International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

DEFENDING DEEP CNNS AGAINST ADVERSARIAL ATTACKS: RECENT TECHNIQUES AND TRENDS.

Mr. Sharifnawaj Yakub Inamdar*[1], Dr. Sonali Kothari[2]

[1]Research Scholar, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India. sharifnawaz12@gmail.com

[1]Assistant Professor, Artificial Intelligence and Data Science Department, AGTI's Dr. Daulatrao Aher College of Engineering, Karad, Maharashtra, India.

[2]Associate Professor, Department of Computer Science and Engineering, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, India. sonali.kothari@sitpune.edu.in

***Corresponding Author:**

Dr. Sonali Kothari,

Associate Professor, Department of Computer Science and Engineering, Symbiosis Institute of Technology, Pune Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India.

E-mail: sonali.kothari@sitpune.edu.in

Abstract:

Adversarial attacks are a big problem for deep neural networks (DNNs), especially when it comes to image classification tasks where even small changes can lead to big mistakes. This weakness makes deep convolutional neural networks (DCNNs), generative adversarial networks (GANs), and new architectures like Vision Transformers less reliable. In the last few years, there have been a lot of new ways to attack, from gradient-based methods like FGSM and PGD to optimization-based, patch, and backdoor attacks. As a result, defense mechanisms have developed in many areas, such as adversarial training, input transformations, detection frameworks, and generative purification methods. Even with these improvements, no one defense has been shown to be strong enough to work in all situations. Most methods have to choose between accuracy, scalability, and resistance to adaptive adversaries. To solve this problem, researchers are looking into optimization-driven methods to make systems more robust and improve training efficiency. These methods include biologically inspired metaheuristics like Sparrow Search and Cuckoo Optimization, as well as modern deep learning optimizers. Benchmark datasets like MNIST, Fashion-MNIST, and CIFAR-10 have become very important for testing both attacks and defenses. They give us a standard way to measure accuracy, precision, recall, and sensitivity when things go wrong. This review brings together changes from 2018 to 2023, with a focus on optimization-enhanced training and the different types of adversarial strategies and defense paradigms. Moreover, we suggest a hybrid defense framework that combines adversarial training, metaheuristic optimization, and lightweight input filtering to protect against both white-box and black-box attacks at multiple levels. The

Received: August 04, 2025

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

goal is to get a full picture of the adversarial landscape, find research gaps that keep coming up, and find ways to make deep learning models for image classification that are more resistant to attacks.

Keywords: Adversarial attacks; Deep learning; Convolutional neural networks; Robustness; Metaheuristic optimization

1. Introduction

Deep learning models have done very well at recognizing images, and in some tests, they have done better than people. But it is now clear that even very accurate models, like deep convolutional neural networks (CNNs), are very easy to attack. Adversarial attacks are small changes to inputs that cause misclassification [1]. These changes are often too small for people to notice, but they can have a big effect on the model's prediction. Figure 1 shows a well-known example: when a CNN sees a panda image with a small noise pattern (scaled by a tiny factor ε), it is very sure that it is not a panda [2][3]. Adversarial examples reveal a core security vulnerability in DNNs: an attacker can readily deceive a classifier by manipulating its gradient-based learning mechanism [1].

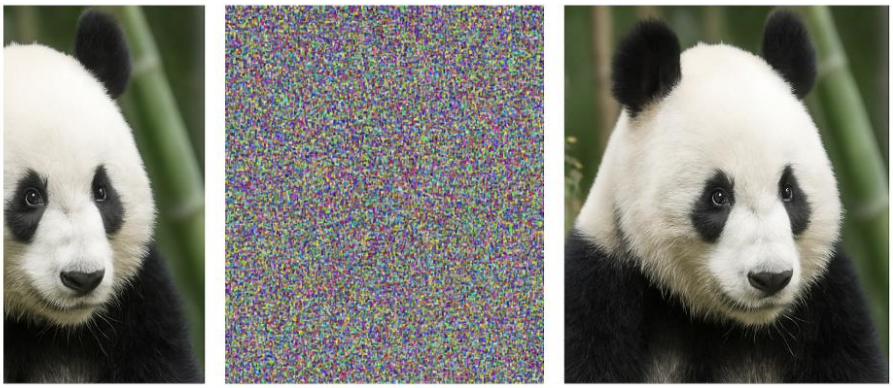

Figure 1. A well-known example of an adversarial situation (Goodfellow et al. 2015):[4]

An original image (a panda, left) is altered with a small amount of noise (middle) to make an adversarial image (right) that the CNN thinks is a gibbon with 99% confidence [2]. The perturbation is exaggerated for demonstration; the real differences are not noticeable.

