COMBINING STRENGTHS TO IMPROVE THE ROBUSTNESS AND GENERALIZATION OF TRAFFIC SIGN RECOGNITION

ROWANNE TRAPMANN

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2066283

COMMITTEE

dr. D. Hendrickson

dr.N. Rankovic

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 13, 2023

Word count: 8177

ACKNOWLEDGMENTS

I would like to show my appreciation to my supervisor, Dr Drew Hendrickson, for his support and critical discussions. I also want to thank my mentor, Jesse van der Zweep, and my colleagues within NavInfo Europe for always being available for feedback, brainstorming sessions, and resources. In addition, I would like to extend my special thanks to my partner, Lowie, my mother, and my fellow thesis student, Kiki, for their support and occasional necessary motivational speeches.

COMBINING STRENGTHS TO IMPROVE THE ROBUSTNESS AND GENERALIZATION OF TRAFFIC SIGN RECOGNITION

ROWANNE TRAPMANN

CONTENTS

Abstract

According to research, neural networks (NNs) and deep neural networks (DNNs) are particularly sensitive to adversarial attacks, which arbitrarily modify the network's output (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2017). These so-called altered outputs can be of great danger. This study aimed to investigate the impact of data augmentation techniques and adversarial training on the classification accuracy and resistance against adversarial attacks in the recognition of traffic signs. By evaluating the performance of two CNN models, ResNet18 and ResNet50, on a dataset of traffic signs using several data augmentation techniques and adversarial attacks. The results showed that integrating data augmentation methods and adversarial training can improve the robustness of the models against adversarial attacks. The ResNet18 model trained with adversarial training and data augmentation techniques achieved an average attack rate of 0.45 and 0.77 on the DeepFool and I-FGSM attacks, respectively. However, it is also notable that the loss over the epochs was high, indicating that the models may still be vulnerable to other types of attacks. The results show that the models are less resistant to the I-FGSM attack, specifically the ResNet50 model. Additionally, it is noteworthy that the model's accuracy decreases when data augmentation techniques, which aim to simulate real-world scenarios, are applied. Overall, this research highlights the importance of considering the robustness of models in the context of computer vision tasks and the need for further research to improve the robustness of CNN models against adversarial attacks.

1 DATA SOURCE, CODE, ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The German Traffic Sign Recognition Benchmark (GTSRB) data set is made publicly available by Stallkamp, Schlipsing, Salmen, and Igel (2011). The specific files used in this research are collected from the Kaggle user Mykola (2018) under the Public Domain CC0 1.0 license. The base code for the adversarial attacks (I-FGSM, C&W, DeepFool) and training is created using the publicly available examples by Kim (2020) and Nicolae et al. (2018) under the MIT License. The author of this thesis acknowledges that they do not have any legal claim to this data. When not produced by the author, images used in this thesis were licensed under creative commons. The code that is used for this research is available on GitHub: https://github.com/RTrapmann/Thesis_project

2 INTRODUCTION

The research goal of this Master Thesis is to develop a combination of advanced techniques that can be applied to improve the robustness of models against adversarial attacks. This research will focus on applying these techniques in computer vision, specifically in the recognition of traffic signs in the automotive sector. The ultimate goal is to create models that are not only accurate in their predictions but also resistant to adversarial attempts to manipulate outputs or corrupt their performance.

2.1 *Problem statement*

Artificial intelligence (AI) is all about computational algorithms that can match human intelligence. Machine Learning (ML) and Deep Learning (DL) are a form of AI where machines learn from their surrounding environments, with or without human interference. Nowadays, there are several fields where these ML and DL-based techniques are applied. However, for safety-critical tasks (i.e., where a mistake is very costly) such as medical imaging or self-driving, the adoption of these techniques, in particular neural networks (NNs), has been cautious due to concerns about their reliability and resistance towards adversarial attacks (Waskom, 2021). For the use cases, in the centre of the security-critical systems, it is therefore vital that the ML components can be robust towards generalization and resist adversarial attacks. For the area of autonomous driving, in specific, it is easy to imagine situations where the model needs to be reliable and robust. A slight difference in environmental conditions or camera position could mean, e.g., that the car does not recognize a pedestrian or a stop sign and cannot correctly calculate the depth estimation for a braking vehicle. Therefore this Master's Thesis is focused on recognising traffic signs, which is a crucial task for self-driving (vehicles), traffic mapping, and traffic surveillance.

2.2 *Relevance*

According to research, neural networks (NNs) and deep neural networks (DNNs) are particularly sensitive to adversarial attacks, which arbitrarily modify the network's output (Madry et al., 2017). These so-called altered outputs can be of great danger. In the physical world, the adversarial examples can be a small change in the camera view or putting a sticker on a traffic sign. There are several models and techniques available in the literature that can measure the robustness of a model and the resistance toward adversarial attacks. However, they need specific data augmentations to reproduce the published results. These techniques often come hand in hand with a degree of inductive bias introduced into the training data set, making it not generalizable to other subsets of the data.

The main objective of this Master's thesis is to study and combine robust techniques for computer vision, specifically in the area of traffic sign recognition for the automotive industry. The focus of the project will be on evaluating the robustness of model architectures such as ResNet18 and ResNet50 against general white-box adversarial attacks, including the Iterative Fast Gradient Sign Method (I-FGSM) and DeepFool. The goal is to gain insights into creating generalizable and robust models that can be applied in safety-critical scenarios, such as self-driving vehicles. The outcome of this research will contribute to the scientific community and help establish trust in the use of Machine Learning and Deep Learning models in the automotive industry. After all, the future of automotive lies in self-driving.

2.3 *Research questions*

Based on the identified research gap and the established significance to society, the primary research question for this study is formulated as follows:

> *How do the integration of data augmentation methods and adversarial training impact the classification accuracy and resistance against adversarial attacks in recognition of traffic signs?*

The following sub-questions will help to answer the main research question:

RQ 1A *To what degree do data augmentations to the training data enhance the robustness and resistance of the trained CNN models ResNet18 and ResNet50, specifically against adversarial attacks, utilizing the I-FGSM method, on the GTSRB data set?*

RQ 1B *To what degree do data augmentations to the training data enhance the robustness and resistance of the trained CNN models ResNet18 and ResNet50, specifically against adversarial attacks, utilizing the DeepFool method, on the GTSRB data set?*

The sub-research questions focus on the effectiveness of using data augmentation techniques on the training data set to improve the robustness and resistance of two specific convolutional neural networks (CNN) models, ResNet18 and ResNet50. The specific adversarial attack methods being used in this case are the Iterative Fast Gradient Sign Method (I-FGSM) and the DeepFool method. To answer the sub-research questions, we will first train the models on the normal data set without data augmentations to establish a baseline for the model's performance. Afterwards, the training data will be augmented with various data transformations, such as adding a Gaussian blur and rotating the image, to measure their effect on the robustness of the CNN models. The robustness will be evaluated using the attack success rate (ASR) and accuracy per iteration as the primary metric.

RQ 2 *To what extent does incorporating adversarial training using the I-FGSM method's attack samples affect the robustness of the ResNet18 CNN model?*

The research question is focused on understanding the impact of incorporating adversarial training on the robustness of convolutional neural network (CNN) models, specifically the ResNet18 model. Research by Zhang et al. (2019) indicates that 'simple' models are more receiving toward the adversarial training process. Adversarial training is a method of training machine learning models to be more robust against adversarial attacks. The specific adversarial attack used is the I-FGSM. The goal is to understand how well the ResNet18 model performs against this attack when trained with adversarial samples. The robustness of the models will be measured using the attack success rate (ASR) and accuracy per iteration.

RQ 3 *How does the model's capacity affect the accuracy of recognizing stop signs and its ability to resist adversarial attacks?*

The research by Zhang et al. (2019) suggests that expanding a model's capacity enhances its robustness. The technique will be demonstrated through the use of ResNet18 and ResNet50 models. While other research questions centre around attacking the model and implementing adversarial training, this question centres around stopping signs' classification accuracy, which is a crucial road safety feature necessary for regulating traffic and preventing collisions. The reliability and accuracy of a model for detecting stop signs are vital for the safe operation of autonomous vehicles that use computer vision for navigation. A stop sign classifier that can handle various lighting, weather conditions, and stop sign designs is crucial for successfully deploying autonomous vehicles in real-world environments. The primary metric for evaluating the accuracy will be the attack success rate (ASR) and accuracy per iteration. The I-FGSM attack will be used to test the model's accuracy. This will give us a better understanding of the robustness of the model against an adversarial attack, which will help make the model more secure and reliable for the real-world use case.

The following chapters summarise the literature and related work on the GTSRB - data set. The methods and experiments used to answer the above-mentioned questions will be discussed extensively in chapter 4 and chapter 5. The results will be presented subsequently in chapter 6.

