INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

A General Defense Framework for Defend against Adversarial Attacks and Physical World Adversaries on Autonomous Driving

Key Words – Security of Intelligent Systems, Robustness of Machine Learning, Adversarial Machine Learning

A Project Proposal Document by

Mr. K.T.Yasas Mahima

W1742097 - 2018362

Supervised by

Mr. Guhanathan Poravi

Submitted in partial fulfilment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster.

November 2021

I.Table Of Contents

II. List Of Table

III. List Of Figures

IV. List Of Abbreviations

Acronym	**Description**
AI	Artificial Intelligence
AV	Autonomous Vehicle
C&W	Carlini and Wagner
CMOS	Complementary Metal Oxide Semiconductor
DL	Deep Learning
Epsilon	A small value which used to control the perturbation size of an adversarial attack.
FGSM	Fast Gradient Sign Method

FIR	Far-infrared
GAN	Generative Adversarial Networks
JSMA	Jacobian-based Saliency Maps
ML	Machine Learning
Perturbation	Anything that humans visually feel to be the same as the input
PGD	Projected Gradient Descent

1. Introduction

In this research, the author tries to find a novel general approach to defend against adversarial attacks and physical world adversaries to improve the robustness of the computer vision-based Deep Learning (DL) models by accessing classification models use in autonomous vehicles (AVs). This document describes the problem, which is going to be addressed, the novelty of the research idea, and the research approach using necessary pieces of evidence, in addition to the feasibility of the research questions. Finally, this describes the timeline of the research in the next few months.

2. Problem Domain

2.1. Adversarial Attacks

Present research studies show that we cannot guarantee the outputs of the Machine Learning (ML) models because an attacker is able to fool, misclassify and limit the performance of these ML models by purposely designed adversarial perturbations, These adversarial perturbations are undetectable by humans but highly confident to fool the models. The main logic behind these adversarial examples is the distribution gap between the training and real-world inputs to the model (Goodfellow et al., 2015). In an ML pipeline, adversarial attacks can happen in both training and inference phases(Jarmul, 2019; I. Moisejevs, 2019). From these testing phase attacks get increased attention due to their strength (llja Moisejevs, 2019). Based on the attackers' knowledge of the DL network these attacks could be classified as black-box, white-box, and grey-box attacks(Ren et al., 2020). The basic theory behind the testing-phase attack is that the attacker generates adversarial perturbations on the input to the model and fools the model(Qiu et al., 2019).

2.2. Physical World Adversaries

Various studies have shown that when using machine learning models outdoors, physical conditions in the world are making adversaries naturally and simultaneously. When concentrating on the computer vision models noise, glare, and fading effects limit the performance. In most cases, developers will not evaluate the model for these conditions, due to the practice of the cross-validation and train test split methods (Kurakin et al., 2017; Temel et al., 2019). This can be identified as the main reason for this vulnerability.

Figure 1- Sample of physical world corruptions (Temel et al., 2019)

2.3. Autonomous Vehicles And Adversarial Machine Learning

When concentrating on adversarial attacks happening domains AVs, the financial sector, and the healthcare sector get a substantial level of attention(Ma et al., 2021). In this research, the author mainly focuses on the AV domain because the number of adversarial defense methods introduced on AVs is relatively low when compared to the number of adversarial attacks introduced(Qayyum et al., 2020). In AVs, computer vision-related DL techniques such as object detection, image classification, and semantic segmentation are used to perform various safety-critical tasks.(Gupta et al., 2021; Morgulis et al., 2019). This has raised a question, can we trust these safety-critical tasks on AVs? Due to the vulnerability for the adversarial attacks. The author's review paper summarizes those adversarial attacks and defense methods introduced on AVs.

Apart from these man-made adversarial attacks, the researchers found that when AVs are driving through a diverse set of single and simultaneous physical unintended conditions such as illumination changes, noise, contrast, pixel L_∞ perturbations.etc. affect the performance of the ML models(Pavol Bielik et al., 2020). Qayyum et.al summarize several reported accidents on AVs due to the physical world adversaries (Qayyum et al., 2020). The risk of adversarial attacks and physical world unintended adversaries would be a critical issue in the field of AVs because in the future with the arrival of AVs the road safety will be dependent on the robustness of the ML and DL models in AVs. In this research, the author mainly considers the DL-based image classification networks on AVs.

3. Problem Definition

Each and every defense method introduced in AVs for adversarial attacks up to the present is robust for specific attack types and can be fooled by unknown attacks(Qayyum et al., 2020). A recent survey paper shows that still, no existing defense method can effectively defend against adversarial attacks(Ren et al., 2020). Moreover, while AVs are moving through diverse physical conditions single or multiple corruptions could appear at the same time (Qayyum et al., 2020). Hence the performance of the DL models could be degraded. Therefore when developing the defense models against the adversarial perturbations for AVs we have to concentrate on both human synthesized and physical adversaries. In particular, when developing such a defense approach for AVs, two requirements should be satisfied according to the literature. They are, since present DL networks in AVs are highly accurate for non-adversarial inputs, improving the robustness of the given DL network without changing the architecture is ideal(Loukmane et al., 2020). In addition since AVs have high resource consumptive tasks, the defense approach should use a minimal amount of computation power unless it gives superior robustness (Deng et al., 2020; Qayyum et al., 2020). However, in reality, it cannot guarantee 100% resilience from any defense approach in security.

3.1. Problem Statement

A general defense framework that makes existing image classification DL models robust against adversarial attacks and physical world adversaries without changing the existing model architecture or without using any auxiliary tool in the inference is required on AVs owing to the fact that those two adversaries are potential security threats for AVs.

4. Motivation

In an era that companies like Tesla, Uber, and Google are going to introduce commercialized AVs the risk of physical and human synthesized adversarial assaults is an essential problem that has to discuss in the autonomous vehicle domain which results in decreasing road safety which ends up in unnecessary accidents, injuries, cost wastages(Li et al., 2021). As mentioned in the review papers (Zhang et al., 2021) and (Deng et al., 2020), implementing a unified approach of several defense technologies as a general defense method to defend against both man-made and physical world unintended adversarial perturbations would be an important step in the right direction in maximizing defense against adversaries on AVs motivated the author to conduct this research.

