Proceedings of CONF-CIAP 2026 Symposium: International Conference on Atomic Magnetometer and Applications
DOI: 10.54254/2753-8818/2026.HZ30283

Low-cost Backdoor Poisoning Attack on the CNN Model

Jingyuan Liu

Department of Computer Science and Technology, University of Nanfang College Guangzhou, Guangzhou, China
liujingyuan@nfu.edu.cn

Abstract. With the widespread use of deep learning in fields such as computer vision, its security issues have also garnered attention. Backdoor Poisoning, a representative data poisoning attack method, is characterized by strong concealment and significant harm. This paper selects the CIFAR-10 dataset as the experimental subject and employs two typical convolutional neural networks, ResNet-18 and VGG-16, as benchmark models to systematically test the effectiveness of low-cost data-level poisoning. The experimental results indicate that, even with a poisoning ratio of only 1%, there is almost no impact on the validation set accuracy of the two models (with a decrease of less than 1%), but the attack success rate (ASR) remains high at 95.1% and 90.0%, respectively. This suggests that a minimal number of poisoning samples can achieve efficient and persistent backdoor implantation, indicating that Convolutional Neural Networks (CNN) models are vulnerable to attacks in an open environment. The research findings confirm the threat of low-cost backdoor attacks and provide a reference for designing targeted defense mechanisms in the future.

Keywords: Convolutional Neural Network, Backdoor Poisoning, Data Poisoning Attack, Attack Success Rate, Model Security

1. Introduction

In the era of rapid development of computer vision, Convolutional Neural Networks (CNNs) have become a key tool, dominating image recognition and classification tasks and playing a significant role in fields such as autonomous driving, face recognition, and video surveillance. They exhibit excellent feature extraction capabilities through deep learning models, significantly improving the quality of recognition tasks. However, due to their high dependence on training data, there is also a potential for intentional abuse, posing security risks and even reliability issues to the entire system. In recent years, a new type of attack called "Backdoor Poisoning" has gradually gained attention and become one of the important topics in academic research and industry applications due to its characteristics of being difficult to detect and potentially causing great harm. Intruders deliberately add a small number of harmful instances with specific activation markers to the training set, which can make the model exhibit predetermined markers at inference time, thus exerting invisible control over the model's decision-making and subtly changing the model's judgment logic without affecting the normal task performance [1-3].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.HZ30283

Unlike traditional adversarial example attacks, backdoor poisoning strategies are covertly implanted during model training. Their characteristic is that they do not require direct intervention or access to model parameters afterward, making the impact more covert and persistent [4-6]. Nowadays, researchers often focus on some high-cost attack methods, which often require a large number of "poisoned" samples or carefully designed activation methods [7, 8]. In the real world, due to limited resources, such strategies may not be fully achievable [9, 10]. It is necessary to explore the effects of low-cost attack strategies on deep learning models, which is of great significance for us to understand whether these models are reliable in real harsh environments and can help us develop more effective protection measures.

This study focuses on the CIFAR-10 dataset, using ResNet18 and VGG16 as control models, aiming to determine whether an attack success rate (ASR) exceeding 90% can be achieved under a 1% contamination ratio. The experiment shows that the accuracy decreases by a maximum of 1% on contaminated data, indicating that even when the degree of contamination is very small, attack strategies can still play a significant role. The study also demonstrates that it is feasible to launch covert attacks on neural network models using simple and low-cost methods, posing a significant threat to convolution-based neural network models. Against this background, this paper delves into the design ideas of lightweight triggers and their potential applications in multi-client scenarios, thereby pioneering new approaches to security considerations in the field of deep learning and providing new directions for reflection.

