

Systemic Adversarial Vulnerability in Deep Learning Classifiers: A Cross-Model Experimental Analysis

Rahul J. Teradal
Design Verification Trainee
Sion Semiconductors Pvt Ltd
Bengaluru, India
rahulteradal45@gmail.com

Misbah Mansoor Nabiwale
Department of ECE
Jain College of Engineering
Belagavi, India
misbahnabiwale@gmail.com

Abstract—Adversarial attacks demonstrate the existence of severe weaknesses in deep learning image classifiers due to misclassification, caused by a small, precisely-designed input perturbation. This work aims to experimentally test the adversarial robustness and the extent to which adversarial vulnerability can be model-dependent or systemic. There are two controlled experiments. First, on a pretrained ResNet-18 model, Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and Carlini-Wagner (CW) attacks are applied to discuss the impact of attack sophistication. Second, cross-model robustness is compared by means of eight pretrained image classification models to evaluate FGSM for cross-model robustness. The findings indicate that optimization-based attacks are much better than single-step attacks and more complex architectures are more robust and lightweight models are more sensitive. These results suggest that adversarial vulnerability is architecture-dependent and exists irrespective of a model. The paper also discusses defense and recovery processes and points out how the principles of Responsible AI can be used to construct high-quality and trustworthy adversarial defenses.

Index Terms—Adversarial Attacks; Deep Learning; Image Classification; Model Robustness; Responsible AI

I. INTRODUCTION

Deep learning models, particularly Convolutional Neural Networks (CNNs), have achieved state-of-the-art performance in image classification and are widely deployed in real-world applications such as autonomous systems, medical imaging, and biometric security. Despite their high accuracy on clean benchmark data, these models are vulnerable to adversarial attacks, which introduce small, carefully crafted perturbations to input images that cause incorrect predictions while remaining imperceptible to human observers.

Adversarial attacks are intentional ways of manipulating the model inputs to provoke misclassification at inference time. The initial demonstration of systematic failure of deep neural networks to small perturbations in inputs was done by [1], who showed that they are not due to random occurrence but instead intrinsic to the behavior of the model. The next work of [2] demonstrated that this weakness is due to the roughly linear behavior of neural networks operating in high-dimensional spaces of inputs, which makes it easy to attack it using gradients.

Fast Gradient Sign Method (FGSM) is among the well-researched adversarial approaches that perturb a single input image in one step along the loss gradients and is compu-

tationally efficient and simple to implement [2]. Projected Gradient Descent (PGD) builds on top of FGSM by adding repeated perturbations subject to a constraint and generates much stronger attacks, as well as being a standard benchmark in adversarial robustness testing [3]. Optimization-based attacks such as the Carlini-Wagner (CW) attack further demonstrate that adversarial vulnerability persists even under strict perturbation budgets and can bypass many existing defense mechanisms [4].

Adversarial attacks are particularly critical in safety-sensitive and security-critical domains, where incorrect predictions can lead to physical harm, financial loss, or privacy violations. Inference-time evasion attacks are especially concerning because they do not require access to training data and can be executed on deployed systems. Large-scale benchmarking studies have confirmed that adversarial vulnerabilities persist across a wide range of ImageNet-trained architectures, even when modern training practices and defense strategies are applied [5].

Several defense strategies have been proposed to mitigate adversarial threats, including adversarial training, input pre-processing, ensemble methods, and robust optimization [3], [6]. Although adversarial training enhances resistance to well-known attacks, it is computationally intensive and is in many cases inapplicable to novel or evading adversaries. Additionally, other defenses are based on gradient obfuscation [7], giving them false strength with respect to enhanced attacks. Consequently, there is no single defense strategy that can be considered comprehensive adversarial robustness.

Outside technical aspects, adversarial robustness has significant consequences to Responsible AI (RAI). Any form of adversarial failure may compromise robustness, fairness, transparency, and accountability and make AI-based decision-making systems less trustworthy, creating a barrier to their application in society [8]–[10]. Handling the adversarial threats thus needs more than effective technical defenses but governance-conscious and human-oriented approaches, as well.

Motivated by these problems, this paper presents an experimental research study of adversarial attacks inflicted on deep learning image classifiers with the following contributions:

- An experimental comparison of FGSM, PGD and CW attacks on a pretrained ResNet-18 to investigate the

impact of attack sophistication.

