Adversarial Attacks on Neural Networks: A Comprehensive Survey, Analysis, and Experimental Test Cases

Abhishek Sharma, Neha, Urvashee, Monika, Deepak Sharma
Department of Master of Computer Applications (MCA)
R.D. Engineering College, Dr. A.P.J. Abdul Kalam Technical University (AKTU)

***Abstract*—Neural networks have achieved remarkable performance across diverse machine learning tasks, yet they remain vulnerable to adversarial attacks—carefully crafted perturbations that cause misclassification or erroneous outputs. This paper presents a comprehensive survey of adversarial attacks on deep neural networks, covering theoretical foundations, attack taxonomies, benchmark datasets, experimental test cases, and state-of-the-art defenses. We analyze major attack paradigms including the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Carlini-Wagner (C&W) attacks, DeepFool, and physical-world attacks. We further present structured test case scenarios across image classification, natural language processing, and audio domains. Our analysis provides insights for researchers aiming to develop robust, trustworthy AI systems.**

Keywords — Adversarial Examples, Neural Networks, FGSM, PGD, Robustness, Deep Learning, Machine Learning Security, Perturbation, Defense Mechanisms

I. INTRODUCTION

Deep neural networks (DNNs) have demonstrated exceptional capability in tasks such as image recognition, natural language processing (NLP), speech synthesis, and autonomous navigation. However, the discovery of adversarial examples by Szegedy et al. (2013) exposed a fundamental vulnerability: imperceptible perturbations to input data can mislead even high-accuracy networks with high confidence.

An adversarial example is an input to a machine learning model that has been intentionally modified to cause the model to make a mistake. These perturbations are often imperceptible to human observers but cause neural networks to misclassify inputs with high confidence. This phenomenon raises critical concerns for the deployment of AI in safety-sensitive domains including autonomous vehicles, medical diagnosis, facial recognition, and financial systems.

This paper aims to:

- Provide a structured taxonomy of adversarial attack methods
- Analyze theoretical underpinnings of adversarial vulnerability
- Present comprehensive test cases with expected behaviors
- Survey current defense strategies and their limitations
- Discuss open challenges and future research directions

II. BACKGROUND AND THEORETICAL FOUNDATIONS

2.1 Neural Network Basics

A deep neural network $f: X \rightarrow Y$ maps an input $x \in X$ to an output $y \in Y$ through a composition of parameterized transformations. For classification, $f(x)$ returns a probability distribution over classes. The model is trained by minimizing a loss function $L(f(x), y)$ over a training dataset.

2.2 Adversarial Examples Defined

Given a natural example x with true label y, an adversarial example x' satisfies:

- $f(x') \neq y$ (the model misclassifies x')
- $\|x' - x\|_p \leq \varepsilon$ (the perturbation is bounded by epsilon under some Lp norm)

Common distance metrics include L0 (number of changed pixels), L2 (Euclidean distance), and $L\infty$ (maximum per-feature change). The adversarial goal is to find $x' = x + \delta$ where δ is the adversarial perturbation.

2.3 Threat Models

The adversarial threat model defines attacker capabilities and goals. Key dimensions include:

Threat Dimension	White-Box	Black-Box	Gray-Box
Model Access	Full access to weights & gradients	Output (scores/labels) only	Partial (architecture known)
Attack Knowledge	Complete	None / limited	Architecture only
Computation Cost	Low-Medium	High	Medium
Transferability	Not required	Required	Optional
Typical Methods	FGSM, PGD, C&W	Square Attack, NES	Transfer-based

III. TAXONOMY OF ADVERSARIAL ATTACKS

3.1 Classification by Goal

- Untargeted Attack: Cause any misclassification ($f(x') \neq y$)
- Targeted Attack: Cause specific misclassification ($f(x') = t, t \neq y$)
- Confidence Reduction: Lower the probability of correct label
- Source/Target Attack: Misclassify specific class as another specific class