2. Relevant theoretical methods

A Convolutional Neural Network (CNN) is a classic deep learning model primarily composed of convolutional layers, pooling layers, and fully connected layers. When processing image information, it starts with low-level features such as edges and gradually extracts them, continuously generating higher-level semantic representations as the number of layers increases. Then, by compressing dimensions, it enhances adaptability to translation invariance, thereby further strengthening the overall generalization performance. Finally, with the help of the fully connected structure, it maps information under different scales to the target output space required by specific tasks. Due to its excellent performance in the entire image classification field, it has become one of the indispensable and important supporting technologies in computer vision research.

Backdoor Poisoning is a novel form of attack targeting the training process of deep learning models. Its implementation involves adding some data samples with specific triggering conditions to the original dataset and intentionally altering the original label information content of these samples, thus achieving a predetermined goal. During inference, if the input contains the same triggering features, the infected model will be forced to give a response of a specified category. Unlike traditional adversarial example attacks, which require continuous generation of perturbations and real-time queries to the model, this method implants malicious logic modules into the training process early on, so it does not require continuous interaction to achieve covert control effects. This makes Backdoor Poisoning highly covert, often manifesting as small geometric shapes or low-frequency noise interference, which appears normal to the naked eye but can activate important convolutional layer units in deep neural networks to control classification results without affecting overall performance.

DOI: 10.54254/2753-8818/2026.HZ30283

3. Experimental design and results

3.1. Experimental setup

3.1.1. Low-cost backdoor poisoning method

This study focuses on data-level backdoor poisoning, which involves randomly mixing a portion of samples with triggers and tampered labels into a clean training set, thereby embedding backdoor logic into the neural network structure during model training. Compared to model-level poisoning, which requires opening the model code to modify parameters to change decision outcomes, it is easier to operate, adapts to more unknown black-box environments, and is more harmful in real life. Therefore, this work adopts different ratios (1%-5%) for "low-cost" verification experiments. Systematic experiments are conducted on two different networks, ResNet 18 and VGG16, and the impact ratio is observed based on different split datasets to examine how the trade-off relationship between ASR and ACC changes for both attackers and defenders.

Trigger design is the key to achieving a covert and efficient backdoor. This uses three different types of triggers: fixed geometric (like the 3x3 color squares in the lower right corner), geometric variants (similar to L-shaped stripes), and local noise triggers (superimposing random noise within a small area). The choice of these types has its own advantages and disadvantages: fixed triggers are easy to build and have strong replicability, while noise triggers are visually more covert and can better avoid being detected. This paper selects different sizes and colors of triggers at different locations, which greatly affect whether the backdoor can be successfully learned and transferred. Generally, for convolutional networks, deeper hidden features are more susceptible to the influence of a small number of high-contrast features, so this study chooses relatively small but higher-contrast local patches as trigger conditions to facilitate the learning and transfer process of the backdoor; meanwhile, a random trigger method can make it harder for manual inspection or simple statistical purification methods to locate it.

In terms of sample selection and labeling strategy, this paper adopts the "target-out-of-class poisoning" strategy: that is, prioritize injecting triggers into non-target classes and uniformly tampering them with labels that attack the target class, in order to minimize the impact on the original distribution of the target class and improve ASR. To prevent excessive contamination of the target class in the training set from causing direct drift, it is also possible to exclude target class samples from poisoning. Randomly selecting a certain proportion (uniformly sampled from the training set) of poisoning samples not only simulates passive data pollution in the real world, but also ensures that the experimental results are reproducible and statistically significant.

In practice, this paper encapsulates the poisoning logic into a dataset wrapper: during training, dynamically apply triggers to the input at the front end of the Dataloader through the PoisonedDataset and change its label; when evaluating ASR performance, this paper overlays the same trigger (without changing the label) on the entire test dataset and then performs statistics. This approach facilitates unified comparison of the effectiveness of various poisoning/anti-poisoning methods across different models or hyperparameters, and is also conducive to conducting various ablation experiments (such as examining the impact of different types of triggers or positions individually).

