electronics

MDPI

Article

One Possible Path Towards a More Robust Task of Traffic Sign Classification in Autonomous Vehicles Using Autoencoders

Ivan Martinović [1], Tomás de Jesús Mateo Sanguino [2,*], Jovana Jovanović [3], Mihailo Jovanović [4] and Milena Djukanović [1]

[1] Faculty of Electrical Engineering, University of Montenegro, Dzordza Vasingtona bb, 81000 Podgorica, Montenegro; ivanma@ucg.ac.me (I.M.); milenadj@ucg.ac.me (M.D.)
[2] Research Centre for Technology, Energy and Sustainability, University of Huelva, Avda. de las Artes S/N, 21007 Huelva, Spain
[3] Faculty of Civil Engineering and Management, University Union Nikola Tesla, 11120 Belgrade, Serbia; jjovanovic@unt.edu.rs
[4] Faculty of Management Herceg Novi, University Adriatik, Zemunska 143, 85348 Meljine, Montenegro; mihailo.jovanovic@fm-hn.com
* Correspondence: tomas.mateo@diesia.uhu.es

Abstract: The increasing deployment of autonomous vehicles (AVs) has exposed critical vulnerabilities in traffic sign classification systems, particularly against adversarial attacks that can compromise safety. This study proposes a dual-purpose defense framework based on convolutional autoencoders to enhance robustness against two prominent white-box attacks: Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). Experiments on the German Traffic Sign Recognition Benchmark (GTSRB) dataset show that, although these attacks can significantly degrade system performance, the proposed models are capable of partially recovering lost accuracy. Notably, the defense demonstrates strong capabilities in both detecting and reconstructing manipulated traffic signs, even under low-perturbation scenarios. Additionally, a feature-based autoencoder is introduced, which—despite a high false positive rate—achieves perfect detection in critical conditions, a tradeoff considered acceptable in safety-critical contexts. These results highlight the potential of autoencoder-based architectures as a foundation for resilient AV perception while underscoring the need for hybrid models integrating visual-language frameworks for real-time, fail-safe operation.

Keywords: traffic sign; classification task; adversarial attack; FGSM; PGD; autoencoder

Academic Editor: Piotr Borkowski

Received: 16 May 2025
Revised: 4 June 2025
Accepted: 6 June 2025
Published: 11 June 2025

Citation: Martinović, I.; Mateo Sanguino, T.d.J.; Jovanović, J.; Jovanović, M.; Djukanović, M. One Possible Path Towards a More Robust Task of Traffic Sign Classification in Autonomous Vehicles Using Autoencoders. *Electronics* **2025**, *14*, 2382. https://doi.org/10.3390/electronics14122382

1. Introduction

The advancement of autonomous vehicles (AVs) has shifted from merely developing self-driving capabilities to emphasizing their safe operation. Deep learning plays a vital role in AV development, particularly in handling perception tasks such as steering, decision-making, and traffic sign recognition. However, these AI-based systems are vulnerable to adversarial attacks—subtle, often imperceptible perturbations to input data that can mislead models and lead to potentially dangerous outcomes [1].

Adversarial attacks can be categorized into digital and physical types. Digital attacks involve modifying input data within the digital domain, while physical attacks manipulate real-world objects to deceive perception systems [2]. These attacks can target various AV sensors, including cameras, LiDAR, radar, and multi-sensor fusion systems. Among these, camera-based perception is particularly vulnerable, especially in the task of traffic sign recognition.

Currently, the traffic sign recognition task is highly vulnerable to adversarial attacks because it relies solely on camera input. Unlike other perception tasks, it cannot be cross-verified or corrected using additional sensors like LiDAR, making it more susceptible to errors. However, in the future, as most vehicles become autonomous, traffic sign recognition will not depend solely on cameras, as traffic signs will be equipped with RFID and Bluetooth technologies [3].

Traffic sign recognition is a multi-class classification task typically structured in two stages: detection and classification [4]. It is especially susceptible to adversarial attacks such as the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and adversarial patches. These attacks can be executed in white-box or black-box settings, depending on the attacker's knowledge of the model. The consequences of such attacks are severe, as they can cause AVs to misinterpret regulatory signs, leading to unsafe behavior.

Although various defense mechanisms have been proposed in the literature, most fall into two categories: detection of adversarial inputs and recovery (or denoising) of perturbed samples. However, existing methods often struggle to detect low-perturbation attacks and typically address detection and recovery separately. Moreover, few approaches are tailored specifically to the traffic sign classification task in AVs, despite its critical safety implications.

This study addresses these gaps by proposing a dual-purpose defense framework based on convolutional autoencoders. The proposed system is capable of both detecting and reconstructing adversarial traffic sign images, with a particular focus on white-box attacks such as FGSM and PGD. Additionally, we introduce a feature-based autoencoder designed to improve detection performance under low-perturbation conditions—an area where conventional methods are notably weak.

The following section outlines the specific hypotheses and objectives of this research.

1.1. Research Hypotheses and Objectives

This study hypothesizes that the robustness of traffic sign classification systems in AVs can be significantly improved through the use of convolutional autoencoders. These models are designed to serve a dual defensive function: (1) detecting adversarial inputs by analyzing reconstruction errors when trained exclusively on clean data, and (2) recovering unperturbed images from adversarially modified inputs, thereby acting as a pre-classification defense mechanism.

The research addresses specific gaps in the literature, including the limited effectiveness of conventional defenses against low-perturbation attacks and the lack of integrated detection-recovery frameworks tailored to traffic sign recognition. To validate this hypothesis, the study sets out to do the following:

- Analyze the vulnerability of traffic sign classification systems to two widely used white-box adversarial attack methods (i.e., FGSM and PGD).
- Design and evaluate a dual-mode autoencoder-based defense system capable of both detecting and restoring images affected by such attacks.
- Introduce a feature-based autoencoder that enhances detection performance in low-perturbation scenarios, addressing a known limitation of conventional autoencoders.
- Contribute to the development of safer and more reliable AV perception systems, in alignment with safety standards such as ISO 21448 [5], by proposing a defense strategy that prioritizes robustness, even at the cost of higher false positive rates in safety-critical contexts.

To this end, the manuscript is organized as follows. Section 2 reports the history, technological achievements, and ethical and regulatory considerations of AVs; Section 3

proposes the materials and methods; Section 4 presents an analysis of the experimental results and discussion; and Section 5 concludes the paper.

1.2. List of Abbreviations

To improve clarity and ensure consistency throughout the manuscript, a list of acronyms used in this study is provided below.