It's scary that adversarial manipulations don't just work on digital pixels; they can also work in the real world. Attackers have shown that putting small stickers on things or changing their textures can trick vision systems. As an example, Figure 2 shows a stop sign with a small sticker trigger that makes a traffic sign classifier think it is a speed limit sign[5][6]. These kinds of physical attacks and backdoor triggers show how adversarial vulnerability can have real-world effects on things like self-driving cars and surveillance. These observations have prompted rigorous research aimed at comprehending attack methodologies and formulating resilient defense strategies for deep learning models. [7][8].

Received: August 10, 2025

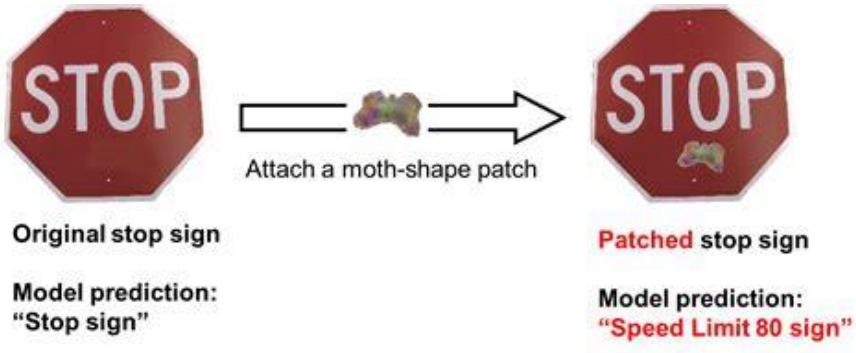

Figure 2. A physical adversarial attack on image classification

A CNN can mistake a stop sign with a small adversarial patch (sticker) for a different sign. [5][6]. These backdoor triggers show that models can be fooled by adversarial perturbations, even in real-world situations.

2. Related Work

2.1 Adversarial Attack Strategies:

In the last ten years, researchers have found different types of attacks on DNNs[9][10]. Attacks are frequently classified based on the attacker's familiarity with the model. In white-box attacks, the attacker can see everything about the model, including its architecture and weights. In black-box attacks, the attacker can only see inputs and outputs. The objective may involve indiscriminate misclassification (any erroneous label) or specific misclassification (a particular erroneous label). Szegedy et al. (2014) and Goodfellow et al. (2015) demonstrated that minuscule perturbations calculated through gradient ascent on the loss function can consistently deceive CNNs. Goodfellow's Fast Gradient Sign Method (FGSM) was one of the first and easiest attacks. It added a small step in the direction of the gradient[14]. Stronger iterative attacks quickly came after that. Kurakin et al. turned FGSM into Basic Iterative Method, and Madry et al. (PGD attack) used perturbations to make very effective adversarial examples[1][15]. The Carlini& Wagner (C&W) attack is another important attack. It optimizes an objective to make the smallest Lp perturbations that can't be found[16]. DeepFool, on the other hand, finds the nearest decision boundary to push the input across[8][17]. Attackers have also come up with ways to do physical and patch attacks (like putting stickers on stop signs to trick people, as shown in Figure 2) and backdoor attacks, in which a model is poisoned during training to respond to certain triggers[5][6]. The ongoing advancement of attack algorithms, ranging from basic gradient-based techniques to complex optimization and even patch-based assaults on Transformers[18][19], highlights the necessity for correspondingly progressive defenses.

2.2 Ways to protect against adversarial attacks:

A variety of defense strategies have been suggested in response to these threats. In general, defense strategies fall into two categories: proactive (making the model more robust during training) and reactive (finding or lessening attacks during inference) [20][21]. Figure 3 shows a list of common defense strategies[22][23], which we will talk about below along with some examples from recent years.

