Project Report for Robustness in Autonomous Driving Steering

Janani Karthikeyan
7010329

Nischal Maharjan
7058343

Swathi Suhas
7057329

Abstract

Autonomous driving technology is revolutionizing the automotive industry, offering the promise of vehicles capable of navigating and making decisions independently. However, ensuring the robustness of autonomous driving systems across diverse scenarios remains a critical challenge. This project investigates various failure scenarios in autonomous driving steering systems and develops algorithms to enhance their robustness. Leveraging the CARLA Steering Dataset for Self-Driving Cars, the study explores classification-based steering angle prediction. Through literature review and experimentation, the project examines adversarial attacks, adversarial training, ensemble methods, and model explainability techniques. Results demonstrate the efficacy of adversarial training and ensemble methods in improving model resilience against adversarial attacks, highlighting the importance of developing robust autonomous driving systems for safety and reliability.

1 Introduction

Autonomous driving technologies represent a paradigm shift in the automotive industry, promising a future where vehicles can navigate and make decisions independently. This transformative potential has spurred significant research and development efforts aimed at realizing the vision of fully autonomous vehicles. However, achieving robustness in autonomous driving systems remains a critical challenge.The ability of autonomous vehicles to operate seamlessly and safely across diverse scenarios is paramount to their widespread adoption and acceptance. Failures in autonomous driving systems, particularly in steering mechanisms, can have serious implications for both passengers and pedestrians. Ensuring the reliability and resilience of these systems is thus a pressing concern for researchers and industry stakeholders alike.

In response to this challenge, our project investigates various failure scenarios in autonomous driving steering systems and develops algorithms to enhance their robustness. Leveraging the CARLA Steering Dataset for Self-Driving Cars as a foundation, the study delves into classification-based steering angle prediction. Through a comprehensive literature review and rigorous experimentation, the project explores adversarial attacks, adversarial training, ensemble methods, and model explainability techniques. The findings of this research shed light on the efficacy of different approaches in improving the resilience of autonomous driving systems against adversarial attacks. By identifying vulnerabilities and proposing mitigation strategies, the project contributes to the ongoing efforts to develop robust autonomous driving technology for enhanced safety and reliability on the roads of tomorrow.

Preprint. Under review.

2 Literature Review

Visual perception and control has been one of the major sectors in the concept of autonomous driving. The susceptibility of perception models to failures highlights the importance of robustness. As mentioned by Tian et al. [1] DNN models often lead to incorrect unexpected behaviors that might cause fatal accidents. Shen et al. [2] proposed an iterative min-max training that improves robustness of learning-based steering. The difference between real-world and simulated data is measured using the metric Fréchet Inception Distance (FID). The model is then trained on a combination of real data and degraded data. Mean Accuracy is used as the evaluation metric for the model. The author tests on clean data, single-perturbation data, multi-perturbation data and previously unseen data. The author also proposes a method for measuring degradation in images and its effects on output using sensitivity analysis.The proposed method for sensitivity analysis in Tunkiel et al. [3] uses a numerical approach to estimate how the model's output responds to changes in the model's input. Sensitivity evaluation is done using one-at-a-time sensitivity analysis where one input parameter is varied at a time while others are kept constant.

Robustness to distribution shift is critical when it comes to autonomous driving. Wiles et al. [4] discussed different approaches including data augmentation and weighted re-sampling. Pretraining proved to be an extremely powerful tool. Out-of-distribution scenarios need to be tackled with care and Filos et al. [5] proposed a new autonomous car novel-scene benchmark called CARNOVEL which is created on top of the CARLA simulator and helps train the model with a suite of tasks with distribution shifts. Other research like in Paul et al. [6] shows that along with pretraining methods, an ensemble-based classifier can make the model more robust.

Dietterich [7] provides a comprehensive review of ensemble learning, particularly in the context of deep learning. It discusses various strategies for successful ensemble methods, factors influencing their efficacy, and reviews recent research efforts in multiple domains. Additionally, it analyzes trends in ensemble learning research and explores the foundational principles and characteristics of ensemble learning systems, including data sampling strategies and training methodologies for baseline classifiers.

