JOURNAL OF BASIC SCIENCE AND ENGINEERING

A Technique for Targeted and Untargeted Adversarial Attack Goal Detection

Prashanth K Y [1], Rohitha Ujjini Matad [2]

[1] Department of Electronics & Communication Engineering, RYM Engineering College, Bellary, Karnataka,India, affiliated to Visvesvaraya Technological University, Belagavi

[2] Department of Electronics & Communication Engineering, Proudhadevaraya Institute of Technology, Hospet, Karnataka, India, affiliated to Visvesvaraya Technological University, Belagavi

DOI: https://doie.org/10.0608/Jbse.2024529242

Abstract— Autonomous vehicles or self-driving cars and semi-autonomous cars are equipped with advanced technology and driver assistance features enabling safe and easier driving experience that abides by the law, rules and regulations and alleviate traffic congestion. However, since these features leverage wireless networks, sensors, and cameras, it also opens windows to threats, vulnerabilities, and hacking attacks. The introduction of machine learning and deep learning in realizing fully autonomous vehicles has provided a massive pace for achieving these autonomous systems. However, the addition of these deep neural networks from sensing, perception, localization, and planning to decision taking stages of autonomous driving is also opening a floodgate for the attackers as they introduce numerous unknown attack surfaces. Some of the vulnerabilities of the deep learning neural networks are getting noticed in the last half-decade, which posing serious questions on the implications of these systems in vehicles, as any exploitation of these results in the fatal on the road. Adversarial attacks are one among the prominent attacks on deep learning-based vehicular autonomous systems, which will fool the neural networks during object detection, object classification, and segmentations. The addition of invisible crafted noise into the images makes them adversarial and have got a capability to deceive the fully trained neural networks during evasion in the driving environment. In retarding the influence of these adversarial images, there are many systems available in detecting the adversarial images. However, they all are incapable of detecting the true goals of the adversary during the adversarial attacks. Hence, an approach is proposed here for detecting the adversarial goal of the attackers when launching the adversarial attacks. The proposed system detects both the untargeted and targeted adversarial attack goals using a trained machine learning model, which predicts the probabilities of all the traffic signs for each traffic pole. The adversarial goal detection machine learning model is trained on the collected large non-adversarial traffic sign probability dataset. Once the traffic pole is detected through navigation system, the difference between the predicted probability of traffic signs from machine learning model and predicted probability of traffic signs from vehicular traffic sign recognition system is found. This positive deference indicates the untargeted attack goal, as it causes the probability of untargeted traffic sign to decrease. Similarly, a negative deference indicates the targeted attack goal, as it causes the probability of targeted traffic sign to increase. Hence, the proposed system detects the adversarial goals precisely, which helps in designing the efficient adversarial defense mechanism. Further to which, the nature of adversarial images, either static or dynamic poses a serious question to the defender to detect an accurate adversarial attack security defense system. Hence, there is a need to detect these static and dynamic images. So, this proposed system also proposes a method to detect the static and dynamic images and detects the adversarial attacker's goal for both static and dynamic adversarial images. The performance of the proposed system is evaluated for the traffic sign recognition system, which uses trained deep learning models. The proposed system detects the adversarial goals with satisfactory accuracy.

Keywords: Adversarial attacks, Targeted Attacks, Untargeted Attacks, Adversarial Goal, Traffic Sign Recognition, Traffic Sign

1. AUTONOMOUS VEHICLE SECURITY

The threat of cyberattacks is burgeoning at an alarming rate, significantly impacting companies across all industries. As the automotive industry makes huge strides towards a fully autonomous future, their cybersecurity needs become even more complex and critical. The very innovation that aims to facilitate the new mobility ecosystem entails first-order cybersecurity challenges with a combination of digital and physical threats. The apparent risks associated with self-driving vehicles, such as hacked autonomous car crashing and abrupt braking, are just the tip of the iceberg. Driverless cars could hold massive amounts of data from oodles of users, making them a lucrative target for cybercriminals. In addition, the vehicle can be compromised, leading to safety issues for drivers and other road users.

Autonomous vehicles or self-driving cars and semi-autonomous cars are equipped with advanced technology and driver assistance features enabling safe and easier driving experience that abides by the law, rules and regulations and alleviate traffic congestion. However, since these features leverage wireless network, sensors, and cameras, it also opens windows to threats, vulnerabilities and hacking attacks. Almost every car on the road today comes with advanced semi-autonomous features such as infotainment system (provides vehicle information and entertainment), Traffic Jam Assist in BMW or Traffic Jam pilot in Audi (coordinates live traffic information with satellites and provides alternate routes using car's navigation system), adaptive cruise control (automatically intervenes brakes when needed to maintain safe driving distance with other cars while driving in cruise control mode), self-parking and lane centering steering, Drive Pilot in Mercedes Benz or Autopilot in Tesla (where the driver can keep their hands off the steering wheel and car can lock lane markings and drive itself within the lanes for up to 30 seconds), etc. The basic underlying principle for engineering and managing these semi-autonomous features in these cars is by configuring car's own computers that has car's operating system software which works behind the scenes collaborating with car's electronics, mechanics, power train, wiring, ignition, chassis, etc. Since, these car computers are code behind the scenes, it exposes all semi-autonomous features of the cars to threats and vulnerabilities. It makes these cars prone to hacking attacks. A car's infotainment system can be hacked to gain access to any unit or component inside the car such as ignition, brakes, drive-train, steering wheel, audio/video systems, parking cameras, door locks, wiper blades, etc., allowing hackers to take full control over the car and wreak havoc while the car is still in operation on the road. This can result in highway mayhems, destruction and casualties.

2. NEED OF ROBUST DEEP LEARNING IN AUTONOMOUS VEHICLES

Recently, with the prevalence of artificial intelligence (AI) and Internet of Things (IoT) technologies, autonomous driving has gained steady improvements, and is getting more and more intelligent to precisely sense environments in the real world, quickly analyze the sensor data, and autonomously make complex decisions. In the foreseeable future, AVs are widely believed to be one of the most popular AI applications in people's daily lives.

Deep learning, the most popular technique of artificial intelligence, is widely applied in autonomous vehicles to fulfill different perception tasks as well as making real-time decisions. Figure 1 demonstrates the workflow and architecture of a deep learning-based ADS. In a nutshell, raw data collected by diverse sensors and high-definition (HD) map information from the cloud are first fed into deep learning models in the perception layer to extract the ambient information of the environment, after which different designated deep/reinforcement learning models in the decision layer kicks off the real-time decisions making process.

