DATAELIXIR: Purifying Poisoned Dataset to Mitigate Backdoor Attacks via Diffusion Models

Jiachen Zhou[1,2], Peizhuo Lv[1,2], Yibing Lan[1,2], Guozhu Meng[1,2*], Kai Chen[1,2*], Hualong Ma[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China
{zhoujiachen, lvpeizhuo, lanyibing, mengguozhu, chenkai, mahualong}@iie.ac.cn

Abstract

Dataset sanitization is a widely adopted proactive defense against poisoning-based backdoor attacks, aimed at filtering out and removing poisoned samples from training datasets. However, existing methods have shown limited efficacy in countering the ever-evolving trigger functions, and often leading to considerable degradation of benign accuracy. In this paper, we propose DATAELIXIR, a novel sanitization approach tailored to purify poisoned datasets. We leverage diffusion models to eliminate trigger features and restore benign features, thereby turning the poisoned samples into benign ones. Specifically, with multiple iterations of the forward and reverse process, we extract intermediary images and their predicted labels for each sample in the original dataset. Then, we identify anomalous samples in terms of the presence of label transition of the intermediary images, detect the target label by quantifying distribution discrepancy, select their purified images considering pixel and feature distance, and determine their ground-truth labels by training a benign model. Experiments conducted on 9 popular attacks demonstrates that DATAELIXIR effectively mitigates various complex attacks while exerting minimal impact on benign accuracy, surpassing the performance of baseline defense methods.

Introduction

Deep neural networks (DNNs) have exhibited remarkable performance in various fields, including autonomous driving (Grigorescu et al. 2020), facial recognition (Wang and Deng 2021) and medical image analysis (Shen, Wu, and Suk 2017), *etc*. Advanced DNN models necessitate large and diverse high-quality data for training. However, collecting high-quality data is an exceedingly resource-intensive task. For instance, ImageNet (Deng et al. 2009) took more than two years to construct and consumed substantial resources. Therefore, it has become customary for individuals to leverage data from external sources, such as data markets and crowd-sourcing platforms (Zeng et al. 2023).

However, except for a few authoritative and community-maintained sources that are well maintained and thereby have high-quality data (HuggingFace 2016), other sources may be poisoned with backdoor samples by miscreants (Li et al. 2022). In particular, backdoor attacks craft poisoned samples by using trigger functions to modify benign images. Victim models trained on these poisoned datasets, although can make accurate predictions for benign inputs, will be controlled to produce wrong outputs given poisoned inputs.

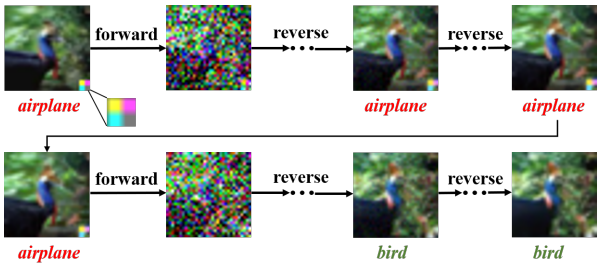


Figure 1: The forward and reverse process in diffusion models. The poisoned image is misclassified as "airplane" due to the trigger in the lower right. The forward process eliminates this trigger by introducing noise, while the reverse process restores the benign features, correcting the model to its ground-truth label "bird".

A straightforward and effective way to defend such backdoor attacks is to identify in the dataset the poisoned samples and further remove or replace them. Prior studies propose dataset sanitization methods to identify poisoned samples by detecting their anomalies in model activations (Chen et al. 2019), spectral signatures (Tran, Li, and Madry 2018), and loss values (Li et al. 2021c; Huang et al. 2022). Subsequently, they remove identified anomalous samples from the dataset to train benign models free from backdoors. However, existing methods have exhibited limited effectiveness, only working on specific types of trigger functions and often resulting in considerable degradation of benign accuracy (Li et al. 2022; Wu et al. 2022). Similar challenges are also encountered in other types of defense approaches, including model reconstruction (Wang et al. 2019; Wu and Wang 2021; Li et al. 2021d; Zeng et al. 2022) and input purification (Doan, Abbasnejad, and Ranasinghe 2020; Shi et al. 2023; May et al. 2023).