- A cross-architecture evaluation using FGSM across eight pretrained image classification models to investigate

whether adversarial vulnerability is model-specific or systemic.

- An examination of defense and recovery mechanisms in the context of adversarial robustness.
- A discussion of adversarial robustness from a RAI perspective.

These contributions provide a concise and experimentally grounded view of adversarial attacks, defenses, recovery strategies, and Responsible AI considerations in deep learning systems.

II. RELATED WORK

Adversarial examples were first identified as a fundamental vulnerability in deep neural networks by [1], who showed that imperceptible input perturbations can reliably cause misclassification. [2] attributed this vulnerability to the approximately linear behavior of neural networks in high-dimensional spaces

and introduced the FGSM as an efficient gradient-based attack.

Subsequent studies proposed stronger adversarial attacks to evaluate robustness under more realistic threat models. Madry et al. [3] introduced PGD, which became a standard first-order adversary for robustness benchmarking. [4] proposed a CW optimization-based attack that bypasses many defenses even under strict perturbation constraints, demonstrating the persistence of adversarial vulnerability.

Research on adversarial transferability showed that adversarial examples can generalize across models and architectures, even in black-box settings [11]. Large-scale benchmarking further confirmed that adversarial vulnerabilities remain prevalent across modern ImageNet-trained architectures despite advances in training and defense strategies [5].

To mitigate adversarial threats, various defenses have been proposed, including adversarial training, input preprocessing, ensemble methods, and robust optimization techniques such as TRADES [3], [6]. However, many defenses rely on gradient obfuscation and fail under adaptive or optimization-based attacks, providing only an illusion of robustness [7]. Recovery mechanisms following successful attacks remain comparatively underexplored.

More recently, adversarial robustness has been discussed within the context of RAI. Ethical and governance frameworks emphasize robustness, transparency, accountability, and human oversight as essential properties of trustworthy AI systems [8]–[10]. While adversarial attacks directly undermine these principles, their explicit integration into empirical robustness studies is still limited.

A concise comparison of prior work and the specific contributions of this study is summarized in Table I.

III. THREAT MODEL AND ADVERSARIAL ATTACKS

An adversarial robustness threat model is a necessity to assess and interpret experimental findings. The work is dedicated

TABLE I
COMPARISON OF PRIOR WORK AND CONTRIBUTIONS OF THIS STUDY

Prior Work	Contribution of this study
Adversarial examples in DNNs [1]	Controlled experimental analysis of attack strength escalation.
Linearity explanation and FGSM [2].	Comparative evaluation of FGSM with PGD and CW attacks.
PGD as robustness benchmark [3].	Systematic study of increasing attack sophistication.
Optimization-based CW attacks [4].	Robustness assessment beyond first-order attacks.
Transferability and black-box attacks [11].	Cross-model analysis of systemic adversarial vulnerability.
Large-scale robustness benchmarking [5].	Complementary controlled experiments across architectures.
Defense mechanisms and adversarial training [3], [6].	Joint evaluation of attacks, defenses and recovery.
Broken defenses and gradient obfuscation [7].	Motivation for recovery-aware robustness analysis.
Responsible AI principles [8]–[10].	Integration of Responsible AI into adversarial robustness study.

to the problem of evasion attacks that are considered at inference time, some of the most realistic and most experimentally studied adversarial cases. All the attacks are assessed through the view of a white-box threat model, where the adversary can access the complete model architecture, parameters, and gradients. This assumption is usually taken as a worst-case standard in assessing adversarial robustness, despite it not being particularly strong, which is typical of the direction taken by [2] and [3].

The adversary attempts to create a perturbed input, i.e., $x' = x + \delta$, where the adversary has a clean input x with true label y , and the goal is to cause the classifier to make a false prediction,

$$f(x') \neq y, \quad (1)$$

and at the same time making the perturbation invisible to human observers. This is imposed by restricting the perturbation magnitude as

$$|\delta|_p \leq \epsilon, \quad (2)$$

where the margin of error is determined by the parameter, ϵ . In line with the convention of image-based adversarial attacks, the study adopts untargeted attacks with an ℓ_∞ -bounded perturbation as proposed in the literature of explaining adversarial attacks on images [2], [3].

