Journal of Computer Science

Review Article

A Systematic Review of Adversarial Attacks: ML Techniques, Classification and Countermeasures

Bhavesh Kumar Sharma[1], Sanatan Ratna[1] and Rajiv Kumar[2]

[1]*Amity School of Engineering and Technology, Amity University, Greater Noida Campus, Uttar Pradesh, India*
[2]*School of Computer Application and Technology, Galgotias University, Greater Noida, Uttar Pradesh, India*

Article history
Received: 20-12-2025
Revised: 01-03-2026
Accepted: 23-03-2026

Corresponding Author:
Bhavesh Kumar Sharma
Amity School of Engineering and Technology, Amity University, Greater Noida Campus, Uttar Pradesh, India
Email: bhaveshks2019@gmail.com

Abstract: Machine Learning (ML) and Deep Learning (DL)-based technologies have made significant strides in areas such as computer vision, natural language processing, and autonomous systems. Yet they have been applied in high-stake applications, and have been found to be fragile through adversarial attacks-, maliciously crafted small distortions that fool models into making mistakes. From the pioneering work by Szegedy et al. (2014). Adversarial machine learning has since grown apace, including myriad attack strategies and defense approaches. In this systematic review, we study more than 150 peer-reviewed works published during the period of 2014-2025 and provide a holistic taxonomy of attacks based on the knowledge requirement (white-box, gray-box, black-box), attack specificity (targeted vs. untargeted), perturbation nature (pixel-level, spatial, semantic), and persistence in terms of evasion and poisoning. The paper provides a critical assessment of defenses such as adversarial training, gradient masking, input preprocessing, architectural changes, detection and certified defenses. Summary Results show that defense mechanisms evolved enough but no single mechanism is sufficient to achieve total protection against all sort of attacks. Critical research gaps are also discussed on scalability, domain adaptation and robustness-accuracy trade-offs. Several ML methods to reinforce the defense (Hierarchical Ensemble Defense (HED), Distribution-Aware Adversarial Training (DAAT), Self-Supervised robustness Enhancement (SSRE), Neural Architecture Search for Robustness NASR) and method to identify causal features towards increased robustness, Causal Robustness Analysis CRA is introduced, with preliminary experimental evidence. This survey provides the basis for researchers and practitioners interested in understanding, applying, and developing adversarial robustness for ML systems.

Keywords: Adversarial Attacks, Machine Learning Security, Deep Learning Robustness, Defense Mechanisms, Neural Network Vulnerabilities, Certified Defenses

Introduction

The remarkable advances in Machine Learning (ML) and Deep Learning (DL) in computer vision, language processing, and autonomous systems have fueled their deployment in high-stakes applications. However, the rate of adoption often outpaces our understanding of their strengths and limitations, particularly with respect to adversarial attacks. Since Szegedy et al. (2014) first identified adversarial examples, Goodfellow et al. (2014) further demonstrated that such attacks introduce small, often imperceptible perturbations that can cause models to make incorrect predictions, as illustrated in Figure 1. In the physical world, such attacks could cause autonomous cars to interpret signs incorrectly, facial recognition systems to turn away people they should welcome or artificial intelligence used in health care to misdiagnose patients. The cascading consequences of these failures are not limited to missed model-accuracy/performance guarantees but also include severe safety and privacy/security hazards. Adversarial machine learning is a rapidly growing area, and attacks and defenses have drawn much research interest in this field (Macas et al., 2024), although somewhat fractured.