3 LITERATURE REVIEW

This literature review aims to provide an overview of current research on adversarial machine learning by examining the existing studies in the field, highlighting their key findings and limitations, and identifying potential areas for future research. The review focuses explicitly on image classification, which is the task of assigning labels or classes to images based on their visual content. It covers the research on algorithms, models, and performance evaluations on various data sets. It also delves into the German Traffic Sign Recognition Benchmark (GTSRB) data set, commonly used for traffic sign recognition and classification research. Additionally, the review covers research on adversarial attacks and defences and the use of state-of-the-art image classification models like CNNs. Furthermore, it provides an understanding of the broader applications of computer vision and its use in fields such as robotics, self-driving cars, and security systems.

3.1 *Image classification*

Image classification is a critical task in computer vision and involves assigning labels or classes to images based on their visual content. The field of computer vision is dedicated to enabling computers to interpret and analyze visual data from the real world through algorithms and models that can identify patterns, objects, and events in images and videos. Research in image classification aims to develop algorithms and models that can accurately classify images into different classes. This involves designing feature extraction methods to extract relevant information from the images, selecting appropriate classifiers to make predictions, and evaluating the performance of the classifiers on different data sets. The German Traffic Sign Recognition Benchmark (GTSRB) data set is widely used in computer vision for traffic sign recognition and classification research. Studies have used the data set to evaluate the performance of different machine learning algorithms, compare the performance of different CNN architectures, and study the impact of preprocessing techniques and data augmentation methods on the performance of traffic sign recognition models. These studies often present high accuracy scores, with some even reaching 99.46% accuracy (He, Nan, Li, Lee, & Yang, 2020).

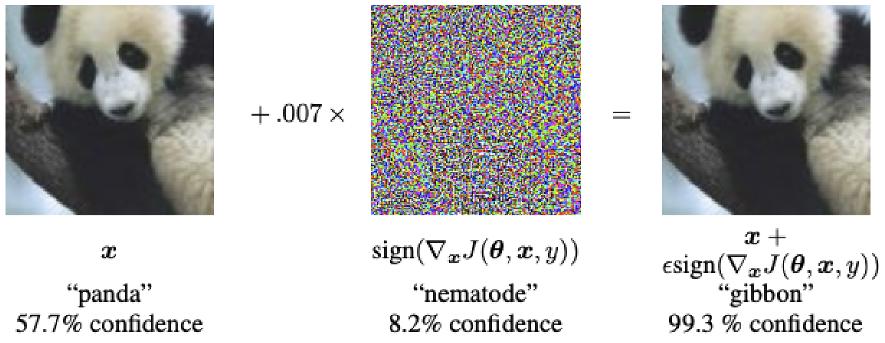
The GTSRB data set has played a significant role in developing and advancing traffic sign recognition and classification. It has contributed to developing more effective and robust traffic sign recognition systems (Sermanet & LeCun, 2011). In addition to traditional image classification tasks, the data set has also been used in research on adversarial attacks and defence. Recently, black-box-oriented attacks such as patch and sticker attacks have been commonly investigated for their ability to simulate real-world scenarios such as stickers or jams on traffic signs (Bayzidi et al., 2022). The state-of-the-art image classification models used in these studies typically include deep neural networks like Convolutional Neural Networks (CNNs), like VGGNet, ResNet, DenseNet, AlexNet, and ShuffleNet V2, among others. The ResNet model, in particular, is a powerful and effective CNN architecture that has significantly impacted the field of computer vision.

3.2 *Computer vision use case*

Computer vision is a field of artificial intelligence that aims to enable computers to interpret and analyze visual data from the world around them (Fei-Fei, Fergus, & Perona, 2006). It involves the development of algorithms and models that can recognize patterns, objects, and events in images and videos and extract meaning from them.

Research in the field of computer vision not only has a wide range of applications, such as object recognition, scene understanding, facial expression analysis, and medical image analysis, but is also applied in many other fields, such as robotics, self-driving cars, and security systems (Fei-Fei et al., 2006). The first Autonomous Land Vehicle project in the U.S.A. started in 1984 (Lipson & Kurman, 2016). It was the first vehicle that used color cameras and laser scanners to find its way on the road. Nowadays, there are a lot of different levels of automation, and each comes with difficulties and set-up requirements. A lot has changed since 1984 regarding the ML components used for these safety-critical tasks. Deep neural networks (DNNs) are becoming more efficient at various challenging machine-learning tasks. They can detect images with near-human accuracy in the image classification area, providing state-of-the-art results for machine-learning tasks. Computer vision components are used for nearly every task necessary for controlling a car: semantic segmentation, object detection, road sign classification, pose estimation, depth estimation, lane line prediction, car trajectory prediction, predicting the intentions of other drivers, etc. Researchers have demonstrated that the computer vision

Figure 1: Fast Gradient Sign Attack (FGSM) example. Source: TensorFlow (CC BY 4.0)


$\boldsymbol{x}$
"panda"
57.7% confidence
$+ .007 \times$
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$=$
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

and machine learning components of current neural networks (NNs) are vulnerable to changes in the input or environment, resulting in faulty or adversarial predictions. These so-called altered outputs can be described as adversarial examples resulting from an adversarial attack. Szegedy et al. (2013) characterizes an adversarial example as, given an input image X, the goal is to find a minimal perturbation η, such that the predicted label for input X is misclassified with respect to the true label (Carlini & Wagner, 2017). This makes the use of neural networks in safety-critical domains challenging. Where research was first solely focused on understanding the architecture of the networks being used, the focus now shifted toward adversarial machine learning. An emerging field of study.

3.3 *Adversarial attacks*

Several kinds of adversarial attacks can trick the machine learning model into making false predictions. Szegedy et al. (2013) detected adversarial examples for the first time in the image classification domain. With these adversarial attacks, it is feasible to distort a picture by a tiny amount, thereby influencing how the image is classified. The overall amount of change applied is sometimes so minimal that it is invisible. The ease by which attackers can find adversarial cases limits the fields in which neural networks can be deployed. Suppose that neural networks are used in self-driving vehicles; adversarial examples might enable a hacker to induce the car to perform undesirable behaviour, e.g., ignoring a stop sign (Carlini & Wagner, 2017). The research conducted by Szegedy et al. (2013) introduces adversarial attacks as worrisome and the models that are affected by them less reliable. In 2013 Szegedy et al. stated that there is a lack of knowledge on why adversarial attacks are so successful and their high complexity levels. More recently, research has been conducted on the success rates of adversarial attacks and defences. However, the high levels of complexity make it more difficult to find one specific defence method, especially for the different attack goals. Following the research from Chakraborty, Alam, Dey, Chattopadhyay, and Mukhopadhyay (2021), a broad classification can be made of the adversarial goals:

- *Confidence reduction*
  The adversary attempts to reduce the confidence of prediction for the target model. For example, making an image of a 'stop' sign can be predicted with lower confidence having a lesser probability of belonging to that class.

- *Misclassification*
  The adversary tries to alter the output classification of an input example to any class different from the original class. For example, a legitimate image of a 'stop' sign will be predicted as any other class different from the class of stop sign.

- *Targeted misclassification*
  The adversary tries to produce inputs that force the output of the classification model to be a specific target class. For example, any input image to the classification model will be predicted as a class of images having a 'go' sign.

- *Source/Target misclassification*
  The adversary attempts to force the classification output for a specific input to be a particular target class. For example, the input image of the 'stop' sign will be predicted as a 'go' sign by the classification model.
  Source: Chakraborty et al. (2021)

Along with the adversarial goals, a distinction can be made between black-box and white-box attacks. This refers to the accessibility of the deployed model and internal gradients. We have full access to the model and internal gradients of white-box attacks, contrary to black-box methods, where we only have access to the model's output.

3.4 *Adversarial training*

Minimizing the vulnerability to adversarial attacks has been one of the top research topics. A well-known empirical defence technique against adversarial attacks, specifically evasive attacks, is adversarial training

Figure 2: Classification of image-based attack. Source: Zixiao Kong et al. (CC BY 4.0)

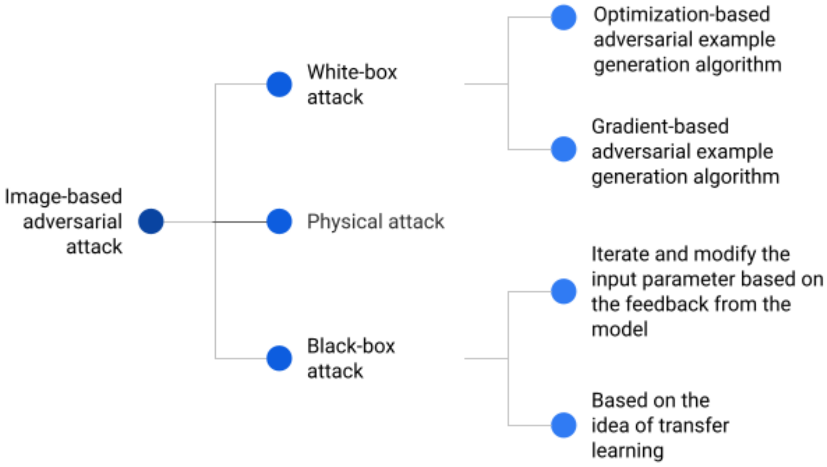
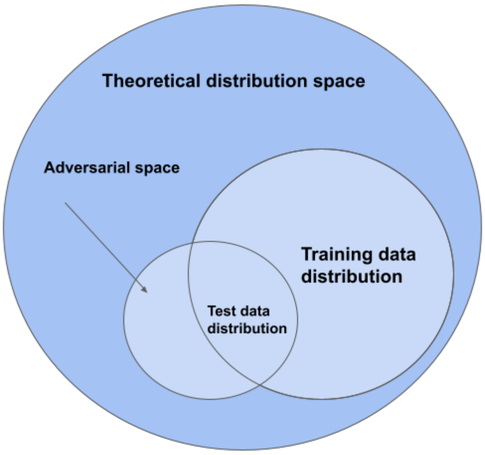