5. Existing Works

Existing works on AVs addressed the vulnerability of image classification networks for human synthesized adversarial attacks and physical world corruptions separately.

5.1. Defense Methods Introduced On AVs For Adversarial Attacks

Even though the autonomous vehicle is a highly concentrated research domain in adversarial machine learning still there are very few defense methods that have been introduced when compared to the number of adversarial attacks introduced.

Citation	Summary	Category	Result	Limitations
(Deng et al., 2020)	Evaluate Adversarial training, Defensive distillation, Anomaly Detection, and Feature squeezing methods on traffic sign classification models	Updating Data, Updating Model, Auxiliary Tools based defense	Models are not completely robust against adversarial attacks	Models are not robust for attacks like PGD and physical world adversaries
(Aung et al., 2017)	FGSM and the Jacobian saliency map method-based adversarial training and Defensive distillation methods were used to make traffic light recognition models adversarially robust.	Updating Data, Updating Model-based defense	Got 91% testing accuracy	Does not robust for physical world adversaries
(Wu et al., 2020)	Proposed a single value decomposition and 5G-based approach to make adversarially robust image classification models. Evaluated using IT-FGSM, JSMA, C&W, and Deep fool attacks.	Auxiliary Tools based defense	Able to defend against the attacks at a sufficient level.	Does not robust for physical world adversaries
(Gan and Liu, 2020)	Implemented an adversarial noise removing method using autoencoders to make adversarially robust image classification models.	Auxiliary Tools based defense	Average of 97% robustness against the FGSM adversarial perturbations	Does not robust for attacks like PGD and physical world adversaries

Table 1-Adversarial attack defense methods introduced on autonomous vehicles

5.2. Defense Methods Introduced On Avs For Physical Distortions

To overcome the problem of physical world distortions researchers proposed both hardware and software-based solution. The below table summarizes the most recent research on the problem.

Citation	Summary	Category	Result	Limitations
(Mohammed et al., 2020)	Using high-cost hardware solutions such as CMOS and FIR cameras	Auxiliary Tools based defense	Marginally successful against a limited number of distortions.	High Cost, Have to integrate several hardware components.
(Porav et al., 2018)	Used a GAN-based physical world perturbation generated approach.	Updating Data based defense	The approach was successful against color changes and weather corruptions	May decrease the existing models' accuracy.
(Shu et al., 2021)	Proposed a physical perturbations generation approach based on PGD attack.	Updating Data based defense	The defense method was successful against unforeseen corruptions as well.	High resource consumption.

Table 2-Defense methods introduced on autonomous vehicles for physical distortions.

5.3. Research Gap

After reviewing the literature several limitations have been identified. Those are summarized as follows.

- The existing research works have been concentrated on the robustness of a specific type of adversaries such as either defense against adversarial attacks or physical adversaries.
- Most of the defense models were used FGSM based adversarial training method.
    - Madry, et .al from the MIT USA showed that FGSM based adversarial training can not increase the resilience for advanced attacks with large epsilon ϵ values and improved resilience can achieve by PGD based adversarial training (Madry et al., 2018).

- None of the approaches were evaluated for simultaneous instances of physical corruption.
- In ECCV 2020 Laugros et al introduced a theoretical general defense approach for both adversaries using FGSM based targeted labeling adversarial training with image on image mix-up method(Laugros et al., 2020). However using the image on image mix-up method for scenarios like traffic sign classification is questionable, because already there could be a similar image to the crafted image with a new label in the dataset. Moreover (Sitawarin et al., 2018) introduced an attack on traffic sign classifiers using the same mix-up technique with viewing angles. The empirical results show there is no considerable performance improvement for L_∞ PGD attack.

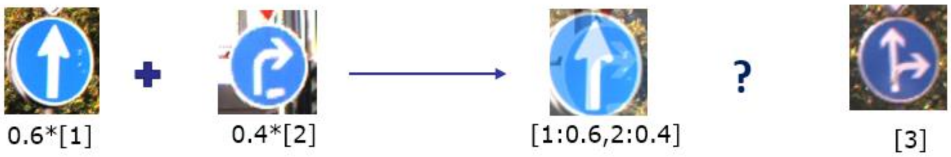

Figure 2- Issue with using the image on image mix-up approach for traffic sign classification scenarios (Self Composed).

As discussed above, the recent literature on implementing a general defense strategy for existing DL networks has both theoretical and performance gaps. In addition, since no general defense approach was tested on robustifying existing safety-critical DL networks in the AVs domain without using any supporting tool in the inference, it has an empirical gap as well. So this research aims to experiment and introduce a novel general defense approach by resolving the identified gaps and evaluate the defense approach mainly on the AVs without changing the existing models' structure or no usage of any auxiliary tool while deployment due to the limited resource constraints in AVs for such defense systems(Deng et al., 2020; Qayyum et al., 2020).

5.4. Research Challenge

The main goal of this research is to address the limitations of literature and enhance the adversarial robustness of the DL models in AVs. From the preliminary review of existing adversarial defense strategies and methods, the author proposes a novel general defense approach to improve the robustness for man-made and physical adversaries. Based on the proposed approach the research challenges can be listed down in the below areas.

1. Adversarial Training
    - Adversarial training is the most efficient defense strategy at present which makes the model robust by re-training it using adversarial examples with their correct labels(Park and So, 2020). However, choosing the optimal attack and epsilon ϵ values to synthesize adversarial examples without degrading the existing model is a challenge(Wang et al., 2019).
2. Random Data Transformation
    - The random data transformation method was chosen to make the models robust against the physical world adversaries. However, choosing and optimally adding random transformations is challenging because it requires mathematical optimization knowledge to ensure the existing models' classification performance for clean samples (Non-adverse images).
3. Integrate the defense methods into the DL model in an optimal way.
    - Effectively integrating the adversarial defense methods is a must because when generating adversarial samples and adding random transformations may cause generalization issues and a poor integration architecture will cause high resource consumption as well.
4. Evaluation of the approach.
    - When evaluating the proposed approach it should have to evaluate for both L_∞ and L_2 adversarial attacks. Moreover, it's essential to evaluate for both singular and mix-up physical world distortions. Thus the evaluation process required lots of resources and time. Moreover, explainable AI methodologies will be used(Linardatos et al., 2021).

