detection in AVs, can be exploited through their sensitivity to objectness scores, making them more vulnerable than traditional classifiers [10].

Patch-based Occlusion-aware Detection (POD)

An effective and computationally simple technique to defend against physical adversarial patch attacks in autonomous vehicle perception systems is patch-based data augmentation, also referred to as Patch-based Occlusion-aware Detection (POD) [14]. This method enhances robustness by introducing synthetic occlusions—randomly generated patch-like patterns—onto training images, simulating potential adversarial attacks. During training, the model learns not only to detect standard objects such as pedestrians or traffic signs but also to ignore or suppress misleading information introduced by the patches. This enables the system to maintain accurate object detection even when facing previously unseen patch types during inference.

Unlike complex architectural modifications or certifiable defenses, POD is lightweight and easy to implement, requiring only the augmentation of training data with patched samples. Moreover, this technique shows strong adaptability by handling patch occlusions of different shapes, sizes, and placements effectively. A 2024 study demonstrated that POD significantly improves the model's resilience against state-of-the-art physical patch attacks while maintaining computational efficiency and clean-image performance [14].

3.4 Sensor Spoofing Attacks

Sensor spoofing attacks trick an autonomous vehicle's perception system by feeding it false sensor data [3]. In LiDAR spoofing, attackers can create fake 3D objects or relay signals to confuse the system, causing it to detect non-existent "ghost" objects or miss real ones—potentially blinding the vehicle or misplacing important obstacles [1], [3]. Even small changes can make critical objects invisible or wrongly located [1]. GPS spoofing involves sending fake GPS signals to the vehicle, misleading it about its location and direction, which can result in incorrect navigation and unsafe driving decisions [3].

Sensor spoofing attacks manipulate the input of one or more sensors to deceive the vehicle's perception system [3]. This can lead to misperception, where objects are wrongly classified or ignored, resulting in incorrect steering, braking, or planning actions [1], [3]. In LiDAR spoofing, attackers can make critical objects disappear or create false ones, hiding nearby vehicles or obstacles [3]. GPS spoofing disrupts the vehicle's sense of position, causing it to follow incorrect routes or make unsafe maneuvers [3]. As a result, the AV's overall system performance degrades. Spoofing reduces the overall quality and trustworthiness of sensor data, weakening the performance of the AV's entire decision-making pipeline [2].

Sensor Fusion-Based GNSS Spoofing Defense

A robust yet straightforward method to mitigate GPS spoofing attacks on autonomous vehicles involves fusing data from multiple onboard sensors—such as GNSS, accelerometers, steering angle sensors, and vehicle speed sensors—and comparing them in real time. In one study, data from these sensors were fused using an LSTM (Long Short-Term Memory) neural network to predict expected changes in vehicle location between time steps. Simultaneously, measurements like turns (via steering angles) were analyzed using k-Nearest Neighbors (k-NN) and Dynamic Time Warping (DTW) to detect motion patterns. By comparing this multi-sensor prediction with actual GNSS readings, the system reliably detects inconsistencies indicative of spoofing attacks [15].

The method was tested using real-world driving datasets, where it successfully identified multiple spoofing scenarios such as false stops, wrong turns, and location shifts—all within acceptable latency limits for vehicle operation [15].

4 Conclusion

The rapid integration of deep learning in autonomous vehicles has undeniably enhanced their perception, planning, and control capabilities; however, it has also exposed these systems to serious adversarial vulnerabilities. Attack methods such as FGSM, PGD, physical patch-based perturbations, and sensor spoofing have been shown to mislead object detection, alter perception, and even hijack navigational decisions, compromising safety and reliability [2], [3], [4], [7].

These attacks exploit weaknesses in both digital models and physical-world perception, affecting components such as traffic sign classifiers, LiDAR point clouds, and GPS systems [1], [3], [4]. Studies show that even small, imperceptible modifications can result in consequences, particularly in multi-task systems where a single compromised module can propagate errors across others [11]. As highlighted in multiple reviews, traditional defenses like adversarial training or preprocessing often fail under real-world conditions due to limited scalability and high computational costs [5], [7], [9].