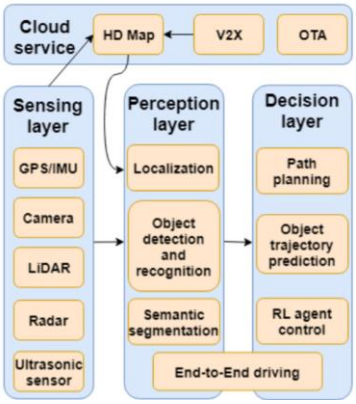


Fig. 1. Autonomous driving system architecture [1]

A deep learning-based Autonomous Driving Systems (ADS) [1] [2] is normally composed of three functional layers, including a sensing layer, a perception layer and a decision layer, as well as an additional cloud service layer as shown in Figure 1. In the sensing layer, heterogeneous sensors such as GPS, camera, LiDAR, radar, and ultrasonic sensors are used to collect real-time ambient information including the current position and spatial-temporal data (e.g. time series image frames). The perception layer, on the other hand, contains deep learning models to analyze the data collected by the sensing layer and then extract useful environmental information from the raw data for further process. The decision layer would act as a decision-making unit to output instructions concerning the change of speed and steering angle based on the extracted information from the perception layer.

Advances in Machine Learning (ML) and Deep Neural Networks (DNNs) bring tremendous potential to make autonomous vehicles a reality. In this setting, sensors such as camera, light detection and ranging sensor (LiDAR), and Infrared (IR) generate streams of real-time data. Envisioned ML applications include: predicting road conditions by interacting with other cars; recognizing risky road conditions; and assisting drivers in taking safer decisions. For this highly-critical application, safety is the major concern, but unfortunately ML algorithms are not traditionally designed and evaluated from this perspective. Tesla also deploys advanced AI models for object detection to implement Autopilot [3]. However, there exist a number of issues against the further development of deep learning-based ADSs adopting this pipeline structure. The AI systems of an autonomous vehicle are working non-stop to recognize traffic signs and road markings, to detect vehicles, estimate their speed, to plan the path ahead. Apart from unintentional threats, such as sudden malfunctions, these systems are vulnerable to intentional attacks that have the specific aim to interfere with the AI system and to disrupt safety-critical functions. First of all, sensors are vulnerable to numerous physical attacks, under which most of the sensors are no longer able to function as normal to collect data in good quality, or they may be adversely instructed to collect fake data, leading to a severe degradation of performance of all learning-based models in the following layers. Furthermore, recent research shows that deep neural networks are vulnerable to adversarial attacks [3] that are designed specifically to induce learning-based models to wrong predictions. The most common adversarial attack is by constructing the so-called adversarial examples that only have slight difference from the original inputs to baffle the neurons in the model.

Due to the necessity and importance of deep learning across all the layers of autonomous driving such as sensing layer, perception layer, and decision layer, there is a need of an robust deep learning system in the all the layers of the autonomous driving, which are free from any security concerns.

3. ADVERSARIAL ATTACKS

Whereas, deep learning performs a wide variety of Computer Vision tasks with remarkable accuracies, Szegedy et al. [4] first discovered an intriguing weakness of deep neural networks in the context of image classification. They showed that despite their high accuracies, modern deep networks are surprisingly susceptible to adversarial attacks in the form of small perturbations to images that remain (almost) imperceptible to human vision system. Such attacks can cause a neural network classifier to completely change its prediction about the image. Even worse, the attacked models report high confidence on the wrong prediction. Moreover, the same image perturbation can fool multiple network classifiers. The profound implications of these results triggered a wide interest of researchers in adversarial attacks and their defenses for deep learning in general. A slightly modified data that lead to incorrect classification as shown in Figure 2 and 3.

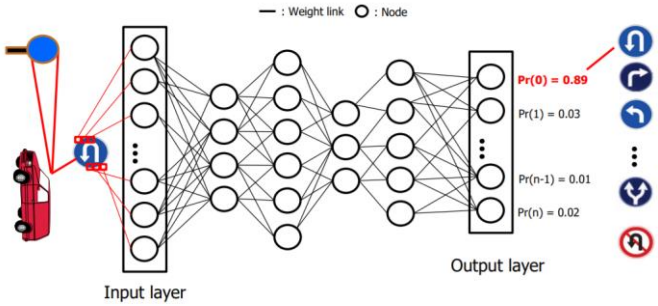


Fig. 2. Original deep learning sign classifier

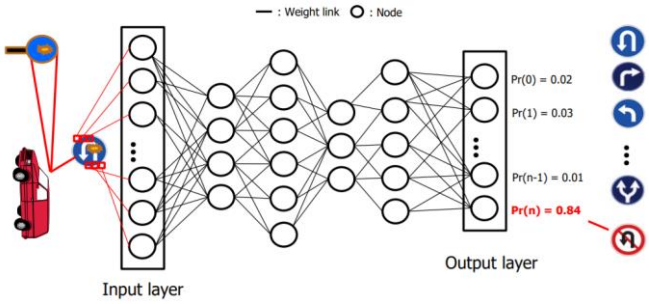


Fig. 3. Perturbed deep learning sign classifier

Adversarial machine learning is a machine learning technique that attempts to fool models by supplying deceptive input [5] [6]. The most common reason is to cause a malfunction in a machine learning model. The machine learning technique are designed to work on training and test data that are generated from the same statistical distribution. These machine learning or deep learning models when applied to real world, adversaries supply data that violates the statistical assumption. The data supplied by the adversary will exploit the specific vulnerabilities and compromise the results of the model. One could often induce a network to change the predicted class of the label, while not changing that how that image is perceived by humans. The particular element that makes these examples adversarial is how little perturbation needed to be applied to get the network to change its mind about the correct classification. Generally, a characteristic of well-trained models is that they're relatively invariant to small amounts of noise. And, when it came to random noise, this was in fact generally the case — experiments have typically confirmed that adding true "white noise" to an image typically doesn't impact the predictions of well-performing models. But when it comes to non-random noise, noise specifically engineered to "fool" the network, a surprisingly small amount of such noise, much less than is perceptible by a human eye, can meaningfully shift the network's final output.

4. EVASION ATTACK THREAT MODEL OF ADVERSARIAL ATTACKS

Adversarial examples are only one possible attack vector on ML systems. Instead of targeting a system at inference time by feeding it an adversarial example, an attacker could also try to compromise the system in the training phase. A first crucial feature for modeling the adversarial attacks is when it occurs. To that end we have two possibilities, evasion and poisoning attacks. Evasion attacks [7] are the ones in the time of inference and assume the model has already been trained. Poisoning attacks [8] in general targets the data, and the training phase of the model. Here, the adversarial attack evasion threat model is considered.