Image-based adversarial attack
White-box attack
Physical attack
Black-box attack
Optimization-based adversarial example generation algorithm
Gradient-based adversarial example generation algorithm
Iterate and modify the input parameter based on the feedback from the model
Based on the idea of transfer learning

(Kong et al., 2021). By adding generated adversarial samples, of the data set, in the training loop, the samples are added to the training set. The altered data set will be used in the training loop. This way, the model learns to ignore the noise and only learns from robust features. Adversarial training is, however, a reasonably delicate defence method since the choices one makes about the hyper-parameters and architecture, among others, can influence the performance of the adversarial training tremendously. Research by Szegedy et al. (2013) suggests that imputing the adversarial examples (with the correct labels) to the first initial training loop of the model might improve the generalization of the resulting models. Several studies have been conducted on this suggestion. However, in research from Ilyas et al. (2019), Tsipras, Santurkar, Engstrom, Turner, and Madry (2018) and Zhang et al. (2019), is a certain trade-off phenomenon between accuracy and robustness established. Specifically when the aim is to find a model that has a high classification accuracy score and is robust. Tsipras et al. (2018) suggests that adversarial training can be used in the same terms as data augmentation. Thus, adding a few adversarial images to the training set. By doing this, the classifier will be trained on the adversarial samples. However, the accuracy and robustness trade-off phenomenon is a downside to this technique. The accuracy will go down as the robustness increases when the amount of adversarial samples is increased. Therefore Rebuffi et al. (2021) conducted research to find a technique to decrease the robust trade-off phenomenon during the adversarial training process.

However, the research does not investigate the generalizability aspect of these models.

Figure 3: Adversarial space representation *"Attackers can exploit pockets of adversarial space between the data manifold fitted by a statistical learning agent and the theoretical distribution space to fool machine learning algorithms" (Freeman & Chio, 2018).*
Source: The author's illustration



3.5 *Data augmentation*

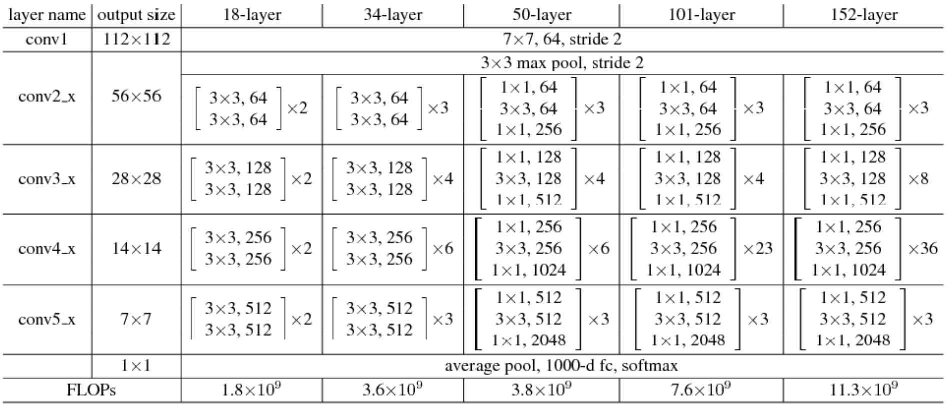
"The essence of data augmentation is to expand the original training set with generated adversarial samples when massive data is lacking, to ensure effective training of the model " (Shi & Han, 2018). Data augmentation techniques are primarily used to avoid overfitting, specifically in fields, such as the medical field, where there is a lack of access to big data. The abilities of these techniques can enhance the size and quality of training data sets. Research by Shi and Han (2018) states that augmentation techniques can make a model more resistant to adversarial attacks and more generalizable. Data augmentation techniques can be used for different purposes, such as making a model more resistant to detecting noise, as stated in the research by Rodriguez, Dokladalova, and Dokládal (2019). Fundamental data transformations such as color space augmentations, horizontal flipping and random cropping are one of the first transformations showing the effect of data augmentations (Shorten & Khoshgoftaar, 2019). The most commonly known data augmentations are based on basic image manipulations, such as *geometric and color space transformations* e.g., flipping, cropping, rotation,

noise injection, and changing the colour space. Concerning the use case, recognition of traffic signs, it is essential to look into kernel filters. These transformations are well-liked for image-processing tasks. Using, e.g., a Gaussian blur filter will generate blurred images, resulting in a model with higher resistance to motion blur. Nevertheless, research by Shorten and Khoshgoftaar (2019) concluded that kernel filters are most effective when used as a layer of the network architecture, contrary to using the filters as a data augmentation technique. A similar result can be achieved by modifying the layers such that the activation layers keep the image pixel values between 0 and 255, contrary to a sigmoid function that will map the pixels to values between 0 and 1 (Shorten & Khoshgoftaar, 2019). More recently, data augmentations have been used to help train models to be more resistant to adversarial attacks.

3.6 *Robust models*

According to the research conducted by Tsipras et al. (2018), it is essential to train models to detect and resist adversarial attacks. However, this process can be computationally expensive and may decrease the model's standard accuracy (Tsipras et al., 2018). The researchers show that mainly focusing on being resistant to adversarial attacks has not been enough to make models more robust. Interestingly, researchers Gokhale, Mishra, Luo, Sachdeva, and Baral (2022) demonstrate that using data augmentation techniques positively affects the robustness of the models toward adversarial attacks. Madry et al. (2017) proposed an adversarial training technique for creating models that are robust and resistant to first-order adversarial attacks, such as the Fast Gradient Sign Attack (FGSM). First-order attacks can only use the first-order derivatives of losses in terms of input. The method states that when a model is adversarially trained with a specific first-order attack, it will also be resistant to other first-order attacks. Madry et al. (2017) simplifies the robustness problem by stating that it is a min-max problem, where the goal is to find a model with minimal loss on any adversarial attack while also being resistant to any adversary. Furthermore, Madry et al. (2017) identifies a relationship between model capacity and adversarial robustness in the MNIST data set (Deng, 2012). Adversarial examples in a non-robust model can change the decision boundary to a more complicated one. Increasing the model capacity can positively affect the robustness of one-step adversarial attacks while only training with the original input data (Zhang et al., 2019). However, the field of model robustness, specifically in regard to models that can withstand adversarial attacks, is still an emerging area of study that requires further research and understanding.

Figure 4: Resnet architecture Source: Pytorch (CC BY 3.0)



layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		[3×3, 64 3×3, 64] ×2	[3×3, 64 3×3, 64] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3	[1×1, 64 3×3, 64 1×1, 256] ×3
conv3_x	28×28	[3×3, 128 3×3, 128] ×2	[3×3, 128 3×3, 128] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×4	[1×1, 128 3×3, 128 1×1, 512] ×8
conv4_x	14×14	[3×3, 256 3×3, 256] ×2	[3×3, 256 3×3, 256] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×6	[1×1, 256 3×3, 256 1×1, 1024] ×23	[1×1, 256 3×3, 256 1×1, 1024] ×36
conv5_x	7×7	[3×3, 512 3×3, 512] ×2	[3×3, 512 3×3, 512] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3	[1×1, 512 3×3, 512 1×1, 2048] ×3
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

4 METHODOLOGY

This chapter will elaborate on the methods, as well as the experimental setup and the experiments themselves that are used in this research.

4.1 *Models*

The ResNet model architecture was selected for this study based on the research of Harisubramanyabalaji, Nyberg, Gustavsson, et al. (2018), which demonstrated the model's wide use and success in various computer vision tasks. The ResNet model is a deep convolutional neural network (CNN) characterized by its skip connections, which enable it to bypass certain layers in the network and directly access the final output. It consists of a series of blocks, each containing multiple convolutional layers and an optional shortcut connection. These skip connections allow the model to learn residual functions, which are the difference between the desired output and the output of the layers in the block.

According to a study by Zhang et al. (2019), increasing the capacity of a neural network can improve its robustness to adversarial samples. To test these findings, the ResNet18 and ResNet50 models will be used in this study. The ResNet18 model is a smaller version of the ResNet model, while the ResNet50 model is a deeper and broader version of the architecture that won the ImageNet Large Scale Visual Recognition Challenge in 2015. The ResNet50 model consists of 50 layers, including convolutional layers, max-pooling layers, and fully connected layers. The specific parameter differences can be seen in figure 4.