Adversarial example generation may be considered as an optimization problem that maximizes the loss of the model as a function of the input subject to a norm constraint. Gradient information allows to efficiently approximate this goal in white-box environments, which are the foundation of the attack techniques analyzed in this study.

A. Gradient-Based and Optimization-Based Adversarial Attacks

This study compares three typical adversarial attack strategies, which use gradient information based on the specified

threat model. The attacks of the choice vary in their formulation and strength and are selected to represent the rising scales of adversarial sophistication. Particularly, FGSM and PGD are first-order gradient-based adversaries, and the Carlini–Wagner (CW) attack is an optimization-based adversary.

1) *Fast Gradient Sign Method (FGSM)*: FGSM is a first-order, single-step adversarial attack that pixel-wise perturbs the input in the direction of the gradient of the loss function with respect to the input-vectors [2]:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y)). \quad (3)$$

FGSM is computationally inexpensive and serves as a standard benchmark attack in testing adversarial vulnerability. Although simple, FGSM is effective to find vulnerabilities in deep learning classifiers.

2) *Projected Gradient Descent (PGD)*: PGD is a generalization of FGSM that uses successive small gradient updates and projects the perturbed input back into the authorized area within the permissible range ϵ -bounded of the gradient, which serves as an iterative first-order attack on the model [3]. PGD is well-known as a powerful first-order adversary and is usually employed as a standard in adversarial robustness testing. Here, PGD is employed to illustrate how iterative gradient-based attacks can significantly enhance attack effectiveness as opposed to those that are single-step.

3) *Carlini–Wagner (CW) Attack*: The CW attack posits adversarial example generation as a clear maximization issue that minimizes the magnitude of perturbations and imposes the misclassification constraint on the adversarial example generation process [4]. The CW attack is a direct optimization of a constrained objective function, in contrast to FGSM and PGD that make use of fixed perturbation steps. CW is optimized using gradient information; though it can also be described as an optimization-based adversarial attack as opposed to a first-order gradient-sign method. CW attacks can be found to overcome many of the existing defenses, and thus make a good adversarial baseline.

B. Alignment with Experimental Design

The experimental design of this study is directly informed by the threat model and attack selection: in the first experiment, the target is a single pretrained model to evaluate all three attack types (FGSM, PGD and CW) in order to determine whether the vulnerability to adversarial examples is systemic or model-specific. In the second experiment, FGSM attack was conducted on multiple pretrained models to determine if adversarial vulnerability is model-specific or systemic. This correspondence will provide a consistent assessment of the efficacy of attacks as well as the cross-model robustness within a single threat model.

IV. EXPERIMENTAL SETUP

This section describes the experimental design used to evaluate adversarial vulnerability in deep learning image classifiers. Two complementary experiments are conducted: (i) a

single-model analysis to assess the impact of increasing adversarial attack sophistication, and (ii) a cross-model analysis to examine whether adversarial vulnerability is model-specific or systemic.

A. Models and Dataset

All experiments are conducted using ImageNet-pretrained CNN models, which are widely deployed in real-world vision applications [12]. ImageNet provides a large-scale and standardized benchmark commonly used for evaluating adversarial robustness.

In the first experiment, a pretrained ResNet-18 model is used as the target classifier due to its widespread adoption and strong baseline performance [13]. In the second experiment, eight ImageNet-pretrained models with diverse architectural characteristics are evaluated to study cross-model adversarial vulnerability. These models differ in depth, connectivity, and computational complexity, enabling a systematic assessment of architecture dependence.

To ensure a controlled and reproducible evaluation, experiments are performed on a validation subset of Imagenette, a curated subset of ImageNet. Thirty validation images that are correctly classified under clean conditions are randomly selected. Restricting evaluation to cleanly classified samples ensures that any observed misclassification is attributable to adversarial perturbations rather than baseline model errors, consistent with standard adversarial evaluation practices [2], [3].

B. Experiment 1: Single-Model Attack Comparison

The first experiment evaluates the relative effectiveness of two first-order gradient-based attacks, FGSM and PGD, and one optimization-based attack, CW, on the pretrained ResNet-18 model.

All attacks are performed under an untargeted, inference-time, white-box threat model, which is a standard setting for adversarial robustness evaluation. FGSM serves as a baseline single-step attack, PGD represents a strong iterative first-order adversary, and CW is an optimization-based attack designed to induce misclassification with minimal perturbation [4]. This setup enables a controlled comparison of how increasing attack sophistication affects model robustness.