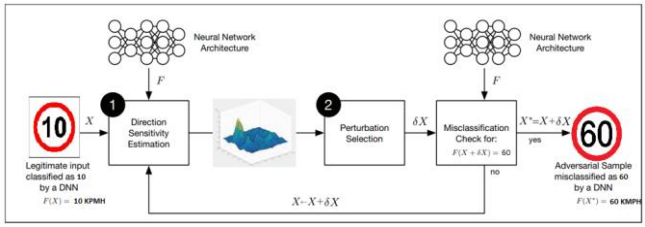


Fig. 4. Evasion Adversarial Attack Threat Model [9]

Potential attacks include having malicious content like malware identified as legitimate or controlling vehicle behavior. In the above figure 4 the deep learning model is predicting the digits of the speed limit traffic signal. A non-adversarial image is correctly classified as 10 by the digit classifier. However, we can add a small amount of calculated perturbation to generate an adversarial image. This adversarial image is now getting misclassified by the DNN classifier as 60. In order to generate the calculated perturbation, the adversary uses Direction Sensitivity Estimation. As shown in above Figure 4, the adversary evaluates the sensitivity of the change to each input feature by identifying directions in the data manifold around sample X in which the model F is most sensitive and is likely to result in class change.

Then adversary performs perturbation selection as shown in above figure 4. The adversary then exploits the knowledge of the sensitivity information to select a perturbation δX in order to obtain an adversarial perturbation which is most efficient. This process of the adversary is a cyclic process and it continues until the disturbed sample image satisfies the adversary's goal.

5. ADVERSARIAL ATTACKER'S GOAL

The attackers may have different reasons to target a specific algorithm. But mostly the attacker has either a specific goal, and needs the algorithm to output a specific output, case in which it is a targeted attack, or just wants to reduce the reliability of the algorithm by forcing a mistake. In the latter, we have an untargeted attack.

- Non-targeted Adversarial Goal: The goal of the non-targeted attack is to slightly modify source image in a way that image will be classified incorrectly by generally unknown machine learning classifier.

- Targeted Adversarial Goal: The goal of the targeted attack is to slightly modify source image in a way that image will be classified as specified target class by generally unknown machine learning classifier.

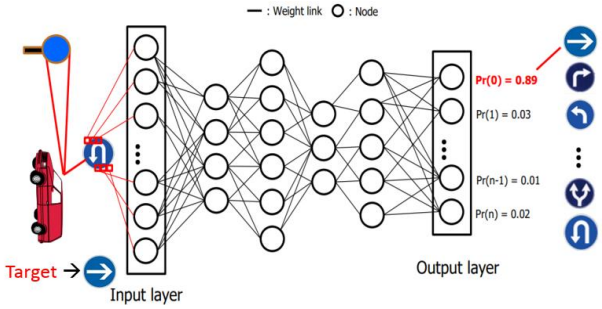
The targeted attack has a target class, Y, that it wants the target model, M, to classify the image I of class X as. Hence, the goal of the targeted attack is to make M misclassify by predicting the adversarial example, I, as the intended target class Y instead of the true class X. On the other hand, the untargeted attack does not have a target class which it wants the model to classify the image as. Instead, the goal is

simply to make the target model misclassify by predicting the adversarial example, I, as a class, other than the original class, X.

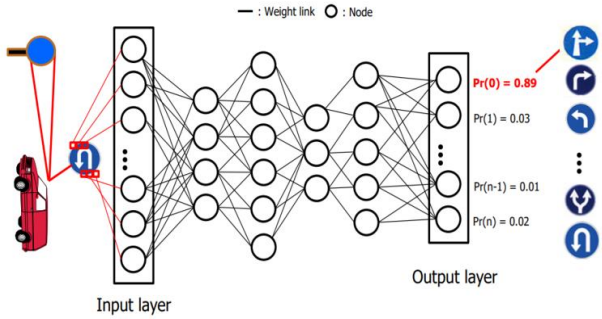
The distinction above considers the outputs of the ML algorithm that the attacker is interested in. An orthogonal distinction can be made by considering inputs the attacker is interested in: the attacker's goal may be to simply misclassify any input, but it may also be to misclassify a specific input, or an input from a specific set (for example, those inputs that should be classified as some specific class). This distinction is also called attack specificity. In most cases, adversarial examples search for perturbations specific to one input drawn from the data generation distribution.

6. NEED TO IDENTIFY ADVERSARIAL ATTACKER'S GOAL

The goal of the attacker plays a crucial role launching the attack from the perspective of launching the attack, the targeted attack, where the attackers targets a specified output class of the classifier, whereas, in the untargeted attack the attacker will not target a specified class in outputting the classifier output. In the threat model of the adversarial attacks, the knowledge of the target system is should be known in the form of an white box and black box attacks, to execute the attack to fulfill his/her attack goal in targeting specified output class of the classifier. In the below example given, the targeted scenario on a traffic sign recognition system is given [10], where a compromised traffic sign board is present as a traffic sign. The added perturbation is marked in a different color in the input sign image for the representation purpose. In Figure 5 (a) the targeted attack is done on the system for the straight sign. In the Figure 5 (b), an untargeted attack is launched on the system, so the output category is other than U turn signal.



(a)



(b)

Fig. 5. Adversarial Attacker's Goal in Traffic Sign Recognition System, (a) Targeted Attack Goal, (b) Untargeted Attack Goal.

The identification of the goal of the attacker is required to identify which one of the output class of the classifier is targeted, so that the output for the targeted output class is analyzed for providing the better defense mechanisms. The existing adversarial attack detection systems as a part of defense mechanism identify the incoming image from the camera as an adversarial image or not. They do not possess any information about the goal of the adversary in launching the attack. Hence, there is need to detect and understand the goal of the attacker who launched the specified adversarial attack on the specified compromised deep learning-based system through evasion.

7. PROPOSED TARGETED AND UNTARGETED ADVERSARIAL GOAL DETECTION SYSTEM

An approach is proposed here to detect the adversarial goal of the attacker, that is whether the launched attack is targeted or untargeted. This is shown in Figure 6. The defense mechanisms of the existing systems detect the adversarial attacks using a sophisticated statistical mechanism, which can detect the added noise in the adversarial images. Hence, the proposed adversarial goal detection system resides after the detection of the adversarial images. There are many types of detection systems available such as adversarial training, defensive distillation, noise cancellation etc.