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

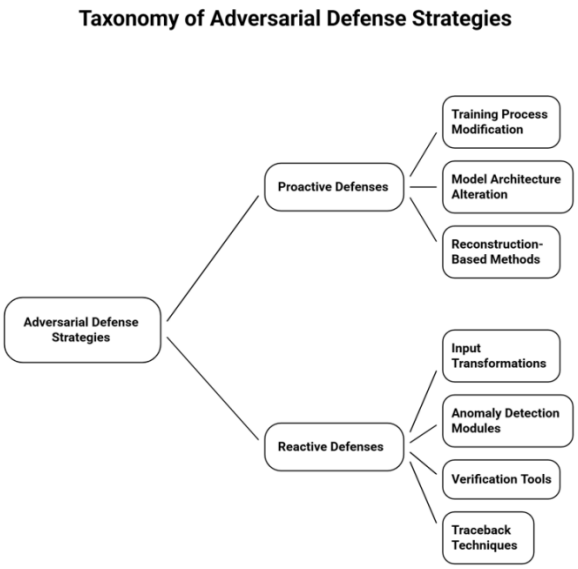

Figure 3. A list of different types of adversarial defense strategies[22][23].

Proactive defenses try to make the model stronger before it goes live (for example, by changing the training or the model itself). Reactive defenses, on the other hand, add external mechanisms while the model is running (for example, input transformations or anomaly detectors).

2.3 Adversarial Training (AT):

Adversarial training is widely considered to be one of the best ways to protect against adversarial attacks [24]. Adversarial training involves retraining the model using adversarial examples created in real time during training to learn parameters that can withstand attacks [24]. Goodfellow et al. first came up with this idea along with FGSM (adding FGSM perturbed examples to training data) [4]. Madry et al. (2018) subsequently formalized adversarial training (AT) as a robust optimization problem, demonstrating that training on the most potent attacks (e.g., PGD adversaries) significantly enhances robustness in both theory and practice [25]. Adversarial training is still the best way to protect against attacks. It makes models more accurate on adversarial inputs (robust accuracy), but it takes longer to train and sometimes makes them less accurate on clean inputs. Recent research has enhanced the efficiency and breadth of adversarial training (AT). For instance, Shafahi et al. (2019) introduced "Free" adversarial training to optimize gradient computations, while Zhang et al. (2019) developed TRADES to achieve a superior equilibrium between robustness and accuracy through a regularization term. AT-based methods directly deal with the attack while learning, and they are often used as a baseline for robustness[24].

2.4 Gradient Masking and Input Transformations:

A lot of defenses try to "mask" or hide the gradients that attackers use[26]. These encompass gradient masking techniques that render the model's decision function non-differentiable or stochastic. For instance, Xie et al. (2018) suggested random resizing and padding of inputs during inference to prevent gradient-based attacks[27][28]. This method (a randomization defense) tries to mess up the attacker's gradient calculations by changing the input in a random

way before sending it to the CNN. Guo et al. (2018) also looked into input transformations like JPEG compression, bit-depth reduction, and image cropping, which get rid of or lessen the adversarial perturbation before classification[28]. You can think of these preprocessingdefenses as reactive add-ons that clean up inputs without changing the model. They are easy to use and often work against weak attacks, but strong adaptive attacks can often get around them by taking the change in the attack model into account[29]. A 2018 study by Athalye et al. found that many gradient-masking defenses, like randomization and compression, could be beaten if the attacker changed their strategy[29]. This made the community look more closely at how well defenses work against adaptive attacks[26].

2.5 Network Architecture and Regularization:

Another proactive strategy is to make the model's architecture or regularization better so that it naturally resists changes. Xie et al. (2019) developed feature denoising networks by incorporating additional layers, such as non-local means and denoising blocks, into CNNs to eliminate adversarial noise from intermediate feature maps[30]. When used with adversarial training, feature-denoisingResNets made ImageNet much more robust[31]. Some methods change activation functions or add normalization layers to make the model less sensitive to changes in the input[32]. Using gradient regularization methods like input gradient smoothing or the Jacobian regularizer, the model is punished for being too responsive to small changes in input[33]. Initially, these methods (like Papernot et al.'s defensive distillation in 2016) looked promising because they made gradients small[33]. However, later attacks broke many of them. Architectural changes, like CapsNets, Bayesian neural networks, or strong Vision Transformer designs, are still being looked into as a way to naturally make systems more robust.

2.6 Finding Adversarial Examples:

Some reactive defenses focus on finding out when an input is adversarial instead of (or in addition to) making the classifier strong. These techniques regard the attack as an anomaly detection issue, identifying inputs that cause atypical internal activations or deviate from the manifold of standard data. Feinman et al. (2017) and others employed statistical tests on layer activations, such as Bayesian uncertainty or kernel density estimates, to identify adversarial inputs. Xu et al. (2018) introduced Feature Squeezing, which involves diminishing input color depth or spatial resolution while monitoring the stability of the classifier's output[34]. If the output of the model changes a lot after squeezing, the input is probably adversarial[34]. Guan and Zhao (2022) recently came up with an inner-class cosine similarity metric that can find adversarial images by looking at how they group together in feature space[35]. Detection can effectively thwart numerous simplistic attacks; however, adaptive attackers frequently devise perturbations to circumvent particular detectors[36][37]. Carlini and Wagner (2017) notably demonstrated that various detection schemes could be circumvented by optimizing perturbations to deceive the detector as well. So, detection is a race to see who can find things first, but when combined with strong training, it can make things even safer.