Both the ResNet18 and ResNet50 models were originally pre-trained on the ImageNet data set, but for the experiments in this study, it was decided not to use the pre-trained models. Instead, the weights and expected input and output sizes were modified to fit the GTSRB (German Traffic Sign Recognition Benchmark) data set. This allows for a more customized and accurate model for the specific task at hand. Overall, the ResNet model is a powerful and effective CNN architecture that has significantly impacted the field of computer vision.

4.2 *Adversarial attacks*

As demonstrated in section 3.3, there are several kinds of adversarial attacks. In this research experiment, only white box attacks with white box settings are used. To be more precise, gradient-based attacks. These attacks identify as perturbing a clean image x for several iterations I with a small step size in the direction of the model's loss function's gradient (Tuna, Catak, & Eskil, 2022). For this research, the choice has been made to focus solely on white box attacks to gather the most information about the results, such that they can be analyzed and the parameters can be changed. The following gradient-based attacks are used:

$$x.adv = x + \epsilon * sign(xJ(\theta, x, y)) \tag{1}$$

I-FGSM

The Fast Gradient Sign Method (FGSM) is a simple and practical adversarial attack method introduced by Goodfellow, Shlens, and Szegedy (2014). It works by linearizing the loss function in the L∞ norm of an original image and finding the maximum of the linearized function (Kurakin et al., 2018). This results in an output image that looks indistinguishable from the original to the human eye but causes the neural network to make an incorrect prediction. The FGSM method generates adversarial samples in just one iteration, which makes it fast, but does not guarantee full misclassification (Goodfellow et al., 2014).

To address this limitation, the iterative fast gradient method (I-FGSM) was introduced. The I-FGSM attack works by iteratively applying the FGSM method to the input data, with each iteration increasing the magnitude of the perturbation with a small step size, such as α=1 (Kurakin et al., 2018). This attack has been shown to be effective at generating adversarial examples that are difficult for machine learning models to detect, as the perturbations it introduces are often small and imperceptible to humans, as seen in figure 1 (Kurakin et al., 2018).

The I-FGSM adversarial attack is an important tool for researchers studying the robustness of machine learning models to adversarial examples. It allows them to evaluate the vulnerability of different models to attacks and to develop methods for improving the robustness of these models (Kurakin et al., 2018).

DeepFool

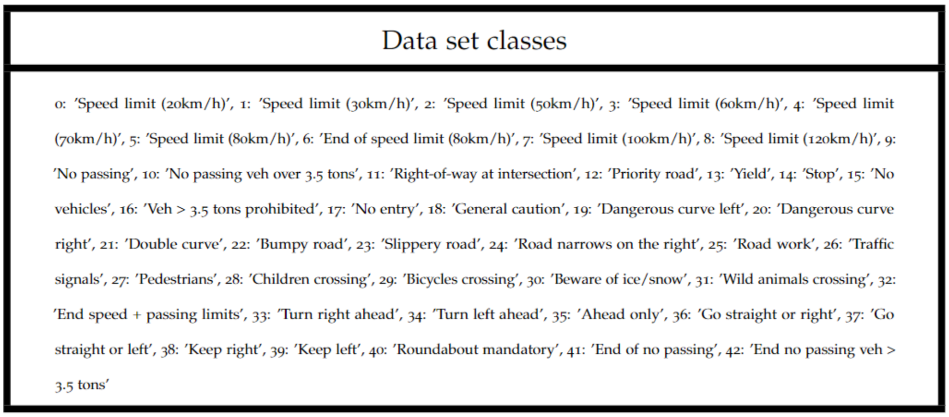
The DeepFool attack, developed by Moosavi-Dezfooli, Fawzi, and Frossard (2016), generates adversarial samples that fool state-of-the-art classifiers. The DeepFool attack uses an iterative optimization process to search for the optimal perturbation, which is then added to the original image to create the adversarial example. At each iteration, the attack estimates the gradient of the decision boundary of the neural network and moves the input image in the direction that will maximize the distance between the input and the decision boundary. This process is repeated until the input is classified as a different class than the original

The attack is based on an iterative linearization of the classifier to generate minimal perturbations sufficient to change classification labels. It effectively evaluates the specific classifier's robustness and enhances its performance by fine-tuning parameters. While computationally efficient, the method can be used to estimate the robustness of complex deep neural networks (DNN) on large-scale data sets. The DeepFool method imagines the classifier's decision space being divided by linear hyperplane boundaries that divide the decision to select different classes. It then tries to shift the image's decision space location directly towards the closest decision boundary. Nonetheless, the decision boundaries are often non-linear, so the algorithm completes the perturbation iteratively until it passes a decision boundary.

4.3 *Data set*

The data set that was selected for this Master Thesis is the German Traffic Sign Recognition Benchmark (GTSRB) data set. The German Traffic Sign Recognition Benchmark (GTSRB) is a data set of traffic sign images widely used in research on traffic sign recognition and classification. The data set was initially published to host a multi-class, single-image classification challenge at the International Joint Conference on Neural Networks (IJCNN) 2011, (Stallkamp et al., 2011). The data set consists of more than 50,000 images of traffic signs belonging to 43 different classes, including speed limit signs, yield signs, and stop signs. The images were collected from various locations in Germany and are annotated with the class label of the traffic sign depicted.

Figure 5: Classification of image-based attack. Source: The author's illustration



Data set classes

0: 'Speed limit (20km/h)', 1: 'Speed limit (30km/h)', 2: 'Speed limit (50km/h)', 3: 'Speed limit (60km/h)', 4: 'Speed limit (70km/h)', 5: 'Speed limit (80km/h)', 6: 'End of speed limit (80km/h)', 7: 'Speed limit (100km/h)', 8: 'Speed limit (120km/h)', 9: 'No passing', 10: 'No passing veh over 3.5 tons', 11: 'Right-of-way at intersection', 12: 'Priority road', 13: 'Yield', 14: 'Stop', 15: 'No vehicles', 16: 'Veh > 3.5 tons prohibited', 17: 'No entry', 18: 'General caution', 19: 'Dangerous curve left', 20: 'Dangerous curve right', 21: 'Double curve', 22: 'Bumpy road', 23: 'Slippery road', 24: 'Road narrows on the right', 25: 'Road work', 26: 'Traffic signals', 27: 'Pedestrians', 28: 'Children crossing', 29: 'Bicycles crossing', 30: 'Beware of ice/snow', 31: 'Wild animals crossing', 32: 'End speed + passing limits', 33: 'Turn right ahead', 34: 'Turn left ahead', 35: 'Ahead only', 36: 'Go straight or right', 37: 'Go straight or left', 38: 'Keep right', 39: 'Keep left', 40: 'Roundabout mandatory', 41: 'End of no passing', 42: 'End no passing veh > 3.5 tons'

5 EXPERIMENTAL SETUP

5.1 *Algorithms and software*

Due to a lack of sufficient GPU power in local resources, a Secure Shell (SSH) connection was made. The SSH connection makes it possible to remotely connect to a computer connected to the SSH server. In this case, it is connected to a computer that operates with the Linux OS at NavInfo Europe. The storage of data and modelling were done remotely on this computer. All programming was done in Visual Studio (VS) Code with the programming language Python 3.8.10. This programming environment is specifically chosen for its ability to support debugging activities and version control. The following libraries were installed: Pandas (McKinney, 2010), and NumPy (Harris et al., 2020), Matplotlib (Hunter, 2007), PIL (Umesh, 2012), Pytorch/torchvision (Paszke et al., n.d.), Seaborn (Waskom, 2021), Adversarial Robustness Toolbox (ART) (Nicolae et al., 2018) and SKlearn (Pedregosa et al., 2011).

5.2 *Dataloader*

The (GTSRB) data set was initially split into an 80% training set and a 20% test set. However, to evaluate the generalizability of the results, the training set was further divided into an 80% training set and a 20% validation set. The test set was left unchanged, and all adversarial attacks were only performed on this set. The data was pre-processed by normalizing the RGB channels to fit the [0,1] scale, using the equation 2, and was loaded into a custom Pytorch data loader, as shown in figure 6. The data loader

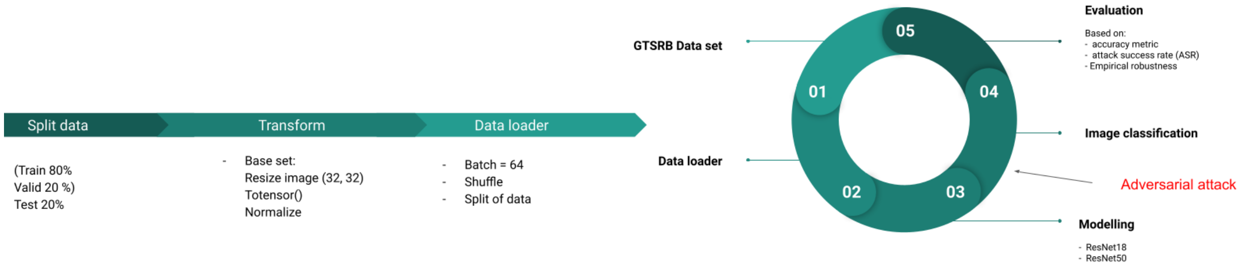


Figure 6: Data pipeline. Source: The author's illustration

contains the specific data set (train, valid, or test), a flag indicating whether the set needs to be shuffled, and a batch size of 64.