6. Research Contribution

The attempts to construct robust ML/DL models increase gradually within the domain of security of intelligent systems. With the arrival of AVs, this has gained higher attention. The contributions of this research could be classified as theoretical contributions and domain contributions.

6.1. Theoretical Contributions

- Experiment and introduce a general defense approach for man-made and physical adversaries using a collaborative approach of PGD L_∞ based adversarial training and a novel image transformation method in the training phase. The proposed approach will not use any auxiliary tool in the inference phase. It will only output an improved version of the given DL model without changing the model architecture.

- Introduce a mathematically optimized novel randomized data transformation approach which mix-up transformations to a single image (Transformation Overlapping) to improve the robustness for both singular and simultaneous instances of the physical corruptions. The author uses the knowledge gained from the literary works which used the input randomization approach in the inference phase(Qiu et al., 2020) as a defense approach for adversarial attacks.

6.2. Domain Contributions

- The defense method will be mainly evaluated on DL image classification models used in AVs. (Traffic Sign Classification, Emergency Vehicle Classification) And it will output robust versions of the given DL classification models without any change of the model architecture or no use of additional auxiliary tools in the inference because low resource consumption in the deployment is essential in AVs.

6.3. Other Contributions

- For the evaluation, the author will use Explainable AI technologies in addition to the evaluation metrics. Using Explainable AI can determine how the proposed approach improves the way of determining the essential pixel attributes to make the predictions under adverse conditions.

7. Research Questions

RQ1 – How to improve the resilience for both adversarial attacks and physical corruptions without updating the network's architecture or the classification pipeline (Naturally robust) ?

RQ2 – What are the main requirements when introducing an adversarial defense approach for AVs?

RQ3 – How can an adversarial training approach be effectively used to increase robustness?

RQ4 – How to integrate adversarial training and adding random transformations as a defense framework without decreasing the existing models' accuracy?

8. Research Aim

This research project aims to design, develop and evaluate a general defense framework that is able to make the existing DL-based classification models in autonomous vehicles robust for both man-made adversarial attacks and physical world adversaries which limits the performance and causes for a security threat.

This research project will simulate and implement a general defense framework for AVs to make the existing DL models adversarially robust for man-made and physical world adversaries. This could be achieved as a combined approach of state-of-the-art adversarial defense approaches (Adversarial re-training with input transformations) with a proper integration architecture and math optimizations.

The final defense models will be evaluated by performing a series of adversarial attacks and the inputs that are affected by the physical adversaries. Moreover to evaluate the performance the author will use Explainable AI as well. The prototype will be presented as a web application where the users can perform attacks on the existing DL models and robust DL models from the proposed framework and evaluate its performance.

9. Objectives

9.1. Research Objectives

The below table summarizes the research objectives or the initial steps necessary to finish the research successfully and the learning outcomes of those objectives.

Objective	**Summary**	**Learning Outcomes**
Literature Review	Do a depth review of required areas for the research **RO1**-To analyze the domain of adversarial attacks and defense against adversarial attacks. **RO2-**To analyze what are the adversarial attacks/physical world adversaries, defense methods on AVs and do a critical review on the advantages and disadvantages of those defense methods. **RO3-**To identify research gaps in the domain of defending adversarial attacks on AVs	LO1 LO4 LO5

	RO4-To identify possible methods to improve the robustness of the DL models. **RO5-**To determine technical skills such as programming languages, development frameworks, and evaluation frameworks/methods required in the research domain.	
Requirement Analysis	Gather user requirements and critically analyze them **RO1-**To determine the awareness of the risk of adversarial machine learning among colleagues, industry experts, and academic experts via meetings and questionnaires. **RO2-**To gather requirements of an adversarial robust DL system and analyze how users/researchers expect the outcome of the research. **RO3-**To get industry/academic experts' feedback and analyze an effective way to present the outcomes.	LO3 LO6 LO7
Design	Designing the architecture of the proposed fusion defense framework **RO1-**To design initial DL (Classification) models required to build and evaluate the proposed framework. **RO2-**To design, a separate component that generates adversarial examples with their correct labels by performing PGD L_∞ attack. **RO3-**To design a method to add random transformations to the dataset and optimize them by a novel mathematical approach: Skewed distribution-based transformations. **RO4-**To design a method that can effectively join the proposed defense methods as a combined solution without reducing the accuracy of the original model.	LO2 LO5 LO7
Development	Developing the proposed general defense framework according to the designed architecture **RO1-**To develop initial classification models which will be used to implement and evaluate the defense framework.	LO1 LO5 LO6 LO7

	RO2-To develop the core functionalities using appropriate hardware and software requirements. **RO3-**To re-train and present robust classification models in AVs as a web application gained from the proposed algorithm.	
Testing and Evaluation	Testing and evaluating the performance of the proposed defense method **RO1-**To create an appropriate test plan for unit and functional testing. **RO2-**To perform various attacks and physical corruptions and verify the robustness of the models after using the proposed defense framework using the appropriate evaluation metrics **RO3-**To get feedback for the research from academic and industry experts.	LO5 LO6 LO7 LO8

Table 3- Research Objectives

10. Project Scope

The primary aim of this study is to simulate and propose a novel general defense framework on AVs. Thus the already hardware solutions like CMOS and FIR cameras for physical world adversaries are not pointed up in this project. Based on the objectives and survey of the literature, the scope of the requirements can be listed down as below.

10.1. In-Scope Requirements

- A general defense framework makes the given classification model robust for the man-made and physical world adversaries - A collaborative approach of adversarial training and random transformations with required optimizations.
- The defense method shouldn’t change the existing models' architecture and shouldn’t use any auxiliary tools in the inference – AVs have complicated tasks, thus they cannot afford so much computational power for a super complex defense approach(Deng et al., 2020).
- DL models robustness evaluation – Evaluating the vulnerability of the classification models before and after using the proposed defense approach using the appropriate evaluation

metrics used in the adversarial ML domain and get the interpretability of the models' performance using Explainable AI.

- Web application for presenting and evaluating the proposed defense approach.