2.7 Defenses for Generative Models:

Using generative models like GANs and autoencoders for defense is another interesting idea. Samangouei et al. (2018) presented Defense-GAN, which employs a trained Generative

Adversarial Network to cleanse inputs[39]. During inference, Defense-GAN looks for a clean image in the GAN's learned distribution that is similar to the input image. It then sends the reconstructed image to the classifier[40]. This process eliminates adversarial perturbations, predicated on the assumption that the GAN can precisely model the distribution of authentic images[41]. Yang et al. (2018) also used a denoisingautoencoder to put adversarial examples back onto the natural image manifold[42][43]. These input reconstruction defenses work on the idea that a model that knows the valid data distribution can filter out adversarial artifacts. Defense-GAN and its variations exhibited enhanced resilience to specific attacks without altering the classifier[44]. However, later research found that their success often depends on how powerful the generative model is and that iterative attacks can still sometimes find perturbations that survive the purification step[45][46]. Integrating generative models (GANs, VAEs) for pre-processing or jointly training them with classifiers continues to be a vibrant area of research.

2.7 Defenses that work together and mix different types of defenses:

No single defense is completely safe, so researchers have looked into using more than one strategy at the same time. For instance, ensemble defenses combine several base classifiers or detectors in the hope that an attack that works on one will be caught by another[32]. Meng et al. (2020) put forth a framework named "ATHENA" that integrates a variety of weak defenses (such as various input transformations and adversarially trained models) and strategically selects the appropriate defense for specific inputs [47][48]. There are also hybrid methods that use different defenses at different points in the pipeline. For example, a new system for detecting network intrusions (Barik et al., 2025) uses adversarial training with an optimizer during model training and a filtering method during inference[49][50]. They specifically added Projected Gradient Descent attacks to the training (adversarial training) along with a pigeon-inspired optimization algorithm to make the model better at learning from those attacks [51]. Then, at test time, they used a simple spatial smoothing filter on the inputs as an extra layer of protection [50]. These multi-step defenses show how using different methods together can make strong models that stay very accurate (over 99% in their case) while lowering the chances of an attack succeeding by a lot[52].

Table 1 compares some of the most important defense techniques for image classification that have been used in the last few years. We enumerate the principal methodologies and results of each study, highlighting the extensive variety of concepts from altered training protocols to external input preprocessing. Even with these advancements, a distinct deficiency persists: numerous defenses that are effective against a specific attack or dataset are inadequate against adaptive attacks or in different tasks[36][37]. Also, defenses usually come with a cost in terms of model complexity or accuracy on clean data. This drives continuous investigation into more broadly applicable, effective defenses that can protect models without compromising performance.

Table 1: Comparison of notable adversarial defense methods for image classification. Each approach has strengths and limitations, and attacks often evolve to target specific defenses. The

ongoing challenge is to develop defenses that are effective against a wide range of adaptive adversaries while retaining high benign accuracy.