From the perspectives of attack feasibility and defense countermeasures, low-cost backdoors exhibit three characteristics: First, they are highly efficient - only a small proportion of poisoning can achieve high ASR on the target class; second, they are covert - the ACC of clean data decreases very little, making them difficult to detect through rough quality checks; third, they are resistant to

DOI: 10.54254/2753-8818/2026.HZ30283

existing defenses - simple data cleaning or adversarial training may only partially reduce ASR, often accompanied by significant performance losses in the main task. In subsequent experiments and discussions, it is necessary to report both ACC and ASR, and design combined defenses (data purification + fine-tuning/architecture improvement, etc.) to evaluate the feasible defense costs and effects in reality.

3.1.2. Dataset and partitioning

This experiment utilizes the CIFAR-10 dataset, which comprises 10 categories of color images with a size of 32×32. Following common practice, this paper allocates 5,000 images from the training set as the validation set, while the official test set is employed as the test set. This paper performs image normalization based on the mean and variance of CIFAR-10 and employs data augmentation techniques during training, including random cropping and random horizontal flipping. During evaluation, this paper only performs ToTensor and normalization.

3.1.3. Model structure

The two types of convolutional networks used in this experiment are ResNet-18, which is adapted for CIFAR-10, and the classic VGG-16 (including the version with BatchNorm). The former is composed of stacked residual blocks (BasicBlock) and utilizes shortcut connections to alleviate the problem of vanishing gradients during deep network training. In this implementation, the input layer uses 3×3 convolution to retain the spatial resolution of small-sized images, which is then sequentially passed through four residual layer groups and subjected to global average pooling. Finally, a linear layer is used to output the class prediction results. The latter uses multi-layer 3×3 convolution and periodic max pooling as the main structure, and adds adaptive average pooling and a single-layer linear classifier to complete the final mapping task. Both networks adopt the He/Kaiming initialization method and perform regular initialization on the batch normalization layers.

3.1.4. Training and evaluation process

The training process follows the standard supervised learning paradigm, with cross-entropy loss used in a mixed dataset with/without poisoning. In terms of the optimizer, SGD with momentum=0.9 is selected, and weight decay is added for regularization. The learning rate is scheduled using CosineAnnealingLR to achieve a smooth decrease. Important hyperparameters such as batch size, initial learning rate, epochs, and random seed are determined before the experiment begins to maintain comparability. To ensure the reproducibility of experimental results, all experiments are conducted on a single RTX-3090 GPU. The specifics are shown in Table 1.

DOI: 10.54254/2753-8818/2026.HZ30283

Table 1. Experimental setup

item	value retrieval	Instructions
model	ResNet-18/VGG-16	Selected model
optimizer	SGD (lr=0.1, momentum=0.9)	weight_decay=5e-4
Batch-Size	128	
Epoch	50	
Poisoning ratio	1%	Random sampling from non-target samples
trigger	Square(3×3), pos=br	
random seed	42	Used for dataset partitioning and replication

During training, record the training/validation loss and accuracy for each epoch, and include the ASR curve when poisoning is enabled. Finally, export the training history as CSV and curve graphs for result presentation and further analysis.

3.2. Experimental results and analysis

3.2.1. ResNet-18 model

This section first compares the training dynamics and final metrics of two groups on ResNet-18: the baseline (without poisoning) and the poisoning group (poison ratio = 1%). All experiments were rerun using the same hyperparameters, namely SGD with lr=0.1, batch=128, and epochs=50. The validation/training accuracy was recorded using ACC. In the case of overlaying triggers on the full test set during backdoor poisoning, ASR was also used to measure and evaluate both the main task and the strength of the backdoor simultaneously.

In the absence of poisoning, ResNet-18 achieves rapid convergence within 50 epochs: the training accuracy is almost 100%, and the best accuracy on the validation set is 88.76% (bestvalacc=0.8876) and remains stable; the corresponding ASR is consistently 0 (poisoning is not enabled), indicating that the model behaves normally and has a good baseline.