- ATC, Automated and connected transport.
- AI, Artificial intelligence.
- ATC, Automatic train control.
- AV, Autonomous vehicle.
- CNN, Convolutional neural network.
- C&W, Carlini and Wagner.
- EATA, European Automotive-Telecom Alliance.
- ETSI, European Telecommunications Standards Institute.
- FDAV, Framework on Automated/Autonomous and Connected Vehicles.
- FGSM, Fast Gradient Sign Method.
- FPR, False positive rate.
- GDPR, General Data Protection Regulation.
- GTSRB, German Traffic Sign Recognition Benchmark.
- IJCNN, Joint Conference on Neural Networks.
- IoT, Internet of things.
- ISO, International Organization for Standardization.
- ITS. Intelligent transportation system.
- IVHS, Intelligent vehicle highways system.
- JSMA, Jacobian-based saliency map attack.
- LiDAR, Light detection and ranging.
- LLM, Large language model.
- MSF, Multi-sensor fusion.
- NHTSA, National Highway Traffic Safety Administration.
- PGD, Projected gradient descent.
- SAE, Society of Automotive Engineers.
- US, United States.
- V2I, Vehicle-to-infrastructure.
- V2V, Vehicle-to-vehicle.
- V2X, Vehicle-to-everything.
- WHO, World Health Organization.

2. Background and Related Work

This section addresses the development and evolution of AVs, highlighting the application of artificial intelligence (AI), the benefits in terms of road safety and efficiency, and the challenges related to robustness, cybersecurity, and the regulation necessary for their safe implementation. The historical stages, technological achievements, and ethical and regulatory considerations that determine the future of this technology are analyzed.

2.1. Artificial Intelligence and Safety Challenges in AVs

Nowadays, research in the field of AVs heavily relies on the application of AI. However, early developments in AVs did not involve AI. One of the first applications of AI in AVs occurred in 1995 when researchers from Carnegie Mellon University developed the autonomous vehicle NavLab5 [6]. NavLab5 traveled approximately 2797 miles from

Pittsburgh to San Diego. In this project, AI was integrated into image processing and steering control.

Today, the application of AI in AVs depends on the level of autonomy. The Society of Automotive Engineers (SAE) defined six levels (L0–L5) of autonomy [7], ranging from no driving automation to full driving automation, in the SAE J3016 standard [8]. The use of AI is crucial for higher levels (L3–L5), which rely heavily on machine learning and deep learning algorithms. At these levels, AI is primarily used for perception tasks, such as traffic sign detection and classification, as well as for complex decision-making and route planning [9]. The development and deployment of AI models for these tasks follow a systematic process that includes the following stages: data collection and pre-processing, model training, model generation, code refinement and optimization, quality assessment, and integration and deployment. Studies show that the trend of developing and deploying AI models in AVs follows the exponential trend of AI technology development [10]. The main trigger for this exponential increase since 2013 was the achievement of AlexNet in the ImageNet 2012 Challenge, which reached a top-5 error rate of 15.3%, more than 10.8 percentage points lower than that of the runner-up [11].

Studies show that the development of AVs is practically justified. Various statistics and analyses indicate that passenger vehicles are often considered one of the most unsafe modes of transportation [12]. According to the World Health Organization (WHO), road traffic injuries are a leading cause of death worldwide, particularly among young people [13]. According to the National Highway Traffic Safety Administration (NHTSA), human error is responsible for approximately 94% of serious crashes [14]. However, it is worth noting that other sources suggest that this figure may vary depending on the methodology and definitions used to classify crash causes [15]. For example, some incidents categorized as human error may involve complex interactions with environmental or systemic factors, and not all of them are necessarily preventable. Despite these nuances, the statistics remain a powerful indicator of the potential benefits of reducing human involvement in driving tasks. This is one of the key motivations behind the development of AVs, which aim to enhance road safety through automation. Nevertheless, AVs introduce new challenges, particularly in ensuring the robustness of AI-based perception systems. Among these, traffic sign recognition is especially critical, as it relies solely on visual input and is highly vulnerable to adversarial attacks. These attacks can subtly manipulate input data to mislead the system, potentially resulting in dangerous misclassifications. Therefore, improving the resilience of AV perception systems against such threats is not only a technical imperative but also a foundational step toward realizing the safety benefits that AVs are expected to deliver.

Scientists see the deployment of AVs as a promising approach to reducing these alarming statistics. They claim that, in addition to improving road safety, AVs can enhance accessibility and provide independent mobility, especially for those unable to drive, such as the elderly, disabled, and youth [9]. Other benefits of AV deployment include reduced congestion, optimized routing, saving time otherwise spent on driving and searching for parking, and reduced direct emissions. Moreover, AVs could reduce individual vehicle ownership and free up land used for parking by establishing vehicle-sharing models. Although there are undoubtedly great benefits to potentially deploying AVs, the existence of significant challenges prevents their broad application. These challenges are, in general, related to AI safety. This means that the deployment of AI models in AVs should be characterized by good robustness and cybersecurity. Robustness is particularly crucial because AVs are highly complex systems and operate in challenging environments consisting of countless situations that are only partially captured by the datasets on which the models are trained. The impact of reduced robustness can lead to very serious consequences. One

situation where a decision-making error occurred happened in 2023 in San Francisco [16]. A pedestrian was hit by two cars: first, a regular vehicle threw her into the path of an autonomous vehicle taxi, which then ran her over and stopped on top of her. From this described situation, we can conclude that reduced robustness caused the system to fail to effectively manage the emergency, where immediate action was required to prevent harm. Cybersecurity is also crucial for AVs, as they are cyber-physical systems vulnerable to malicious actors driven by profit or intent to cause physical harm [7,17,18]. This challenge is significant even for conventional vehicles, as demonstrated in a study where, in 2014, researchers first showed how well-known regular vehicles could be attacked over the Internet to gain remote control and induce crashes [7]. The study identified two main types of attacks on AVs: evasion attacks and adversarial examples and data poisoning. Adversarial example attacks pose a primary threat, as they can be executed without detailed knowledge of the AI system's internal design, simply by monitoring the outputs of AI components. Practically, this can involve altering the perceived environment through sensor manipulation, such as placing stickers on signs to mislead the vehicle into exceeding speed limits. In contrast, data poisoning attacks target communication channels between vehicles and manufacturers during the update process. Malicious actors use this method to inject corrupted data during the training phase, potentially altering AI components that could then be deployed on a large scale. The cyber vulnerabilities of AVs also adds to the complexity of software packages, ranging from around 100MB for Level 2 vehicles to hundreds of gigabytes to terabytes for higher levels of autonomy [10].