The images in the GTSRB data set were collected in 2011 under real-life conditions, meaning that the quality of the images may not be as high as current standards. However, the decision was made not to include a step in the preprocessing cycle to improve the quality of these images to keep the model more realistic. Even if the images were collected today in 2023, the quality would not be consistent across all 50,000 images. In addition, all images are in the RGB (Red, Green, Blue) color space and are normalized by transforming the components to fit the [0,1] scale by subtracting the mean from the channel and dividing the output by the standard deviation.

$$Output[channel] = (input[channel] - mean[channel])/std[channel] \tag{2}$$

5.3 *Data transformers*

Four different sets of data augmentation techniques are made to execute the experiments. The augmentation techniques are divided to get a valid insight into the effect different techniques can have on the robustness and accuracy of the models.

- *Set 1* Base pre-processing set + RandomRotation of 1 degree. The RandomRotation transformation randomly rotates an image by a specified angle. In this case, the angle is set to one degree. This transformation was chosen to mimic the real-world scenario where the camera may not perfectly align with the traffic signs. According to research by Rodriguez et al. (2019), rotating an image by one degree can significantly impact the classification accuracy of a model.

Therefore, this transformation was included in the data augmentation process to test the robustness of the model under such conditions.

- *Set 2* Base + Gaussian Blur with a kernel size of (5, 9). A Gaussian blur is well known for its blurring effect. It is often used to reduce noise and reduce the amount of detail in an image, as well as to mimic foggy weather conditions. By applying a Gaussian blur to the images in the data set, we can better simulate the effects of foggy weather on the performance of the model

- *Set 3* Base + RandomRotation + Gaussian Blur + ColorJitter. The ColorJiiter transform randomly changes an image's brightness, contrast, and saturation in the "RGB" space. It takes a value of 0.5 brightness and 0.3 hue. This can help the model learn to recognize the underlying features of the traffic signs rather than just memorizing the exact colors of the images in the training set. This can be useful in situations where the model may encounter traffic signs with different lighting or color conditions than those in the training set.

- *Set 4* Base + ElasticTransform, $\alpha 1$. The ElasticTransform adds a random distortion to the image using a displacement field generated from Gaussian noise. This distortion can help to mimic certain real-world conditions, such as rainy weather, as the resulting image appears stretched or distorted. It is applied with a small magnitude, 1.0 α, to avoid introducing too much noise into the image

5.4 *Experiments*

The experiments are designed to evaluate the robustness of the ResNet18 and ResNet50 models to adversarial attacks, focusing on the I-FGSM and DeepFool attacks. The results of these experiments are analyzed and discussed in chapter 6 and 7 to provide insights into the robustness of the ResNet models to adversarial attacks and to identify the most effective methods for improving robustness. The following steps are taken to give a substantiated answer to the research question:

- *Experiment 1* The ResNet18 and ResNet50 models are trained on the GTSRB data set using the standard training procedure. The models are evaluated on the test set to determine their baseline performance. The I-FGSM and DeepFool adversarial attacks are applied to the test set, and the models are evaluated on the adversarial examples to determine their performance under attack. The results of the attacks are compared to the baseline performance to assess the robustness of

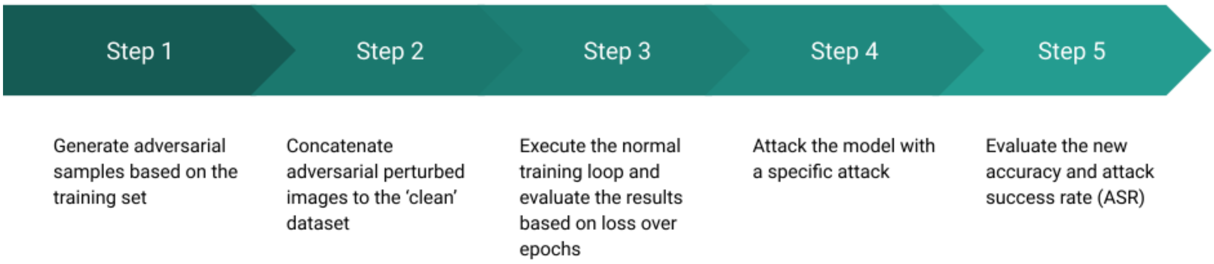

Step 1
Step 2
Step 3
Step 4
Step 5
Generate adversarial samples based on the training set
Concatenate adversarial perturbed images to the 'clean' dataset
Execute the normal training loop and evaluate the results based on loss over epochs
Attack the model with a specific attack
Evaluate the new accuracy and attack success rate (ASR)

Figure 7: Data pipeline, experiment 2. Source: The author's illustration

the models to adversarial attacks. Figure 6 provides details on the data pipeline.

- *Experiment 2* The purpose of this experiment is to evaluate the robustness of the ResNet18 and ResNet50 models to adversarial attacks, specifically focusing on the I-FGSM. To do this, adversarial training will be applied to the models to improve their robustness. Adversarial examples will be generated and added to the training data set, and the models will be trained on this augmented data set. The models will then be tested on the original test data set, and their performance will be evaluated. The results of the experiments will be analyzed and discussed in terms of the model's robustness to adversarial attacks and the effectiveness of adversarial training in improving robustness, as visualized in figure 7.

- *Experiment 3* This experiment focuses on answering the error analysis research question, to what extent the model's capacity influences the accuracy of classifying Stop-signs and resistance toward adversarial attacks. To do this, the images with the stop-sign label will be labelled and act as a separate 'data set'. After separating the images, experiment 1 will be repeated for the I-FGSM attack to answer the research question. The results of these experiments will be compared to determine the effect of the model's capacity on the accuracy of stop-sign classification and resistance to adversarial attacks.

5.5 *Evaluation metrics*

The experiments will be evaluated based on the following metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

This metric measures the accuracy of the model's predictions. In this case, it shows how well the model can correctly classify the images. The score is calculated by dividing the number of correct predictions by the total number of predictions.

Attack Success Rate (ASR), Empirical Robustness Measure Empirical robustness measures the ability of a machine learning model to perform well on unseen data and can be evaluated by assessing the model's generalization performance. In other words, it measures how well a model generalizes to new, unseen data (Moosavi-Dezfooli et al., 2016). Empirical robustness can be evaluated by calculating the percentage of test examples that are misclassified. The attack success rate (ASR) represents the proportion of adversarial examples that are misclassified by a model compared to the total number of adversarial examples generated. For example, an ASR of 80% would mean that out of 100 adversarial examples, 80 were misclassified by the model (Goodfellow et al., 2014).

$$ASR = \frac{Number.successful.adv.examples}{Number.adv.examples} * 100\% \tag{4}$$

6 RESULTS

This section will present the key outcomes of the experiments. The emphasis will be on the attack success rate (ASR) and accuracy, indicating how well the model performs on unseen data. Additionally, the effects of adversarial attacks will be analyzed to evaluate the model's robustness. The adversarial attacks simulated different scenarios in which the model may encounter adversarial perturbed inputs. The results of these attacks can provide insights into the model's ability to handle such inputs. The findings from this analysis will provide valuable insights into the performance and robustness of the models and can be used to identify areas for improvement.

6.1 *Baseline models*

The ResNet18 and ResNet50 models were initially trained on the standard data set without any data augmentation techniques to establish a benchmark for the model's performance. Both models were trained for 30 epochs using the Adam optimizer, with cross-entropy as the loss function. This was done to ensure that the model's performance is not affected by any data augmentation techniques and to provide a point of reference for comparison with the results obtained when data augmentation is used.

Table 1: Baseline model accuracy scores, based on 30 epochs

Models	Accuracy
ResNet18	0.83
ResNet50	0.85

The results obtained from the initial training provide a baseline for the model's performance. They will be a starting point for comparing the results obtained when data augmentation techniques are applied.