3.2 Classification by Knowledge

- White-Box Attacks: Full access to model parameters and gradients
- Black-Box Attacks: Access to model outputs only
- Gray-Box Attacks: Partial knowledge of model architecture
- Physical-World Attacks: Perturbations applied to physical objects

3.3 Classification by Perturbation Type

- Lp Norm-Bounded Perturbations (L0, L2, L∞)
- Semantic Adversarial Examples (color, rotation, texture changes)
- Universal Perturbations (single perturbation fools multiple inputs)
- Patch Attacks (localized visible perturbations)

IV. MAJOR ADVERSARIAL ATTACK METHODS

4.1 Fast Gradient Sign Method (FGSM)

Proposed by Goodfellow et al. (2014), FGSM is the foundational single-step gradient-based attack. It generates adversarial examples using:

$$x' = x + \varepsilon \cdot sign(\nabla x\, L(f(x), y))$$

where ε controls perturbation magnitude. FGSM is computationally efficient but generates relatively weak adversarial examples. It operates in the L∞ norm space.

Property	Value
Attack Type	White-Box, Untargeted/Targeted
Norm	L∞
Steps	Single step
Time Complexity	O(1) gradient computation
Typical ε (images)	0.01 to 0.3 (normalized [0,1])
Success Rate (MNIST)	~85-95% at ε=0.3
Success Rate (ImageNet)	~60-80% at ε=0.05

4.2 Projected Gradient Descent (PGD)

Madry et al. (2018) proposed PGD as an iterative extension of FGSM and the strongest first-order white-box attack. It performs multiple gradient steps with projection back onto the epsilon-ball:

$$x(t+1) = \Pi B(x,\varepsilon)[\, x(t) + \alpha \cdot sign(\nabla x\, L(f(x(t)), y))\,]$$

where Π is the projection operator and α is the step size. PGD with random restart is considered the de facto standard for adversarial training evaluation.

Property	Value
Attack Type	White-Box, Untargeted/Targeted
Norm	L∞, L2
Steps	10–100+ iterations
Step Size (α)	ε/steps or 2.5ε/steps
Success Rate vs. Normal Model	>99% (CIFAR-10, ε=8/255)
Success Rate vs. Adv. Trained Model	~40-60% (CIFAR-10, ε=8/255)

4.3 Carlini & Wagner (C&W) Attack

Carlini and Wagner (2017) proposed a powerful optimization-based attack that formulates adversarial example generation as:

minimize $||x' - x||p + c \cdot f(x')$

subject to x' ∈ [0,1]^n. The C&W attack uses a change of variables to handle box constraints and optimizes a surrogate loss. It achieves near-100% success rate on distillation-defended networks and is highly effective against many defenses.

Variant	Norm	Success on MNIST	Distortion (L2)
C&W-L2	L2	~100%	0.3-2.0
C&W-L0	L0	~100%	Minimal pixels changed
C&W-L∞	L∞	~100%	0.1-0.3

4.4 DeepFool

DeepFool (Moosavi-Dezfooli et al., 2016) finds the minimum perturbation to move a sample across the decision boundary by iteratively linearizing the classifier. It typically produces smaller perturbations than FGSM while maintaining high fooling rates. DeepFool is used as a benchmark for minimum perturbation analysis.

4.5 Universal Adversarial Perturbations (UAP)

Moosavi-Dezfooli et al. (2017) demonstrated that a single image-agnostic perturbation δ can fool a classifier on most natural images. Universal perturbations generalize across different inputs, posing a severe real-world threat as attackers only need to compute the perturbation once.

4.6 Jacobian-Based Saliency Map Attack (JSMA)

JSMA (Papernot et al., 2016) uses the Jacobian of the output with respect to inputs to identify the most influential pixels. It perturbs only a small set of pixels (L0 norm attack), making it especially relevant for scenarios where feature sparsity matters.