In the previous section, we discussed several challenges related to AI safety. However, it is equally important to address the ethical and regulatory challenges associated with AVs [19]. Ethical considerations in AI involve ensuring that these vehicles are developed and utilized in a fair, transparent, and respectful manner while upholding human values. Regulations play a crucial role in governing the deployment of AI technologies, ensuring safety, accountability, and adherence to legal standards in AVs. Currently, there are various regulatory documents guiding the development and deployment of AVs. These include a comparative international report [20], which covers regulatory advancements, driverless testing, emerging technologies, and key influencers across nine countries: Canada, China, Germany, Hungary, India, Poland, South Korea, the United Kingdom, and the United States (US). Additionally, the Framework on Automated/Autonomous and Connected Vehicles (FDAV) [21], developed at the intergovernmental level, seeks to harmonize global regulations for automated driving and promote innovation. Moreover, a 2021 international regulatory overview outlines testing and deployment frameworks for AVs across 27 countries, encompassing China and countries across Europe, Asia-Pacific, and North America. In parallel with these regulatory efforts, it is important to consider how the deployment of AVs and ACT systems impacts different stakeholder groups. These impacts are not uniform, as the benefits and challenges vary significantly depending on whether the perspective is that of drivers, governments, or cities. Table 1 provides a structured summary of these differences, offering a broader societal context to the technical and regulatory challenges discussed above.

Overcoming all challenges in the development and deployment of AVs is only possible through a multidisciplinary approach involving stakeholders from all relevant fields (Figure 1). These challenges highlight the need for robust perception systems, particularly in tasks such as traffic sign recognition, which are critical for AV safety and highly susceptible to adversarial manipulation.

Table 1. Benefits and challenges of ACT systems.

Group	Benefits	Challenges
Drivers	■ Increased safety: reduced accidents, fewer fatalities and injuries, safer roads for all users. ■ Improved comfort: smoother and less stressful rides, reduced time behind the wheel, access to real-time traffic and weather information, personalized route suggestions, and other infotainment functions. ■ Enhanced efficiency: shorter trips, reduced fuel consumption, less time wasted in traffic, optimized parking, and streamlined toll payments. ■ New mobility services: access to on-demand ride-hailing, car-sharing, and AV services.	■ Data privacy: concerns about the collection, use, and sharing of personal data, including location, driving habits, and preferences. ■ Cybersecurity: risk of hacking into vehicles, which could lead to accidents, identity theft, or manipulation of driving behavior. ■ Cost: the cost of equipping vehicles with ACT systems may be high, and there may be ongoing fees for data and service subscriptions. ■ Acceptance and trust: concerns about the reliability and accuracy of ACT systems, as well as potential biases or discrimination in their algorithms. ■ Digital divide: ensuring equitable access to ACT systems and benefits for all drivers, regardless of income, location, or technical literacy.
Governments	■ Better traffic management: reduced congestion, improved traffic flow, optimized use of road infrastructure, dynamic traffic signal control, real-time incident detection and response, and proactive congestion mitigation. ■ Increased road safety: reduced accidents, fatalities, and injuries, safer roads for all users, enforcement of traffic laws, identification and intervention in high-risk driving behaviors, and accident prevention through driver warnings and alerts. ■ Reduced environmental impact: lower greenhouse gas emissions, improved air quality, reduced noise pollution, promotion of fuel-efficient and eco-friendly driving practices, and congestion reduction to minimize idling and fuel consumption. ■ New mobility services: facilitation of new mobility services (e.g., self-driving cars, car-sharing, and micro-mobility options) to reduce reliance on private vehicles and expand transportation choices. ■ Economic benefits: job creation in the ACT industry, increased productivity due to reduced travel times, improved supply chain efficiency, and stimulation of economic activity in urban and rural areas. ■ Data-driven decision-making: collection and analysis of real-time traffic data to inform transportation planning, infrastructure investment, and policy development.	■ Interoperability: lack of global standards for ACT, making it difficult for systems from different manufacturers to communicate with each other, hindering widespread adoption and data sharing. ■ Cost: significant investment required to upgrade existing infrastructure, deploy ACT systems, and establish data centers and communication networks. ■ Liability and legal issues: establishing clear legal frameworks for the use of ACT systems, addressing issues of liability in accidents involving AVs and defining data ownership and privacy rights. ■ Public acceptance and trust: addressing public concerns about privacy, safety, and job displacement, building trust in ACT systems and ensuring their responsible and ethical implementation. ■ Cybersecurity threats: protecting ACT infrastructure and data from cyberattacks, ensuring the integrity and reliability of communication networks, and developing robust cybersecurity protocols. ■ Ethical considerations: addressing ethical concerns related to the use of AI in ACT systems, ensuring fairness, transparency, and non-discrimination in decision-making algorithms.

Table 1. *Cont.*

Group	Benefits	Challenges
Cities	■ Improved quality of life: cleaner air, less noise, safer streets, reduced traffic congestion, improved accessibility for all residents, and enhanced livability in urban areas. ■ Economic development: attraction of investment, creation of jobs in the ACT sector, growth of businesses that rely on efficient transportation, and stimulation of local economies. ■ Increased sustainability: reduced reliance on private vehicles, promotion of sustainable transportation modes (e.g., public transit, cycling, and walking) and support for green initiatives to reduce carbon emissions. ■ Smarter and more efficient cities: use of data and information to optimize traffic flow, manage parking, reduce energy consumption, improve waste collection, and enhance overall city operations. ■ Enhanced citizen engagement: provision of real-time traffic and transportation information to citizens, enabling informed travel choices, reducing commuter stress, and improving overall quality of life. ■ Transformation of urban spaces: reallocation of road space for pedestrian-friendly areas, parks, and green spaces, creating more livable and vibrant urban communities.	■ Infrastructure investment: significant investment required to upgrade existing infrastructure (e.g., roads, bridges, and parking facilities) to accommodate ACT systems and support new mobility services. ■ Data management and privacy: establishing robust data governance frameworks to protect citizen privacy, ensuring secure data storage and access, and complying with data protection regulations. ■ Digital divide: ensuring equitable access to ACT systems and benefits for all residents, regardless of income, location, or digital literacy, to prevent widening socioeconomic disparities. ■ Community engagement and acceptance: actively involving the public in the planning, implementation, and operation of ACT, addressing their concerns, and building trust to ensure successful adoption and utilization. ■ Equity and social justice: ensuring that ACT systems are designed and implemented in a fair and equitable manner, avoiding discrimination and protecting the rights of all users, including marginalized or disadvantaged groups. ■ Ethical considerations: addressing ethical concerns related to the use of AI in ACT systems, ensuring fairness, transparency, and non-discrimination in decision-making algorithms. ■ Future-proofing: ensuring that ACT systems are scalable, adaptable, and resilient to future technological and societal changes, to maintain their relevance and usefulness in the long term.