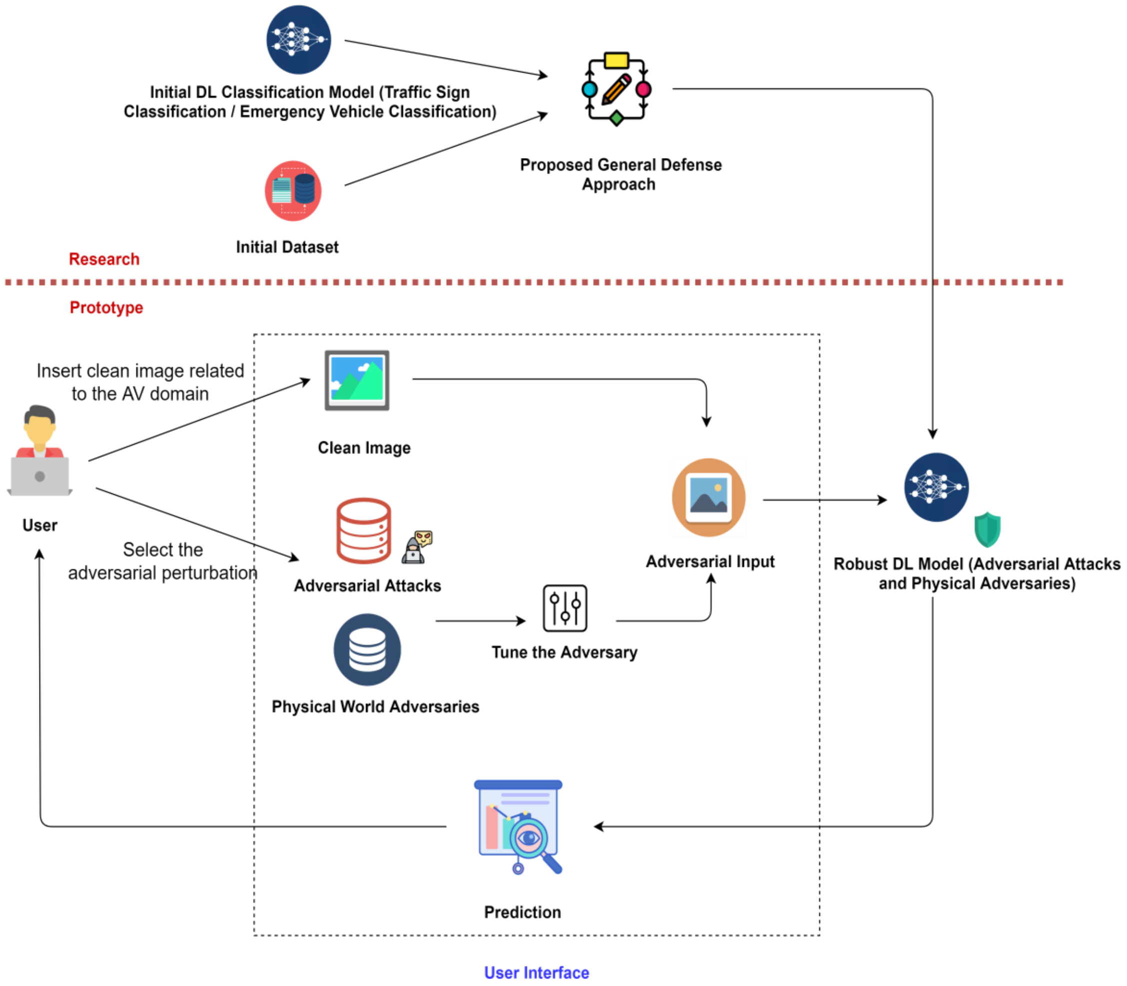

Figure 3-Prototype feature diagram of the research. (Self-Composed)

10.2. Out Of Scope Requirements

- The proposed defense framework cannot be used to increase the robustness of the models which use annotation types like semantic segmentation and object detection.
- The proposed system may not be robust for Grey-Box attacks and attacks performed by sign/sticker embedding.
- This approach cannot be used as a rain-removing application (de-raining) because the proposed study concentrates on improving the resistance against transposing illuminations, blur, hue, and noise...etc. due to the physical world conditions.
- The proposed defense framework is not an automated approach. It’s a piece of an experiment of a novel general defense approach to defending man-made and physical adversarial perturbations as a unified solution.

11. Methodology

11.1. Research Methodology

In (Boaz and Ashby, 2003) described the quality of research based on four main aspects; "*methodological quality, quality of reporting, appropriateness of strategy, and relevance to policy and practice*". The below table summarizes the selected scientifically methodized research process which suits this project.

Scientific Research Methodology	Philosophy	The **pragmatism approach** was chosen among the positivism, pragmatism, realism, and interpretivism approaches because the authors will **investigate** and **experiment** various methodologies as a combined approach to identify which works best for achieving the research goal.
	Approach	This research intends to **experiment and prove** a hypothesis that can improve the general resilience of the DL networks against human synthesized adversarial attacks and physical adversaries. Among the possible methods of deductive and inductive, the **deductive approach** was chosen because the research aims at applying a combination of **existing theories extensively**. As the data analysis approach, both **qualitative and quantitative** methods were chosen.
	Strategy	The strategy of research defines how you implement the methodology by answering the research questions. Among the possible candidates, **surveys**, **interviews,** and **experiments** based on evaluation metrics will be used.
	Choice	For this research, the **mixed method** was chosen among the possible techniques of mono, mixed, and multi-methods because both **quantitative** and **qualitative** data will be used for the research such as surveys, performance values, and responses from interviews.
	Time Horizon	The collection of data have to be done at one point in the evaluation phase of the authors' research. Thus a **cross-**

		sectional method was selected among the longitudinal and cross-sectional methods.
	Procedures	For collecting the data mainly the records and libraries of the organizational dataset will be used. Apart from that surveys, interviews, reports, and statistics will be used.

Table 4-Research methodology

Research Hypothesis: The research hypothesizes that using an optimized form of adversarial training approach and random data transformation method makes existing classification models robust against the man-made and physical adversarial perturbations.

Research Process: Optimize the existing classification models using a general defense method of adversarial training and random transformations.

Prototype Input: Images related to the classification models in AVs will be used for the investigations including some physical world constraints.

Research Output: Optimizing the existing classification model and outputs an adversarially robust model (An improved version of the given model).

11.2. Development Methodology

Since the project will be designed, built, and evaluated until the desired/acceptable output is achieved as the development methodology **Prototyping** method will be chosen.