Luo and Liu [8] explores the efficacy of ensemble methods in improving classification accuracy by leveraging the wisdom of the crowd concept. It discusses how ensemble methods, initially proposed to enhance supervised and semi-supervised learning by amalgamating multiple learning models, have gained traction in machine learning. The majority voting rule, exemplified by algorithms like Random Forest, is the prevailing method for aggregation in ensemble techniques. However, the assumption that the majority answer is always correct may not hold in certain scenarios, challenging the effectiveness of traditional ensemble methods. Inspired by Bayesian Truth Serum (BTS), which incentivizes truthful human judgments, the paper proposes Machine Truth Serum (MTS) to augment ensemble learning. MTS aims to detect minority answers that are more likely to be correct, even when the majority is wrong, by mimicking the concept of "surprising" popularity. The proposed methods, Heuristic MTS (HMTS) and Discriminative MTS (DMTS), utilize regression and classification models, respectively, to estimate peer prediction information and make informed decisions on minority answers. The paper's contributions include the introduction of novel ensemble methods to complement existing techniques, promising experimental results across real-world datasets, and analytical validation of the proposed approaches' correctness.

3 Methodology

3.1 Dataset and Base Classifier

For this project, we have leveraged the CARLA Steering Dataset for Self-Driving Cars introduced in Hassan [9]. This dataset is comprised of road images and corresponding steering angles captured using the CARLA simulator. The data was collected by simulating a Tesla Model 3 vehicle, in a diverse range of road scenarios and weather conditions (Figure 1). The original dataset is provided as

a regression problem with floating steering values for each image. For simplicity of study for this project, we have processed the labels into classes and considered it as a classification problem.

For our base classifier, we have considered a simple CNN for ease of interpretability and training complexity. The network consists of a series of convolutional layers followed by maxpool layers and at the end stacked with the fully connected layers as a classification head with 3 output units as shown in Figure 2 (Left).

Left Steer Straight Steer Right Steer

Figure 1: Sample images from Dataset

3.2 Adversarial Attack

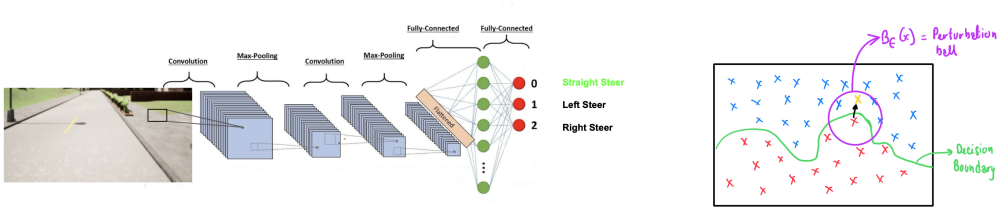


Figure 2: (Left):Base classifier and (Right) Adversarial attack

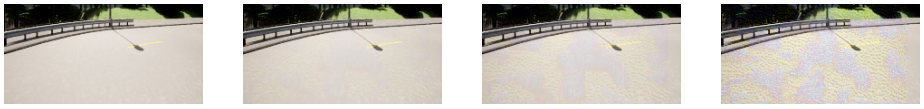
An adversarial attack is a technique employed in the field of machine learning to intentionally manipulate or deceive a model's predictions or behavior. Based on training data the model learns a decision boundary which it uses later during inference time to predict unseen samples. The goal of an adversarial attack is to cause the model to misclassify or produce an incorrect output by introducing carefully crafted perturbations into the input data. The addition of perturbation in such a way that the sample is pushed beyond the decision boundary to another region results in the misclassification of the sample. As shown in figure 2(Right) the yellow sample is a perturbed sample of the red class which is now predicted as the blue class. These perturbations are often imperceptible to humans but can significantly alter the model's decision-making process. So the amount of perturbation should be restricted such that it maintains the originality of samples. This is defined by the perturbation ball denoted by ϵ. One of the adversarial attacks is the Projected Gradient Descent Attack(PGD) which we have experimented with in this project described in detail below sections.

3.2.1 Untargeted PGD Attack

PGD attack is an adversarial attack where we add noise to the image iteratively using the image gradients as guidance. The noise is injected to increase the loss which is done by moving in the direction of the gradient of the loss function as expressed in 1. We want to maximize the loss w.r.t. the true class thus misclassifying the sample.

$$\max_{x' \in B_\epsilon(x)} L(\theta; x', y) \quad \Rightarrow \quad x_{t+1} = clip_{[-\epsilon,\epsilon]} \left[x_t + \alpha \cdot sgn \left(\frac{\partial}{\partial x} L(\theta; x_t, y) \right) \right] \tag{1}$$

As mentioned the amount of perturbation is dependent upon the hyperparameter ϵ. We have experimented with different values of ϵ as shown in Figure 3. As we increased the value of ϵ the artifacts are more distinct in the images.