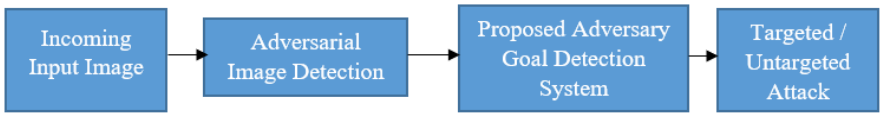


Fig. 6. Proposed Adversarial Goal Detection system in Adversarial Defense Flow

$$adv\_x = x + \epsilon * \text{sign}(\nabla_x J(\theta, x, y)) \quad \text{.....(1)}$$

Where,

- adv_x : Adversarial image.
- x : Original input image.
- y : Original input label.
- ϵ : Multiplier to ensure the perturbations are small.
- θ : Model parameters.
- J : Loss.

It involves three steps in this order:

- Calculate the loss after forward propagation,
- Calculate the gradient with respect to the pixels of the image,
- Nudge the pixels of the image ever so slightly in the direction of the calculated gradients that maximize the loss calculated above.

In the case of untargeted adversarial attack, the first step is to choose the class for which untargeted attack need to be launched. For the specified input image, the error function or the loss function of the classifier is found. Then the gradient of the cost function is found with respect to the input pixels using forward propagation. Once the gradient of the cost or error function is found then the input pixels are updated using the pixel update formula given in formula 1. The updated perturbations are generated. In the case of untargeted attacks, the error function is maximized by travelling in the same direction of the gradient for the input image. In the case of the targeted attacks, the target output class for which the

perturbations are generated by decreasing the error for the loss or error function by travelling in the opposite direction of gradient.

In the case of Untargeted adversarial goal, the Probability (Output Class) = Maximized
In the case of Targeted adversarial goal, the Probability (Output Class) = Minimized
Untargeted Attack: The target model to recognize the adversarial input image to the classifier as a class other than the original class.
Targeted Attack: The target model to recognize the adversarial input image as a particular intended class.

In the case of traffic sign recognition system in vehicles as shown in Figure 7, there are three signs are considered. The considered signs are left (L), Right (R), and Straight (S). Without the adversarial attacks, the probability predictions for all the classes of the neural network based deep learning classifier are given as below

P (S) = 0.9
P (L) = 0.05
P (R) = 0.05

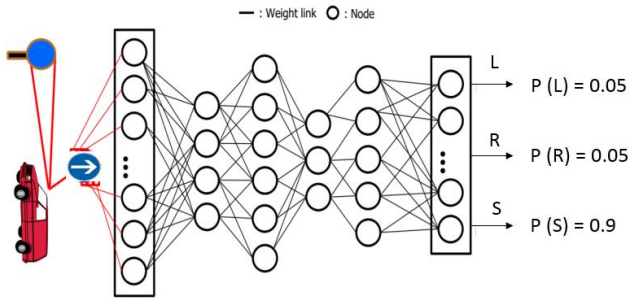
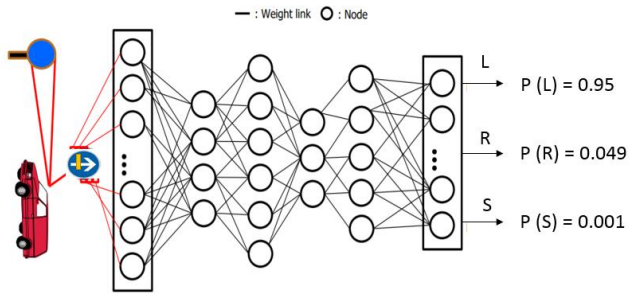


Fig. 7. Probability predictions for the traffic sign classifier

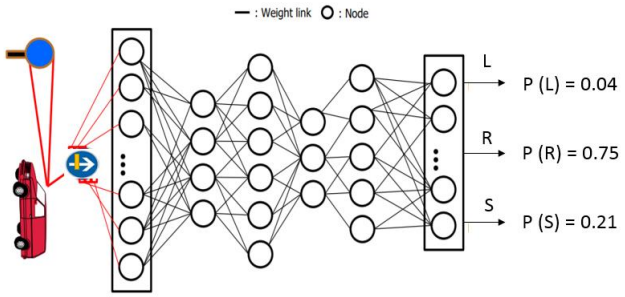
For the input traffic sign of S, the classifier predicted a maximum probability for class S in the output and predicted minimum probabilities for the remaining two classes L and R.

In case of the untargeted adversarial goal, for the perturbed input traffic sign of Sper, the P(S) is minimized. And next large probable class is considered as the output. In case of the targeted adversarial goal for a target of R, for the perturbed input traffic sign of Sper, the P(R) is maximized. And the targeted output class R, which has the maximum probability is considered as output. These are shown in Figure 8 (a) and (b) respectively.



(a)

Vol. 21, No. 1, (2024)
ISSN: 1005-0930



(b)

Fig. 8. Probability predictions for the traffic sign classifier, (a) untargeted goal, (b) targeted goal

The proposed solution for the detection of adversarial attackers goals either as targeted or untargeted utilizes the traffic infrastructure and navigation map data for the detection of adversarial goals. To illustrate this scenario, consider a traffic signal, where the attacker has launched the adversarial attack on the traffic signals by physically or remotely placing a perturbed adversarial traffic sign in place of the legitimate traffic sign in the traffic signal. This is shown in Figure 9, where a marked vehicle is approaching the traffic signal.

The traffic signs which are encountered in the traffic infrastructure can be categorized as

a) Static traffic signs
b) Dynamic traffic signs

The Static traffic signs are those, which are attached to the traffic pole and will be available for a longer period of time and are considered permanent in any geographical location. The traffic infrastructure will be monitored by the governed legal bodies for any change in the traffic infrastructure such as traffic poles, road lanes, and etc.

The Dynamic signs are the ones, which are present either for the short duration of time or they are periodically changing from one location to another on a random basis. Some of those signs, which are considered as dynamic are the construction signs, which will be available till the completion of construction, and sudden traffic deviation signs, etc.

The algorithmic steps of the proposed system are given below for the detection of targeted and untargeted attacks.

1. When the vehicle is approaching the traffic signal, the front camera of the vehicle detects the traffic signals and fed into the vulnerable deep learning-based traffic sign recognition system. Here the considered traffic signs are L, R and S. The traffic sign recognition system detects the traffic signs using a neural network-based classifier, where it predicts the probability of the detected being L, R, and S using a probability value.