C. Experiment 2: Cross-Model FGSM Vulnerability

The second experiment investigates whether adversarial vulnerability is inherent to specific architectures or represents a systemic weakness of deep learning classifiers. To isolate architectural effects from attack-specific optimization strength, FGSM is applied uniformly across eight ImageNet-pretrained models.

Each model is evaluated using the same perturbation budget and attack configuration on the same set of 30 cleanly classified validation images. Models are tested independently to ensure a fair and controlled cross-model comparison. This experiment is designed to assess the generality of adversarial vulnerability under conservative threat assumptions [2].

D. Evaluation Metrics

Attack effectiveness is measured using two metrics. Attack Success Rate (ASR) is defined as the percentage of samples for which an adversarial perturbation causes a change in the predicted label. In addition, mean confidence drop is computed as the average reduction in the model's predicted class confidence between clean and adversarial inputs and is reported as a percentage.

Alongside aggregate metrics, per-sample predictions and confidence scores are recorded to facilitate detailed failure analysis. Qualitative visual inspection of representative adversarial examples is also performed to confirm that perturbations remain imperceptible to human observers.

E. Summary of Experimental Design

In summary, the experimental setup consists of:

- 1) A single-model comparison of FGSM, PGD, and CW attacks to evaluate the effect of attack sophistication on robustness.
- 2) A cross-model FGSM evaluation to assess whether adversarial vulnerability generalizes across architectures.

Together, these experiments provide complementary insights into adversarial attack effectiveness and the systemic nature of adversarial vulnerability, directly supporting the analyses presented in subsequent sections.

V. RESULTS AND ANALYSIS

This section gives the experimental results in accordance with two objectives of this study: (i) examining the effects of making adversarial attacks more sophisticated, and (ii) assessing the adversarial vulnerability as model specific or systemic.

A. Experiment 1: Single-Model Attack Comparison

The first experiment assesses FGSM, PGD, and CW attacks to a pretrained ResNet-18 model on 30 randomly selected validation images correctly classified under clean conditions. This controlled experimental design allows a direct comparison of attack strengths.

TABLE II

COMPARISON OF ADVERSARIAL ATTACK EFFECTIVENESS ON A PRETRAINED RESNET-18 MODEL EVALUATED ON 30 RANDOMLY SELECTED, CLEANLY CLASSIFIED IMAGENETTE VALIDATION IMAGES.

Attack	ASR (%)	Mean Confidence Drop (%)
FGSM	63.33	36.50
PGD	73.33	36.40
CW	96.67	48.41

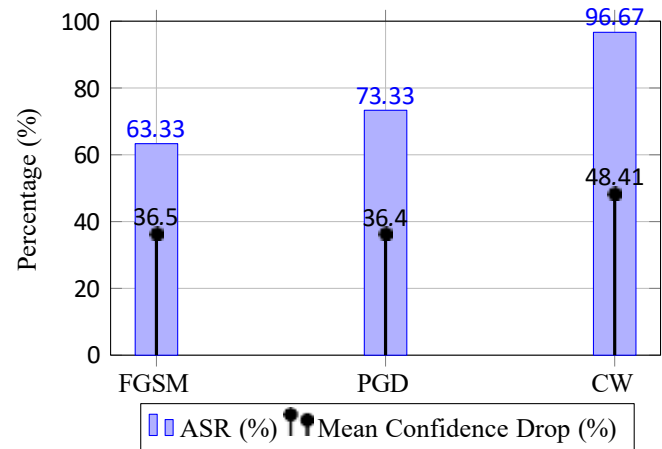


Fig. 1. Combined comparison of attack success rate (ASR) and mean confidence drop (both expressed as percentages) for FGSM, PGD, and CW attacks on a pretrained ResNet-18 model.



Fig. 2. Representative FGSM adversarial example on a pretrained ResNet-18 model, where an imperceptible perturbation induces a label change and reduces prediction confidence.

The rates of escalation in the effectiveness of the attacks with the increasing weight of the optimization strength are evidenced by Table II and Fig. 1. Although FGSM is a one-step attack, it still provides a high ASR, which proves that simple gradient-aligned perturbations can be used to cause misclassification. PGD also enhances ASR and this indicates that iterative first-order optimization is more certain to cross decision boundaries, which is in line with the fact of its role as a robustness benchmark [3].