Reference (Year)	**Defense Approach**	**Key Technique & Results**
Goodfellow *et al.* [2015]	Adversarial Training (FGSM)	Introduced training on FGSM adversarial examples; improved robustness but not against stronger attacks.
Papernot *et al.* [2016]	Defensive Distillation	Used knowledge distillation to smooth gradients; initially reduced attack success, but later defeated by adaptive attacks.
Xie *et al.* [2018]	Randomization Defense	Random resize-padding of inputs at inference; helped obscure gradients, ranked #2 in NIPS'17 defense challenge, but adaptive attacks circumvented it.
Guo *et al.* [2018]	Input Transformations	Applied image compression (e.g. JPEG) and bit-depth reduction to remove adversarial noise; effective on simple attacks, but less so on optimized attacks.
Samangouei *et al.* [2018]	Defense-GAN (Input Purification)	Trained GAN to project inputs onto natural image manifold, filtering perturbations; improved robustness without modifying classifier (especially for MNIST/CIFAR), but limited by GAN capacity.
Xie *et al.* [2019]	Feature Denoising CNNs	Modified ResNet architecture with feature-denoising blocks and adversarial training; achieved state-of-the-art robustness on ImageNet (with ~27% robust accuracy at ϙ=4).
Tramèr *et al.* [2020]	Ensemble of Defenses (ATHENA)	Combined multiple weak defenses (transforms + adversarially trained models) in an ensemble; demonstrated enhanced security through diversity.
Wang *et al.* [2021]	Adversarial Training + Augmentations	Incorporated additional data augmentations (e.g. adversarial noise + random crops) during training; improved robustness and generalization to unforeseen attacks.
Guan & Zhao [2022]	Adversarial Example Detection	Developed inner-class cosine similarity detector to flag out-of-distribution inputs; detected many adversarial samples on CIFAR-10, with low false positive rate on clean data.
Barik*et al.* [2025]	Hybrid Training+Filtering Defense	Used PGD-based adversarial training enhanced by a bio-inspired optimizer (PIOA) for robust model training, and applied spatial smoothing at inference; achieved >99% accuracy and sharply reduced attack success on intrusion detection data.

International Journal of Applied Mathematics
Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3. Research Gap Analysis:

The preceding survey indicates that numerous defenses can mitigate particular attacks; however, they frequently exhibit constrained generality and substantial computational expense. Adversarial training is still a key part of the process, but it takes a lot longer to train and might fit too well to certain types of attacks[29]. Preprocessing methods provide minimal protection but can be circumvented by sophisticated attacks that emulate the defense[26]. Detection is useful, but it can't stop things from happening and might even cause false alarms. Additionally, the majority of research has concentrated on modifying training or incorporating inference-time modules, with a limited number of studies integrating multi-stage defenses. There is a distinct deficiency in methodologies that integratively amalgamate training robustness, input filtering, and model optimization. Another gap is making defenses work with new architectures like Vision Transformers and with adaptive attacks. Initial research suggested that Transformers might be more robust than CNNs[18][19], but later work (Patch-Fool) showed that Transformers can be uniquely vulnerable to patch-based attacks[53][54]. This means that defenses need to change as model architectures do. Finally, the literature shows that modern optimization algorithms could be useful in defense design. Numerous metaheuristic and bio-inspired algorithms, such as particle swarm, genetic algorithms, and simulated annealing, have demonstrated efficacy in optimizing hyperparameters or architectures of deep networks[55][56]. However, their use in defending against attacks is still new. Using these kinds of algorithms could lead to new defenses that look for the best configurations or strong features that standard training can't find. In short, the research community needs defense strategies that are more flexible, based on groups, and focused on finding the best solutions to deal with the growing number of threats.

4. Results and Discussion

To fill the gaps we found, we propose a hybrid defense framework that combines ideas from the literature into a single, strong learning system. Figure 4 shows the suggested method, which is aimed at image classification models (DCNNs) but can also be used with other deep networks. There are two main parts to the framework:

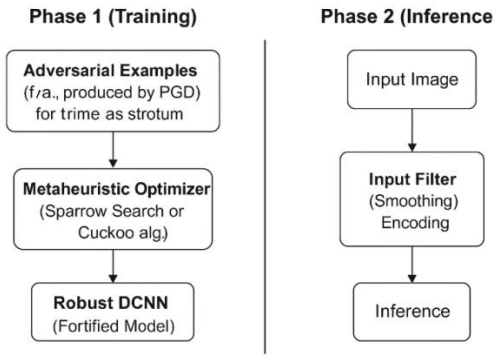

Figure 4. Proposed two-phase defense model (inspired by Barik et al. 2025)[57][58]

Received: August 10, 2025

In Phase 1 (Training), adversarial examples, such as those produced by the PGD attack, are created in real time and used to train a strong DCNN. A metaheuristic optimizer (like the Sparrow Search or Cuckoo algorithm) adjusts the training process (learning rates, class weights, etc.) to make it converge faster and find better minima. The outcome is a fortified model. In Phase 2 (Inference), incoming images go through a light input filter (like smoothing or encoding) to get rid of any possible problems before being classified by the robust model. This hybrid has both proactive and reactive defenses to make it safer.