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

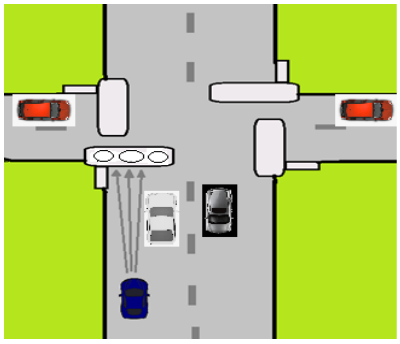
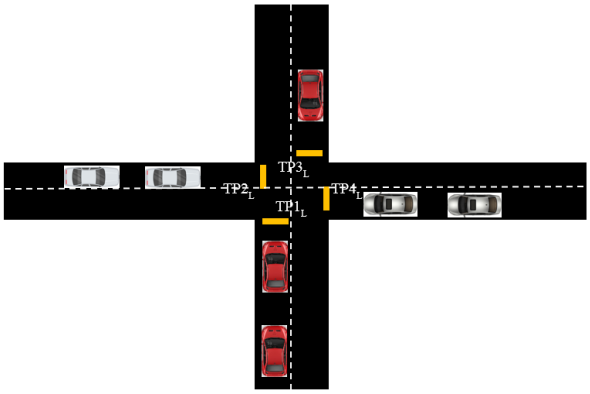
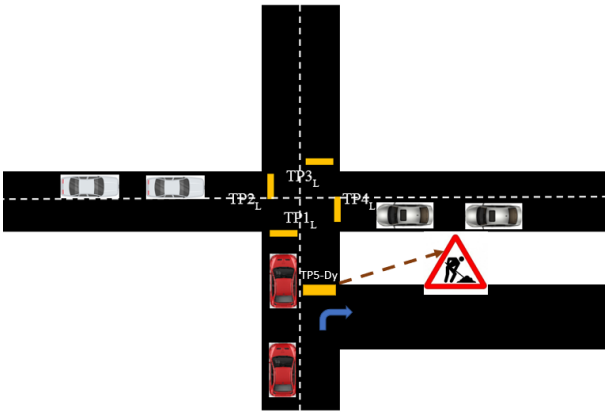


Fig. 9. Traffic signal classifier for the illustrated traffic pole

2. Consider a Traffic intersection scenario, where the adversarial attack is launched by the adversary. Before launching the adversarial attacks, numerous vehicles moving in the traffic poles such as TP1, TP2, TP3, and TP4 in case of the static traffic signs shown in Figure 10 (a) and pole TP5-Dy in case of the dynamic traffic sign shown in Figure 10 (b).



(a)



(b)

Fig. 10. Traffic Intersection scenario for static and dynamic traffic poles. (a) Traffic intersection scenario for adversarial attacks evasion threat model for a static traffic pole, (b) Traffic Diversion scenario for an adversarial attacks evasion threat model for a dynamic traffic pole

3. During the regular traffic movements, the traffic sign recognition system, which is present in all the autonomous vehicles facing the traffic poles TP1, TP2, TP3, TP4, and TP5-Dy predicts the probability of the displayed traffic sign TS for their respective traffic poles.

4. The predicted traffic sign classification probability for all the displayed signs for all the vehicles moved over a certain duration is gathered for all the traffic poles. It is assumed that traffic sign classifier accuracy is almost same across all the vehicles.

   For Example: For the Traffic Sign Recognition System in one of the traffic poles, which predicts the probabilities for all the classes of the 3-class classifier, the predicted probabilities, when sign S is displayed are,

   For vehicle 1,

$$P(S)_{\text{Vehicle-1}} = 0.9$$
$$P(L)_{\text{Vehicle-1}} = 0.05$$
$$P(R)_{\text{Vehicle-1}} = 0.05$$
$$\text{Traffic Sign Recognition Output} = \text{Max}(P(S), P(L), P(R)) = S$$

   For vehicle 2,

$$P(S)_{\text{Vehicle-2}} = 0.89$$
$$P(L)_{\text{Vehicle-2}} = 0.06$$
$$P(R)_{\text{Vehicle-2}} = 0.05$$
$$\text{Traffic Sign Recognition Output} = \text{Max}(P(S), P(L), P(R)) = S$$

   .
   .
   .
   .

   For vehicle N,

$$P(S)_{\text{Vehicle-N}} = 0.92$$
$$P(L)_{\text{Vehicle-N}} = 0.01$$
$$P(R)_{\text{Vehicle-N}} = 0.07$$
$$\text{Traffic Sign Recognition Output} = \text{Max}(P(S), P(L), P(R)) = S$$

   Similarly, the predicted probabilities, when the traffic sign L is displayed are

   For vehicle 1,

$$P(S)_{\text{Vehicle-1}} = 0.05$$
$$P(L)_{\text{Vehicle-1}} = 0.9$$
$$P(R)_{\text{Vehicle-1}} = 0.05$$
$$\text{Traffic Sign Recognition Output} = \text{Max}(P(S), P(L), P(R)) = L$$

   For vehicle 2,

$$P(S)_{\text{Vehicle-2}} = 0.06$$
$$P(L)_{\text{Vehicle-2}} = 0.89$$
$$P(R)_{\text{Vehicle-2}} = 0.05$$
$$\text{Traffic Sign Recognition Output} = \text{Max}(P(S), P(L), P(R)) = L$$

   .
   .
   .
   .

   For vehicle N,

$$P(S)_{\text{Vehicle-N}} = 0.05$$
$$P(L)_{\text{Vehicle-N}} = 0.85$$
$$P(R)_{\text{Vehicle-N}} = 0.1$$

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Traffic Sign Recognition Output = Max (P (S), P (L), P (R)) = L

Similarly, the predicted probabilities, when the traffic sign R is displayed are

For vehicle 1,

$$P(S)_{Vehicle\text{-}1} = 0.1$$
$$P(L)_{Vehicle\text{-}1} = 0.06$$
$$P(R)_{Vehicle\text{-}1} = 0.84$$

Traffic Sign Recognition Output = Max (P (S), P (L), P (R)) = R

For vehicle 2,

$$P(S)_{Vehicle\text{-}2} = 0.20$$
$$P(L)_{Vehicle\text{-}2} = 0.10$$
$$P(R)_{Vehicle\text{-}2} = 0.70$$

Traffic Sign Recognition Output = Max (P (S), P (L), P (R)) = R

.
.
.
.

For vehicle N,

$$P(S)_{Vehicle\text{-}N} = 0.11$$
$$P(L)_{Vehicle\text{-}N} = 0.20$$
$$P(R)_{Vehicle\text{-}N} = 0.69$$

Traffic Sign Recognition Output = Max (P (S), P (L), P (R)) = R

5. The gathered data is when the traffic signs are legitimate, and it also includes multiple visits for the same vehicle numerous times.