In this paper, we propose DATAELIXIR, a novel sanitization approach tailored to purify poisoned datasets. We utilize the forward and reverse process of diffusion models to

*Corresponding Authors.

turn poisoned samples into benign ones, thus training benign models devoid of backdoors. As shown in Figure 1, the forward process introduces noise to the image, thus eliminating the trigger features and preventing the victim model from making predictions based on them. While the reverse process restores the benign features of the image, making the victim model to reclassify it as its ground-truth label.

Specifically, for each sample in the training dataset, we perform iterative rounds through the forward and reverse process and extract intermediary images from each round to construct the candidate set. Based on the presence of label transition in their candidate sets, we categorize samples into three types: the Benign Set, Poisoned Set and Suspicious Set, of which the latter two are regarded as anomalous. Then we detect the target label by quantifying its distribution discrepancy in the candidate sets of anomalous samples. Ultimately, for these anomalous samples, we select their purified images from their respective candidate sets considering pixel and feature distance, and determine their ground-truth labels through the training of a benign model.

We evaluate the efficacy of DATAELIXIR against 9 popular backdoor attacks. Compared to 4 baseline defense methods aimed at dataset sanitation, DATAELIXIR demonstrates superior performance by effectively mitigating diverse trigger functions, while maintaining the benign accuracy of the model. In particular, in regard to the detection rate of poisoned samples, our approach achieves considerably increases spanning from 37.02% to 56.77% on CIFAR10 and from 1.85% to 62.10% on Tiny ImageNet, as well as the decrease in false positive rates from 2.11% to 47.46% on CIFAR10 and from 0.38% to 35.60% on Tiny ImageNet. Moreover, the benign models obtained by DATAELIXIR also exhibit remarkable enhancements, with an average increase in accuracy from 2.81% to 17.27% on CIFAR10 and from 0.20% to 36.90% on Tiny ImageNet, and the decrease in backdoor attack success rate from 10.73% to 84.81% on CIFAR10 and from 8.45% to 97.56% on Tiny ImageNet.

Our contributions are summarized as follows:

- We propose DATAELIXIR[1], which is able to accurately identify poisoned samples as well as their target labels without prior knowledge of trigger functions and patterns. The poisoned samples can be further purified and replaced with the benign through diffusion models.
- Experiments against 9 popular attacks highlight the effectiveness and generalizability of DATAELIXIR in mitigating backdoors, while preserving the benign accuracy, and outperforming mainstream defense methods.

Related Work

Diffusion Model

Diffusion models (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021; Rombach et al. 2022) have exhibited excellent performance in generating diverse and high-quality images. The original DDPM consists of two Markov Chains, wherein the forward process adds noise to the image, and the reverse process restores the image from Gaussian noise using neural networks. Recent studies (Yoon, Hwang, and Lee 2021; Nie et al. 2022; Xiao et al. 2023; Carlini et al. 2023) have attempted to employ diffusion models for adversarial purification. Different from these methods primarily target the purification of adversarial perturbation in inputs during the inference stage, DATAELIXIR harnesses the potential of diffusion models to counter backdoor triggers, and aims at purifying datasets poisoned by backdoor attacks to serve for training benign models.