4.7 Zeroth-Order Optimization (ZOO) / Black-Box Attacks

ZOO (Chen et al., 2017) estimates gradients using finite differences without access to model internals. More recent black-box methods include:

- Square Attack: Score-based attack using random search in L∞ or L2 ball
- Natural Evolution Strategies (NES): Gradient estimation via population-based sampling
- Boundary Attack: Decision-based attack starting from the target class
- HopSkipJumpAttack: Improved decision-based method with gradient estimation at boundary

4.8 Physical-World Attacks

Physical attacks add adversarial perturbations to real-world objects. Notable examples:

- Adversarial Patches (Brown et al., 2017): Printable patches that fool classifiers in physical environments
- Stop Sign Attacks: Stickers that cause misclassification of traffic signs by autonomous driving systems
- Eyeglass Frame Attacks (Sharif et al., 2016): Glasses that fool facial recognition systems
- 3D Adversarial Objects: 3D-printed objects that fool classifiers from multiple viewpoints

- Acoustic Attacks (Carlini et al., 2018): Audio adversarial examples that fool speech recognition

V. EXPERIMENTAL TEST CASES

This section provides structured test cases covering image classification, NLP, and audio domains. Each test case defines attack parameters, expected inputs/outputs, and evaluation metrics.

5.1 Test Case 1: FGSM on MNIST Digit Classification

Field	Detail
Dataset	MNIST (28×28 grayscale, 10 classes)
Model	LeNet-5 (99.1% clean accuracy)
Attack	FGSM (Untargeted)
Epsilon Values	0.05, 0.1, 0.2, 0.3, 0.4, 0.5
Input	Clean digit image x with label y
Adversarial Input	$x' = x + \varepsilon \cdot \text{sign}(\nabla L)$
Expected Output (clean)	Correct digit label, confidence >0.99
Expected Output (ε=0.1)	Misclassification rate ~20-30%
Expected Output (ε=0.3)	Misclassification rate ~85-95%
Evaluation Metrics	Attack Success Rate (ASR), L∞ distortion, SSIM
Success Criterion	ASR > 80% at ε=0.3, L∞ perturbation ≤ 0.3

5.2 Test Case 2: PGD Attack on CIFAR-10

Field	Detail
Dataset	CIFAR-10 (32×32 RGB, 10 classes)
Model	ResNet-18 (93.5% clean accuracy)
Attack	PGD-20 (20 iterations, L∞)
Epsilon (ε)	8/255 ≈ 0.031
Step Size (α)	2/255 per step
Restarts	10 random restarts
Input	Clean image x with label y
Expected Output (clean)	Correct class label, confidence >0.9
Expected Output (adversarial)	Misclassification rate >95%
Evaluation Metrics	ASR, Robust Accuracy, Average L∞ perturbation
Adversarial Training Comparison	Robust model: ~50% robust accuracy at ε=8/255

5.3 Test Case 3: C&W L2 Attack on ImageNet

Field	Detail
Dataset	ImageNet (224×224 RGB, 1000 classes)
Model	InceptionV3 (78.8% Top-1 clean accuracy)
Attack	C&W-L2 (Targeted)
Target Class	10 randomly selected target classes per image
Optimization Steps	1000 Adam steps, learning rate 0.01
Binary Search Steps	9 (for constant c)
Expected Distortion (L2)	< 2.0 on normalized [0,1] images
Expected ASR (targeted)	>95% for top-1 classes
Imperceptibility Test	Human study: <5% detection rate
Evaluation Metrics	Targeted ASR, L2 distortion, SSIM, LPIPS

5.4 Test Case 4: Black-Box Transfer Attack

Field	Detail
Dataset	CIFAR-10
Surrogate Model	VGG-16 (92% accuracy)
Target Model	ResNet-50 (93% accuracy, unknown to attacker)
Attack on Surrogate	PGD-50, ε=8/255
Transfer Metric	Transfer Success Rate (TSR)
Expected TSR (same arch family)	~40-60%
Expected TSR (different arch)	~25-40%
Input/Output Pair	Input: clean image → Surrogate adv. example → Target model
Test Criterion	TSR > 25% demonstrates cross-model transferability
Defense Impact	Adversarial training reduces TSR by ~10-15%