6. The gathered traffic sign recognition system data for the displayed traffic signs, indicates the most probable values associated with the traffic sign classifiers for the respective traffic poles. When the adversarial attacks are launched on these traffic poles during evasion, the predicted probabilities take a significant drift from their average probabilities values predicted by the recognition systems, when the traffic pole is free from adversarial attacks. Hence, this serve as a measure for the detection of adversarial goals.

7. Based on the gathered vehicular traffic recognition data, the data is collected by the vehicular communication system and is updated to the respective traffic poles into the navigation system.

8. The traffic poles and their associated traffic sign probability data predicted by the traffic sign recognition system of all the vehicles visited the poles are collected and used to train a machine learning model for predicting the traffic sign probabilities.

9. The proposed machine learning system is a regression model for predicting the traffic sign probabilities for each of the detected poles. The trained machine learning model is used to detect the adversarial goal either targeted or untargeted.

   The steps followed in training the machine learning model for predicting the traffic signal probabilities for each traffic pole, which contains multiple signs are given below.

a) Supervised Adversarial Goal Detection data preparation for Each Traffic Pole

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

i. Traffic sign recognition system data is collected for all the vehicles for the respective traffic poles.
ii. The data is pre-processed to extract the engineered feature points for the pattern determination.
iii. Data is grouped into three different categories of L, R and S for each traffic poles in case of a three sign signal pole.
iv. Labelling of all the probabilities is done for the respective classes L, R and S.
v. Supervised learning data is prepared.

b) Machine Learning based Modelling of Adversarial Goal Detection for Each Traffic Pole

i. Supervised traffic sign probability data is collected from all the vehicles.
ii. Collected probability and its associated classes are assigned are labels from L, R and S.
iii. Hyperparameters are chosen for the regression deep learning model for extracting and learning the patterns of the collected data.
iv. Best neural network architecture is selected for modelling the relationship between the probability value and respective traffic sign class when no adversarial attack is observed.
v. The AI model Hyperparameters are tuned recurrently until a minimum prediction error for the prediction of traffic sign probability for the displayed signs in the traffic poles.
vi. Trained AI model is prepared after fine-tuning to predict the traffic sign probabilities for the displayed signs.
vii. The trained AI model is termed as the Adversarial Goal detection model.

c) Traffic Sign Probability Evaluation using a Trained Adversarial Goal Machine Learning Model for each Traffic Pole

i. The algorithmic steps followed in the detection of traffic pole and its kind such as static or dynamic is identified is given below

1) The traffic poles latitude and longitude are updated into the navigation system during the process of data preparation from step 7.

2) The navigation system is also updated with the vehicular traffic recognition data, the data is collected by the vehicular communication system and is updated to the respective traffic pole, which are identified using their IDs and location into the navigation system.

3) If the traffic pole is recognized by the traffic sign recognition system in vehicle, then the recognized traffic pole location data and its ID are mapped to the corresponding traffic poles.

Recognized_Traffic_pole = Mapping (TP, Latitude, Longitude)

4) The mapping of the traffic pole data with the recognized traffic sign recognition system output is a search problem.

Detected_Traffic_pole = Search (TP_ID, Latitude, Longitude)

5) If the detected pole is identified in the search mechanism from the navigation data then it is considered as a static pole. Otherwise, it is considered as a dynamic traffic pole, which is added into the traffic infrastructure temporarily.

The block diagram for the detection of the kind of traffic pole either static or dynamic is shown in Figure 11.

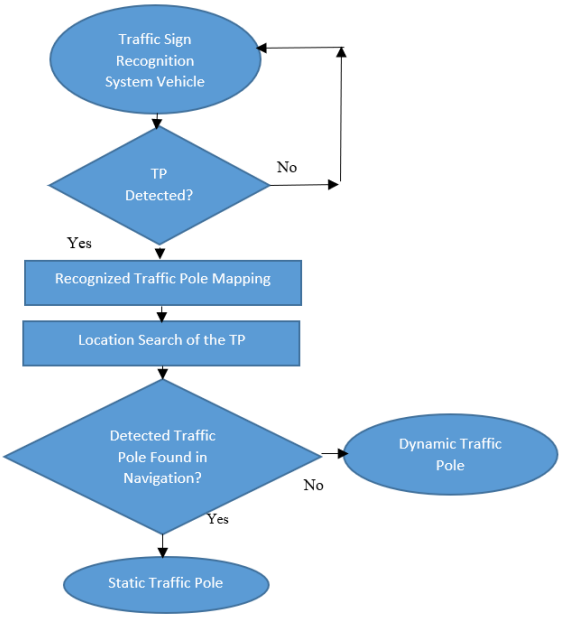


Fig. 11. Block diagram for the detection of static and dynamic traffic poles

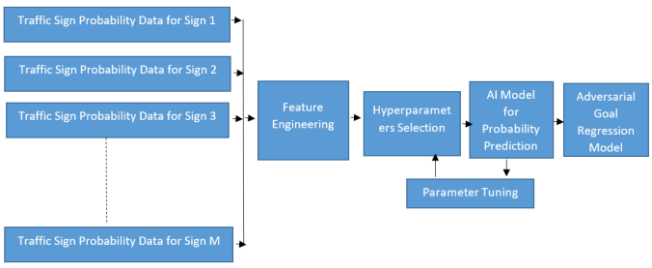
If the traffic pole detected is dynamic, then following steps are followed.

A. The detected dynamic traffic pole's details are added into the navigation system.
B. The dynamic pole details such as ID, latitude, Longitude, Probability value, and signs are collected as a data collection mechanism.
C. The collected data serve as a new data for the added additional dynamic traffic pole.
D. The newly added dynamic traffic pole data from the other vehicles are also collected, and is added to the model preparation for the Adversarial Goal detection model given in Step 9.
E. The traffic pole predicted probability value is considered for the detected dynamic pole and moved to Step 10.

If the traffic pole is identified as static then the following steps are performed.

ii. Hyperparameters are chosen to predict the probability of the traffic signs using received information about the traffic pole.
iii. Trained AI model is imported to evaluate the traffic sign probability for the displayed traffic pole for all the signs of the pole.
iv. The AI model predicts the probability value of the traffic signs.

Machine Learning based Modelling of Adversarial Goal for Each Traffic Pole



Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Fig. 12. Machine Learning based Modelling of Adversarial Goal

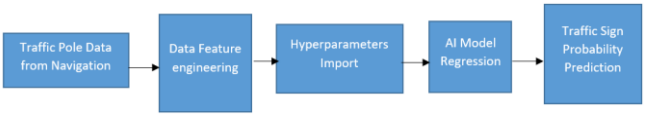


Fig. 13. Traffic Sign Probability Evaluation using a Trained Adversarial Goal Machine Learning Model

10. The predicted probabilities for all the traffic signs from the Adversarial goal detection AI model are considered as $P_{AI\_Adv\_Goal}$ (L), $P_{AI\_Adv\_Goal}$ (R), and $P_{AI\_Adv\_Goal}$ (S).