6.2 *Adversarial attacked models*

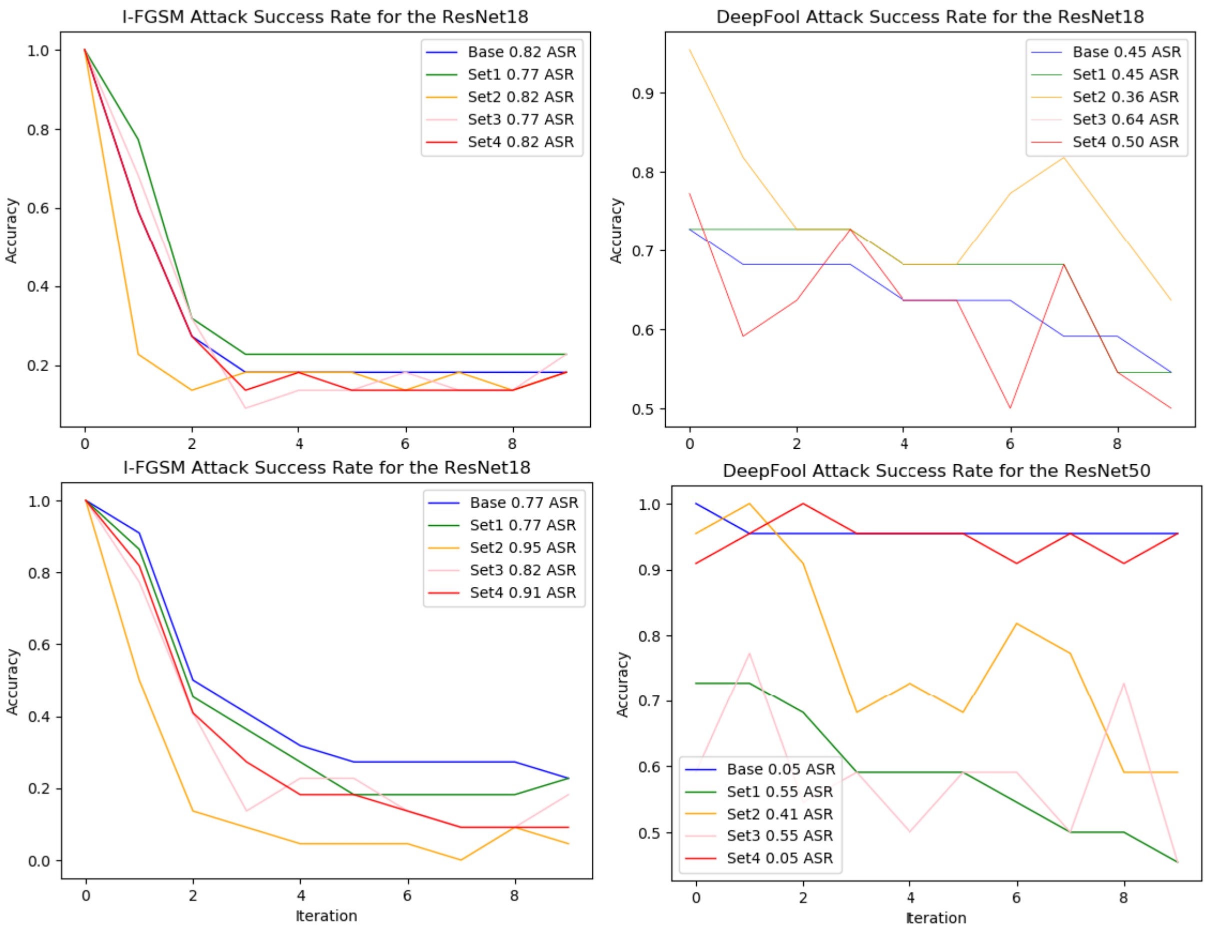
The results show that adversarial training of the ResNet18 model has helped to make it more resistant to adversarial samples. Specifically, when the model was tested against the DeepFool attack, the average attack rate was 0.45, and when tested against the I-FGSM attack, the average attack rate was 0.77. This suggests that the model's robustness has improved due to adversarial training, as it can correctly classify a higher percentage of adversarial samples. However, It is important to note that this does not guarantee that the model is robust to all types of adversarial examples and attack methods, as the loss over epochs is noticeably high. This goes in line with the research by, Tsipras et al. (2018).

The I-FGSM and DeepFool attack

The results of the experiment show that set 2 has a distinct trend in terms of attack success rate, with a lower average rate for the I-FGSM attack (0.41 and 0.36) but a higher rate for the DeepFool attack (0.95 and 0.82). This is concerning as it suggests that the model's defences are easily bypassed by the DeepFool attack. It, in theory, means that the model can only successfully classify 5% of the test data when using the ResNet50 model.

It is interesting that set 3 performs the worst on the I-FGSM attack, with an average success rate of 0.91. This could be linked to the ColorJiiter transform, as it randomly changes an image's brightness, contrast, and saturation in the "RGB" space. It would mean that the model would train on the underlying features, such as the shapes of the traffic signs instead of the colors. The same trend can be seen with the DeepFool method, with an average success rate of 0.32. Set 4 focuses on the ElasticTransform. This transformation technique adds a random distortion to the image

Figure 8: I-FGSM and DeepFool attack, ASR represents the average success rate



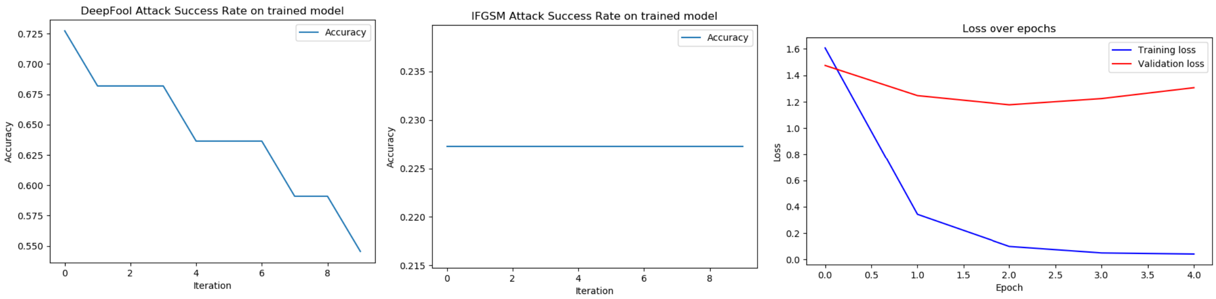
using a displacement field generated from Gaussian noise. This distortion can help to mimic certain real-world conditions, such as rainy weather, as the resulting image appears stretched or distorted. The effect of this transformation can be seen by the randomness in the average success rates. There is a tremendous difference between the DeepFool average success rate of 0.05 and 0.91 for the I-FGSM attack. It is remarkable to see that the model's accuracy rate decreases when data augmentation techniques to mimic real-world scenarios are applied, except for the DeepFool attack on the ResNet50 model, where the success rate is 0.05. This means that the model correctly classifies 95% of the images.

Table 2: Accuracy stop signs. In this context, Y stands for the actual label of the stop sign within the data set and is compared to the adversarial prediction. At the same time, Ypred represents the label predicted by the model and is also compared to the adversarial prediction.

	Accuracy score	
Models	**Y**	**Ypred**
ResNet18	0.01	0.01
ResNet18	0.98	0.98

6.3 *Adversarial trained models*

Figure 9: I-FGSM and DeepFool attack on the adversarially trained ResNet18



The results show that adversarial training of the ResNet18 model has helped to make it more resistant to adversarial samples. Specifically, when the model was tested against the DeepFool attack, the average attack rate was 0.45, and when tested against the I-FGSM attack, the average attack rate was 0.77. This suggests that the model's robustness has improved due to adversarial training, as it can correctly classify a higher percentage of adversarial samples. However, It is important to note that this does not guarantee that the model is robust to all types of adversarial examples and attack methods, as the loss over epochs is noticeably high. This goes in line with the research by Tsipras et al. (2018).

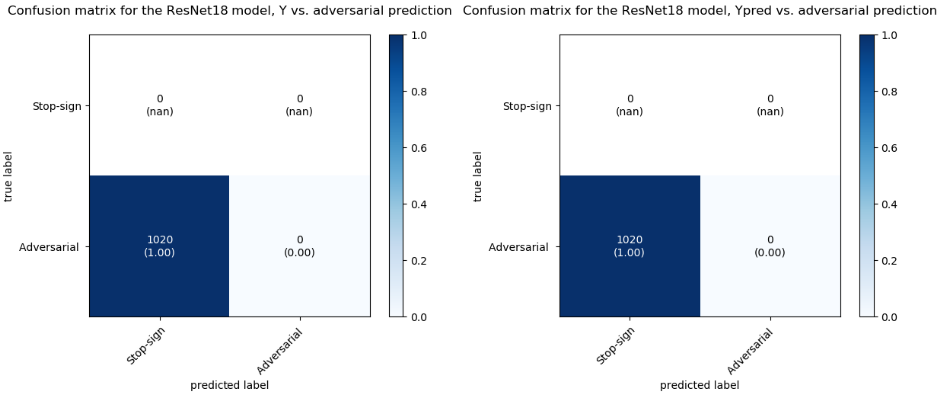
6.4 *Stop sign*

The results from experiment 3 show that the precision and recall of the ResNet50 model when comparing true labels to adversarial labels is 0. Similarly, the precision and recall of the ResNet50 model when comparing

Table 3: Performance Metrics

	Precision	Recall
ResNet50 (Y vs Adv)	0	0
ResNet50 (Ypred vs Adv)	0	0
ResNet18 (Y vs Adv)	0	0
ResNet18 (Ypred vs Adv)	0	0

Figure 10: Confusion matrix. Source: The author's illustration

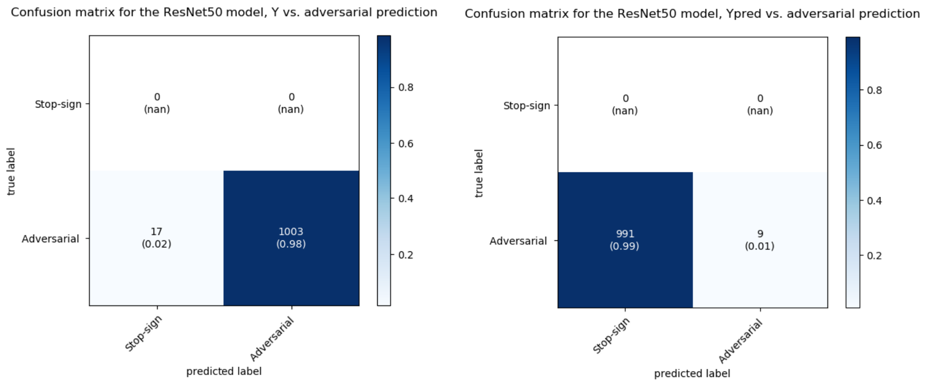


predicted labels to adversarial labels is also 0. The same results can be seen for the ResNet18 model when comparing true labels to adversarial labels and predicted labels to adversarial labels, with precision and recall both being 0. These results suggest that the models are not very robust to adversarial examples, and they can be improved. It is important to note that high precision and recall values are desirable and are often used as performance metrics, low values, like 0, indicate that the model is not performing well.