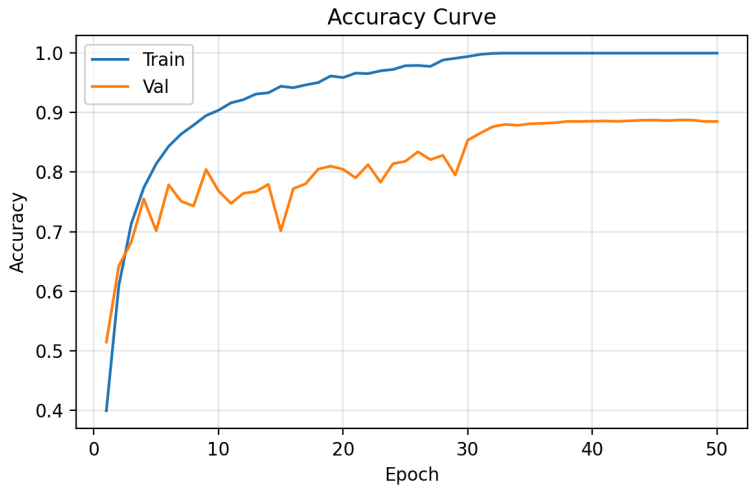


Figure 1. Change in accuracy during ResNet-18 training without poisoning (picture credit: original)

DOI: 10.54254/2753-8818/2026.HZ30283

Figure 1 shows the Change in accuracy during ResNet-18 training without poisoning. Subsequently, with 1% poisoning (target class=airplane, trigger=3×3square) setting, the convergence trend of the training curve is similar to that without poisoning: the training accuracy is also almost infinitely close to 100%, and the validation accuracy eventually stabilizes at approximately 87.9% (val_acc ≈ 0.8766; best_val ≈ 0.879). This shows that for the main task accuracy, the impact of poisoning can be completely ignored (drop < 1%). As for the backdoor strength, ASR immediately spikes to nearly 25% from the first epoch, then quickly rises and remains at a high level after approximately 5-10 epochs (ASR in most epochs is approximately equal to 0.90-0.96), and converges to approximately 95.1% towards the end (final ASR ≈ 0.9508). The results indicate that only a 1% poisoning ratio can enable the backdoor to be efficiently learned and retained in the model for a long time.

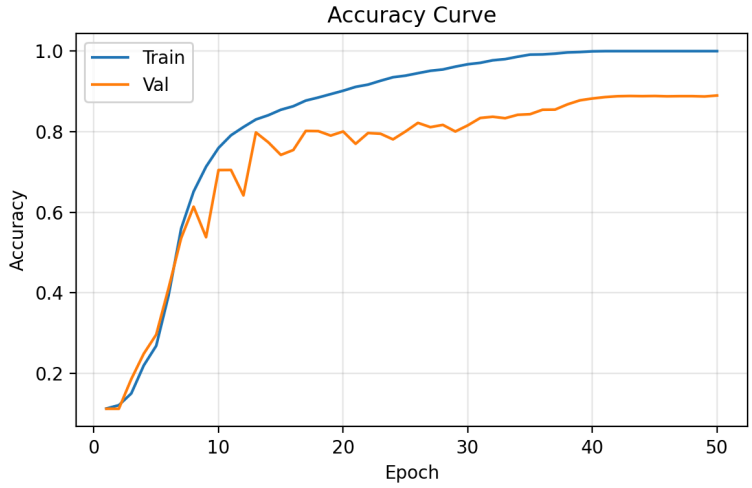


Figure 2. Change in accuracy during poisoning training of ResNet-18 (picture credit: original)