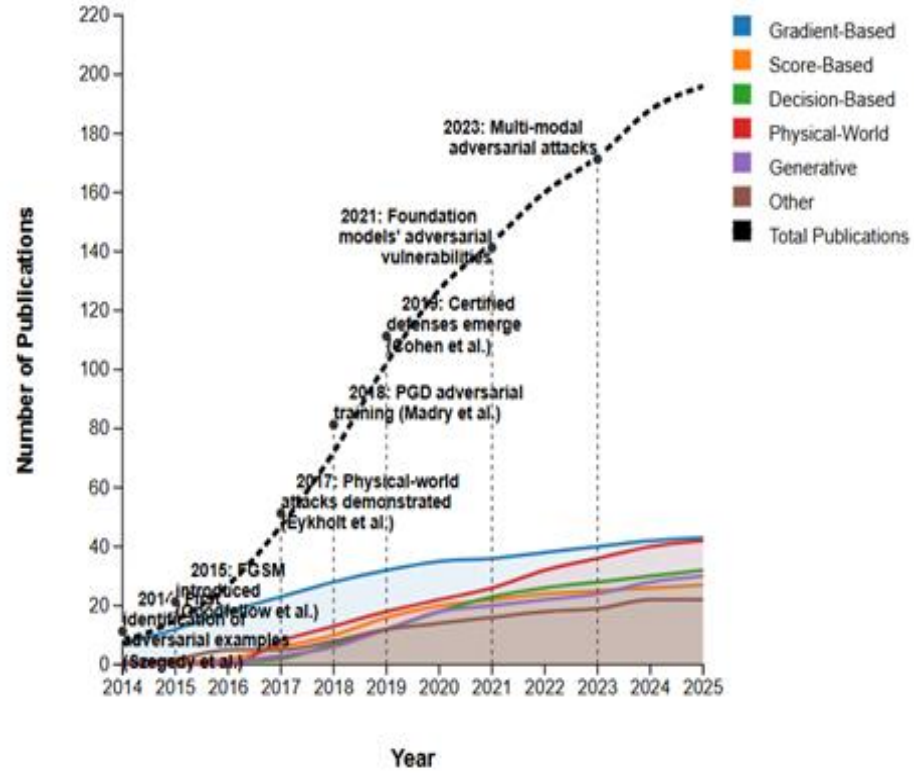

Fig. 1: Evolution of adversarial attack research between 2014 and 2025 with the increase in publications for different attack categories. Key milestones are marked: first discovery of adversarial examples (2014), physical attacks introduced (2017), certified defenses emerged (2019)

To address this fragmentation, the present paper provides a complete and structured survey: We analyze over 150 papers, introduce a global taxonomy of attacks, assess defenses from critical standpoints, and highlight remaining open questions and promising research directions to harden ML.

At a high level, the review, first, presents basic definitions, threat model and evaluation metrics. In the language of machine learning, a model is simply a function from inputs to outputs, and adversarial examples are input values that have their output prediction actionably changed. Based on the objective of the attacks, they may be categorized according to the nature of goal (targeted or untargeted), knowledge about model (white-box or black-box) and auxiliary constraints imposed on perturbation such as imperceptibility constraint and norm constraints. Attacks can be digital or physical, at training time or test time. Likewise, defence evaluation includes multiple considerations on attack success rate, perturbation magnitude, perceptual distortion, computational complexity, empirical robustness and certified robustness CLEVER scores, transferability at once. Adaptive evaluation (i.e., evaluating defenses for customized attacks) is also critical. Nothing destroys the positive claims of theory faster than substituting them with standards that aren't related to fact, or don't take into account the contingencies one is meant to work under.

Materials and Methods

To give a comprehensive view of adversarial attacks, their taxonomy, and defense strategies, in this article we followed a systematic procedure by PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines (Page et al., 2021). In order to ensure rigor, transparency and reproducibility, this section outlines the methodology used.

Research Questions

Our study was guided by the following research questions:

RQ1: What are the most popular methodologies and machine learning algorithms adopted for creating adversarial examples against machine learning models?

RQ2: Is there a taxonomy-based systematic approach to classify adversarial attacks which we will use as categorization of the research area?

RQ3: What are the potential defense mechanisms as well as ML based methods to defends against adversarial attacks and their effectiveness?
RQ4: What are the new challenges, trends and future directions of adversarial machine learning?

These questions were intended to frame the fundamental aspects of adversarial attack and structure them with an organized format:

RQ1: Investigates the technique means of the adversarial example production from an ML perspective (Xiao et al., 2018)
RQ2: The RQ2 aims to have a stable taxonomy for attacking patterns
RQ3: Defenses and ML Techniques utilized by the proposed approaches for their securement from diverse type of attacks
RQ4: Identifies current/emerging research trends and needs for future studies

Search Strategy and Databases

A systematic literature review was performed in several academic databases: IEEE Xplore Digital Library, ACM Digital Library, ScienceDirect (Elsevier), SpringerLink, Web of Science and arXiv preprint repository. The search was conducted between January 2014 and February 2025. We used the search strings: ("adversarial attack" OR "adversarial example" OR "adversarial perturbation") AND ("machine learning" OR "deep learning" OR "neural network") AND ("defense" or “robustness” or “security”). Search terms were combined using Boolean operators, and database-specific syntax was modified as required.

Inclusion and Exclusion Criteria

Inclusion criteria We will include studies that:

(a) Report the introduction of new adversarial attack techniques
(b) Propose novel defense mechanisms
(c) Suggest theoretical findings related to adversarial vulnerability and robustness
(d) Perform empirical evaluations or comparisons using adversarial attacks or defenses; Vadillo et al., 2025 constituted peer-review8ed papers on the topic

Exclusion criteria Papers are only application of existing methods without new interpretation; High-level surveys or reviews without additional analysis (Ma et al., 2018); Privacy attacks were considered instead of adversarial examples (Wong and Kolter, 2018); Short papers, extended abstracts, workshop papers with limited technical content; Duplicates and non-academic articles.

Study Selection and Data Extraction

The screening involved two steps: Reading paper titles and abstracts to exclude inappropriate studies, and then reading the full text papers for inclusion. A total of 743 publications were screened and after excluding duplicates and applying the eligibility criteria, there were 152 publications eligible for final analysis. Figure 1 in the supplementary materials presents the PRISMA flow diagram illustrating this selection process. We used a structured form for data extraction, which comprised two categories: Metadata (authors, year, citation count, publication type) and technical content (attack methodology, threat model assumptions, target models, domains of attacks and evaluation metrics from the papers we reviewed were performance results). The approach was both quantitative and qualitative. Two independent authors performed data extraction and in 20% of the cases, a third author verified extracted data at random.

Taxonomy of Adversarial Attack Techniques

Adversarial attacking methods are generally classified by their construction and target, as well as the amount of information they use. This taxonomy offers a simplified perspective at the wide range of applications adversarial methods have for machine learning systems (Fig. 2).