Backdoor Attack

Backdoor attacks aim to induce the victim model to accurately predict labels for benign inputs, while intentionally misclassifying poisoned inputs to the target label. Among various techniques of backdoor attacks, poisoning the training datasets is the most common method. Attackers craft poisoned samples by employing trigger functions to modify benign images and altering their labels to the target one. Since the pioneering backdoor attack BadNets (Gu et al. 2019) that simply stamped patterns (*e.g.*, 3×3 white square) on benign images, various trigger functions have been proposed to enhance the attack effectiveness and imperceptibility, including Invisible triggers (Chen et al. 2017; Li et al. 2020), Clean Label triggers (Shafahi et al. 2018), Sample Specific triggers (Nguyen and Tran 2020, 2021; Doan et al. 2021; Li et al. 2021b), and Frequency triggers (Zeng et al. 2021; Wang et al. 2022), *etc.*

Besides poisoning-based backdoor attacks, there exist other methods to embed backdoors without poisoning training data, including manipulating model weights (Garg et al. 2020; Rakin, He, and Fan 2020) or altering model structures (Tang et al. 2020; Li et al. 2021a). The primary focus of our study is on purifying the poisoned training datasets to defend backdoor attacks as 1) data poisoning is the most generalized way to implant backdoors due to its agnosticism to model structures. 2) it takes only one-time effort but can benefit an unlimited number of model training tasks.

Backdoor Defense

To defend against poisoning-based backdoor attacks, several approaches have been proposed to sanitize the poisoned datasets. These methods focus on filtering out poisoned samples based on their anomalies in model activations (Chen et al. 2019), spectral signatures (Tran, Li, and Madry 2018), and loss values (Li et al. 2021c; Huang et al. 2022). Then identified samples are removed to avoid potential injection of backdoors and enable the training of benign models.

Except for dataset sanitation, model reconstruction aims to mitigate the impact of backdoors in the infected model by pruning poisoned neurons (Wu and Wang 2021), fine-tuning model weights (Li et al. 2021d; Zeng et al. 2022), and reversing backdoor triggers (Wang et al. 2019). While input purification seeks to purify poisoned inputs using generative adversarial networks (Doan, Abbasnejad, and Ranasinghe 2020; Cheng et al. 2023) and diffusion models (Shi et al. 2023; May et al. 2023) during the inference stage, thus preventing them to activate backdoors in the infected model.

[1]https://github.com/Manu21JC/DataElixir

Unlike previous dataset sanitization methods, our approach aims to purify identified poisoned samples into benign instances instead of outright removal, including the determination of their purified images and ground-truth labels. While some input purification methods also utilize diffusion models to eliminate the impact of triggers, DATAELIXIR is different from them as 1) the purification of inputs occurs during the inference stage, aimed at preventing the activation of backdoors in the infected model rather than clearing them from existence. 2) these methods focus solely on image purification, while DATAELIXIR encompasses a broader scope, including identifying the poisoned samples and determining their ground-truth labels.

Methodology

In this section, we present the threat model and methodology of our proposed approach. As shown in Figure 2, DATAELIXIR consists of four stages: Candidate Set Construction, Anomalous Samples Identification, Target Label Detection and Purified Dataset Generation. Comprehensive technical details are introduced in the following subsections.

Threat Model

Given the untrustworthy data from external sources, the defender needs to determine whether it is poisoned, and if poisoned, identify and purify the poisoned samples for training benign models free from backdoors. This defense mechanism operates in a black-box manner, without prior knowledge of the attack settings (*e.g.*, the target label, poison rate and trigger functions), as well as any known benign or poisoned samples from the given data. We assume that the defender has access to pre-trained diffusion models representing identical or similar distributions (consistent with prior studies (Nie et al. 2022; Xiao et al. 2023; May et al. 2023; Shi et al. 2023; Dolatabadi, Erfani, and Leckie 2023; Jiang et al. 2023)), which can be sourced as off-the-shelf models or trained using data provided from authoritative and trustworthy communities and institutions. Beyond this assumption, our experiments also validate the effectiveness of DATAELIXIR using diffusion models trained on disparate data and even poisoned data.