Original Clean sample $\epsilon = 0.01$ $\epsilon = 0.02$ $\epsilon = 0.05$

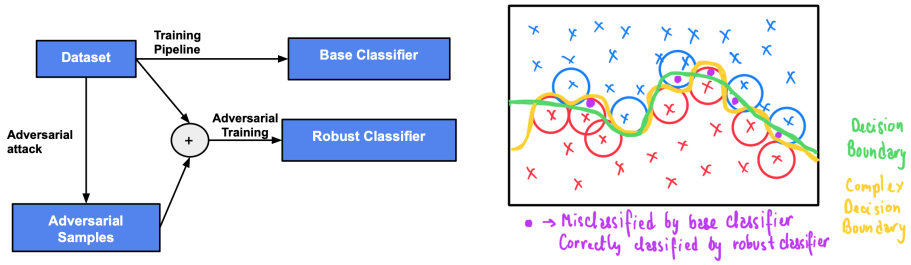
Figure 3: PGD attack with different Perturbation balls

3.2.2 Targeted PGD Attack

In an untargeted PGD attack, the goal was to increase the loss function concerning the true class, given that the class it is being misclassified into was not the concern. Whereas in the targeted PGD attack we specifically want a particular sample to be misclassified to target class. Thus here instead of maximizing the loss w.r.t to the true class we now minimize the loss w.r.t the target class. This results in an update rule as shown in Equation 2.

$$\min_{x' \in B_\epsilon(x)} L(\theta; x', y_{targ}) \quad \Rightarrow \quad x_{t+1} = clip_{[-\epsilon,\epsilon]} \left[x_t - \alpha \cdot sgn \left(\frac{\partial}{\partial x} L(\theta; x_t, y_{targ}) \right) \right] \tag{2}$$

3.3 Adversarial Training



Block diagram Effect of Adversarial Training

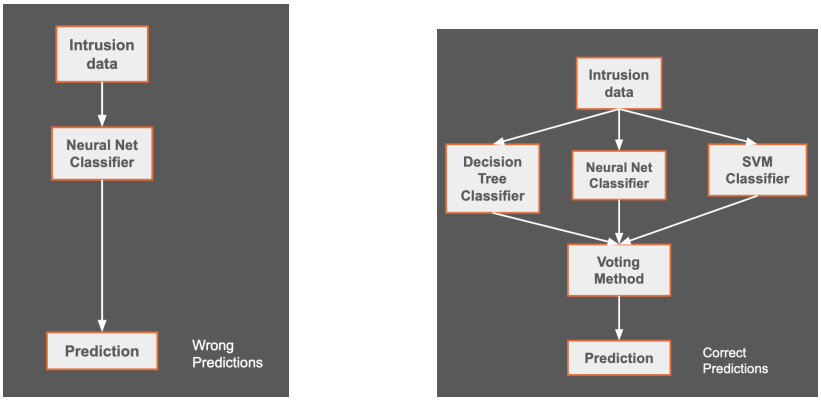
Figure 4: Adversarial Training

$$\min_{\theta} E_{(x,y)\ \mu} \left[\max_{x' \in B_\epsilon(x)} L(\theta; x', y) \right] \tag{3}$$

The adversarial attacks are a prominent issue to the robustness of the model. And one of the solutions for such attacks is adversarial training. We first generate the adversarial samples using the adversarial attack then include those samples as well in the training set. This results in min-max optimization as shown in equation 3. This allows the model to be robust to such perturbations as well. The effect of adversarial training is shown in figure 4. The adversarial training allows the model to learn complex boundaries that are immune to adversarial attacks. The perturbed samples misclassified previously are now correctly classified as shown in the figure.

3.4 Ensemble Method

Besides adversarial training ensemble methods are also one of the prominent ways to tackle adversarial attacks. We train multiple models to solve the same problem. In our experiments, we have considered Decision trees and SVM in addition to the neural networks as shown in Figure 5. Since a PGD attack is a white box attack that uses model parameters to compute the gradient, the adversarial samples are generated concerning the neural network and thus can fool it. However, the learning of decision trees and SVM is different than that of neural networks thus the adversarial samples that fool neural networks, decision trees, and SVM should obtain better performance on them as well. Thus taking into account the prediction from multiple models and generating the final prediction leads to better results and it allows the model to be immune to adversarial attacks to a certain extent.