5.5 Test Case 5: Universal Adversarial Perturbation

Field	Detail
Dataset	ImageNet validation set (50,000 images)
Model	VGG-19
Training Set for UAP	10,000 random images
Perturbation Constraint	$L\infty \leq 10/255$
Algorithm	DeepFool-based iterative UAP
Expected Fooling Rate (training set)	>85%

Field	Detail
Expected Fooling Rate (holdout set)	>80% (generalization)
Computation	~1 hour on GPU for full UAP
Evaluation	Fooling Rate, per-class fooling rates, perturbation norm

5.6 Test Case 6: Adversarial Attack on Text Classification (NLP)

Field	Detail
Dataset	SST-2 (Sentiment Analysis), IMDb
Model	BERT-base fine-tuned (93% accuracy)
Attack Method	TextFooler (word substitution based)
Perturbation Type	Synonym replacement preserving semantics
Max Words Changed	20% of total words
Input Example	'The movie was absolutely wonderful' → Class: Positive
Adversarial Example	'The movie was absolutely terrific' → Class: Negative (fooled)
Expected ASR	>80% at 20% word change rate
Grammaticality	Perplexity increase < 50%
Evaluation Metrics	ASR, Semantic Similarity (USE score >0.8), Perplexity

5.7 Test Case 7: Physical Adversarial Patch

Field	Detail
Task	Object Detection (Stop Sign Classification)
Model	YOLOv5
Attack	Adversarial patch (printable, 10cm × 10cm)
Placement	Affixed to stop sign surface
Test Conditions	Day/Night, 3m-15m distance, various angles (0°, 30°, 60°)
Input	Video frames from dashcam
Expected Output (no patch)	Stop sign detected, confidence >0.95
Expected Output (with patch)	Stop sign missed or misclassified as speed limit sign
Expected ASR	>70% across all test conditions
Metrics	Detection Rate, False Negative Rate, Confidence Score

5.8 Test Case 8: Audio Adversarial Examples

Field	Detail
Task	Automatic Speech Recognition (ASR)

Field	Detail
Model	DeepSpeech2
Dataset	LibriSpeech (clean test set)
Attack	Imperceptible audio perturbations (Carlini 2018)
Target	Arbitrary target transcription
SNR Constraint	> 30 dB (imperceptible to humans)
Input	Natural speech audio clip
Adversarial Input	Slightly modified audio (indistinguishable by ear)
Expected Output (original)	Correct transcription
Expected Output (adversarial)	Specific target transcription injected by attacker
ASR	>95% targeted transcription success at SNR >30dB
Evaluation	Word Error Rate (WER), SNR, Human Perception Test

VI. COMPARATIVE ATTACK PERFORMANCE SUMMARY

Attack	Type	Norm	ASR (Normal)	ASR (Robust)	Perturb. Cost	Transferable
FGSM	White-box	L∞	60-95%	10-30%	Very Low	Moderate
PGD-20	White-box	L∞	95-99%	40-60%	Low	High
C&W-L2	White-box	L2	99-100%	20-40%	High	High
DeepFool	White-box	L2	90-98%	15-35%	Medium	Moderate
UAP	White-box	L∞/L2	80-90%	30-50%	High (once)	High
ZOO	Black-box	L2	70-90%	20-35%	Very High	N/A
Square Atk	Black-box	L∞	75-92%	30-50%	High	N/A
TextFooler	Black-box	Discrete	80-90%	30-50%	Medium	High

VII. DEFENSE MECHANISMS

7.1 Adversarial Training

Adversarial training (Madry et al., 2018) augments training data with adversarial examples. It is the most empirically validated defense. However, it requires significant computational overhead (~10x normal training) and reduces clean accuracy by 3-10%.