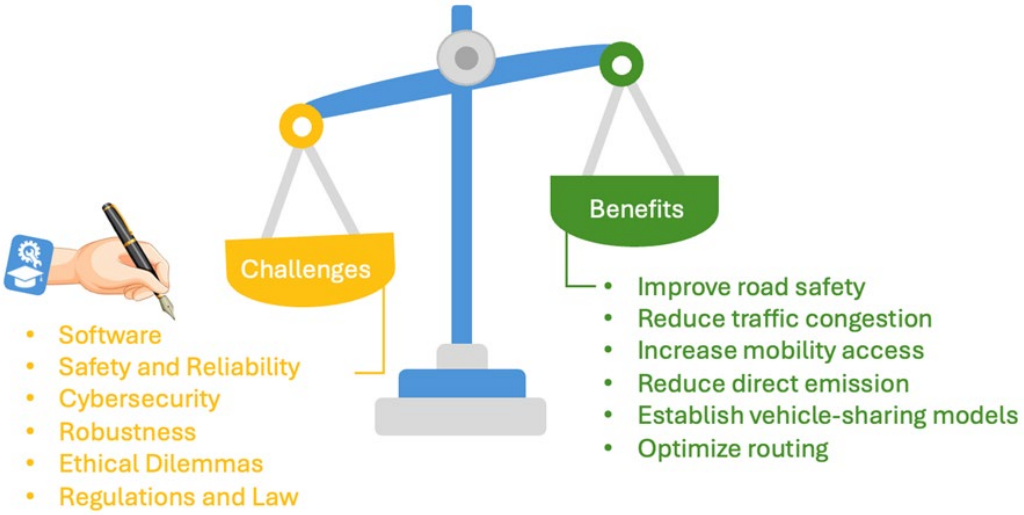

Figure 1. The tradeoff of benefits and challenges in the development and deployment of AVs. This article focuses on the primary challenges that currently hinder the widespread deployment of AVs. The hand with the pen in the image symbolizes that stakeholders from all pertinent fields are working on overcoming these challenges.

2.2. Adversarial Attacks in Traffic Sign Recognition

The traffic sign recognition task is highly vulnerable to adversarial attacks because it relies solely on camera input. Unlike other perception tasks, it cannot be cross-verified or corrected using additional sensors like LiDAR, making it more susceptible to errors. Traffic sign recognition is a multi-class classification task that follows a two-stage pipeline: traffic sign detection and traffic sign classification [2]. Adversarial attacks can be carried out in either a white-box or black-box manner. A white-box attack occurs when the attacker has full access to the model's architecture, weights, input, and output. Well-known white-box adversarial attack techniques include FGSM, PGD, Jacobian-based Saliency Map Attack (JSMA), and the Carlini and Wagner (C&W) attack. In a black-box attack, the attacker attempts to introduce perturbations—either digitally or within the physical environment—without direct access to the model. One well-known attack is the adversarial patch, which does not modify all pixels of the input image but instead alters a specific region within the image. Figure 2 illustrates how adversarial perturbations can visually alter traffic signs in a way that deceives classification models, even though the changes may be imperceptible to human drivers.

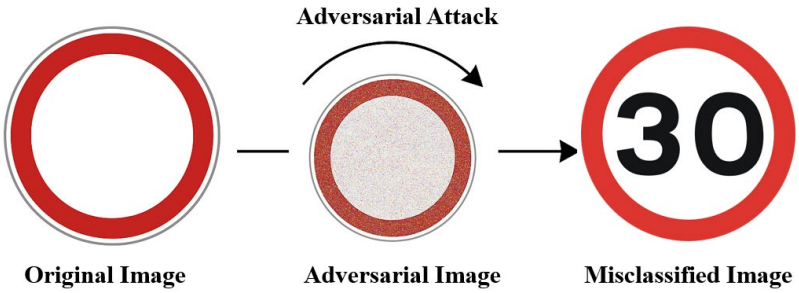

Figure 2. Adversarial attack on traffic sign classification task.

In the current literature, researchers have proposed various defense mechanisms against adversarial attacks. These approaches generally fall into two categories: attack detection and recovery of adversarial examples. Attack detection methods aim to identify whether an input has been adversely perturbed, while recovery-based defenses attempt to reconstruct or denoise the adversarial input to restore the original sample. This study focuses on developing a defense mechanism against state-of-the-art FGSM and PGD adversarial attacks in the context of traffic sign classification. The proposed defense leverages a generative autoencoder model, which can be utilized in both detection and recovery paradigms.

2.3. Limitations of Existing Defense Mechanisms

Although various defense mechanisms have been proposed to counter adversarial attacks, most fall into two categories: detection of adversarial inputs and recovery (or denoising) of perturbed samples. However, these approaches often suffer from key limitations. Many are not effective against low-perturbation attacks, which are more difficult to detect but equally dangerous. Others treat detection and recovery as separate tasks, lacking an integrated framework. Moreover, few studies focus specifically on traffic sign recognition in AVs, despite its critical role in road safety. These gaps motivate the development of a dual-purpose autoencoder-based defense system, as proposed in this study [22].

3. Materials and Methods

In this section, we first discuss the dataset, followed by an explanation of the types of adversarial attacks relevant to our research. Finally, we present the details of our proposed method.

3.1. Data Collection

The dataset used in this research is sourced from the German Traffic Sign Recognition Benchmark (GTSRB) repository [23]. It was originally created for a single-image classification challenge held at the International Joint Conference on Neural Networks (IJCNN) in 2011 [24]. The raw data were constructed using approximately 10 h of video footage recorded during daytime driving on various road types across Germany. Data collection was conducted in March, October, and November of 2010 to capture a variety of seasonal and lighting conditions. A Prosilica GC 1380CH camera (Stadtroda, Germany) with automatic exposure control was used to record the videos at a frame rate of 25 frames per second. The original video frames, from which the traffic sign images were extracted, had a resolution of 1360 × 1024 pixels. Through postprocessing of the recorded driving sequences, researchers created the final dataset with approximately 50,000 images, categorized into 43 distinct traffic sign classes. These classes represent a broad range of sign types, shapes, and visual conditions. The dataset is divided into a training set of about 39,000 images and a test set of approximately 12,000 images. Image resolutions vary from 15 × 15 to 250 × 250 pixels; therefore, preprocessing techniques were applied to standardize the input size (50 × 50) for consistent model training. Additionally, augmentation techniques were applied, including color jittering, random equalization, random horizontal flip, random vertical flip, Gaussian blur, and random rotation.