Base Classifier Multiple classifier

Figure 5: Ensemble Method

3.5 Explainability of the model using Visualization

3.5.1 Gradient Visualization

The neural network is a complex and black box model we are not able to interpret the result directly and how the model is generating the predictions. It might be that the model is predicting using some irrelevant features. Hence for explainability, we have visualized gradients to have a better understanding of which features are being used to generate the predictions.

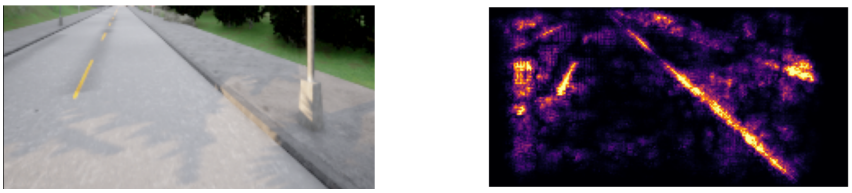


Figure 6: Visualization of gradients

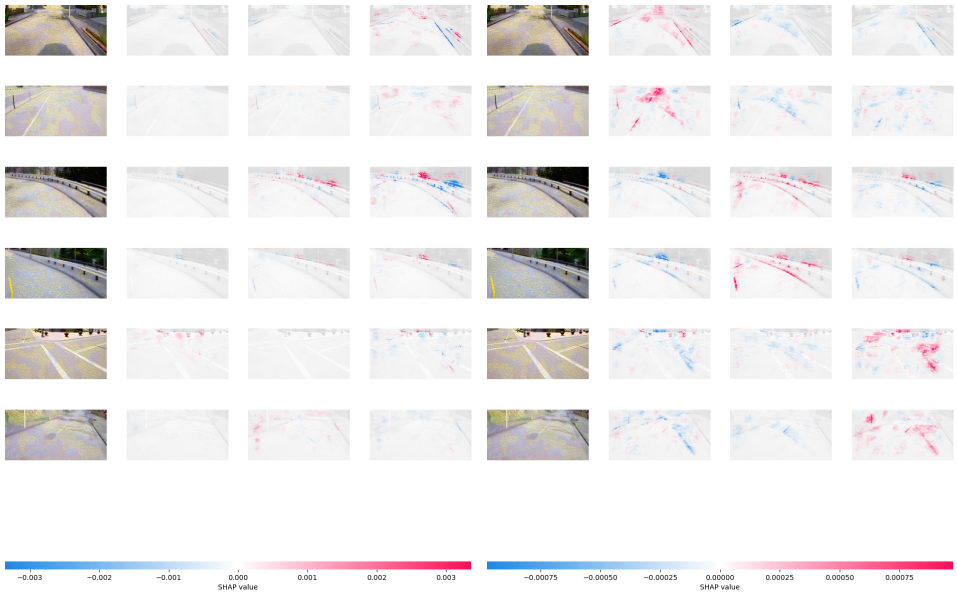
As shown in Figure 6 we can see that for the image provided the lane lines and the edge of roads have high activations hence model is using these features to generate the predictions.

3.5.2 SHAP - DeepExplainer

As mentioned earlier, understanding the inner workings of machine learning models can be challenging due to their complexity. Fortunately, there are now various tools available to explain model behavior. One such tool is SHAP (Shapley Additive exPlanations), which offers global interpretability by quantifying the importance of specific features. It achieves this by calculating the average marginal contribution of a feature value across all possible combinations.

In our project, we utilized SHAP's DeepExplainer module, specifically designed for deep neural network architectures. Upon applying DeepExplainer, we obtained insightful results, as depicted in Figure 7. Each point on the graph is color-coded to represent the value of the corresponding feature, with red indicating high contributions and blue indicating low contributions.

Initially, when visualizing SHAP outputs from the base classifier, discerning which features contribute to the model's output was challenging. However, upon applying the tool to the robust classifier, trained on perturbed data, a stark contrast emerged. Clear red lines delineated significant regions of the image that strongly influenced the model's output. This enhancement in interpretability was particularly evident when comparing SHAP visualizations for straight, left, and right data samples.