7.2 Certified Defenses

Certified defenses provide provable robustness guarantees:

- Interval Bound Propagation (IBP): Propagates input bounds through the network
- Randomized Smoothing (Cohen et al., 2019): Smooths predictions using Gaussian noise, provides certified L2 radius
- SDP-based Verification: Semidefinite programming for exact certified bounds

7.3 Detection Methods

- Feature Squeezing: Reduce input complexity (color depth reduction, smoothing) to detect adversarial examples

- MagNet: Autoencoders to detect and reconstruct adversarial inputs
- Local Intrinsic Dimensionality (LID): Statistical test based on neighborhood manifold structure

7.4 Input Preprocessing Defenses

- JPEG Compression: Removes high-frequency perturbations
- Bit-Depth Reduction: Quantizes pixel values
- Spatial Smoothing: Gaussian or median filtering
- PCA-based Denoising: Projects inputs to clean manifold

Note: Most preprocessing defenses are broken by adaptive attacks that account for the preprocessor in the attack objective.

7.5 Summary of Defenses

Defense	Robustness Level	Clean Accuracy Impact	Certified?	Adaptive Attack Resistant?
Adversarial Training (PGD)	High	Moderate loss (3-10%)	No	Partially
Randomized Smoothing	Medium-High	Low-Moderate	Yes	Yes
Feature Squeezing	Low-Medium	Low	No	No
Input Transformation	Low	Low	No	No
Certified IBP	Low-Medium	High loss	Yes	Yes
Ensemble Methods	Medium	Minimal	No	Partially

VIII. EVALUATION METRICS FOR ADVERSARIAL ROBUSTNESS

8.1 Attack-Side Metrics

- Attack Success Rate (ASR): Fraction of adversarial examples that fool the model
- Average Perturbation Norm: Mean L0/L2/L∞ distortion across adversarial examples
- Perceptual Similarity: SSIM, LPIPS scores measuring visual similarity to clean input
- Query Efficiency (Black-box): Number of model queries required per adversarial example

8.2 Defense-Side Metrics

- Robust Accuracy: Accuracy on adversarial examples generated by specified attack
- Certified Radius: Guaranteed robustness radius under given norm
- Clean Accuracy: Model performance on natural (unperturbed) data
- Security Evaluation Error (SEE): Difference between standard and robust accuracy

8.3 Standard Benchmarks

- RobustBench Leaderboard: Standardized evaluation under AutoAttack
- AutoAttack: Parameter-free ensemble attack combining APGD-CE, APGD-T, FAB-T, Square Attack
- MNIST-C, CIFAR-10-C: Corruption robustness benchmarks

IX. ADVERSARIAL ATTACKS ACROSS DOMAINS

Domain	Attack Target	Notable Attack	Real-World Risk	Key Paper
Computer Vision	Image classifiers, detectors	PGD, C&W, Patch	Autonomous vehicles, surveillance	Madry et al. 2018
NLP	Text classifiers, NMT, QA	TextFooler, BERT-Attack	Misinformation, content filter bypass	Jin et al. 2020

Domain	Attack Target	Notable Attack	Real-World Risk	Key Paper
Speech/Audio	ASR, speaker ID	Audio C&W, Devil Whisper	Voice assistants, authentication bypass	Carlini et al. 2018
Malware Detection	Binary classifiers	Gradient-free evasion	Malware bypassing AV systems	Grosse et al. 2017
Reinforcement Learning	RL agents, game players	Policy manipulation	Robotics, game AI, trading	Huang et al. 2017
Medical Imaging	Diagnosis models (X-ray, MRI)	FGSM, targeted attacks	Misdiagnosis, treatment errors	Finlayson et al. 2019
Graph Neural Nets	Node classifiers, link pred.	Metattack, Nettack	Fraud detection bypass	Zugner et al. 2018

X. OPEN CHALLENGES AND FUTURE DIRECTIONS

10.1 The Accuracy-Robustness Trade-off

Empirical and theoretical results suggest a fundamental tension between clean accuracy and adversarial robustness. Tsipras et al. (2019) proved that for some data distributions, robust classifiers must sacrifice accuracy on natural examples. Overcoming this trade-off remains a central open problem.

10.2 Certified Robustness at Scale

Existing certified defenses scale poorly to large models and datasets like ImageNet. The gap between certified and empirical robustness remains significant, limiting the practical applicability of formal verification methods.