To counter these challenges, recent research has proposed a variety of advanced and practical defense strategies tailored to real-world AV systems. Hybrid methods incorporating ensembling and OCR for traffic sign recognition [6], memory-augmented autoencoders for filtering adversarial noise [8], and Patch-based Occlusion-aware Detection (POD) for physical patch attacks [14] have demonstrated notable success in improving robustness. Additionally, sensor fusion-based frameworks that leverage LSTM and k-NN models to detect GNSS spoofing offer lightweight, onboard protection for navigation systems [15].

Yet, despite these advancements, most solutions remain task-specific, and comprehensive, system-level defenses that integrate perception, planning, and control remain limited. The future of secure AV deployment thus lies in developing scalable, real-time defense systems that combine robust learning models with redundancy, anomaly detection, and context-aware decision-making strategies, ensuring resilience against evolving adversarial threats [1], [3], [12].

References

1. Ibrahum A.D.M., Hussain M., Hong J.E., “Deep Learning Adversarial Attacks and Defenses in Autonomous Vehicles: A Systematic Literature Review from a Safety Perspective”, Artificial Intelligence Review, 2025, 58 (1), 1–53.

2. Girdhar M., Hong J., Moore J., “Cybersecurity of Autonomous Vehicles: A Systematic Literature Review of Adversarial Attacks and Defense Models”, IEEE Open Journal of Vehicular Technology, 2023, 4, 417–437.

3. Deng Y., et al., “Deep Learning-Based Autonomous Driving Systems: A Survey of Attacks and Defenses”, IEEE Transactions on Industrial Informatics, 2021, 17 (12), 7897–7912.

4. Mahima K.T.Y., Ayoob M., Poravi G., “Adversarial Attacks and Defense Technologies on Autonomous Vehicles: A Review”, Applied Computer Systems, 2021, 26 (2), 96–106.

5. Almutairi S., Barnawi A., “Securing DNN for Smart Vehicles: An Overview of Adversarial Attacks, Defenses, and Frameworks”, Journal of Engineering and Applied Science, 2023, 70 (1), 16.

6. Khan Z., Chowdhury M., Khan S.M., “A Hybrid Defense Method Against Adversarial Attacks on Traffic Sign Classifiers in Autonomous Vehicles”, arXiv preprint, 2022. https://arxiv.org/abs/2205.01225

7. Badjie B., Cecílio J., Casimiro A., "Adversarial Attacks and Countermeasures on Image Classification-Based Deep Learning Models in Autonomous Driving Systems: A Systematic Review", ACM Computing Surveys, 2024, 57 (1), 1–52.

8. Shibly K.H., et al., "Towards Autonomous Driving Model Resistant to Adversarial Attack", Applied Artificial Intelligence, 2023, 37 (1), 2193461.

9. Gupta S., Maple C., Passerone R., "An Investigation of Cyber-Attacks and Security Mechanisms for Connected and Autonomous Vehicles", IEEE Access, 2023, 11, 90641–90669.

10. Choi J., Tian Q., "Adversarial Attack and Defense of YOLO Detectors in Autonomous Driving Scenarios", IEEE Intelligent Vehicles Symposium, 2022.

11. Sobh I., et al., "Adversarial Attacks on Multi-Task Visual Perception for Autonomous Driving", arXiv preprint, 2021. https://arxiv.org/abs/2107.07449

12. Jiao R., et al., "End-to-End Uncertainty-Based Mitigation of Adversarial Attacks to Automated Lane Centering", IEEE Intelligent Vehicles Symposium, 2021.

13. Chen C., et al., "Revisiting Adversarial Perception Attacks and Defense Methods on Autonomous Driving Systems", arXiv preprint, 2025. https://arxiv.org/abs/2505.11532

14. Strack L., et al., "Defending Against Physical Adversarial Patch Attacks on Infrared Human Detection", IEEE International Conference on Image Processing, 2024.

15. Dasgupta S., Rahman M., Islam M., Chowdhury M., "Sensor Fusion-Based GNSS Spoofing Attack Detection Framework for Autonomous Vehicles", IEEE International Conference on Image Processing, 2024.