Visualization on base classifier Visualization on robust classifier

Figure 7: SHAP Visualization

4 Results

4.1 Success of PGD attack

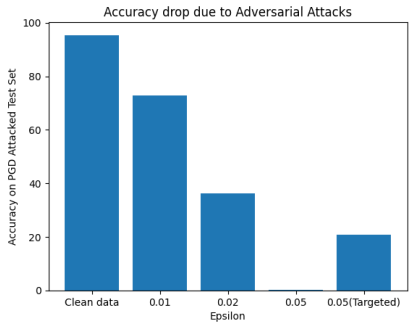


Figure 8: Accuracy Drop using PGD Attack

Perturbation	Test Accuracy
Clean Data	95.39%
0.01	72.76%
0.02	36.16%
0.05	0.16%
0.05(Targeted)	20.75%

Table 1: Accuracy on Test set for different value of epsilon

As shown in Figure 8 the accuracy on the test set decreased when the samples were subjected to the PGD attack. As the value of perturbation is increased the accuracy drop was proportional to it.

Also the effect of targeted attack and untargeted attack is shown here. Since the targeted attack is more constrained we need to perturb in direction of targeted class the success rate of targeted PGD attack is less than that of untargeted attack. Using the ϵ value of 0.05, untargeted attack we achieve accuracy of 0.16% whereas in case of targeted attack 20.75% accuracy is still retained. So the model misclassifies almost all samples in untargeted attack but it find difficult to misclassify certain samples in targeted attack.

4.2 Robustness due to Adversarial Training

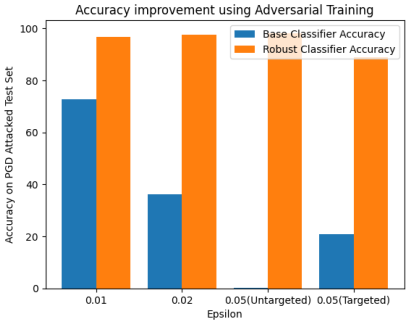


Figure 9: Accuracy Improvement due to Adversarial Training

Perturbation	Base Classifier Accuracy	Robust Classifier Accuracy
0.01	72.76%	96.62%
0.02	36.16%	97.59%
0.05	0.16%	98.23%
0.05(Targeted)	20.75%	88.92%

Table 2: Accuracy of Base Classifier vs Robust Classifier

Figure 9 shows the result of adversarial training on the test set. We see that the adversarial trained robust classifier outperforms base classifier trained on clean dataset when subjected to adversarial attacks.

4.3 Robustness due to Ensemble Method

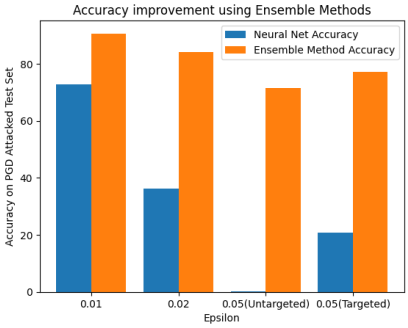


Figure 10: Accuracy Improvement due to Ensemble Method

Perturbation	Neural Network Accuracy	Ensemble Method Accuracy
0.01	72.76%	90.58%
0.02	36.16%	84.10%
0.05	0.16%	71.48%
0.05(Taargeted)	20.75%	77.04%

Table 3: Accuracy of Neural Net Classifer vs Ensemble Method

Figure 10 shows the result of ensemble methods on the test set. Compared to neural network stand alone there is increment in performance of ensemble method. Thus ensemble method is comparatively immune to the adversarial attacks.

5 Future Enhancement

Our project is relatively small in scope due to the time constraint we had but there is immense possibility of its future enhancement. To begin with we have experimented with l_∞ norm. We can also experiment with the l_2 norm and see its effect on adversarial attack and training. As mentioned in section 3.1 we have simplified the problem statement into a classification problem to study the algorithm. Integrating regression models into the existing framework offers opportunities to improve the system's steering angle estimation. By incorporating regression techniques, the system could aim to enhance its control mechanisms, enabling finer control and adaptability across diverse driving scenarios. This integration would allow the vehicle to respond more accurately to subtle changes in road conditions, ultimately enhancing overall safety and efficiency.

Another major issue that occurs in Autonomous driving is due to distribution shift. Distribution shifts occur when the data distribution during training differs from that encountered during deployment. Future research could focus on investigating the impact of distribution shifts on system performance and resilience. Understanding and mitigating these shifts are crucial for developing more robust algorithms capable of handling diverse and evolving scenarios effectively.