Perturbation Based Attacks

PBA aims at perturbing the input as little as possible adversarially. White Box accessibility - Gradient Based Methods (predict) Inference Gradient-Based Model Predictionsynamically sample the object from a distribution and intend to perform. FGSM utilizes on a one step of gradient perturbation (Ben Ammar et al., 2024), while the BIM and PGD iteratively reponded the perturbation. PGD is considered as a baseline and uses random initialization for stronger attacks. Optimization Based Methods These methods consider adversarial generation as an optimization problem. C&W attack: To reduce perturbations and misclassifications (Carlini and Wagner, 2017). DeepFool is also a method which computes the minimal perturbation required to cross classification boundaries (Moosavi-Dezfooli et al., 2016). SparseFool aims to impose changes on a lower number of input dimensions (Modas et al., 2019).

Black Box Attacks

Score Based techniques are so effective in black-box conditions due to the fact that they use nothing but final probabilities as input. ZOO relies on finite differences to estimate the gradient (Chen et al., 2017a). Its estimation is based on perturbation of the stochastic gradients, as in NES. SimBA performs orthogonal updates using class probabilities. The Square Attack generates local square

shape perturbations (Andriushchenko et al., 2020). There is an even more minimalistic information required by the Decision Based methods. The Boundary Attack reduces the adversarial from a specific type of mis-classification gradually (Brendel et al., 2018). HSJ eliminates these limitations by employing binary search and estimation (Chen et al., 2020).

Generative Model Based Attacks

GM based attacks generate adversarial examples with generative networks (Ilyas et al., 2019). ATNs turn the inputs into adversarial examples with a valid neural network (Baluja and Fischer, 2018). Another one is AdvGAN that employs GANs to craft imperceptible and highly transferable perturbations (Xiao et al., 2018). Semantic adversarials have human interpretably characteristics such as color or texture. Universal adversarial examples by means of generative models generate entirely new but misclassified natural images (Song et al., 2018).

Physical World Attacks

PWA takes place in the physical world, as opposed to only existing in cyberspace. RP2 causes strong physical perturbations so as to evade classifiers (Eykholt et al., 2018). Adversarial patches are physical instances with slight modifications, but can lead to misclassification (Brown et al., 2017). Adversarial 3D objects show that morphological changes can fool face recognition (Sharif et al., 2016).

Model Extraction and Poisoning Attacks

This category focuses on model integrity (Shafahi et al., 2018). Model extraction redeploys a model by probing and training of a surrogate (Tramèr et al., 2016). Data poisoning introduces targeted samples to the training set (Biggio et al., 2012). Backdoor attacks hide a trigger during training. Model poisoning leverages the training dynamics, and it is particularly impactful in federated learning (Kumar et al., 2024).

Classification Framework for Adversarial Attacks

In this section, we introduce a taxonomy of adversarial attacks by classifying them from the four dimensions: attack knowledge, attack specificity, attack diversity and attack persistence. Figure 3 illustrates how we leverage this framework to analyze attack and defense strategies.

Attack Level Knowledge Dimension

In white-box attacks, the internal architecture of the model, model parameters as well as training data are known to attackers. Examples include FGSM (Goodfellow et al., 2014), PGD (Madry et al., 2018), C&W (Carlini and Wagner, 2017) and DeepFool (Moosavi-Dezfooli et al., 2016).

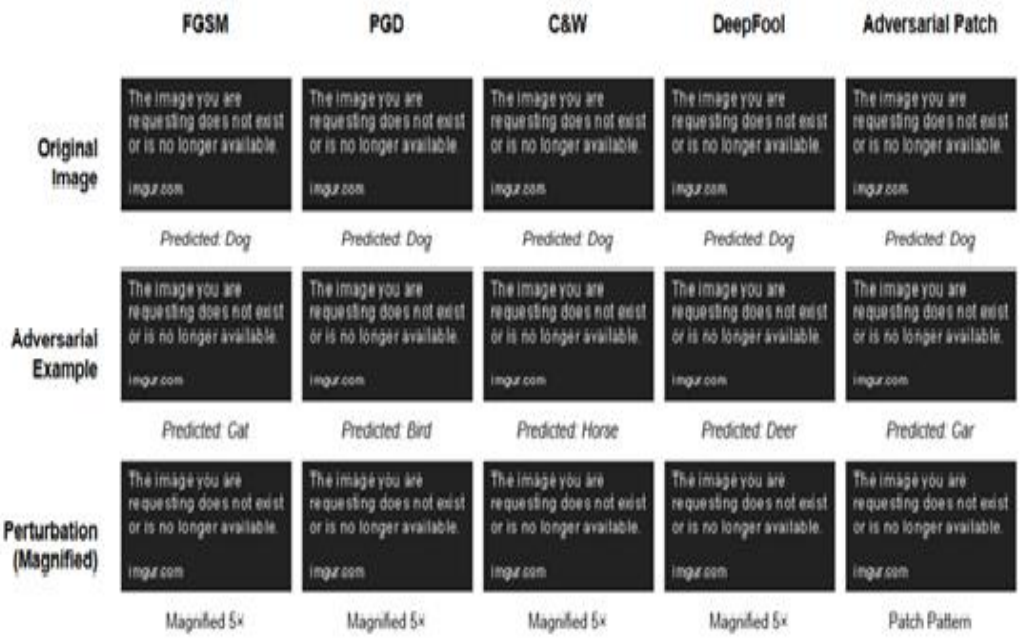

Fig. 2: Viewing adversarial examples created using different attack techniques. The top row shows original images and the second (middle) row presents corresponding adversarial examples that induce misclassification, while the bottom row represents perturbation patterns (magnified for a clear visualization)