11.3. Requirement Elicitation Methodology

To gather the requirements for the prototype the author will use **interviews questionnaire** and knowledge gained from the **literature.**

11.4. Design Methodology

Since this research is focusing on implementing a novel algorithmic approach, as the design methodology the author chose Structured Systems Analysis & Design Method (SSADM) from the among possible approaches of SSADM and OODA (Object-oriented analysis and design).

11.5. Evaluation Methodology

Evaluation plays a critical part in a research study. Proper evaluation emphasizes the solidity of the research findings. For the evaluation of the proposed general defense approach authors chose both evaluation metrics and benchmarking evaluation approaches.

11.5.1. Evaluation Metrics

For the quantitative evaluation of the research, the authors chose evaluation metrics used in the adversarial machine learning domain. They are as follows(Madry and Kolter, 2018).

1. **Accuracy** – Calculates the ratio of the correctly predicted samples to the total samples in the testing set.
2. **Adversarial Error** – Calculated the ratio of the in-correctly predicted samples to the total samples when the samples are perturbed by adversaries.

$$adv_{error} = \frac{1}{n}\sum_{i=1}^{n}(y_i - h_\theta(x_{i_{adv}}))$$

Where y_i is the correct label, h_θ is the DL network, $x_{i_{adv}}$ is the adversarial input and n is the number of samples.

3. **Adversarial Loss** – Calculate the CrossEntropy Loss (Rose Wambui, 2021) of the testing sample when the testing set is perturbed by the adversaries.

$$adv_{loss} = \frac{1}{n}\sum_{i=1}^{n}\ell(h_\theta(x_{i_{adv}})), y_i)$$

Where ℓ is the loss function

In contrast in a case of an imbalanced dataset, the author chose the following evaluation metrics for the evaluation(Wardhani et al., 2019).

1. **Cohen's Kappa Score** - This calculates the agreement between two evaluators (Groud-Truth and Prediction). Since this uses, probabilistic methods, the literature shows this is one of the appropriate evaluation metrics for imbalance data(Maarit Widmann, 2020).

$$\mathcal{K} = (\mathcal{P}_0 - \mathcal{P}_e)/(1 - \mathcal{P}_0)$$

Where $\mathcal{P}_0$ is the predicted agreement and $\mathcal{P}_e$ is the expected agreement result (Ground Truth) when both evaluators assign labels randomly.

2. **Area under Curve (AUC)** - This gives a combined measure of the performance across all possible thresholds in the classification problem.

11.5.2. Benchmarking

Previous research on general defense approaches and defenses for physical world adversaries used ImageNet:*150GB* (Deng et al., 2009) and ImageNet-C:*66GB* (Hendrycks and Dietterich, 2019) datasets for benchmarking. Due to the limited resources, the author won’t be able to do a benchmark on those datasets and those are not specific to the research domain AVs. However, the author will conduct a logical benchmark of the proposed approach with the general defense work introduced by (Laugros et al., 2020) as an evaluation of resilience for particular attacks or physical world distortions due to the unavailability of the code-bases of their work. Moreover, the complexity of the final defense artifact could be benchmarked as the number of models/tools deployed in the inference(Deng et al., 2020). In addition, the author will benchmark the capturing of the important features by the DL network before and after using the proposed approach under adverse conditions using Explainable AI algorithms.

11.6. Project Management Methodology

For managing the project, the **Agile Prince-2** method will be used because this research is simulation-based from the possible approaches of Agile, Kanban, Waterfall, and PRINCE2...etc.

11.6.1. Resource Management

Based on the proposed objectives and functionalities the identified software, hardware, data, and skills required to fulfill the research project are as follows.

11.6.1.1. Hardware Requirements

- Core i5 processor or above – Minimum quad-core processor which can perform highly resource consumption computer vision tasks.
- 50GB Disk Space or More – To store the datasets and ML models.
- 8GB RAM or More – A ram that can manage large datasets with an average of 1GB-4 GB in size and trained DL models.

Note – If hardware resources are not sufficient propose to use Google Colab free tire version.

11.6.1.2. Software Requirements

- Operating System – An OS that can perform huge computational tasks and be able to run required DL-related tools without any compatibility issues.
- Python/R – The main programming language used to develop the DL models.
- Flask/Django – Web framework for the prototype.
- Keras / Pytorch – The DL libraries use to develop classification models. Have to benchmark the performance and developer friendliness of each library since Keras has advanced functionalities and PyTorch is a research-friendly library.
- OpenCV/Scikit Image – Image processing libraries which used to add transformations to the datasets.

- Jupyter NoteBook / Pycharm – Integrated development environments for programming.
- GitHub/GitLab – Version controlling and code backup tool.
- Zotero/Mendeley – The reference management tool for thesis and research papers.
- MS Word/Google Docs – For documentation purposes.
- Google Drive/DropeBox- To back up required documents.

11.6.1.3. Data Requirements

- Datasets for implementing CNN models – Kaggle and Google Dataset Search.

11.6.1.4. Skills Requirements

- Experience in understanding mathematical optimizations and evaluations.
- Deep understanding of the neural networks and evaluation methods.
- Deep understanding of the adversarial attacks and defense strategies.
- Creative writing skills.

11.6.2. Risk Management

When doing a project it's always associated with the risks such as technical, theoretical, or any other unpredictable vulnerabilities. To finish a project with the expected outcome a well-managed risk management plan is essential.

Risk	Risk Level	Risk Frequency	Mitigation Plan
Deep knowledge of adversarial attacks and math. – Adversarial attacks and defending is mostly a theoretical and hypothesis-based research area. Therefore first it should have proper knowledge about theoretical aspects and equations.	Medium	High	There is a limited number of tutorials by some universities which can get a basic knowledge of the domain. Moreover reviewing more literature should help in solving the theoretical risks.
Lack of experts who worked with adversarial machine learning – In Sri Lanka, there is a limited number of experts who are working in this	Medium	Medium	Find out some international Ph.D. students and professors who are working

domain. Therefore to get feedback and evaluate the project, the author should have to put a hard effort to find domain experts.			on the adversarial attacks domain.
Updating the requirements of the research – Based on the theoretical knowledge and the hardware resources the initial requirements of the project could be changed.	Medium	High	Using prototyping-based development approaches help with this, Moreover can get feedback from the supervisor and other experts as well.
Insufficient hardware resources – Research in adversarial machine learning takes lots of hardware resources and if the hardware resources are not sufficient it may cause increasing the execution time.	High	High	Use free cloud-based solutions like Google Colab or IBM Watson
Privacy and Policies – The field of adversarial machine learning is an ongoing research area Sometimes these attacks are used by the forces as well. So when implementing a defense model first it should have a clear understanding of those policies	Low	Low	Review more literature and ask from domain experts.
Any unpredictable risk – There can be unpredictable issues such as the Covid-19 pandemic, natural disasters which can affect the project.	High	Medium	Manage the work according to a timetable and always try to keep daily or weekly goals.