4.1 Phase 1:Improved Adversarial Training

We add adversarial training to the learning loop during the training phase to make a strong DCNN model. During each training epoch, adversarial examples are intentionally created from the existing model (utilizing a robust attack such as PGD or FGSM) and incorporated into the training batch. This makes the model learn from hard, messed-up examples, which makes it more robust[24]. To avoid problems like gradient vanishing and class imbalance during this process, we change the training hyperparameters on the fly. This is where a metaheuristic optimizer comes in. We use a mix of the Sparrow Search Algorithm (SSA) and Cuckoo Search to change the learning rate schedule, the weights of the network layers, or the rates at which data is sampled. This is based on how sparrows and cuckoos act in nature. These algorithms iteratively improve parameters by searching the space and staying away from bad local minima[55][56]. Combining SSA's ability to search the whole world with Cuckoo's ability to use the best solutions makes training faster and finds model parameters that are the most resistant to attack. This leads to a gradient-based DCNN (GMDCN) that is stronger against adversarial changes because it has been exposed to adversarial data and its training regime has been smartly optimized.

4.2 Phase 2: Filtering and sorting of input.

After training the strong model, we add a light preprocessing step for input when we deploy it. Before a new input image can be used, it goes through a filtering process to get rid of any noise that could be harmful. A basic spatial smoothing filter, for instance, can average pixel values in nearby areas, which reduces high-frequency changes[50]. As mentioned earlier, other possible filters are bit-depth reduction or reconstruction based on an autoencoder. The trained GMDCN then classifies the filtered image. The model is already strong enough to handle most perturbed inputs, and the filter adds an extra layer of protection by cleaning up any strange inputs that the model might not have seen. This extra step costs very little during inference (for example, a small blur filter) but makes the system much more reliable when it is under attack, as shown in similar hybrid systems[52].

5. Expected Results

A defense-in-depth is created by combining adversarially-hardened training with input filtering. The suggested framework is tested on well-known vision datasets like MNIST, Fashion-MNIST, and CIFAR-10, which are used in research on adversarial robustness [17]. We use accuracy, precision, recall (sensitivity), and other metrics to measure performance on both clean and adversarially attacked test sets. We expect faster training convergence and maybe even higher clean accuracy than standard adversarial training when we add a metaheuristic optimizer. This is because the optimizer looks for a better trade-off solution. The sparrow–

Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

cuckoo hybrid optimization should also make the model more stable, which will stop it from overfitting to certain types of attacks. In general, this proposed model aims to improve the state of the art by providing a DCNN that keeps high classification reliability in a wide range of adversarial conditions. This is a step toward making deep learning models safer for use in the real world.

6. Conclusion

Adversarial attacks on deep learning models have spurred an intense research initiative to safeguard DNNs in image classification. In this review, we looked at new ways to attack and grouped ways to defend against them, such as adversarial training, input transformation, detection, and hybrid strategies. There are good things about each defense, but none of them is perfect. We talked about a new integrated framework that combines strong training with metaheuristic optimization and input filtering to fill in the gaps that exist right now. This multi-faceted strategy, coupled with novel concepts such as Transformer-specific defenses and generative model enhancements, signifies a promising avenue for future investigation. To make deep learning systems that are not only accurate but also strong and safe, we need to keep coming up with new defense strategies that are based on careful testing against adaptive attacks.

References

1. Goodfellow, I. J., Shlens, J., &Szegedy, C. (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
2. Akhtar, N., &Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. IEEE Access, 6, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385
3. Wang, J., Wang, C., Lin, Q., Luo, C., Wu, C., & Li, J. (2022). Adversarial attacks and defenses in deep learning for image recognition: A survey. Neurocomputing, 514, 162–181. https://doi.org/10.1016/j.neucom.2022.09.103
4. Macas, M., Wu, C., & Fuertes, W. (2023). 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
5. Khamaiseh, S. Y., Alsmadi, M., Serrano, W., & Bajaj, D. (2022). Adversarial deep learning: A survey on adversarial attacks and defense mechanisms on image classification. IEEE Access, 10, 102266–102291. https://doi.org/10.1109/ACCESS.2022.3208754
6. Abomakleb, A., Alazab, M., Wang, K., Wang, S., Wang, Y., & Wang, Z. (2025). A comprehensive review of adversarial attacks and defense strategies in deep neural networks. Technologies, 13(5), 202. https://doi.org/10.3390/technologies13050202
7. Miller, D. J., Xiang, Z., &Kesidis, G. (2020). Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. Proceedings of the IEEE, 108(3), 402–433. https://doi.org/10.1109/JPROC.2020.2973367
8. Xie, C., Wu, Y., Maaten, L., Yuille, A. L., & He, K. (2018). Mitigating adversarial effects through randomization. In International Conference on Learning Representations (ICLR).