ResNet18

The confusion matrix, 10, illustrates that the model outputs are similar, with a high number of false positives (FP) when comparing the labels to the adversarial prediction. This indicates that the ResNet18 model is incorrectly classifying some non-stop sign images as stop signs. This phenomenon, known as a type 1 error, suggests that the model is not robust to adversarial examples and can be improved.

Figure 11: Confusion matrix. Source: The author's illustration



ResNet50

The confusion matrix, 11, illustrates that the model outputs are not consistent. When comparing the true label to the adversarial prediction, there are more true negatives (TN), which means that the model correctly identifies that the image is not a stop sign. However, when comparing the predicted label to the adversarial label, there are more false positives (FP), indicating that the model is incorrectly classifying some non-stop sign images as stop signs. This phenomenon, known as a type 1 error, suggests that the model is not robust to adversarial examples and can be improved.

7 DISCUSSION

7.1 *Error analysis*

The results of the confusion matrices indicate that it is impossible to make a simple conclusion about the effect of data augmentation techniques and adversarial attacks on the model's performance. The error analysis aimed to examine the impact of these factors by focusing on one traffic sign, the stop sign, and evaluating how well the model can classify it when it is presented with variations in color. The results showed an imbalance in the number of stop-sign images, with only 10 compared to the 1000 images used from the test data. However, when examining the effects of the other experiments, it is clear that the model struggles to correctly classify images when presented with variations that force it to focus on the underlying characteristics of the signs. This further emphasizes the need for robustness in deep learning models and the importance of considering different variations of data, including adversarial examples, when training and evaluating models.

7.2 *Expanding model capacity*

Research by Zhang et al. (2019) suggests that expanding a model's capacity enhances its robustness. Expanding the model's capacity, such as using a larger or deeper neural network, may not necessarily be the solution for improving the robustness of CNN models against adversarial attacks. While increasing the capacity of a model can potentially improve its overall performance, it may not address the underlying issues that make the model vulnerable to adversarial attacks. The research findings suggest that a combination of data augmentation techniques and adversarial training can be more effective in improving the robustness of CNN models. However, it is important to note that this study only evaluated two specific models, and it is possible that other models may have different results. Therefore, further research is needed to investigate the robustness of different models and to develop more robust models for recognizing traffic signs in the real-world

7.3 *Adversarial attacks*

The Attack Success Rate (ASR) is a commonly used measure in the literature to evaluate the effectiveness of adversarial attacks. It is defined as the proportion of adversarial examples that can successfully fool the model, i.e., the percentage of adversarial examples that are classified differently from their original class Carlini and Wagner (2017). In other words, it is the ratio of the number of adversarial examples that successfully evade the model's defences to the total number of adversarial examples generated. In this research, the ASR is calculated and reported as the primary evaluation metric to measure the robustness of the ResNet18 and ResNet50 CNN models against the DeepFool and I-FGSM attack. Goodfellow et al. (2014) describes the I-FGSM attack as being able to generate high-quality adversarial examples that are visually similar to original inputs, making them difficult for humans and machines to detect. The results of the research support this, as the I-FGSM attack has a higher success rate on both the ResNet18 and ResNet50 models compared to the DeepFool attack. Specifically, the average success rate per iteration of the DeepFool on the ResNet18 model is 0.48, and the average success rate of the I-FGSM is 0.80. On the ResNet50 model, the average success rate per iteration for the DeepFool attack is 0.32, while the I-FGSM has an average success rate of 0.84. The research also explores the effect of data augmentation techniques on the robustness of the models against adversarial attacks. The results show that the models are less resistant to the I-FGSM attack, specifically the ResNet50 model. Additionally, it is noteworthy that the model's accuracy decreases when data augmentation techniques, which aim to simulate real-world scenarios, are applied. This is observed with the exception of the DeepFool attack on the ResNet50 model, where the fourth set has an average success rate of 0.05. As previously stated in section 5.3, research by Rodriguez et al. (2019) suggests that rotating an image by one degree significantly affects the classification accuracy of a model. The results of the experiments confirm that the transformation does indeed impact the accuracy, but when compared to the average attack rate from the baseline, it is not as significant as suggested by Rodriguez et al. (2019).

Overall, the research results indicate that the ASR is a useful measure for evaluating the robustness of CNN models against adversarial attacks and that data augmentation techniques can significantly impact the models' robustness. Further research is needed to better understand the relationship between data augmentation techniques and the robustness of CNN models against adversarial attacks.

7.4 *Limitations*

One of the main limitations of this research is the limited scope of the models, data set, and adversarial attacks used. The research aims to find a combination of techniques that can improve the accuracy of traffic sign recognition, but to truly evaluate the effectiveness of these techniques, they should be tested on a wider range of models, data sets, and adversarial attacks. Additionally, the current project focuses mainly on white-box attacks, where the attacker has full knowledge of the model's architecture and parameters. It would be beneficial to also test the robustness of the models against black-box attacks, where the attacker does not have access to this information. Another limitation of this research is that it only focuses on the task of image classification, whereas expanding the scope to other tasks within the computer vision field, such as object detection, would give a more comprehensive understanding. Furthermore, the research lacks diversity in evaluation metrics, which makes it difficult to make conclusive statements about the robustness of the models.

7.5 *Further research*

To build on this research, future studies could investigate the effects of black-box and white-box attacks, as well as the use of transfer learning and a wider variety of models. For example, in the case of adversarial attacks, black-box attacks are those where the attacker does not have access to the internal workings of the model. In contrast, white-box attacks assume the attacker has full access. Comparing the results of these two types of attacks would provide a better understanding of the robustness of the model under different threat scenarios. Testing the research by Madry et al. further to see if the adversarial training actually works against other attacks would be a good next step to further validate the findings. Additionally, transfer learning, which is a technique to use the knowledge learned from one task to improve the performance of a different but related task, could be useful for improving the model's generalisation. Furthermore, testing the model on different data sets and a wider variety of models would allow us to understand how the model behaves and performs under different conditions. It would help to improve the robustness and generalization of the model.

7.6 *Scientific and societal relevance*

The findings and conclusions of this Master's thesis have the potential to benefit both the scientific community and specific industries. Developing a model that is reliable, robust, and capable of generalization across different subsets and tasks can have a positive impact on these communities. For the scientific community, the results of this project can provide valuable insights into the creation of generalizable and robust models and further aid in advancing the safety aspect. In the specific case of traffic sign recognition, this project's results can help build models that are more robust to real-world scenarios. This can help build trust in the implementation of ML and DL models in autonomous vehicles, which is crucial for the future of the automotive industry. Overall, the results of this project can be applied to improve the safety, efficiency and reliability of autonomous vehicles, which are an important part of our future transportation.

8 CONCLUSION

> *How do the integration of data augmentation methods and adversarial training impact the classification accuracy and resistance against adversarial attacks in recognition of traffic signs?*

In conclusion, the integration of data augmentation methods and adversarial training has a significant impact on the classification accuracy and resistance against adversarial attacks in recognition of traffic signs. The results of the experiments indicate that the use of data augmentation techniques can improve the robustness of the models, making them more resistant to different types of data variations. Additionally, the use of adversarial training, specifically on the ResNet18 model, has been shown to improve the model's resistance to adversarial attacks, as seen in the overall low average attack rate on the DeepFool and I-FGSM attacks. However, it is important to note that the training process resulted in a high loss over the epochs, and more research needs to be done to understand the trade-off between the robustness and generalization of the model. Furthermore, the results of this research are limited by the scope of the models, data set, and adversarial attacks used, and further research is needed to test the results on a more diverse set of models, data sets, and attacks

REFERENCES

Bayzidi, Y., Smajic, A., Hüger, F., Moritz, R., Varghese, S., Schlicht, P., & Knoll, A. (2022). Traffic sign classifiers under physical world realistic sticker occlusions: A cross analysis study. In *2022 ieee intelligent vehicles symposium (iv)* (pp. 644–650).

Carlini, N., & Wagner, D. (2017, March). Towards evaluating the robustness of neural networks.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, *6*(1), 25–45.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, *29*(6), 141–142.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, *28*(4), 594–611.