Table 5-Risk Mitigation Plan

11.6.3. Schedule

11.6.3.1. Deliverables

Deliverable	Date
Project Proposal	1st Nov 21
Literature Review Document - A survey of both existing works on the domain and subdomains.	18th Oct 21
Review Paper - Publish a research paper that reviews the existing attacks, defense methods, and identified research gaps in the domain.	15th Oct 21
Software Requirement Specification - Requirements to be satisfied in the research prototype.	22nd Nov 21
System Design Document - A document specifying the architecture of the proposed defense framework based on identified techniques from the LR.	6th Dec 21
Prototype - Prototype of the research which integrates the proposed defense framework.	25th Apr 22
Thesis - The final documentation of the research work discusses the research process, findings, and decisions.	25th Apr 22
Project Research Paper - A research paper that summarizes the proposed defense framework and its performance.	1st May 22
Extended research paper- Extended version of the conference paper	1st June 22

Table 6- Project deliverables and dates

11.5.3.2. Gantt Chart

TASK	START	END	DAYS
1. Project Initiation	**1 Aug 2021**	**4 Nov 2021**	**96**
1.1 Selection of research domain,problem and a gap	1 Aug 2021	13 Aug 2021	13
1.2 Mentor approval of the research problem and gap	15 Aug 2021	23 Aug 2021	9
1.3 Getting domain knowledge of the research	1 Aug 2021	4 Nov 2021	96
1.4 Working on initial Literature Review	7 Aug 2021	31 Oct 2021	86
1.5 Identifying research aims , requirments and the objectives	1 Sep 2021	4 Nov 2021	65
1.6 Formative Project Initiation Document (PID) submission	1 Sep 2021	12 Sep 2021	12
1.7 Mentor review of formative PID	12 Sep 2021	18 Sep 2021	7
1.8 Self evaluation of formative PID	18 Sep 2021	4 Nov 2021	48
1.9 PID submission	4 Nov 2021	4 Nov 2021	1
1.10 Ethics form submission	2 Nov 2021	2 Nov 2021	1
2. Literature Review	**1 Aug 2021**	**14 Mar 2022**	**226**
2.1 Categorize the components that have to cover in the LR	1 Aug 2021	7 Aug 2021	7
2.2 Gathering realated papers	7 Aug 2021	18 Sep 2021	43
2.3 Reading and summarize the research papers	7 Aug 2021	18 Oct 2021	73
2.4 Reviewing gathered research data	1 Sep 2021	18 Oct 2021	48
2.5 Preparing draft LR	2 Sep 2021	18 Oct 2021	47
2.6 Mentor review of draft LR	3 Sep 2021	30 Oct 2021	58
2.7 Self evaluation of draft LR	4 Sep 2021	30 Oct 2021	57
2.8 LR submission	5 Sep 2021	31 Oct 2021	57
2.9 Extended LR	6 Sep 2021	14 Mar 2022	190
3. Requirement Gathering	**1 Oct 2021**	**4 Jan 2022**	**96**
3.1 Analyze current processes	1 Oct 2021	20 Dec 2021	81
3.2 Preparing questionnaires	12 Oct 2021	6 Dec 2021	56
3.3 Distributing questionnaires to target audiences	26 Oct 2021	20 Dec 2021	56
3.4 Preparing interview questions	12 Oct 2021	6 Dec 2021	56
3.5 Conducting interviews	26 Oct 2021	20 Dec 2021	56
3.6 Aggregating data from questionnaires	1 Nov 2021	4 Jan 2022	65
3.7 Aggregating data from interviews	1 Nov 2021	4 Jan 2022	65
3.8 Analyzing received data	23 Nov 2021	4 Jan 2022	43
3.9 Preparing data for Software Requirement Specification (SRS)	1 Dec 2021	4 Jan 2022	35
3.10 Preparing draft SRS	1 Nov 2021	27 Dec 2021	57
3.11 Mentor review of draft SRS	14 Dec 2021	3 Jan 2022	21
3.12 Self evaluation and correction of draft SRS	14 Dec 2021	4 Jan 2022	22
3.13 SRS submission	4 Jan 2022	4 Jan 2022	1
4. Designing of the System	**12 Oct 2021**	**25 Jan 2022**	**106**
4.1 Identifying components of existing adversarial robuest systems in Avs	12 Oct 2021	10 Jan 2022	91
4.2 Deciding the scope and boundaries of the prototype	19 Oct 2021	10 Jan 2022	84
4.3 Designing the high level architecture of the prototype solution	21 Dec 2021	10 Jan 2022	21
4.4 Mentor review of high level architecture	5 Jan 2022	17 Jan 2022	13
4.5 Self evaluation and refining of high level architecture	5 Jan 2022	17 Jan 2022	13
4.6 Creating system design	5 Jan 2022	25 Jan 2022	21
4.7 Designing system design diagram	5 Jan 2022	25 Jan 2022	21
4.8 Mentor review of the system design diagram	18 Jan 2022	25 Jan 2022	8
5. Selection of Tools and Technology	**15 Oct 2021**	**25 Jan 2022**	**103**
5.1 Reviewing defense technologies on man-made adversarial attacks	15 Oct 2021	25 Jan 2022	103
5.2 Reviewing defense technologies on physical adversarial puterbations	15 Oct 2021	25 Jan 2022	103
5.3 Reviewing quantitative evaluation methods	15 Oct 2021	25 Jan 2022	103
5.4 Selection of appropriate technologies for the collaborative solution	1 Dec 2021	25 Jan 2022	56
5.5 Mentor review of selected tools and technologies	4 Jan 2022	25 Jan 2022	22
6. Prototype Implementation	**14 Dec 2021**	**11 Apr 2022**	**119**
6.1 Define prototype deliverables	14 Dec 2021	10 Jan 2022	28
6.2 Implementing core features	1 Jan 2022	1 Mar 2022	60
6.3 Integrating compartments of the app	4 Jan 2022	7 Mar 2022	63
6.4 Review code and best practices	15 Feb 2022	31 Mar 2022	45
6.5 Demo and presentatin of prototype to mentor	1 Apr 2022	11 Apr 2022	11
6.6 Refine the prototype	29 Mar 2022	11 Apr 2022	14
7. Testing and Evaluation	**1 Mar 2022**	**25 Apr 2022**	**56**
7.1 Identify Unit components to test	1 Mar 2022	11 Apr 2022	42
7.2 Identifying functional and nonfunctional requirements	1 Mar 2022	11 Apr 2022	42
7.3 Creating test plans	1 Apr 2022	25 Apr 2022	25
7.4 Writing unit tests	1 Apr 2022	25 Apr 2022	25
7.5 Writing integration test	1 Apr 2022	25 Apr 2022	25
7.6 Writing usability test	1 Apr 2022	25 Apr 2022	25
7.7 Getting feedback from the academic and industry experts	1 Apr 2022	25 Apr 2022	25
7.8 Final evaluation of research component implementation	25 Apr 2022	28 Apr 2022	4
8. Dissertation and Documentation	**1 Mar 2022**	**26 Apr 2022**	**57**
8.1 Drafting the dissertation	29 Mar 2022	18 Apr 2022	21
8.2 Mentor review dissertation	5 Apr 2022	26 Apr 2022	22
8.3 Refining dissertation	5 Apr 2022	26 Apr 2022	22
8.4 Submission of final dissertation	29 Apr 2022	29 Apr 2022	1
8.5 Writing documentation for the product	12 Apr 2022	26 Apr 2022	15
8.6 Project presentation and viva	13 May 2022	30 May 2022	18
9. Extra Tasks	**1 Aug 2021**	**30 May 2022**	**303**
9.1 Writing a review paper	1 Sep 2021	31 Oct 2021	61
9.2 Mentors evaluation of the review paper and submission	24 Oct 2021	31 Oct 2021	8
9.3 Write the dissertation	21 Dec 2021	26 Apr 2022	127
9.4 Writing a research paper	1 Mar 2022	25 Apr 2022	56
9.5 Mentors evaluation of the research paper and submission	25 Apr 2022	30 Apr 2022	6
9.6 Designing the public version of the system	15 Mar 2022	30 May 2022	77
9.7 Develop system for the public release	27 Apr 2022	23 May 2022	27