Candidate Set Construction

The raw dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ may contain meticulously crafted poisoned samples that inject backdoors into the victim model M trained on it. As shown in Figure 1, we utilize the forward and reverse process of diffusion models for the purification of poisoned samples. Specifically, the forward process q adds Gaussian noise to the original image x^0 for T steps (from x^0 to x^T) according to a variance schedule $\{\beta^t \in (0,1)\}_{t=1}^{T}$, yielding the noised image x^T:

$$\begin{aligned} q(x^{1:T}|x^0) &:= \prod_{t=1}^{T} q(x^t|x^{t-1}) \\ q(x^t|x^{t-1}) &:= \mathcal{N}(x^t; \sqrt{1-\beta^t}x^{t-1}, \beta^t \boldsymbol{I}) \end{aligned} \tag{1}$$

While training on the poisoned dataset, the model M is extremely overfitting to the trigger features. This prompts M to predict the poisoned image as the target label, instead of its ground-truth label based on its benign features. During the forward process, the introduction of noise eliminates these trigger features thus would change the prediction of M. However, the benign features are also degraded by the noise. We use the reverse process p_θ to restore benign features (from x^T back to x^0), yielding the purified image $\widetilde{x}$:

$$\begin{aligned} p_\theta(x^{0:T}) &:= p(x^T) \prod_{t=1}^{T} p_\theta(x^{t-1}|x^t) \\ p_\theta(x^{t-1}|x^t) &:= \mathcal{N}(x^{t-1}; \mu_\theta(x^t, t), \Sigma_\theta(x^t, t)) \end{aligned} \tag{2}$$

The mean $\mu_\theta(x^t, t)$ is learned by a neural network parameterized by θ, and the variance $\Sigma_\theta(x^t, t)$ can either be time-step dependent constants (Ho, Jain, and Abbeel 2020) or learned by a neural network (Nichol and Dhariwal 2021). The reverse process restores the image with guidance from the remaining benign features. Since the diffusion model represents the benign distribution, the benign features degraded during the forward process can be well-recovered whereas the trigger features cannot. Consequently, the purified image $\widetilde{x}$ only contains the benign features of x, thus making M correctly classify it as its ground-truth label.

During the purification of each image, to effectively eliminate all trigger features, we iterate through the forward and reverse process for n rounds, where the reverse process of each round consists of T steps, thereby producing $n \times T$ intermediary images. If an image is poisoned, the labels of its intermediary images transit from the target label back to the ground-truth label as the trigger features are eliminated. In contrast, the purification of the benign image does not involve such label transition. Thus, the presence of transition in the labels of the intermediary images provides a strong indication that the sample is likely poisoned. Specifically, for the ith sample (x_i, y_i) in D, we perform the forward and reverse process for n rounds, and extract intermediary images from the last m steps in the reverse process of each round to construct its candidate set $C_{(x_i,y_i)} = \{(x_j, y_j)\}_{j=1}^{n \times m}$. We can then determine whether the sample (x_i, y_i) is anomalous based on the presence of label transition in $C_{(x_i,y_i)}$.

Anomalous Samples Identification

Label Transition Analysis. When assessing the presence of label transition in the candidate set, we consider the second-highest count of labels as the transition coefficient η. If η surpasses the predefined threshold τ, we consider it indicative of label transition. The motivation behind η is as follows. Ideally, the label distribution of the candidate set for benign sample should exhibit a unimodal distribution, wherein all labels align with the ground-truth label. While the label distribution transits to a bimodal distribution with $\eta \geq \tau$, it indicates that the sample is poisoned. The appearance of the second peak in the distribution is most likely due to the transition from the target label to the ground-truth label, a direct result of the effective elimination of trigger features.