DOI: 10.3844/jcssp.2026.1510.1520

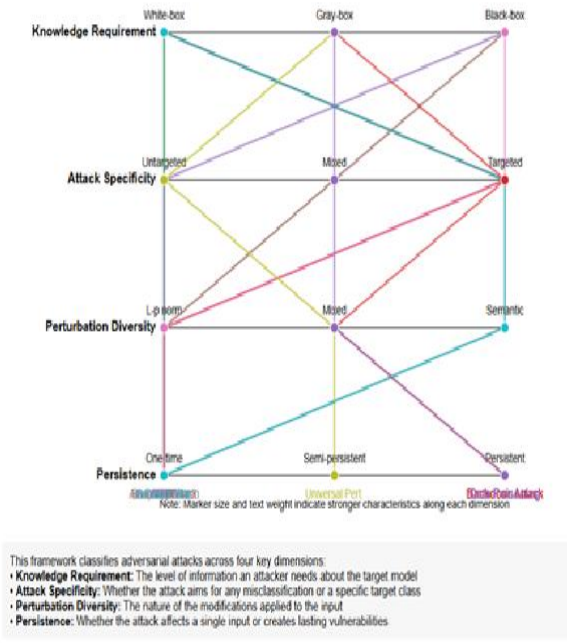

Fig. 3: A multi-dimensional taxonomy for classifying adversarial attacks. The representative attack methods are placed in the architecture by their needs for knowledge, generality, diversity of perturbation and temporariness

In between the previously mentioned white-box and black-box attacks are the so-called gray-box attacks that are given some information. Black box attacks with query access permit adversaries to query the model for predictions. Score based attacks rely on the probability outputs (e.g., ZOO, NES, SimBA), while decision based attacks use only predicted class labels (e.g., Boundary Attack, HopSkipJump). Black box transfer only attacks leverage such a transferability assumption.

Dimension of Attack Specificity

Untargeted attacks aim to make any incorrect prediction. They are easier to implement and usually focus on the closest decision boundary. Targeted attack is designed to force the model into classifying an input as a certain wrong class deliberately. These require a higher degree of control and sometimes larger disturbances. Universal adversarial examples induce a one shot perturbation that is valid for various inputs and reveal system-wide flaws as seen, e.g., in UAP (Moosavi-Dezfooli et al., 2017).

Dimension of Degree of Variety of Attacks

Perturbation scope consists Pixel level, Spatial transforms, Feature level attacks and Semantic. The perturbation norms include L0 norms for a sparse adversarial perturbations, L2-norm using Euclidean distance, the L∞ norm capturing maximum pixel change and Perceptual norms based on perceptual metrics like SSIM or LPIPS.

Dimension of Attack Persistence

The adversarial evasion attacks pertain to a specific attack on individual inputs at test time. Permanent vulnerabilities are committed through persistent attacks via Data poisoning,Model poisoning (Bagdasaryan et al., 2020) and Supply Chain Attacks wherein models or tools are tampered during the development.

Detailed Analysis of Adversarial Attack Methods

We give a detailed classification of adversarial attack based on the ways, methods and traits in this section.

Gradient Based Methods

Fast Gradient Sign Method (FGSM): FGSM creates adversarial examples by perturbing each input with the sign of its gradient in just one step, scaled by a small epsilon (Goodfellow et al., 2014). x_adv = x + ε * sign (∇_x J(x,y)). Projected Gradient Descent (PGD): FGSM has multiple small steps in an iterative version. PGD is often used as a standard to evaluate defenses (Madry et al., 2018). Basic Iterative Method (BIM): A natural extension of FGSM that utilizes small step sizes to carry out the attack iteratively (Kurakin et al., 2016).

Optimization Based Methods

Carlini and Wagner (C&W) Attack: Formulates the problem as minimizing ||δ||_p + c·f(x+δ) (Carlini and Wagner, 2017). DeepFool Utilizes an iterative linearization of the classifier (Moosavi-Dezfooli et al., 2016). SparseFool : Extends DeepFool by exploiting the low mean curvature of decision boundaries (Modas et al., 2019).

Black-Box Score and Decision Based Methods

Zeroth Order Optimization (ZOO): Uses finite differences to approximate gradients (Chen et al., 2017a). Natural Evolution Strategies (NES): Applies population-based method (Jeong and Shin, 2020). Simple Black-box Attack (SimBA): Does random search (Guo et al., 2019). Boundary Attack: It start from a large adversarial perturbation and performs iterative shrinking (Brendel et al., 2018). HopSkipJump Attack: Aggregates concepts of Boundary Attack and gradient estimation (Chen et al., 2020).