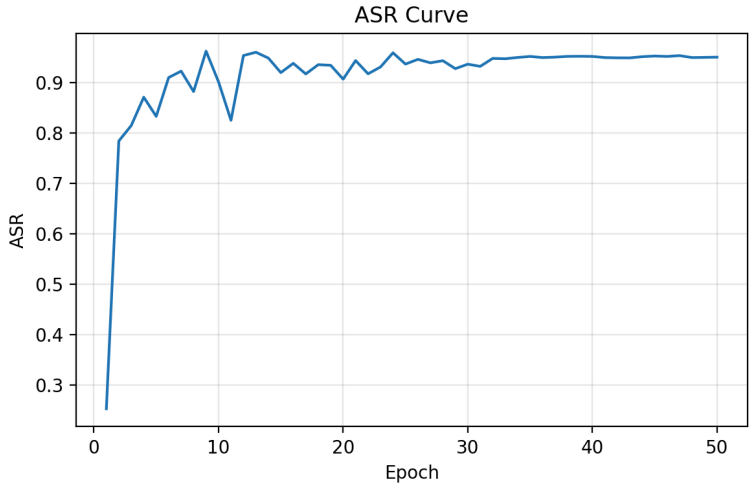


Figure 3. Variation diagram of ASR during poisoning training with ResNet-18 (picture credit: original)

Figure 2 shows the change in accuracy during poisoning training of ResNet-18. Figure 3 shows the variation diagram of ASR during poisoning training with ResNet-18. When 1% data-level

DOI: 10.54254/2753-8818/2026.HZ30283

poisoning is enabled under the same training configuration conditions, the sensitivity of ResNet-18 becomes very strong: the model's validation accuracy on the clean test set is almost unaffected (from ≈88.8% to ≈87.9%), but the attack success rate rapidly increases and stabilizes at about 95% after the trigger is superimposed on the full test set, indicating that its backdoor pattern is effectively learned by the network and can be retained for a long time. This suggests that in a standard pipeline without additional protection, a small proportion can achieve imperceptible and highly efficient successful backdoor implantation. On the one hand, it emphasizes the vulnerability of CNNs to data poisoning attacks, and on the other hand, it indicates that maintaining the performance of the main task cannot be used as a methodological basis for judging model security.

3.2.2. VGG-16 model

This section presents the performance of VGG-16 in two experiments: clean and poisoned, under the same training configuration (SGD, lr=0.1, batch=128, epochs=50; the remaining hyperparameters are consistent with those of ResNet-18). The metrics remain the same: training/validation accuracy (ACC) and attack success rate (ASR) after applying a trigger to the entire test set. Figure 4, Figure 5, and Figure 6 depict the ACC curves for clean training, poisoned training, and the ASR curve under poisoning conditions, respectively.

In the absence of poisoning, the training process of VGG-16 is stable. In each epoch, the training accuracy monotonically increases and tends to saturate, and the accuracy on the validation set also remains around 89% from the middle, indicating that this network can learn well as a baseline on CIFAR-10.

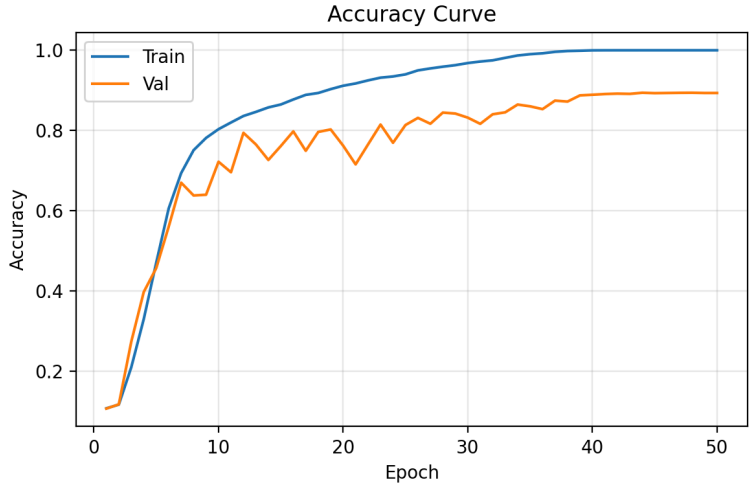


Figure 4. Change in accuracy during training without poisoning for VGG16 (picture credit: original)