The CW attack has a near-perfect attack success and largest degradation of confidence, indicating the weakness of the defenses that are based on gradient masking or basic pre-processing. This observation supports earlier results that optimization based attacks can bypass most existing vulnerability mechanisms [4], [7]. Fig. 2 provides a qualitative example demonstrating that, with high confidence, imperceptible perturbations can cause changes in semantic labels.

On the whole, Experiment 1 proves strong attacks are likely to significantly underestimate adversarial risk, which is why it is important to test with strong threat models.

B. Experiment 2: Cross-Model FGSM Vulnerability

The second experiment applies FGSM uniformly across eight ImageNet-pretrained architectures to examine whether adversarial vulnerability is architecture-dependent or systemic.

TABLE III
CROSS-MODEL EVALUATION OF FGSM VULNERABILITY ACROSS EIGHT
IMAGENET-PRETRAINED ARCHITECTURES USING 30 RANDOMLY
SELECTED, CLEANLY CLASSIFIED VALIDATION IMAGES.

Model	FGSM ASR (%)	Mean Confidence Drop (%)
ResNet-18	73.33	51.47
VGG-16	73.33	46.91
MobileNet-V2	56.67	44.28
DenseNet-121	46.67	40.21
EfficientNet-B0	36.67	30.31
SqueezeNet	86.67	50.93
GoogLeNet	53.33	49.06
ShuffleNet-V2	56.67	41.10

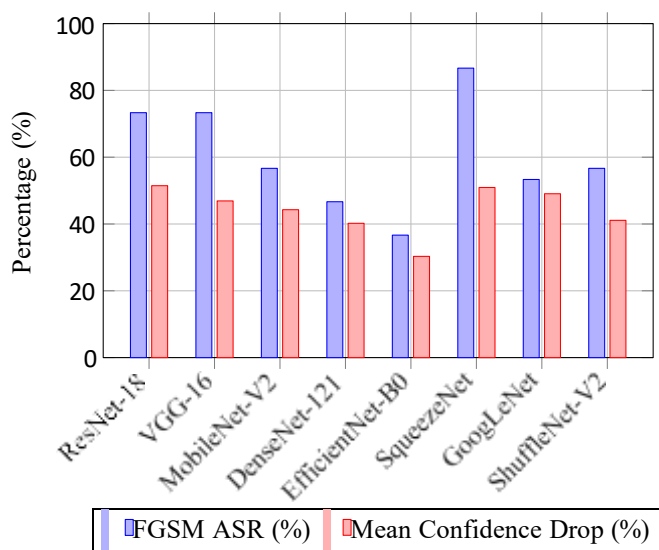


Fig. 3. Cross-model comparison of FGSM attack success rate and mean confidence drop (both expressed as percentages) across ImageNet-pretrained architectures.

As shown in Table III and Fig. 3, FGSM induces misclassification and confidence degradation across all evaluated models, with ASR values ranging from moderate to high. Although architectural differences influence the severity of the attack, no model exhibits inherent resistance to adversarial perturbations. Notably, models with strong clean-data performance do not demonstrate improved robustness, confirming that clean accuracy is not a reliable indicator of adversarial resilience.

The consistent reduction in prediction confidence further indicates that adversarial perturbations systematically degrade model reliability, even when top-1 predictions remain unchanged. These findings align with prior studies on adversarial transferability and large-scale robustness benchmarking [5], [11].

Taken together, Experiment 2 confirms that adversarial vulnerability is a systemic weakness of contemporary deep learning classifiers rather than a property of specific architectures.

VI. DISCUSSION

The experimental findings support the contributions summarized in Table I and explain the essence of adversarial vulnerability in the deep learning image classifiers. The comparative

analysis of FGSM, PGD and CW attacks indicated that there is an evident upward trend in the effectiveness of attack with a rise in the strength of optimization. This validates the idea

that robustness measurements that solely use weak or single-step attacks can be highly inaccurate measures of adversarial risk and clearly shows the importance of measuring robustness models against strong threat models.