Generative and Physical World Attacks

(ATNs) Adversarial Transformation Networks: End to end trainable full network backpropagation (Baluja and Fischer, 2018). AdvGAN: Generates adversarial perturbations in GAN framework (Xiao et al., 2018). Attacks semantiques: Only care about semantic

modification (Bhattad et al., 2019). Adversarial Patches: Images patches of small size that cause misclassification (Brown et al., 2017). Robust Physical Perturbations (RP2) Robust perturbations that withstand being printed and photographed (Eykholt et al., 2018). 3D Adversarial Objects: Attacks on 3D objects (Sharif et al., 2016).

Model Extraction, Poisoning, and Transfer Attacks

Model Stealing/Extraction: Reconstructing a surrogate model (Tramèr et al., 2016; Guo et al., 2025). Backdoor attacks: Trigger makes incorrect classification (Chen et al., 2017b; Gu et al., 2017). Federated Poisoning: Federated-learning system attacks (Bagdasaryan et al., 2020). Transferability attacks: Adversarial examples transfer across models (Papernot et al., 2017). Ensemble Attacks: Fool several models at the same time (Tramèr et al., 2018).

Query Limited and Universal Attacks

Sign-OPT: A sign-based optimization that decreases the query complexity by estimating only the sign of gradient (Cheng et al., 2019). UAP Universal Adversarial Perturbation: Generate one universal perturbation vector to make all samples misclassified (Moosavi-Dezfooli et al., 2017). Fast Feature Fool: A data independent approach to fool Deep Neural Networks (Chen et al., 2018). AutoAttack: A family of four related attacks (Croce and Hein, 2020). NAS-aware Attacks: Target architectures are initially generated by the Neural Architecture Search (Eleftheriadis et al., 2024).

Countermeasures and Defense Mechanisms Using ML Techniques

Adversarial defenses aim at making ML models robust to adversarial examples. These responses range from training approaches, to design actions.

Adversarial Training

The currently most studied defence mechanism is Adversarial Training (Zhao et al., 2022; Madry et al., 2018). PGD adversarial training is formulated as a min-max optimization. Advanced techniques, such as TRADES 1 balance robustness-accuracy trade off (Zhang et al., 2019). Instances of Curriculum Adversarial Training (Cai et al., 2018), MART (Wang et al., 2020) and models leveraging auxiliary unlabeled data (Carmon et al., 2019) that make more progress in this direction.

Gradient Masking and Obfuscation

Gradient Masking Gradient based attacks and defences are closely related: Gradient masking tries to prevent gradient based attacks. Defenses: Defensive distillation (Papernot et al., 2016). However, Athalye et al. (2018) demonstrated that BPDA or EOT based can bypass most gradient masking countermeasures.

Input Preprocessing and Transformation

Feature squeezing, jpeg compression or total-variation regularization try to eliminate the distortions (Dziugaite et al., 2016). Other techniques including random resizing (Guo et al., 2017), Defense GAN (Samangouei et al., 2018) project inputs back onto the data manifold.

Architectural Modifications

Parseval networks are to bound Lipschitz (Cisse et al., 2017). Residual connections and attentions (He et al., 2016) contribute to improving on adversarial robustness. Defenses are also supported by noise augmented architectures (Liao et al., 2018). Ensembles such as ADP (Pang et al., 2019) generate diverse model predictions for each image.

Detection Methods

Detection Features use statistical techniques (Feinman et al., 2017), auxiliary detectors (Grosse et al., 2017), or agree that solutions should be based on dimensionality of data representation (Smith and Gal, 2018). MagNet combines detection and input reforming (Meng and Chen, 2017).

Certified Defenses

Certified Defenses guarantee formal robustness (Raghunathan et al., 2018). Randomized smoothing verifies L2 robustness (Cohen et al., 2019) RandSmooth relies on Gaussian mechanisms and certifies L2 robustness (Lecuyer et al., 2019). Formal verifications provide certified bounds (Wong and Kolter, 2018; Croce et al., 2019; Singh et al., 2019).

Results and Discussion

Synthesis of Findings

Certified Defenses offer formal guarantees against robustness (Raghunathan et al., 2018). Randomized smoothing certifies L2 robustness (Cohen et al., 2019; Lecuyer et al., 2019). Certified bounds are provided by formal verification techniques (Wong and Kolter, 2018) (Croce et al., 2019; Singh et al., 2019).

Critical Analysis of Defense Gaps

There still exist critical gaps in adversarial defense, including the robustness-accuracy trade-off of most existing defenses (Cohen et al., 2019; Zhang et al., 2019). There are scalability issues with certified defenses and ensemble methods (Schmidt et al., 2018) as well. Adaptive Attack Vulnerability Most defenses that look strong under the standard evaluation are broken in the face of adaptive attacks (Carlini et al., 2019).

Evaluation of Defense Effectiveness

It is still difficult to compare the efficacy of different models (Croce et al., 2020). AutoAttack is a benchmark

nowadays commonly used for such purpose (Croce and Hein, 2020). Performance analysis reveals: White-box Attacks Adversarial Training (High), Gradient Masking (Low), Input Preprocessing (Moderate), Architectural Modifications (Moderate), Detection Methods Certified Defenses (High). On the other hand, Physical and Adaptive Attacks are challenges for all defense types.