When implementing a 1% data-level poisoning attack, with the target category set to "airplane" and the trigger parameter set to a 3×3 square region, the accuracy of the VGG-16 model during training remained stable compared to when it was not contaminated. After many rounds of iterative adjustments, the validation set accuracy gradually converged to 88.8%-89.0%, which is about 1.5 percentage points lower than that of the clean samples. As for the detection performance against adversarial samples, the results exhibited completely different characteristics: initially, there was a phenomenon of numerical fluctuation due to sampling bias and the influence of randomization factors. However, during the 8th to 12th iteration stages, it rose rapidly and eventually stabilized at a

DOI: 10.54254/2753-8818/2026.HZ30283

relatively high level (generally between 0.88 and 0.92). The overall convergence value was almost above the 90% level, indicating that this network architecture can effectively perform its main tasks even in the face of potential security risks and exhibits good anti-interference capabilities.

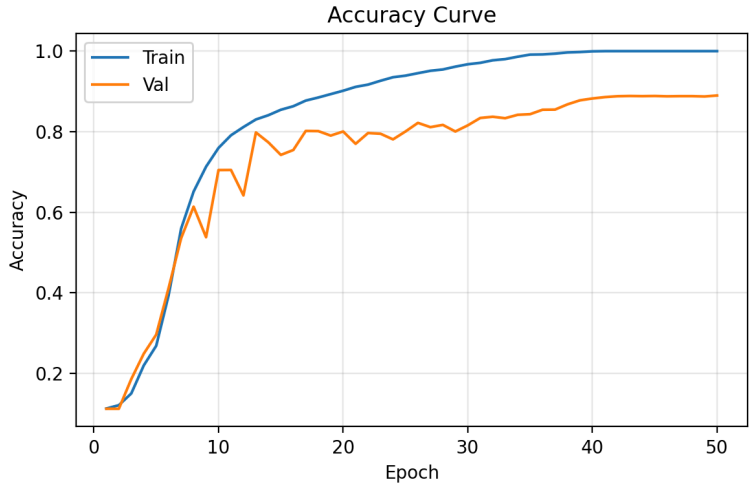


Figure 5. Acc change chart during VGG16 poisoning training (picture credit: original)

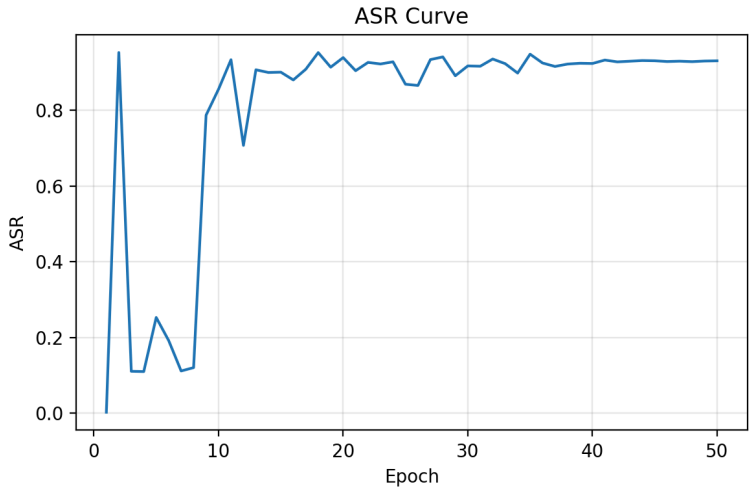


Figure 6. Change graph of ASR during VGG16 poisoning training (picture credit: original)

From the experimental data, it can be observed that under the experimental conditions set in this study, the VGG-16 model exhibits the following characteristics: it achieves a high validation accuracy of approximately 89% on a clean test set and maintains a relatively stable performance level of about 88.9% even after suffering from low-intensity poisoning attacks. When the trigger exerts influence on all test samples, the average misclassification rate significantly increases to nearly 90%, a phenomenon that is basically consistent with the ResNet-18 network. This indicates that both networks possess a certain degree of resistance to perturbations and exhibit abnormal responses in the presence of slight contamination, suggesting that mainstream convolutional neural network architectures are insufficient for low-cost data-level backdoor injection protection.