The cross-model FGSM analysis also indicates uncovering that adversarial vulnerability exists in many different architectures regardless of clean-data performance. Despite the impact of architectural design on the level of vulnerability, there are no tested models that have an intrinsic resistance to adversarial perturbations. This justifies the contribution in Table I related to systemic adversarial vulnerability and is in line with previous results on adversarial transferability and large-scale benchmarking [5], [11].

Besides the decline in accuracy, the identical decrease in prediction confidence of all models is also evidence that model reliability is reduced when attacked. The practical consequences of this are on deployed systems, where lack of confidence can spread the uncertainty or misjudgments on the down-stream activities. Altogether, the results contextualize adversarial robustness as a socio-technical problem, inspiring the need to think about defense and governance problems beyond the single technical solutions.

VII. DEFENSE AND RAI CONSIDERATIONS

Defense mechanisms have been suggested to alleviate adversarial attacks on deep learning models in several different ways. One of the most appropriate methods is adversarial training that involves the use of adversarial examples during training to enhance resistance against known attacks. It is, however, computationally costly, can typically go wrong when presented with unseen or adaptive adversaries, and can have a trade-off between clean accuracy and robustness, and has restricted practical value.

Other defense mechanisms such as input preprocessing, transformation-based mechanisms, ensemble mechanisms and robust optimization mechanisms provide partial security but are often vulnerable to iterative or optimization-based attacks. In addition, most of the suggested defenses are based on gradient obfuscation and offer false strength when tested in adaptive threat models. These restrictions are in line with the experimental observation that powerful attacks always dominate weak defense.

In addition to prevention, recovery in real-world deployments is necessary since adversarial attacks cannot always be avoided. The solutions of adversarial retraining, uncertainty estimation, anomaly detection, and monitoring with human-in-the-loop supervision allow the systems to be gracefully degraded and still have operational reliability. Although of greater significance, recovery strategies have not been widely studied as compared to attack generation and defense design. Adversarial robustness is not only technically mitigated in the Responsible AI viewpoint. The existence of adversarial

failures compromises the fundamental principles of RAI, including robustness, transparency, accountability, and trust towards AI-based systems of making decisions. Incorporation of RAI principles with explainability, accountability mechanisms and human oversight services as a meta-level defense that supplements technical defenses and recovery solutions can help create resilient and reliable AI systems.

VIII. CONCLUSION AND FUTURE WORK

This study discussed why deep learning image classifiers are still vulnerable to inference-time adversarial attacks, even when the accuracy of such devices is high on clean data. This problem is acute due to the fact that it is possible to conduct such attacks on deployed systems and cause critical failures in safety- and security-critical applications. Two controlled experiments were performed to examine this: a white-box threat model, FGSM, PGD, and CW attacks on a pretrained ResNet-18 model, and a cross-model test of FGSM on eight pretrained models. The findings indicate that optimizing at-tack performance with adversarial objectives and adversarial vulnerability is maintained across all of the models under consideration and indicates that clean accuracy is not a strong predictor of robustness. These results stem the significance of evaluating robustness with strong adversaries and cross-model analysis, and not weak or isolated attacks.

Developing further on the discovery of adversarial vulnerability being systemic and increased in the presence of strong threat models, future research will generalize this analysis to larger datasets and other architectures as well as adaptive black-box attacks. The paper will also consider the multimodal adversarial threats and defense mechanisms implemented to integrate governance issues with technical mitigation to increase long-term resilience and trust.

REFERENCES

- [1] C. Szegedy, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
- [2] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *stat*, vol. 1050, no. 9, 2017.
- [4] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy*, pp. 39–57, IEEE, 2017.
- [5] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, and J. Zhu, "Benchmarking adversarial robustness on image classification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [6] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International conference on machine learning*, pp. 7472–7482, PMLR, 2019.
- [7] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International conference on machine learning*, pp. 274–283, PMLR, 2018.
- [8] B. Mittelstadt, "Principles alone cannot guarantee ethical ai," *Nature machine intelligence*, vol. 1, no. 11, pp. 501–507, 2019.
- [9] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.
- [10] OECD, "Recommendation of the council on artificial intelligence." OECD Legal Instruments, OECD/LEGAL/0449, 2019.
- [11] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, ASIA CCS '17*, (New York, NY, USA), p. 506–519, Association for Computing Machinery, 2017.
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
- [13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.