3.2. FGSM and PGD Adversarial Attacks

In this research, our focus is on two common types of white-box adversarial attacks, FGSM and PGD. Figure 3 illustrates how different values of ε visually affect traffic signs under FGSM attacks, highlighting the increasing impact of the disturbance on sign legibility. Similarly, Figure 4 shows the effects of PGD attacks, where a progressive degradation of the image is observed as the value of ε increases, demonstrating the greater capacity of this method to induce classification errors.

Figure 3. FGSM adversarial attacks on the traffic sign classification task for different ε values.

Adversarial attacks such as FGSM and PGD exploit gradients to mislead neural networks (Table 2). FGSM generates adversarial examples by computing the gradient of the loss with respect to the input and perturbing the input in the direction that maximizes the loss, causing misclassification with a single step. PGD, on the other hand, is a stronger, iterative version of FGSM. It repeatedly applies small perturbations to maximize the loss while ensuring that the total noise remains within a defined constraint, typically an L∞ norm. If the perturbation exceeds this constraint, it is projected back onto the allowed

range. While PGD is more effective at fooling models, it is also more computationally intensive [25].

Figure 4. PGD adversarial attacks on the traffic sign classification task for different ε values.

Table 2. FGSM and PGD adversarial attacks [25].

FGSM	PGD
$x' = x + \varepsilon \cdot sign(\nabla_x J(\Theta, x, y))$	$x'_{t+1} = \Pi(x'_t + \varepsilon \cdot sign(\nabla_x J(\Theta, x'_t, y)))$
x′—adversarial sample	Π—projection in the specified norm
x—original sample	x'_t—adversarial image generated at iteration t
ε—parameter controls the amount of noise added to the original sample	ε—parameter controls the amount of noise added to the original sample
Θ—model parameters (weights and biases)	Θ—model parameters (weights and biases)
$\nabla_x J(\Theta, x, y)$—gradient	$\nabla_x J(\Theta, x', y)$—gradient

3.3. Adversarial Attacks Defense Approach with Autoencoder

Autoencoders represent a class of self-supervised neural networks that learn structure within data through a two-phase architecture: an encoder function f that maps input data $x \in R^d$ to a lower-dimensional latent representation $z \in R^p$ (where $p < d$), and a decoder function g that reconstructs the original input space from this compressed representation [26]. Popular types of autoencoders include denoising autoencoders that clean up noisy data, convolutional autoencoders that work especially well with images and variational autoencoders that learn to generate new data similar to what they have seen before.

In this article, we proposed two different approaches to applying autoencoders as a defense mechanism against adversarial attacks in the traffic sign classification task. The first approach involves detecting adversarial attacks by training an autoencoder to reconstruct only clean (non-adversarial) data. In this case, when the autoencoder encounters an adversarial image, the reconstruction error will be significantly higher, which can be used to flag such inputs (Figure 5A). The second approach aims to remove perturbations and recover the original image, acting as a preprocessing defense. These types of autoencoders are trained with adversarial samples as the input and clean images as the output (Figure 5B). Additionally, for adversarial attack detection, we implement a feature autoencoder. Unlike the previous methods, this autoencoder uses feature maps from the last convolutional layer of the traffic sign classifier, rather than clean images, for training [27].

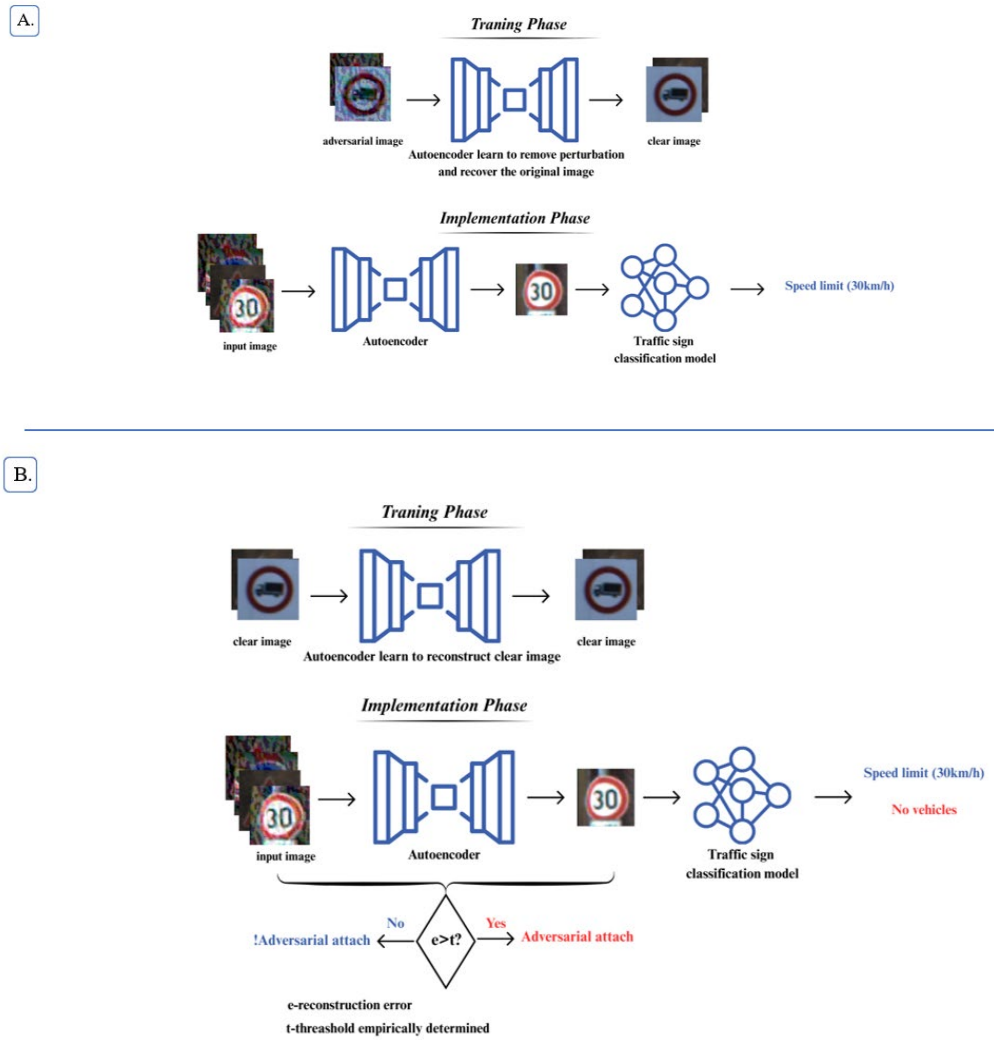

Figure 5. Autoencoder-based defense approaches against adversarial attacks in traffic sign classification. (**A**) Detection of adverse attacks by training a self-chire to rebuild only clean (non-adverse) data. (**B**) Elimination of disturbances and original image recovery, acting as a preprocessing defense.