Environmental factors such as weather conditions, road surfaces, and traffic patterns significantly influence driving dynamics. Future work in this area could aim to develop algorithms that can adapt to real-time environmental changes, ensuring safe and efficient navigation in diverse conditions. By accounting for these variables, the autonomous driving system can optimize its decision-making processes, improving overall performance and reliability.

Model explainability techniques such as SHAP as discussed above offer insights into the decision-making processes of complex machine learning models. Further exploration and refinement of these techniques could improve transparency and interpretability. By gaining deeper insights into the inner workings of the system, research can identify areas for optimization and refinement, enhancing overall system performance and reliability.

6 Reflections

Reflecting on the journey of investigating robustness in autonomous driving steering systems, several key insights emerge. Firstly, the project underscored the critical importance of robustness in ensuring the safety and reliability of autonomous vehicles. As autonomous driving technology continues to advance, the need to develop resilient systems capable of operating seamlessly across diverse scenarios becomes increasingly apparent.

Secondly, the exploration of various scenarios where steering failures may occur highlighted the complexity of the problem domain. From adversarial attacks to distribution shifts, the system must contend with a wide range of challenges that can impact its performance. Understanding these challenges is essential for developing effective mitigation strategies and enhancing system robustness.

Additionally, the integration of regression models and the consideration of environmental factors offer promising avenues for future research. By improving steering angle estimation and adapting to real-time environmental changes, autonomous vehicles can navigate more safely and efficiently in diverse conditions. This highlights the importance of continually refining and optimizing algorithms to meet the evolving needs of autonomous driving technology.

7 Conclusion

In conclusion, our project highlights the crucial need for resilient autonomous driving systems to ensure passenger and pedestrian safety. While we initially achieved promising accuracy rates with our base classifier, our exploration uncovered vulnerabilities to adversarial attacks, emphasizing the necessity for robust defense mechanisms.

To mitigate these vulnerabilities, we employed PGD training and ensemble methods, enhancing our model's resilience and generalization capabilities. Visualizations such as gradient analysis and SHAP provided valuable insights into our model's decision-making processes, aiding in understanding its strengths and weaknesses.

The incorporation of ensemble methods significantly improved accuracy and resilience by leveraging the strengths of diverse classifiers. Looking ahead, integrating regression models and considering environmental factors offer avenues for further enhancement. Our project underscores the dynamic nature of autonomous driving research, emphasizing the importance of innovation and vigilance in addressing emerging challenges and advancing technology.

References

[1] Yuchi Tian et al. "DeepTest: Automated Testing of Deep-Neural-Network-Driven Autonomous Cars". In: ICSE '18. Gothenburg, Sweden: Association for Computing Machinery, 2018, pp. 303–314. ISBN: 9781450356381. DOI: 10.1145/3180155.3180220. URL: https://doi.org/10.1145/3180155.3180220.

[2] Yu Shen et al. "Improving Robustness of Learning-based Autonomous Steering Using Adversarial Images". In: *ArXiv* abs/2102.13262 (2021). URL: https://api.semanticscholar.org/CorpusID:232068823.

[3] Andrzej T. Tunkiel, Dan Sui, and Tomasz Wiktorski. "Data-driven sensitivity analysis of complex machine learning models: A case study of directional drilling". In: *Journal of Petroleum Science and Engineering* 195 (2020), p. 107630. ISSN: 0920-4105. DOI: https://doi.org/10.1016/j.petrol.2020.107630. URL: https://www.sciencedirect.com/science/article/pii/S0920410520306975.

[4] Olivia Wiles et al. "A fine-grained analysis on distribution shift". In: *arXiv preprint arXiv:2110.11328* (2021).

[5] Angelos Filos et al. "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3145–3153.

[6] Rahul Paul et al. "Mitigating Adversarial Attacks on Medical Image Understanding Systems". In: *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. 2020, pp. 1517–1521. DOI: 10.1109/ISBI45749.2020.9098740.

[7] Thomas G. Dietterich. "Ensemble Methods in Machine Learning". In: *Multiple Classifier Systems*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 1–15. ISBN: 978-3-540-45014-6.

[8] Tianyi Luo and Yang Liu. "Machine truth serum: a surprisingly popular approach to improving ensemble methods". In: *Machine Learning* 112.3 (2023), pp. 789–815.

[9] Imtiaz Ul Hassan. *CARLA ,Steering DataSet for Self Driving Cars*. 2021. DOI: 10.34740/KAGGLE/DS/1285878. URL: https://www.kaggle.com/ds/1285878.