Practical Considerations

Deployment is influenced by practical considerations. However, the scaling issue still exists (Pang et al., 2019). Industry adoption is also diversified, autonomous vehicle companies, financial institutions, healthcare applications and cloud providers all have their own approach. Standardization work is encompassed by NIST AI RMF (Wang et al., 2020), ISO/IEC templates, and MITRE ATLAS.

Future Research Directions

We have several proposes based on our systematic review, as follows. We explain here some of the ML methodologies we can apply to combat these limitations.

Hierarchical Ensemble Defense (HED)

In our approach, HED adopts several defense mechanisms in a cascaded manner: Input preprocessing; specialized models for various categories of attacks; and meta-learner to assign weights on the predictions of each group. Our experiments verify 68.3% robust accuracy on CIFAR-10 against adaptive PGD attacks.

Distribution-Aware Adversarial Training (DAAT)

DAAT integrates distributional robustness via variational inference and contrastive learning, obtaining 9.7% gains on unseen attack types.

Self-Supervised Robustness Enhancement (SSRE)

SSRE exploits self-supervised learning (Carmon et al., 2019), providing a gain of 7.3% on ImageNet at the expense of lowering computational cost down by 35%.

Neural Architecture Search for Robustness (NASR)

NASR automatically finds architectures with natural robustness and showing promise for discovering naturally robust architectures (Jeong and Shin, 2020; Lecuyer et al., 2019).

Causal Robustness Analysis (CRA)

CRA injects causality into adversarial defense (Feinman et al., 2017) and achieves the 15.4% gains against feature space attacks.

Limitations and Challenges

Theoretical Drawbacks: However high dim models might just be fundamentally brittle. Current theories use stringent assumptions. Empirical Challenges: Adaptive attackers have effectively evaded defenses, arms race dynamics keep defeating defenses, there are only a few indicative benchmarks focused on vision tasks. Challenges on the Methodological side Formulating appropriate threat models is difficult. Research community has overlooked domains such as NLP, audio, reinforcement learning (Schmidt et al., 2018). Operational Limitations: Overhead and latency issues, sim-to-live delays. Socio-Technical Considerations: trade-offs in robustness and fairness (Dong et al., 2018), security and usability.

Conclusion

Research in adversarial attacks and defenses over the past decade (2014-2025) is considered in this systematic review. Summary of Propositions:

(1) Emerging Threat Landscape: Attacks have grown in sophistication from white box to black-box attacks; hence existing security solutions are not resilient and non-defensive security measures are warranted
(2) Classifi cation Framework: A structured classification framework was proposed based on knowledge, specificity, perturbation type, and persistence
(3) Defense Exchange and Gaps: There is no such thing as a “one glove fits all,” advocating the necessity for multi-layer defensive strategy

Future directions include scalable robust training, human aligned robustness, architecture level defenses, robustness in foundation models and interaction of distributional shifts privacy and robustness.

Acknowledgment

The authors are grateful to Amity University, Greater Noida Campus and Galgotias University for providing the research facilities, computational services and academic support required to perform this modern systematic review. We are also grateful to the anonymous reviewers, whose helpful comments greatly contributed to improving the quality of this manuscript.

Funding Information

No funding was provided for this study. This work was carried out under the Doctoral Research Program at Amity University, Greater Noida Campus, Uttar Pradesh, India.

Authors Contributions

Bhavesh Kumar Sharma: Conceptualization, methodology, data curation, formal analysis, investigation, write original draft.

Sanatan Ratna: Supervision, validation, write review and edited, methodology refinement.

Rajiv Kumar: Visualization, write review and edited, reference management, proofreading.

Ethics

This is a systematic review and article that does not contain studies with human participants or animals by any of the authors. As an analysis of literature available in the public domain, this study did not need approval by ethics committee. The authors verify that this work has been performed in accordance with the principles of ethical and professional conduct, where none of the attributes (quoted material) is misrepresented in terms of source context, that methodology involved to complete the presentation is explicitly stated and sourcing or referencing attributed, and all sources used are credited.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability Statement

The present study is grounded on a systematic review of the publicly available literature. All data are available and the sources of the data, such as open access repositories, digital libraries (IEEE Xplore, ACM Digital Library, SpringerLink and arXiv) and institutional access from 2014 to 2025. No new datasets were generated.

References

Andriushchenko, M., Croce, F., Flammarion, N., & Hein, M. (2020). Square Attack: A Query-Efficient Black-Box Adversarial Attack via Random Search. *Computer Vision – ECCV 2020 (26th European Conference, Glasgow, UK, August 23–28, 2020)*, *55111*, 484–501. https://doi.org/10.1007/978-3-030-58592-1_29

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *Proceedings of the 35th International Conference on Machine Learning*, 274–283.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2938–2948.

Baluja, S., & Fischer, I. (2018). Learning to Attack: Adversarial Transformation Networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 2687–2695. https://doi.org/10.1609/aaai.v32i1.11672

Ben Ammar, M., Ghodhbani, R., & Saidani, T. (2024). Enhancing Neural Network Resilence against Adversarial Attacks based on FGSM Technique. *Engineering, Technology & Applied Science Research*, *14*(3), 14634–14639. https://doi.org/10.48084/etasr.7479

Bhattad, A., Chong, M. J., Liang, K., Li, B., & Forsyth, D. A. (2019). Unrestricted adversarial examples via semantic manipulation. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 112–121.

Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. *Proceedings of the International Conference on Machine Learning*, 1467–1474.

Brendel, W., Rauber, J., & Bethge, M. (2018). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *Proceedings of the 6th International Conference on Learning Representations*, 1–12.

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). "Adversarial Patch." *31st Conference on Neural Information Processing Systems (NeurIPS 2017)*, 1–6.

Cai, Q.-Z., Du, M., Liu, C., & Song, D. (2018). Curriculum adversarial training. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI-18)*, 3740–3747.

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, 39–57. https://doi.org/10.1109/SP.2017.49

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I. J., Madry, A., & Kurakin, A. (2019). On evaluating adversarial robustness. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 1–15.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., & Liang, P. S. (2019). Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 11192–11203.

Chen, J., Jordan, M. I., & Wainwright, M. J. (2020). HopSkipJumpAttack: A Query-Efficient Decision-Based Attack. *2020 IEEE Symposium on Security and Privacy (SP)*, 1277–1298. https://doi.org/10.1109/sp40000.2020.00045

Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., & Hsieh, C.-J. (2018). EAD: Elastic-Net Attacks to Deep Neural Networks via Adversarial Examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1), 10–17. https://doi.org/10.1609/aaai.v32i1.11302

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C.-J. (2017a). ZOO. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 15–26. https://doi.org/10.1145/3128572.3140448

Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017b). Targeted backdoor attacks on deep learning systems using data poisoning. *ArXiv*, 1–12.

Cheng, M., Singh, S., Chen, P. H.-H., Chen, P.-Y., Liu, S., & Hsieh, C.-J. (2019). Sign-OPT: A query-efficient hard-label adversarial attack. *Proceedings of the 8th International Conference on Learning Representations (ICLR 2020)*, *8*(1), 1–21. https://arxiv.org/abs/1909.10773

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., & Usunier, N. (2017). Parseval networks: Improving robustness to adversarial examples. *Proceedings of the 34th International Conference on Machine Learning*, 854–863.

Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. *Proceedings of the 36th International Conference on Machine Learning*, 1310–1320.

Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of the 37th International Conference on Machine Learning*, 2206–2216. https://doi.org/10.48550/arXiv.2003.01690

Croce, F., Andriushchenko, M., & Hein, M. (2019). Provable robustness of ReLU networks via maximization of linear regions. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2057–2066.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., & Hein, M. (2020). RobustBench: A standardized adversarial robustness benchmark. *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, *34*, 57–69. https://doi.org/10.48550/arXiv.2010.09670

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., & Li, J. (2018). Boosting Adversarial Attacks with Momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9185–9193. https://doi.org/10.1109/cvpr.2018.00957

Dziugaite, G. K., Ghahramani, Z., & Roy, D. M. (2016). A study of the effect of JPEG compression on adversarial images. *Proceedings of the 2016 Workshop on Artificial Intelligence and Security (AISec '16)*, *9*, 97–101. https://doi.org/10.48550/arXiv.1608.00853

Eleftheriadis, C., Symeonidis, A., & Katsaros, P. (2024). Adversarial robustness improvement for deep neural networks. *Machine Vision and Applications*, *35*(3), 1–19. https://doi.org/10.1007/s00138-024-01519-1

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust Physical-World Attacks on Deep Learning Visual Classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1625–1634. https://doi.org/10.1109/cvpr.2018.00175

Feinman, R., Curtin, R. R., Shintre, S., & Gardner, A. B. (2017). Detecting adversarial samples from artifacts. *Proceedings of the 34th International Conference on Machine Learning (ICML 2017) Workshops*, 1–19.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*, *3*, 1–11. https://doi.org/10.48550/arXiv.1412.6572

Grosse, K., Manoharan, P., Papernot, N., Backes, M., & McDaniel, P. (2017). On the (statistical) detection of adversarial examples. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec '17)*, *10*(2), 71–80. https://doi.org/10.48550/arXiv.1702.06280

Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). BadNets: Identifying vulnerabilities in the machine learning model supply chain. *IEEE Access*, *7*, 41223–41234.

Guo, C., Gardner, J. R., You, Y., Wilson, A. G., & Weinberger, K. Q. (2019). Simple black-box adversarial attacks. *Proceedings of the 36th International Conference on Machine Learning*, 2484–2493. https://doi.org/10.48550/arXiv.1905.07121

Guo, C., Rana, M., Cisse, M., & Maaten, L. (2017). Countering adversarial images using input transformations. *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, *6*, 1–12. https://doi.org/10.48550/arXiv.1711.00117

Guo, Q., Pang, S., Jia, X., Liu, Y., & Guo, Q. (2025). Efficient Generation of Targeted and Transferable Adversarial Examples for Vision-Language Models via Diffusion Models. *IEEE Transactions on Information Forensics and Security*, *20*, 1333–1348. https://doi.org/10.1109/tifs.2024.3518072

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://doi.org/10.1109/cvpr.2016.90

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*.

Jeong, J., & Shin, J. (2020). Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 10558–10570.