4. Results and Discussion

In this study, we demonstrate how autoencoders can be employed as a defense mechanism against adversarial attacks in the context of traffic sign classification. The first step involved designing a convolutional neural network (CNN) classifier, which was trained on the GTSRB dataset, divided into training and test subsets. The model achieved a training accuracy of 99.8% and a test accuracy of 98.8%. As supported by existing literature, CNN-based classifiers are vulnerable to adversarial attacks. To investigate this vulnerability and quantify its impact on classification performance, we implemented adversarial example generators using both the FGSM and PGD techniques. When applying FGSM attacks to the CNN classifier, performance dropped significantly—from 98.8% to 25.5%—as the perturbation level (ε) increased across values {0, 0.001, 0.006, 0.02, 0.03, 0.04, 0.05, 0.1}. By introducing an autoencoder-based defense mechanism, the negative impact of FGSM attacks was mitigated. The most significant improvement was observed at $\varepsilon = 0.04$, where the defense increased classification accuracy by approximately 10% (Figure 6). However, it is not an adequate improvement for systems such as AVs, which operate in environments where mistakes are not forgiven.

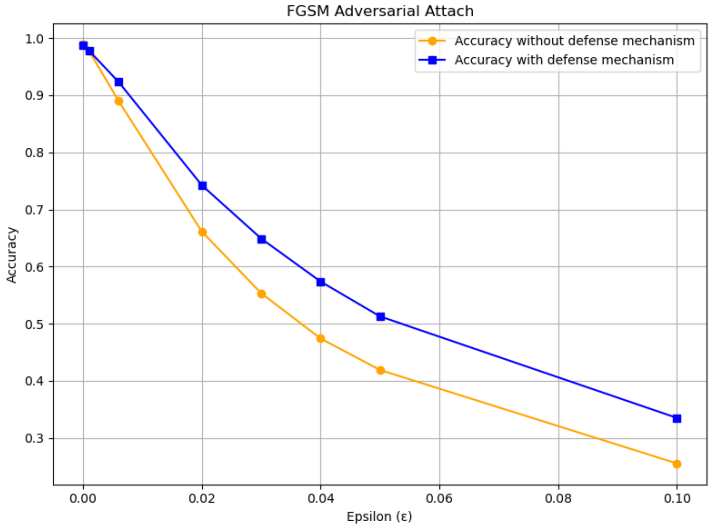

Figure 6. Accuracy of CNN classifier after FGSM adversarial attacks.

The same case is with a PGD attack, which on the CNN classifier is illustrated in Figure 7. Compared to the FGSM attack, PGD causes a more substantial degradation in classification performance as the perturbation level (ε) increases. The proposed defense mechanism improves classifier accuracy to varying degrees depending on ε, with the most notable improvement observed at ε = 0.03, where accuracy increases by approximately 30%. Although the autoencoder-based defense approach enhances classifier robustness under adversarial attacks, it does not provide complete protection. To address this limitation, we propose a second autoencoder-based approach that functions as an adversarial attack detector. Combined with a large language model (LLM), this system can serve as an intelligent defense framework capable of alerting drivers to potential attacks in real time. However, the performance indicates that detection does not work effectively for low values of perturbation levels. The best performance, among ε values treated in this research, is achieved for ε = 0.1, with a precision of 94.9%, a recall of 93.86%, and an F1 score of 94.40%, as shown in Figure 8. While higher ε values were not explored in this study, ε = 0.1 was selected as an upper bound based on prior literature indicating that perturbations beyond this threshold often result in visually unrealistic or easily detectable adversarial examples, which fall outside the scope of practical threat models considered here [28].

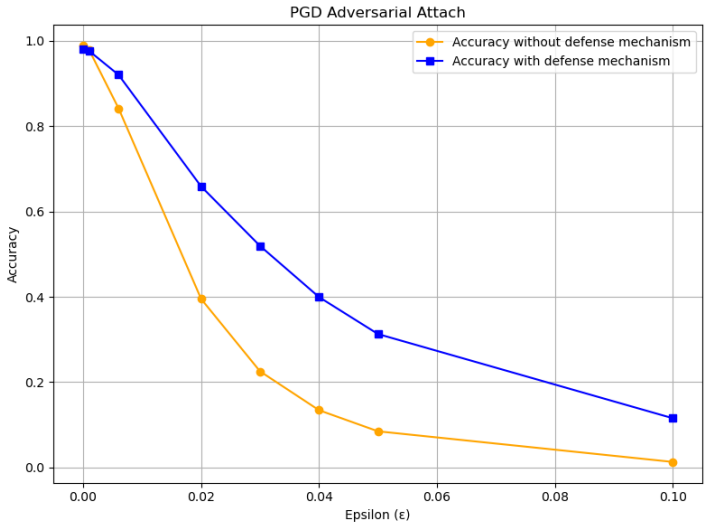

Figure 7. Accuracy of CNN classifier after PGD adversarial attacks.

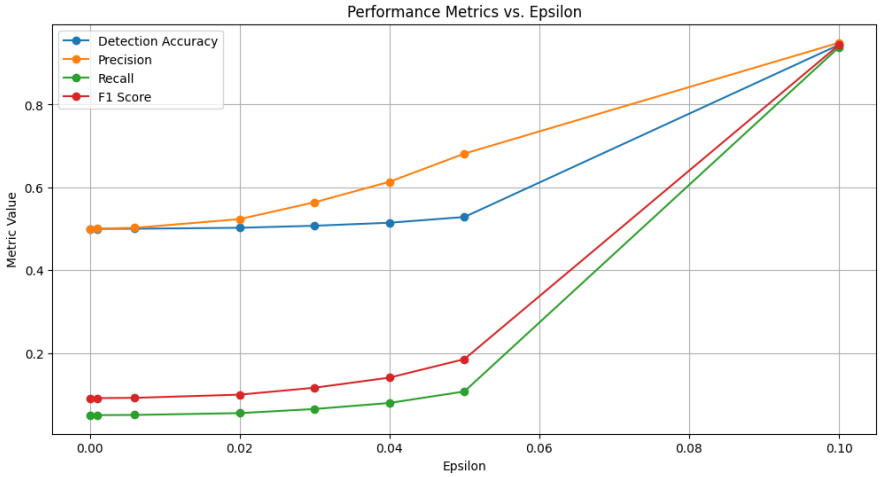

Figure 8. Adversarial attack detection across different epsilon values.