11. During the real-time operation, when the vehicle reaches the traffic pole, the predicted traffic sign probabilities from the traffic sign recognition system present in the vehicle are considered as $P_{Vehicle}$ (L), $P_{Vehicle}$ (R), and $P_{Vehicle}$ (S).

12. During the compromised traffic sign due to evasion adversarial attacks, if the Adversarial goal detection AI model predicted probabilities are

$$P_{AI\_Adv\_Goal}(L) = 0.6$$
$$P_{AI\_Adv\_Goal}(R) = 0.3$$
$$P_{AI\_Adv\_Goal}(S) = 0.1$$

a) *During the compromised traffic sign due to evasion adversarial attacks, if the vehicular predicted probabilities under untargeted adversarial attacks for sign L are*

$$P_{Vehicle}(L) = 0.01$$
$$P_{Vehicle}(R) = 0.3$$
$$P_{Vehicle}(S) = 0.1$$

$$\text{Probability\_Difference (L)} = P_{AI\_Adv\_Goal}(L) - P_{Vehicle}(L) = 0.6-0.01 = 0.59$$
$$\text{Probability\_Difference (R)} = P_{AI\_Adv\_Goal}(R) - P_{Vehicle}(R) = 0.3-0.3 = 0.00$$
$$\text{Probability\_Difference (S)} = P_{AI\_Adv\_Goal}(S) - P_{Vehicle}(S) = 0.1-0.1 = 0.00$$

b) *During the compromised traffic sign due to evasion adversarial attacks, if the vehicular predicted probabilities under targeted adversarial attacks for sign L for the target class R are*

$$P_{Vehicle}(L) = 0.6$$
$$P_{Vehicle}(R) = 0.8$$
$$P_{Vehicle}(S) = 0.1$$

$$\text{Probability\_Difference (L)} = P_{AI\_Adv\_Goal}(L) - P_{Vehicle}(L) = 0.6-0.6 = 0.00$$
$$\text{Probability\_Difference (R)} = P_{AI\_Adv\_Goal}(R) - P_{Vehicle}(R) = 0.3-0.8 = -0.5$$
$$\text{Probability\_Difference (S)} = P_{AI\_Adv\_Goal}(S) - P_{Vehicle}(S) = 0.1-0.1 = 0.00$$

13. The difference between the Adversarial goal detection AI model predicted probabilities and the vehicular predicted probabilities and are calculated. This is termed as Delta.

14. If Delta = is positive then it is detected as untargeted adversarial goal
If Delta = is negative then it is detected as targeted adversarial goal

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

The proposed system for the detection of adversarial attackers goal which is either targeted or untargeted is shown in Figure 14.

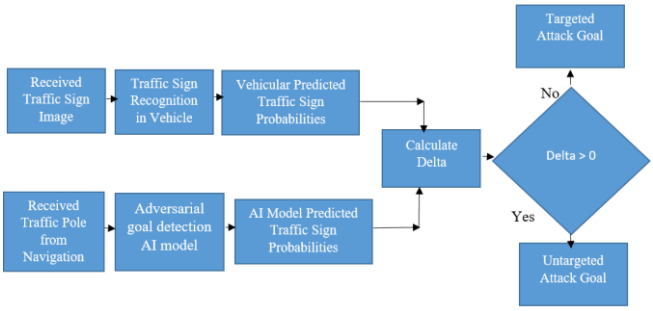


Fig. 14. Proposed system for the detection of adversarial goal in adversarial attacks

8. RESULTS AND DISCUSSION

The proposed system is verified by initially designing the Traffic sign classifier using the available public source traffic sign datasets. The implementation is done using the German Traffic Road Sign Benchmark (GTRSB) dataset [11] and Chinese Traffic Sign Database (CTSD) [12] for designing the traffic sign classifier. The GTRSB dataset contains 43 different traffic sign classes of 50000 images. And the CTSD contains 6164 traffic sign images containing 58 sign categories. The class labels of the both the dataset are shown in Figure 15 and 16. For experimenting the proposed system, a custom dataset is prepared by fetching the images from both the GTRSB and CTSD datasets for the considered 3 traffic signs such as left, right and straight. The prepared custom dataset contains 2800 images of the three traffic signs considered.

Fig. 15. Class Labels of GTSD Dataset

Vol. 21, No. 1, (2024)
ISSN: 1005-0930

Fig. 16. Class Labels of CTSD Dataset

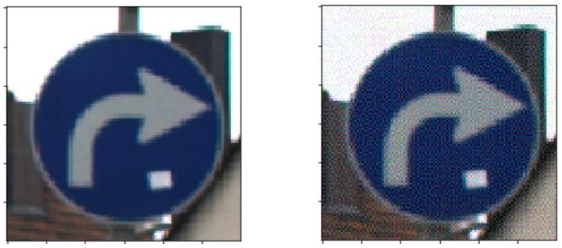
After the preparation of the Traffic Sign Recogntion Custom Dataset, the Convolutional Neural Network based Classifer is designed for classifying the traffic signs. The data abnormality was reduced by discarding some of the data of some classes and adding augmented data to the classes which were falling short of images. The parameters of the designed CNN classifier for classifying the traffic signs are given in Table I.

Table I. Parameters used to train traffic sign classifier.

Data Size	2800
Image Size	60x60x3
Classes	3
Test-Train Split	20:80
Epoch	5
Loss Function	Categorical Cross entropy
Optimizer	Adam
Learning Rate	0.001

The designed traffic sign classifier was observed with a Train and validation accuracy of 99% and 100% respectively.

After the traffic sign classifier is prepared then the adversarial attacks are launched on the traffic signs to create the targeted and untargeted adversarial images of the traffic signs. There are numerous approaches available in the literature for generating the adversarial images. But the simplistic purposes, the FGSM [4] technique is used to generated adversarial perturbations. The idea of FGSM is to perform an action opposite of the gradient descent algorithm in order to maximize the loss. The sample adversarial image generated using the FGSM approach is shown in Figure 17.