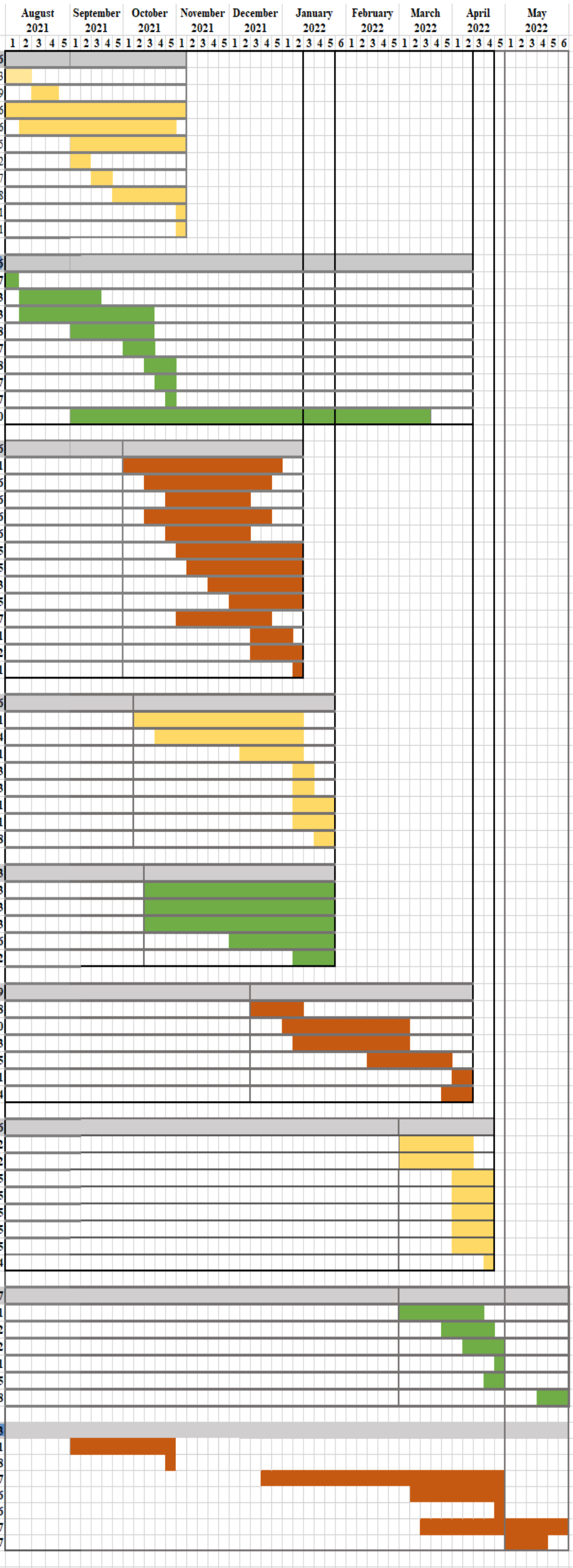

Figure 4-Research plan described in a Gantt chart. (Self Composed)

References

Aung, A.M., Fadila, Y., Gondokaryono, R., Gonzalez, L., 2017. Building Robust Deep Neural Networks for Road Sign Detection. ArXiv abs/1712.09327.

Boaz, A., Ashby, D., 2003. Fit for Purpose? Assessing Research Quality for Evidence Based Policy and Practice.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Deng, Y., Zheng, X., Zhang, T., Chen, C., Lou, G., Kim, M., 2020. An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models. 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom) 1–10.

Gan, H., Liu, C., 2020. An Autoencoder Based Approach to Defend Against Adversarial Attacks for Autonomous Vehicles, in: 2020 International Conference on Connected and Autonomous Driving (MetroCAD). pp. 43–44. https://doi.org/10.1109/MetroCAD48866.2020.00015

Goodfellow, I., Shlens, J., Szegedy, C., 2015. Explaining and Harnessing Adversarial Examples. CoRR abs/1412.6572.