Received: August 10, 2025

9. Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. In Proceedings of the Network and Distributed System Security Symposium (NDSS).
10. Xie, C., Wang, J., Zhang, Z., Ren, Z., &Yuille, A. (2019). Feature denoising for improving adversarial robustness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 501–509).
11. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In Proceedings of the IEEE Symposium on Security and Privacy (S&P) (pp. 39–57). IEEE.
12. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the 35th International Conference on Machine Learning (ICML) (pp. 274–283). PMLR.
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., &Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR).
14. Moosavi-Dezfooli, S.-M., Fawzi, A., &Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2574–2582).
15. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., & Hu, X. (2018). Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 9185–9193).
16. Kurakin, A., Goodfellow, I. J., &Bengio, S. (2018). Adversarial examples in the physical world. In R. V. Yampolskiy (Ed.), Artificial intelligence safety and security (pp. 99–112). CRC Press.
17. Biggio, B., &Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023
18. Samangouei, P., Kabkab, M., &Chellappa, R. (2018). Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In International Conference on Learning Representations (ICLR).
19. Song, Y., Kim, T., Nowozin, S., Ermon, S., &Kushman, N. (2018). PixelDefend: Leveraging generative models to understand and defend against adversarial examples. In International Conference on Learning Representations (ICLR).
20. Guo, H., Zheng, H., & Liu, C. (2018). Countering adversarial images using input transformations. In International Conference on Learning Representations (ICLR).
21. Buckman, J., Roy, A., Raffel, C., &Goodfellow, I. (2018). Thermometer encoding: One hot way to resist adversarial examples. In International Conference on Learning Representations (ICLR).
22. Guan, S., & Zhao, W. (2022). Adversarial detection based on inner-class adjusted cosine similarity. Applied Sciences, 12(18), 9406. https://doi.org/10.3390/app12189406
23. Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (AISec@CCS) (pp. 3–14). ACM.
24. Sadeghi, K. S., Banerjee, A., & Gupta, S. K. (2020). A system-driven taxonomy of attacks and defenses in adversarial machine learning. IEEE Transactions on Emerging Topics in Computational Intelligence, 4(3), 450–467. https://doi.org/10.1109/TETCI.2019.2954874

Received: August 10, 2025

25. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., & Jordan, M. (2019). Theoretically principled trade-off between robustness and accuracy. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 7472–7482). PMLR.
26. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In International Conference on Learning Representations (ICLR).
27. Cheng, X., Niu, Y., Wei, S., Yang, Y., & Liu, X. (2020). Approximating direction-independent perturbations for transferable adversarial examples. IEEE Transactions on Information Forensics and Security, 15, 3103–3118. https://doi.org/10.1109/TIFS.2020.2982333
28. Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS) (pp. 1528–1540). ACM.
29. Jana, S., Ko, C. Y., Mahloujifar, S., & Mittal, P. (2020). Athena: A framework based on diverse weak defenses for building adversarially robust models. arXiv preprint arXiv:1805.04790.
30. Fu, Y., Zhang, S., Wu, S., Wan, C., & Lin, Y. (2022). Patch-Fool: Are vision transformers always robust against adversarial perturbations?.arXiv preprint arXiv:2203.08392.
31. Bhojanapalli, S., Chakrabarti, A., Glasner, D., Li, D., Unterthiner, T., &Veit, A. (2021). Understanding robustness of transformers for image classification. arXiv preprint arXiv:2103.14586.
32. Shao, J., Zhang, Y., Li, J., Li, D., Wu, K., & Yang, Y. (2022). On the adversarial robustness of vision transformers. In Proceedings of the 30th ACM International Conference on Multimedia (MM) (pp. 2774–2782). ACM.
33. Ezugwu, A. E., Shukla, A. K., Nath, R., Akinyelu, A. A., Agushaka, J. O., Chiroma, H., &Muhuri, P. K. (2021). Metaheuristics: A comprehensive overview and classification along with bibliometric analysis. Artificial Intelligence Review, 54(6), 4237–4316. https://doi.org/10.1007/s10462-020-09952-0
34. Gharehchopogh, F. S., Namazi, M., Ebrahimi, L., &Abdollahzadeh, B. (2023). Advances in sparrow search algorithm: A comprehensive survey. Archives of Computational Methods in Engineering, 30(1), 427–455. https://doi.org/10.1007/s11831-022-09804-w
35. Chu, X., Ji, X., Wang, Z., Lu, Y., & Luo, P. (2020). Noisy feature adversarial training. arXiv preprint arXiv:2010.12152. https://arxiv.org/abs/2010.12152
36. Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., & Ma, X. (2021). Backdoor learning: A survey. AI Open, 2, 33–43. https://doi.org/10.1016/j.aiopen.2021.03.002
37. Barik, K., &Misra, S. (2025). A comprehensive defense approach of deep learning-based NIDS against adversarial attacks. Multimedia Tools and Applications, 82. Advance online publication. https://doi.org/10.1007/s11042-025-18794-7
38. Khan, Z., Chowdhury, M., & Khan, S. M. (2022). A hybrid defense method against adversarial attacks on traffic sign classifiers in autonomous vehicles. arXiv preprint arXiv:2205.01225. https://arxiv.org/abs/2205.01225
39. Costa, J. C., Papa, J. P., &Adeli, H. (2024). How deep learning sees the world: A survey on adversarial attacks & defenses. IEEE Access, 12, 61113–61136. https://doi.org/10.1109/ACCESS.2024.3388282