10.3 Real-World and Physical Attacks

Most research focuses on digital attacks. Physical-world attacks involving real cameras, lighting variation, and viewing angles present harder challenges for both attackers and defenders. Robust evaluation protocols for physical-world scenarios are nascent.

10.4 Attacks on Foundation Models

Large language models (LLMs) and vision-language models exhibit unique adversarial phenomena including jailbreaking, prompt injection, and multimodal adversarial examples. Understanding and defending these models is a rapidly evolving frontier.

10.5 Adaptive Attacks

Many proposed defenses were later broken by adaptive attacks that account for the defense in the attack objective. Rigorous evaluation must always include adaptive attacks designed specifically for the defense being evaluated (Tramer et al., 2020).

XI. RELATED WORK

Adversarial robustness has been studied from multiple angles. Goodfellow et al. (2014) first systematically analyzed FGSM. Szegedy et al. (2013) discovered that adversarial examples exist in image classifiers. Madry et al. (2018) framed adversarial training as a min-max optimization problem, providing the PGD attack and adversarial training framework. Carlini and Wagner (2017) demonstrated that many defenses fail against stronger optimization-based attacks. Cohen et al. (2019) introduced randomized smoothing as a scalable certified defense. Croce and Hein (2020) proposed AutoAttack, providing a standardized evaluation benchmark. Recent work has extended adversarial analysis to graph neural networks, transformers, and generative models.

XII. CONCLUSION

This paper has presented a comprehensive analysis of adversarial attacks on neural networks, covering theoretical foundations, attack taxonomies, experimental test cases across multiple domains, defense mechanisms, and evaluation frameworks. Our analysis shows that adversarial vulnerability is a fundamental property of current neural network architectures, not an artifact of specific models or datasets.

The test cases presented provide actionable experimental protocols for researchers evaluating both attack effectiveness and defense robustness. As neural networks are increasingly deployed in high-stakes environments, systematic adversarial

evaluation must become a standard practice alongside accuracy benchmarking.

Future work should focus on certified defenses that scale to large models, better understanding of the robustness-accuracy trade-off, and developing robust evaluation frameworks for physical-world and multi-modal adversarial scenarios.

REFERENCES

[1] Szegedy, C., et al. (2013). Intriguing properties of neural networks. arXiv:1312.6199.
[2] Goodfellow, I., et al. (2014). Explaining and harnessing adversarial examples. ICLR 2015.
[3] Madry, A., et al. (2018). Towards deep learning models resistant to adversarial attacks. ICLR 2018.
[4] Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. IEEE S&P 2017.
[5] Moosavi-Dezfooli, S., et al. (2016). DeepFool: A simple and accurate method to fool deep neural networks. CVPR 2016.
[6] Moosavi-Dezfooli, S., et al. (2017). Universal adversarial perturbations. CVPR 2017.
[7] Papernot, N., et al. (2016). The limitations of deep learning in adversarial settings. EuroS&P 2016.
[8] Chen, P., et al. (2017). ZOO: Zeroth order optimization based black-box attacks. AISec 2017.
[9] Brown, T., et al. (2017). Adversarial patch. NIPS Workshop 2017.
[10] Cohen, J., et al. (2019). Certified adversarial robustness via randomized smoothing. ICML 2019.
[11] Carlini, N., et al. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. IEEE DLS 2018.
[12] Jin, D., et al. (2020). Is BERT really robust? TextFooler. AAAI 2020.
[13] Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with AutoAttack. ICML 2020.
[14] Tramer, F., et al. (2020). On adaptive attacks to adversarial example defenses. NeurIPS 2020.
[15] Tsipras, D., et al. (2019). Robustness may be at odds with accuracy. ICLR 2019.
[16] Zugner, D., et al. (2018). Adversarial attacks on neural networks for graph data. KDD 2018.
[17] Finlayson, S., et al. (2019). Adversarial attacks on medical machine learning. Science 2019.
[18] Sharif, M., et al. (2016). Accessorize to a crime. CCS 2016.