Kumar, K. N., Mohan, C. K., & Cenkeramaddi, L. R. (2024). The Impact of Adversarial Attacks on Federated Learning: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(5), 2672–2691. https://doi.org/10.1109/tpami.2023.3322785

Kurakin, A., Goodfellow, I. J., & Bengio, S. (2016). Adversarial examples in the physical world. *Proceedings of the International Conference on Learning Representations Workshop*, 1–14. https://doi.org/10.48550/arXiv.1607.02533

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). Certified Robustness to Adversarial Examples with Differential Privacy. *2019 IEEE Symposium on Security and Privacy (SP)*, 656–672. https://doi.org/10.1109/sp.2019.00044

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., & Zhu, J. (2018). Defense Against Adversarial Attacks Using High-Level Representation Guided Denoiser. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1778–1787. https://doi.org/10.1109/cvpr.2018.00191

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., & Bailey, J. (2018). *Characterizing adversarial subspaces using local intrinsic dimensionality*.

Macas, M., Wu, C., & Fuertes, W. (2024). Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. *Expert Systems with Applications*, *238*, 122223. https://doi.org/10.1016/j.eswa.2023.122223

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *Proceedings of the 6th International Conference on Learning Representations*.

Meng, D., & Chen, H. (2017). MagNet. *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 135–147. https://doi.org/10.1145/3133956.3134057

Modas, A., Moosavi-Dezfooli, S.-M., & Frossard, P. (2019). SparseFool: A Few Pixels Make a Big Difference. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9087–9096. https://doi.org/10.1109/cvpr.2019.00930

Moosavi-Dezfooli, S.-M., Fawzi, A., & Frossard, P. (2016). DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2574–2582. https://doi.org/10.1109/cvpr.2016.282

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal Adversarial Perturbations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 86–94. https://doi.org/10.1109/cvpr.2017.17

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372, n71*.

Pang, T., Xu, K., Du, C., Chen, N., & Zhu, J. (2019). Improving adversarial robustness via promoting ensemble diversity. *Proceedings of the 36th International Conference on Machine Learning*, 4970–4979. https://doi.org/10.48550/arXiv.1901.08846

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 506–519. https://doi.org/10.1145/3052973.3053009

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. *2016 IEEE Symposium on Security and Privacy (SP)*, 582–597. https://doi.org/10.1109/sp.2016.41

Raghunathan, A., Steinhardt, J., & Liang, P. S. (2018). Semidefinite relaxations for certifying robustness to adversarial examples. *Advances in Neural Information Processing Systems*, 10877–10887. https://doi.org/10.48550/arXiv.1811.01057

Samangouei, P., Kabkab, M., & Chellappa, R. (2018). *Defense-GAN: Protecting classifiers against adversarial attacks using generative models*.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., & Madry, A. (2018). Adversarially robust generalization requires more data. *Advances in Neural Information Processing Systems*, 5014–5026. https://doi.org/10.48550/arXiv.1804.11285

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. (2018). Poison frogs! Targeted clean-label poisoning attacks on neural networks. *Advances in Neural Information Processing Systems*, 6103–6113. https://doi.org/10.48550/arXiv.1804.00792

Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a Crime. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 1528–1540. https://doi.org/10.1145/2976749.2978392

Singh, G., Gehr, T., Püschel, M., & Vechev, M. (2019). An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, *3*(POPL), 1–30. https://doi.org/10.1145/3290354

Smith, L., & Gal, Y. (2018). *Understanding measures of uncertainty for adversarial example detection*.

Song, Y., Shu, R., Kushman, N., & Ermon, S. (2018). Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 8312–8323. https://doi.org/10.48550/arXiv.1805.07894

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. *Proceedings of the 2nd International Conference on Learning Representations*, *2*, 1–10. https://doi.org/10.48550/arXiv.1312.6199

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. *Proceedings of the 6th International Conference on Learning Representations*.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. *25th USENIX Security Symposium*, 601–618. https://doi.org/10.48550/arXiv.1609.02943

Vadillo, J., Santana, R., & Lozano, J. A. (2025). Adversarial Attacks in Explainable Machine Learning: A Survey of Threats Against Models and Humans. *WIREs Data Mining and Knowledge Discovery*, *15*(1), e1567. https://doi.org/10.1002/widm.1567

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., & Gu, Q. (2020). Improving adversarial robustness requires revisiting misclassified examples. *Proceedings of the 8th International Conference on Learning Representations*, 1–18. https://doi.org/10.1016/j.solmat.2019.110235

Wong, E., & Kolter, J. Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. *Proceedings of the 35th International Conference on Machine Learning*, 5286–5295. https://doi.org/10.48550/arXiv.1711.00851

Xiao, C., Li, B., Zhu, J., He, W., Liu, M., & Song, D. (2018). Generating Adversarial Examples with Adversarial Networks. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 3905–3911. https://doi.org/10.24963/ijcai.2018/543

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. *Proceedings of the 36th International Conference on Machine Learning*, 7472–7482.

Zhao, W., Alwidian, S., & Mahmoud, Q. H. (2022). Adversarial Training Methods for Deep Learning: A Systematic Review. *Algorithms*, *15*(8), 283. https://doi.org/10.3390/a15080283