40. Jiang, H., Liu, Z., & Wang, S. (2020). Adversarial attacks and defenses in deep learning: A survey. International Journal of Computer Science and Engineering, 20(2), 117–134.
41. Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., & Ren, K. (2021). Invisible adversarial attack: An adaptive and unnoticeable approach. IEEE Transactions on Dependable and Secure Computing, 18(5), 1474–1488. https://doi.org/10.1109/TDSC.2021.3073743
42. Yang, Y., Zhang, G., Katabi, D., & Xu, Z. (2020). Boundary thickness and robustness in deep models. In Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI) (pp. 3109–3115).
43. Rózsa, A., Günther, M., & Boult, T. E. (2016). Towards robust deep neural networks with BANG. arXiv preprint arXiv:1612.00138.
44. Meng, D., & Chen, H. (2017). MAGNET: A two-pronged defense against adversarial examples. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS) (pp. 135–147). ACM.
45. Cao, Y., & Gong, D. (2022). Recent advances in vision transformer robustness against adversarial attacks. Electronics, 11(3), 406. https://doi.org/10.3390/electronics11030406
46. Deb, S., Tammisetti, N. P., & Malakar, S. (2020). Advances in metaheuristics for deep learning: optimization of hyperparameters and architectures. In 2020 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE. https://ieeexplore.ieee.org/document/9185794
47. Yang, L. T., Shi, Y., Chen, X., & He, L. (2021). Federated defense against adversarial attacks in edge computing. IEEE Transactions on Industrial Informatics, 17(7), 5068–5076. https://doi.org/10.1109/TII.2020.3040699
48. Ren, S., Qin, Z., & Yang, Y. (2021). A survey of adversarial attacks and defenses in deep learning. Cybersecurity, 4(1), 18. https://doi.org/10.1186/s42400-021-00082-w
49. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2805–2824. https://doi.org/10.1109/TNNLS.2018.2886017
50. Liu, H., Lang, B., Liu, M., & Yan, H. (2020). A survey on adversarial machine learning in cyber security. IEEE Access, 8, 67591–67617. https://doi.org/10.1109/ACCESS.2020.2986235
51. Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., & Goldstein, T. (2019). Adversarial training for free! In Advances in Neural Information Processing Systems 32 (NeurIPS) (pp. 3358–3369).
52. Croce, F., & Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Proceedings of the 37th International Conference on Machine Learning (ICML). PMLR.
53. Cohen, J. M., Rosenfeld, E., &Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In Proceedings of the 36th International Conference on Machine Learning (ICML) (pp. 1310–1320). PMLR.
54. Wong, E., &Kolter, J. Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. In Proceedings of the 35th International Conference on Machine Learning (ICML) (pp. 5283–5292). PMLR.
55. 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. In

Volume 38 No. 7s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1625–1634).

56. Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:1712.09665.
57. Uesato, J., Alayrac, J., Huang, P., Stanforth, R., Fawzi, A., &Kohli, P. (2019). Are labels required for improving adversarial robustness? In Advances in Neural Information Processing Systems 32 (NeurIPS) (pp. 12192–12202).
58. Gowal, S., Rebuffi, S., Wiles, O., Stimberg, F., Calian, D. A., & Mann, T. A. (2021). Improving robustness using generated data. In Advances in Neural Information Processing Systems 34 (NeurIPS).

Received: August 10, 2025