Based on the existing literature, we conclude that traditional methods for detecting adversarial attacks are ineffective at low perturbation levels (ε). For this reason, we developed a feature autoencoder that reconstructs feature maps from clean images. Images with minor perturbations closely resemble clean images, which explains why conventional convolutional autoencoders fail to detect these attacks. As shown in Table 3, there are noticeable differences between the feature maps of clean and adversarial images, even at low perturbation levels. The performance of the proposed feature-based autoencoder is illustrated in Figure 9. The results demonstrate that its detection accuracy is significantly superior to that of a traditional autoencoder. Additionally, the feature autoencoder exhibits effective scaling of detection performance with increasing adversarial strength (ε), achieving a detection rate improvement from 72% to 100%. The primary limitation of this approach is a relatively high false positive rate (FPR). However, in the context of AVs, the detection system is designed to prioritize maximum adversarial recall, even at the expense of increased FPR. A false negative—where an adversarial input goes undetected—could result in unsafe or illegal vehicle behavior, such as failing to stop at a manipulated traffic sign. In contrast, a false positive typically triggers a conservative fallback action, such as human intervention, rerouting, or system revalidation, which could be further enhanced through integration with LLMs. Therefore, an elevated FPR of approximately 70% is considered an acceptable tradeoff to ensure robust adversarial detection under strong attack scenarios. This design choice aligns with fail-safe principles in safety-critical systems, where maintaining safety takes precedence over performance metrics. Future work will explore strategies for reducing the FPR without compromising detection capability, thereby improving the applicability of autoencoders as a defense mechanism against adversarial attacks in traffic sign classification tasks.

Table 3. Comparison of feature maps between clean and adversarial samples.

Perturbation Level (ε)	Percentage (%) Average Difference bn3 Layer
0	0%
0.001	5.35%
0.006	29.74%
0.02	80.64%
0.03	105.58%
0.04	124.64%
0.05	138.47%
0.1	175.74%

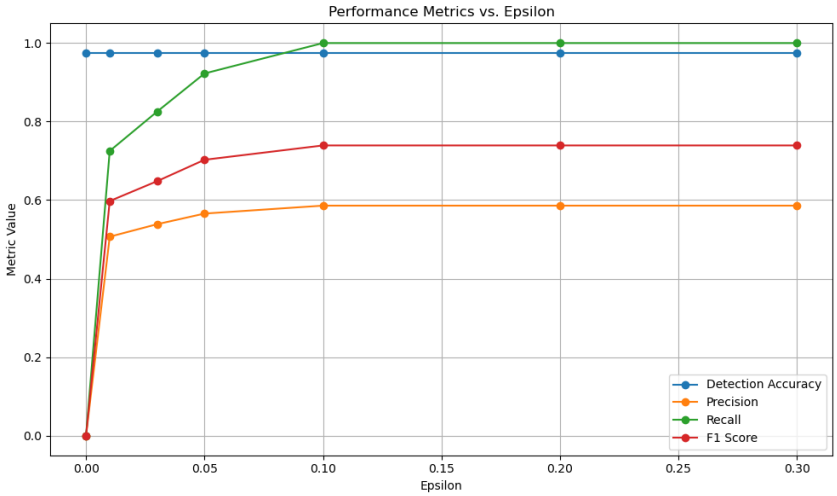

Figure 9. Performance of the proposed feature-based autoencoder across different epsilon values.

5. Conclusions

AVs can be characterized as highly complex technological systems. This complexity emerges from the dynamic environments in which these systems operate. Functioning effectively in such complex environments requires reliable and robust perception capabilities. Traffic sign recognition and classification represent some of the most critical perception tasks, as traffic signs convey regulatory information, warnings, and guidance that dictate how road users should navigate safely through particular sections of public roadways. The manipulation of this information can lead to serious consequences. Over time, various attack mechanisms have been developed to decrease the performance of perception systems. In this article, we focus on proposing autoencoder-based defense approaches for two common types of adversarial attacks: FGSM and PGD. Our results demonstrate that these defense approaches can mitigate adversarial attacks by improving classifier performance and functioning as detectors for adversarial samples. Specifically, the proposed defense recovered up to 30% of classification accuracy under PGD attacks at a perturbation level of $\varepsilon = 0.03$ and achieved a 10% improvement under FGSM at $\varepsilon = 0.04$. In detection mode, the autoencoder reached an F1 score of 94.4% at $\varepsilon = 0.1$. Furthermore, a feature-based autoencoder achieved 100% detection accuracy at $\varepsilon = 0.1$, albeit with a 70% false positive rate—an acceptable tradeoff in safety-critical AV systems.

The literature on adversarial attacks and corresponding defenses is extensive, but establishing practical and efficient protective measures to reinforce the adversarial robustness of AVs remains an unresolved challenge. Although convolutional autoencoders offer some protection against adversarial attacks, they should not be considered a comprehensive solution to this complex challenge. Future research should consider combining autoencoder approaches with visual-language models to enhance defensive capabilities.