Freeman, D., & Chio, C. (2018). *Machine learning and security: Protecting systems with data and algorithms*. O'Reilly Media. Retrieved from https://books.google.nl/books?id=rl2ftgEACAAJ

Gokhale, T., Mishra, S., Luo, M., Sachdeva, B. S., & Baral, C. (2022, March). Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Harisubramanyabalaji, S. P., Nyberg, M., Gustavsson, J., et al. (2018). Improving image classification robustness using predictive data augmentation. In *37th international conference on computer safety, reliability, and security (safecomp), västerås, sweden, 18-21 september, 2018* (pp. 548–561).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., ... Oliphant, T. E. (2020, September). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

He, Z., Nan, F., Li, X., Lee, S.-J., & Yang, Y. (2020). Traffic sign recognition by combining global and local features based on semi-supervised classification. *IET Intelligent Transport Systems*, *14*(5), 323–330.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, *9*(3), 90–95.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, *32*.

Kim, H. (2020). *adversarial-attacks-pytorch.* https://github.com/Harry24k/adversarial-attacks-pytorch.

Kong, Z., Xue, J., Wang, Y., Huang, L., Niu, Z., & Li, F. (2021). A survey on adversarial attack in the age of artificial intelligence. *Wireless Communications and Mobile Computing*, *2021*.

Kurakin, A., Goodfellow, I., Bengio, S., Dong, Y., Liao, F., Liang, M., ... others (2018). Adversarial attacks and defences competition. In *The nips'17 competition: Building intelligent systems* (pp. 195–231). Springer.

Lipson, H., & Kurman, M. (2016). *Driverless: intelligent cars and the road ahead.* Mit Press.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

McKinney, W. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th python in science conference.* SciPy.

Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016, June). Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).*

Mykola. (2018). *Gtsrb - german traffic sign recognition benchmark.* https://www.kaggle.com/datasets/meowmeowmeowmeowmeow/gtsrb-german-traffic-sign.

Nicolae, M.-I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., ... Edwards, B. (2018). Adversarial robustness toolbox v1.2.0. *CoRR*, *1807.01069*. Retrieved from https://arxiv.org/pdf/1807.01069

Paszke, A., Gross, S., Chanan, G., Yang, E., Devito, Z., Lin, Z., ... Research, A. I. (n.d.). *Automatic differentiation in PyTorch.* https://openreview.net/pdf?id=BJJsrmfCZ.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., & Mann, T. A. (2021). Data augmentation can improve robustness. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 29935–29948). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2021/file/fb4c48608ce8825b558ccf07169a3421-Paper.pdf

Rodriguez, R., Dokladalova, E., & Dokládal, P. (2019). Rotation invariant cnn using scattering transform for image classification. In *2019 ieee international conference on image processing (icip)* (pp. 654–658).

Sermanet, P., & LeCun, Y. (2011). Traffic sign recognition with multiscale convolutional networks. In *International joint conference on neural networks (ijcnn)* (pp. 2809–2813). Retrieved from http://yann.lecun.com/exdb/publis/pdf/sermanet-ijcnn-11.pdf

Shi, Y., & Han, Y. (2018). Schmidt: Image augmentation for black-box adversarial attack. In *2018 ieee international conference on multimedia and expo (icme)* (pp. 1–6).

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, *6*(1), 1–48.

Stallkamp, J., Schlipsing, M., Salmen, J., & Igel, C. (2011). Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, *32*, 323–332.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013, December). Intriguing properties of neural networks.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*.

Tuna, O. F., Catak, F. O., & Eskil, M. T. (2022). Exploiting epistemic uncertainty of the deep learning models to generate adversarial samples. *Multimedia Tools and Applications*, *81*(8), 11479–11500.

Umesh, P. (2012). Image processing in python. *CSI Communications*, *23*.

Waskom, M. (2021, April). seaborn: statistical data visualization. *J. Open Source Softw.*, *6*(60), 3021.

Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L. E., & Jordan, M. (2019, 09–15 Jun). Theoretically principled trade-off between robustness and accuracy. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 7472–7482). PMLR. Retrieved from https://proceedings.mlr.press/v97/zhang19p.html

APPENDIX A

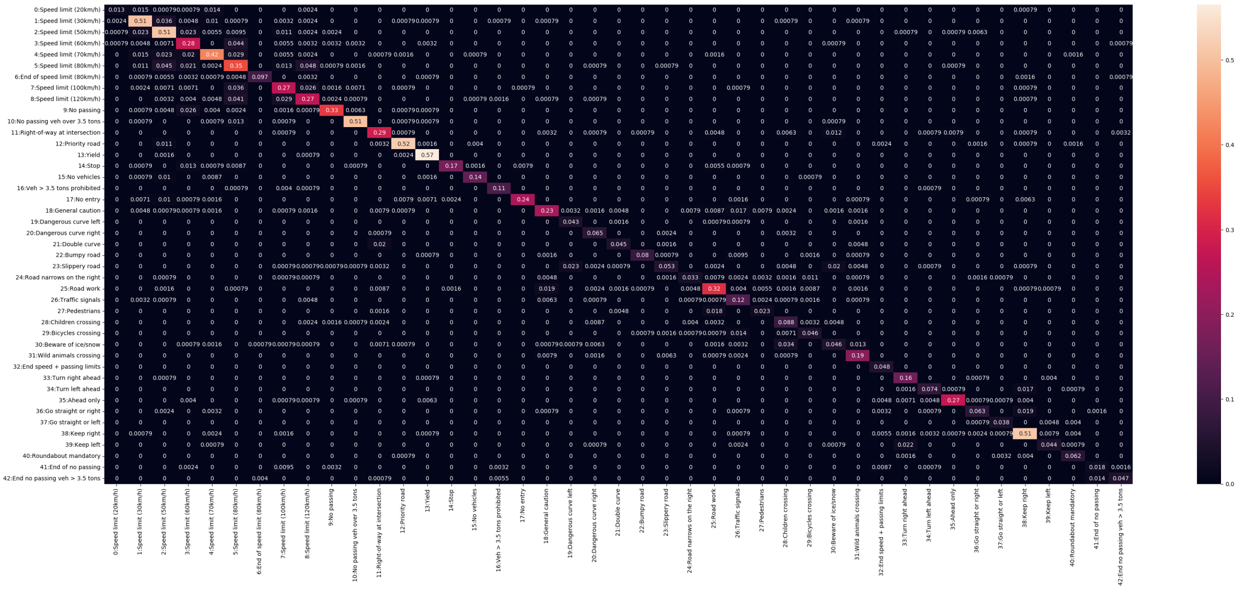


Figure 12: Confusion matrix for the ResNet18 base model

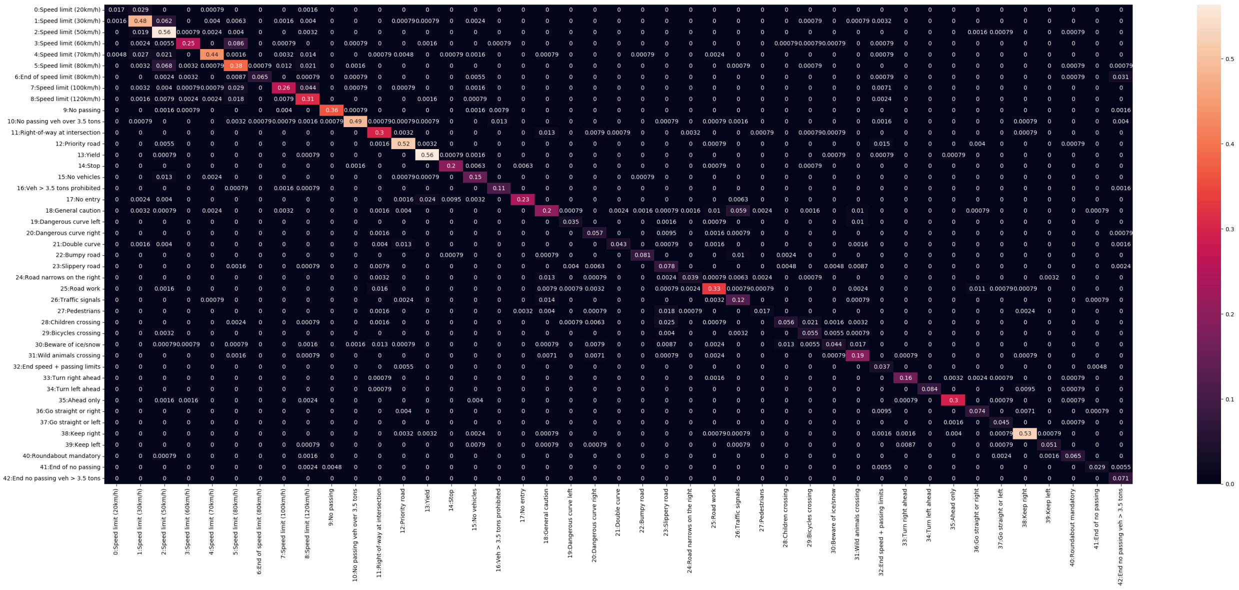


Figure 13: Confusion matrix for the ResNet50 base model