(a) (b)

Fig. 17. Adversarial Targeted Adversarial Image using FGSM Approach. (a) Trained Model Prediction for Original Image: Right Sign (b) Model Prediction for Adversarial Image: Left Sign

After training the successful traffic sign recognition system model, which will be resided in the vehicles, the proposed adversarial goal detection machine learning model is prepared. The legitimate traffic sign probability data, which is the output of the prepared traffic sign recognition system is required for modelling the adversarial goal detection model. Hence, here in this experimentation, the legitimate data is collected by collecting inferred traffic sign probabilities using the traffic sign recognition model in vehicles. Series of the inferencing is done for varied driving scenarios of which causes different visibility scenarios of traffic signs under different conditions. The list of scenarios included in the collection of the legitimate probability predictions from the traffic sign classifier in vehicle are summarized in Table II and Table III.

Table II. Legitimate Traffic Sign Classifier Predicted Probability Collection Scenarios

Number of Signs Considered/Supervised Classes	3
Number of Vehicles for Each TP	1000
Signs Considered in Each Traffic Pole	L, R, S
Number of Traffic Poles	25
Total Probability Values	75000
Probability Values per Class	25000
Different Driving Environments	Morning, Evening, Day Time, Nighttime, Rainy, Foggy, Snow

The collected Legitimate traffic Sign Classifier Probability data is used to design and train the proposed Adversarial Goal Detection Machine Learning Model, which is a single layer neural network with three regression heads each for each traffic sign considered for each traffic pole. The fused information of traffic pole ID is used as an input to the regression model to predict the actual probability values of the traffic signs. This model is termed as the adversarial attack goal detection model. The parameters considered for the design and training of the adversarial goal detection model are given in Table III. For the modelled adversarial goal detection task, the neural network-based regression system is designed by feeding the encoded traffic pole features into the neural network input nodes and predicted the legitimate traffic sign probability values from the trained model.

The hyperparameters used in training the proposed adversarial goal detection model are given in Table III.

Table III. Parameters used in the proposed Adversarial Goal Detection Model.

Data Size	75000
Learning rate	1e-5
Classes	3
Batch size	16
Train-Test Split	80%-20%
Loss Function	Mean Square Error
Optimizer	Adam

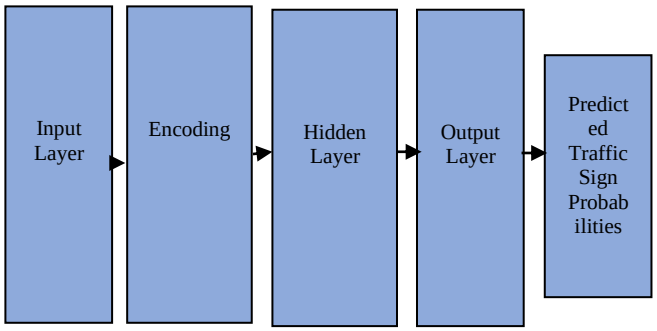


Fig. 18. Proposed Feed Forward based Adversarial Goal Detection Model

The proposed system is evaluated for identifying the generated adversarial image is either targeted or untargeted using the proposed adversarial goal detection model and existing traffic sign recognition model in the vehicle by taking the delta of the predicted probabilities. The performance evaluation of the proposed system in identifying the adversarial goal as targeted and untargeted is done by conducting the inferencing using both the goal detection model and traffic sign recognition model for the below experimentations. The targeted and untargeted images considered for experimentation are given in Table IV.

Table IV. Dataset of Targeted and Untargeted Images

Number of Traffic Poles	25
Number of traffic signs for each pole	3
Total Number of signs	75
Number of untargeted adversarial images using FGSM	75
Number of targeted adversarial images using FGSM	150

The performance of the proposed adversarial goal detection model is evaluated using confusion matrix. It is observed that the True Positive rate for both Targeted and Untargeted adversarial goal detection using the proposed approach after feeding 75 untargeted and 150 targeted adversarial images are 98.67 % and 99.34% respectively. Hence, it is observed that using the proposed approach, we have successfully detected and identified the adversarial goals using the adversarial images with satisfactory accuracy.

CONLUSION

Deep learning, the most popular technique of artificial intelligence, is widely applied in autonomous vehicles to fulfill different perception tasks as well as making real-time decisions. However, the application of deep learning into autonomous driving is still an unanswered question in the artificial intelligence community. The security of deep learning neural networks is showing lots of holes in the past few years as the new vulnerabilities of the neural networks are getting invented. Adversarial attacks is one of the most popular attacks on the deep learning system in the autonomous driving, which is created by adding a human non-noticeable noise into the images which are a part of driving environment. The fooling of the deep learning systems inside the vehicles due to the evasion adversarial images is posing a serious question into the automotive cybersecurity community. In this regard, there are numerous ways of adversarial image detection mechanisms are available in the prior-art. However, they lack the information of the adversary's goals in launching the adversarial attacks on the deep learning system makes this system a special one as this technique proposes a technique for the detection of adversarial goals as either targeted or untargeted. The proposed system detects adversarial goals with satisfactory accuracy. The future work includes identifying the attackers information related to the black box and white box of the target system.

REFERENCES

1. S. D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y. H. Eng, D. Rus, and M. H. Ang, "Perception, Planning, Control, and Coordination for Autonomous Vehicles," Machines, Vol. 5, issue 1, 2017, pp. 1-54.
2. S. Behere and N. Torngren, "A Functional Architecture for Autonomous Driving," in Proc. WASA'15, Montereal, Canada, 2015, pp. 1-7
3. Golson, Jordan; Bohn, Dieter (2016-10-19). "All new Tesla cars now have hardware for 'full self-driving capabilities'". The Verge. Retrieved 2016-10-22.
4. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in International Conference on Learning Representations (ICLR), 2015.
5. Goodfellow, Y. Bengio and A. Courville, Deep Learning, Cambridge, MA, USA:MIT Press, 2016.
6. N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in IEEE Symposium on Security and Privacy, SP. IEEE Computer Society, 2017.
7. Kostyumov, Vasily. "A survey and systematization of evasion attacks in computer vision." International Journal of Open Information Technologies 10.10 (2022): 11-20.
8. Tian, Z., Cui, L., Liang, J., & Yu, S. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. ACM Computing Surveys, 55(8), 1-35.
9. Qiu S, Liu Q, Zhou S, Wu C. Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied Sciences. 2019; 9(5):909. https://doi.org/10.3390/app9050909
10. 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 Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
11. J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The german traffic sign recognition benchmark: A multi-class classification competition," in International Joint Conference on Neural Networks (IJCNN). IEEE, 2011.
12. LinLin Huang,Prof.Ph.D, (2015), Chinese Traffic Sign Database, https://nlpr.ia.ac.cn/pal/trafficdata/index.html