Author Contributions: Conceptualization, I.M., J.J., M.D. and T.d.J.M.S.; methodology, I.M. and M.J.; validation, I.M. and J.J.; formal analysis, I.M. and J.J.; investigation, I.M., J.J., M.J. and T.d.J.M.S.; resources, I.M., M.D. and T.d.J.M.S.; data curation, I.M. and M.J.; writing—original draft preparation, I.M., M.J. and T.d.J.M.S.; writing—review and editing, M.D. and T.d.J.M.S.; visualization, I.M., M.D. and T.d.J.M.S.; supervision, M.D. and T.d.J.M.S.; project administration, J.J. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset available on request from the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Liu, S.; Li, L.; Tang, J.; Wu, S.; Gaudiot, J.L. *Creating Autonomous Vehicle Systems*; Springer: Berlin/Heidelberg, Germany, 2017.
2. 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. *Artif. Intell. Rev.* **2024**, *58*, 28. [CrossRef]
3. Wiseman, Y. Conjoint Vehicle License Plate identification System. *Open Transp. J.* **2020**, *14*, 164–173. [CrossRef]
4. Pavlitska, S.; Lambing, N.; Zöllner, J.M. Adversarial Attacks on Traffic Sign Recognition: A Survey. *arXiv* **2023**, arXiv:2307.08278.
5. *ISO 21448:2022*; Road Vehicles—Safety of the Intended Functionality. Technical Report; International Organization for Standardization: Geneva, Switzerland, 2022.
6. Jochem, T.; Pomerleau, D.; Kumar, B.; Armstrong, J. PANS: A Portable Navigation Platform. In Proceedings of the IEEE Symposium on Intelligent Vehicle, Detroit, MI, USA, 25–26 September 1995. Available online: https://www.cs.cmu.edu/~tjochem/nhaa/navlab5_details.html (accessed on 5 June 2025).
7. Hamon, R.; Junklewitz, H.; Sanchez Martin, J.I.; Fernandez Llorca, D.; Gomez Gutierrez, E.; Herrera Alcantara, A.; Kriston, A. *Artificial Intelligence in Automated Driving: An Analysis of Safety and Cybersecurity Challenges*; JRC127189; European Commission: Brussels, Belgium, 2022. Available online: https://publications.jrc.ec.europa.eu/repository/handle/JRC127189 (accessed on 5 June 2025).
8. Serban, A.; Poll, E.; Visser, J. A Standard Driven Software Architecture for Fully Autonomous Vehicles. *JASEN* **2020**, *1*, 20. [CrossRef]
9. Katiyar, N.; Shukla, A.; Chawla, N.; Singh, R.; Singh, S.K.; Husain, M.F. AI in Autonomous Vehicles: Opportunities, Challenges, and Regulatory Implications. *Educ. Adm. Theory Pract.* **2024**, *30*, 6255–6264. [CrossRef]
10. Garikapati, D.; Shetiya, S.S. Autonomous Vehicles: Evolution of Artificial Intelligence and the Current Industry Landscape. *BDCC* **2024**, *8*, 42. [CrossRef]
11. Witten, I.H.; Frank, E.; Hall, M.A.; Pal, C.J. Deep learning. In *Data Mining*; Morgan Kaufmann: Burlington, MA, USA, 2017; pp. 417–466. [CrossRef]
12. Legal Team. *The Safest Forms of Transportation*; Shaw Cowart Attorneys at Law LLP: Austin, TX, USA, 2025. Available online: https://www.shawcowart.com/what-are-considered-the-safest-modes-of-transportation/#:~:text=Car%20travel%20is%20risky%20and,more%20dangerous%20than%20passenger%20boating (accessed on 5 June 2025).
13. World Health Organization. *Road Traffic Injuries*; WHO: Geneva, Switzerland, 2023. Available online: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries#:~:text=Approximately%201.19%20million%20people%20die,adults%20aged%205%E2%80%9329%20years (accessed on 5 June 2025).
14. Boyce Holleman & Associates. What Percentage of Car Crashes Are Caused by Human Error? Available online: https://boyceholleman.com/car-crashes-are-caused-by-human-error/#:~:text=Worldwide,%20studies%20consistently%20show%20that,staggering%2094%25%20of%20all%20crashes.&text=In%20Mississippi,%20the%20numbers%20tell%20a%20similar%20story (accessed on 5 June 2025).
15. Smith, B.W. *Human Error as a Cause of Vehicle Crashes*; Stanford Center for Internet and Society: Stanford, CA, USA, 2013. Available online: https://cyberlaw.stanford.edu/blog/2013/12/human-error-cause-vehicle-crashes/ (accessed on 5 June 2025).
16. Rose, A.; Miracle, V.; Kopp, J. A woman was found trapped under a driverless car. It wasn't the first car to hit her. *Cable News Network*. 3 October 2023. Available online: https://edition.cnn.com/2023/10/03/tech/driverless-car-pedestrian-injury/index.html (accessed on 5 June 2025).
17. Dede, G.; Naydenov, R.; Malatras, A.; Hamon, R.; Junklewitz, H.; Sanchez, I. *Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving*; ENISA and JRC, ENISA Reports; The Publications Office of the European Union: Luxembourg, 2021. Available online: https://www.enisa.europa.eu/sites/default/files/publications/ENISA-JRC%20Report%20-%20Cybersecurity%20Challenges%20in%20the%20Uptake%20of%20Artificial%20Intelligence%20in%20Autonomous%20Driving.pdf (accessed on 5 June 2025).
18. Mateo Sanguino, T.J.; Lozano Domínguez, J.M.; De Carvalho Baptista, P. Cybersecurity certification and auditing of automotive industry. *Adv. Transp. Policy Plan.* **2020**, *5*, 95–124. [CrossRef]
19. Bonnefo, J.F. *Ethics of Connected and Automated Vehicles Recommendations on Road Safety, Privacy, Fairness, Explainability and Responsibility*; European Commission: Brussels, Belgium, 2020.
20. Dentons. Global Guide to Autonomous Vehicles 2024. Available online: https://www.dentons.com/en/insights/guides-reports-and-whitepapers/2024/may/29/global-guide-to-autonomous-vehicles-2024 (accessed on 5 June 2025).
21. UNECE. *Framework Document for Automated/Autonomous Vehicles (UPDATED)*; United Nations Economic Commission for Europe: Geneva, Switzerland, 2022; Available online: https://unece.org/info/publications/pub/365097 (accessed on 5 June 2025).
22. Wang, J.; Shi, L.; Zhao, Y.; Zhang, H.; Szczerbicki, E. Adversarial attack algorithm for traffic sign recognition. *Multimed Tools Appl.* **2022**. [CrossRef]
23. Mykola. *GTSRB—German Traffic Sign Recognition Benchmark*; Kaggle Inc.: San Francisco, CA, USA, 2011. Available online: https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign/data (accessed on 5 June 2025).

24. Stallkamp, J.; Schlipsing, M.; Salmen, J.; Igel, C. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In Proceedings of the 2011 International Joint Conference on Neural Networks, San Jose, CA, USA, 31 July–5 August 2011; pp. 1453–1460. [CrossRef]
25. Kansal, K.; Krishna, P.S.; Jain, P.B.; Honnavalli, S.R.P.; Eswaran, S. Defending against adversarial attacks on COVID-19 classifier: A denoiser-based approach. *Heliyon* **2022**, *8*, e11209. [CrossRef] [PubMed]
26. Liu, X.; Zhang, F.; Hou, Z.; Mian, L.; Wang, Z.; Zhang, J.; Tang, J. Self-supervised Learning: Generative or Contrastive. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 857–876. [CrossRef]
27. Ye, H.; Liu, X. Feature autoencoder for detecting adversarial examples. *Int. J. Intell. Syst* **2022**, *37*, 7459–7477. [CrossRef]
28. Xu, H.; Ma, Y.; Liu, H.C. Adversarial Attacks and Defenses in Images, Graphs and Text: A Review. *Int. J. Autom. Comput.* **2020**, *17*, 151–178. [CrossRef]