DOI: 10.54254/2753-8818/2026.HZ30283

Table 2. Experimental results

model	Experimental setup	ACC	ASR
ResNet-18	No poisoning	88.8	0.0
ResNet-18	1% poisoning	87.9	95.1
VGG-16	No poisoning	89.0	0.0
VGG-16	1% poisoning	88.9	90.0

As shown in Table 2, the validation accuracy of the two architectures under non-attack conditions is 88.8% and 89.0% respectively, but it drops to 87.9% and 88.9% under attack conditions. Although they are almost unaffected, the success rates reach 95.1% and 90.0%, fully demonstrating the harmfulness brought by low-cost and low-probability attacks.

4. Conclusion

This paper focuses on the research of low-cost backdoor poisoning in CNN models. By adding only 1% of backdoor-poisoned data to the CIFAR-10 dataset, it can be observed that both ResNet-18 and VGG-16 can learn effectively without reducing the validation accuracy of the neural network to near the non-poisoned state, and they learn relatively stable backdoor patterns.

These research conclusions indicate that even with a very small proportion of malicious samples, CNN can lead to severe backdoor phenomena. Traditional methods that rely solely on the accuracy of the main task are difficult to use to identify existing vulnerabilities. How to detect and prevent backdoor attacks while ensuring model accuracy has become a key consideration for the future. This research provides an experimental foundation for the design and evaluation of subsequent protection mechanisms, and it also highlights the vulnerability of deep learning models to security threats in an open environment.

References

[1] Barni, M., Kallas, K., & Tondi, B. (2019, September). A new backdoor attack in cnns by training set corruption without label poisoning. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 101-105). IEEE.
[2] Cinà, A. E., Grosse, K., Vascon, S., Demontis, A., Biggio, B., Roli, F., & Pelillo, M. (2025). Backdoor learning curves: Explaining backdoor poisoning beyond influence functions. International Journal of Machine Learning and Cybernetics, 16(3), 1779-1804.
[3] Rao, Q., Wang, L., & Liu, W. Rethinking CNN's Generalization to Backdoor Attack from Frequency Domain. In The Twelfth International Conference on Learning Representations.
[4] Kolouri, S., Saha, A., Pirsiavash, H., & Hoffmann, H. (2020). Universal litmus patterns: Revealing backdoor attacks in cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 301-310).
[5] Severi, G., Meyer, J., Coull, S., & Oprea, A. (2021). {Explanation-Guided} backdoor poisoning attacks against malware classifiers. In 30th USENIX security symposium (USENIX security 21) (pp. 1487-1504).
[6] Zhao, P., Zhu, W., Jiao, P., Gao, D., & Wu, O. (2025). Data poisoning in deep learning: A survey. arXiv preprint arXiv: 2503.22759.
[7] Kallas, K., Tannous, C., & FARAOUN, H. (2025). A Game-Theoretic Approach to Cost-Aware Backdoor Attacks and Defenses in Deep Learning. Authorea Preprints.
[8] Xia, P., Li, Z., Zhang, W., & Li, B. (2022). Data-efficient backdoor attacks. arXiv preprint arXiv: 2204.12281.
[9] Abbasi, B. H., Zhang, Y., Zhang, L., & Gao, S. (2025). Backdoor Attacks and Defenses in Computer Vision Domain: A Survey. arXiv preprint arXiv: 2509.07504.
[10] Bai, Y., Xing, G., Wu, H., Rao, Z., Ma, C., Wang, S., ... & Kang, J. (2024). Backdoor attack and defense on deep learning: A survey. IEEE Transactions on Computational Social Systems.