Gupta, A., Anpalagan, A., Guan, L., Khwaja, A.S., 2021. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. Array 10, 100057. https://doi.org/10.1016/j.array.2021.100057

Hendrycks, D., Dietterich, T.G., 2019. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ArXiv abs/1903.12261.

Jarmul, K., 2019. Privacy Attacks on Machine Learning Models. Privacy Attacks on Machine Learning Models. URL https://www.infoq.com/articles/privacy-attacks-machine-learning-models/ (accessed 4.5.21).

Laugros, A., Caplier, A., Ospici, M., 2020. Addressing Neural Network Robustness with Mixup and Targeted Labeling Adversarial Training, in: ECCV Workshops.

Li, Y., Xu, X., Xiao, J., Li, S., Shen, H.T., 2021. Adaptive Square Attack: Fooling Autonomous Cars With Adversarial Traffic Signs. IEEE Internet of Things Journal 8, 6337–6347. https://doi.org/10.1109/JIOT.2020.3016145

Linardatos, P., Papastefanopoulos, V., Kotsiantis, S., 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy 23. https://doi.org/10.3390/e23010018

Loukmane, A., Graña, M., Mestari, M., 2020. A Model for Classification of Traffic Signs Using Improved Convolutional Neural Network and Image Enhancement, in: 2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS). pp. 1–8. https://doi.org/10.1109/ICDS50568.2020.9268761

Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F., 2021. Understanding Adversarial Attacks on Deep Learning Based Medical Image Analysis Systems. Pattern Recognit. 110, 107332.

Maarit Widmann, 2020. Cohen’s Kappa: what it is, when to use it, how to avoid pitfalls. URL https://www.knime.com/blog/cohens-kappa-an-overview

Madry, A., Kolter, Z., 2018. Adversarial Robustness - Theory and Practice. URL https://adversarial-ml-tutorial.org/ (accessed 10.4.21).

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. ArXiv abs/1706.06083.

Mohammed, A.S., Amamou, A., Ayevide, F.K., Kelouwani, S., Agbossou, K., Zioui, N., 2020. The Perception System of Intelligent Ground Vehicles in All Weather Conditions: A Systematic Literature Review. Sensors (Basel, Switzerland) 20.

Moisejevs, llja, 2019. Evasion attacks on Machine Learning (or “Adversarial Examples”). Evasion attacks on Machine Learning (or “Adversarial Examples”). URL https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1

Moisejevs, I., 2019. Adversarial Attacks and Defenses in Intrusion Detection Systems: A Survey. International Journal of Artificial Intelligence and Expert Systems (IJAE) 8, 44–62.

Morgulis, N., Kreines, A., Mendelowitz, S., Weisglass, Y., 2019. Fooling a Real Car with Adversarial Traffic Signs. ArXiv abs/1907.00374.

Park, S., So, J., 2020. On the Effectiveness of Adversarial Training in Defending against Adversarial Example Attacks for Image Classification. Applied Sciences 10. https://doi.org/10.3390/app10228079

Pavol Bielik, Tsankov, Dr.P., Krause, Prof.A., Vechev, Prof.M., 2020. Reliability Assessment of Traffic Sign Classifiers. Federal Office for Information Security.

Porav, H., Maddern, W.P., Newman, P., 2018. Adversarial Training for Adverse Conditions: Robust Metric Localisation Using Appearance Transfer. 2018 IEEE International Conference on Robotics and Automation (ICRA) 1011–1018.

Qayyum, A., Usama, M., Qadir, J., Al-Fuqaha, A., 2020. Securing Connected Autonomous Vehicles: Challenges Posed by Adversarial Machine Learning and the Way Forward. IEEE Communications Surveys Tutorials 22, 998–1026. https://doi.org/10.1109/COMST.2020.2975048

Qiu, H., Zeng, Y., Zhang, T., Jiang, Y., Qiu, M., 2020. FenceBox: A Platform for Defeating Adversarial Examples with Data Augmentation Techniques. ArXiv abs/2012.01701.

Qiu, S., Liu, Q., Zhou, S., Wu, C., 2019. Review of Artificial Intelligence Adversarial Attack and Defense Technologies. Applied Sciences 9. https://doi.org/10.3390/app9050909

Ren, K., Zheng, T., Qin, Z., Liu, X., 2020. Adversarial Attacks and Defenses in Deep Learning. Engineering 6, 346–360. https://doi.org/10.1016/j.eng.2019.12.012

Rose Wambui, 2021. Cross-Entropy Loss and Its Applications in Deep Learning. URL https://neptune.ai/blog/cross-entropy-loss-and-its-applications-in-deep-learning

Shu, M., Shen, Y., Lin, M.C., Goldstein, T., 2021. Adversarial Differentiable Data Augmentation for Autonomous Systems.

Sitawarin, C., Bhagoji, A., Mosenia, A., Chiang, M., Mittal, P., 2018. DARTS: Deceiving Autonomous Cars with Toxic Signs. ArXiv abs/1802.06430.

Wang, J., Zhang, T., Liu, S., Chen, P.-Y., Xu, J., Fardad, M., Li, B., 2019. Beyond Adversarial Training: Min-Max Optimization in Adversarial Attack and Defense. ArXiv abs/1906.03563.

Wardhani, N.W.S., Rochayani, M.Y., Iriany, A., Sulistyono, A.D., Lestantyo, P., 2019. Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data, in: 2019 International Conference on Computer, Control, Informatics and Its Applications (IC3INA). pp. 14–18. https://doi.org/10.1109/IC3INA48034.2019.8949568

Wu, F., Xiao, L., Yang, W., Zhu, J., 2020. Defense against adversarial attacks in traffic sign images identification based on 5G. EURASIP Journal on Wireless Communications and Networking 2020, 173. https://doi.org/10.1186/s13638-020-01775-5

Zhang, T., Deng, Y., Lou, G., Zheng, X., Jin, J., Han, Q.-L., 2021. Deep Learning-Based Autonomous Driving Systems: A Survey of Attacks and Defenses. IEEE Transactions on Industrial Informatics 1–1. https://doi.org/10.1109/TII.2021.3071405