We categorize samples in the raw dataset D into three types based on the presence of label transition in the candidate set: the Benign Set B, Poisoned Set P and Suspicious Set S, where samples in P and S are considered anomalous:

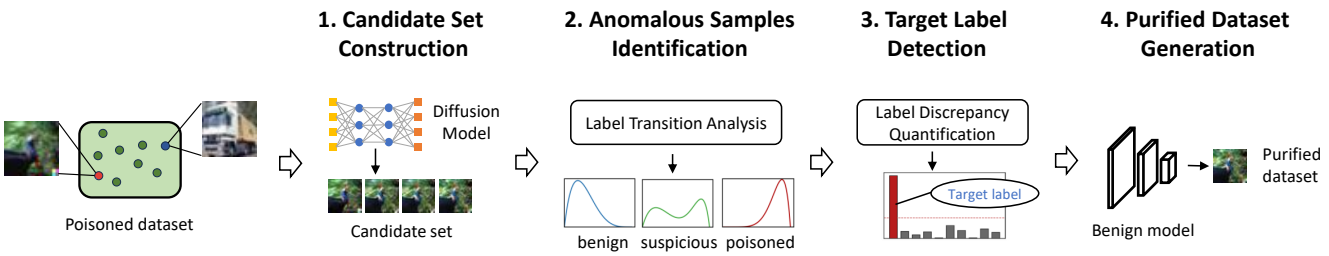


Figure 2: Overview of DATAELIXIR.

$$(x_i, y_i) \in \begin{cases} B, & \text{if } \eta < \tau, \forall (x_j, y_j) \in C_{(x_i,y_i)}, y_j = y_i \\ P, & \text{if } \eta < \tau, \forall (x_j, y_j) \in C_{(x_i,y_i)}, y_j = y_g \\ S, & \text{if } \eta \geq \tau \end{cases} \tag{3}$$
$$B \cup P \cup S = D$$

In particular, for the sample (x_i, y_i) where $\eta < \tau$, if the labels of $C_{(x_i,y_i)}$ remain consistent with y_i, this provides conclusive evidence that the sample is benign, with its benign features remaining intact throughout the entire purification process. Such samples constitute the Benign Set B, with x_i as the purified image and y_i as the ground-truth label, as neither have been poisoned or modified.

Conversely, the Poisoned Set P also comprises samples where $\eta < \tau$, but the labels of $C_{(x_i,y_i)}$ deviate from y_i, indicating that these samples are poisoned, with trigger features effectively eliminated during the initial round of forward and reverse process and then classified as their ground-truth label y_g. For such samples, we need to select their purified images from their candidate set since x_i has been poisoned, and y_g can be directly used as the ground-truth label.

The Suspicious Set S comprises samples where $\eta \geq \tau$, which can arise in two cases. Firstly, trigger features on the poisoned image may not be effectively eliminated until later iterations, causing the labels to exhibit the transition from the target label to the ground-truth label. Secondly, benign features on the benign image may be excessively corrupted, leading to false positives. For samples in S, we next need to further select their purified images from their candidate set and determine their ground-truth label.

Target Label Detection

Upon identifying the anomalous samples in Poisoned Set P and Suspicious Set S, we proceed to determine whether the dataset is poisoned by detecting the target label. If any outlier label is detected, it indicates that the dataset has been poisoned, with this label corresponding to the target label. This detection process leverages a crucial insight: there exists distribution discrepancy between samples with the target label and those with benign labels in $P \cup S$. This discrepancy arises because the samples with the target label in $P \cup S$ mainly consists of poisoned samples, while samples with benign labels are primarily false positives of benign samples. We utilize the candidate sets of samples with each label to characterize such distribution discrepancy, followed by employing outlier detection methods such as *Median Absolute Deviation* (Leys et al. 2013) to detect the target label.

Label Discrepancy Analysis. Specifically, for each label y, we construct $\mathcal{C}_{\mathrm{y}}$ as the union of the candidate sets $C_{(x_i,y_i)}$ for samples (x_i, y_i) in $P \cup S$ with label y as follows:

$$\mathcal{C}_{\mathrm{y}} = \bigcup_{(x_i,y_i) \in P \cup S} \{C_{(x_i,y_i)} \mid y_i = \mathrm{y}\} \tag{4}$$

We provide two metrics to measure label discrepancy for the candidate set. Firstly, we count the number of samples in $\mathcal{C}_{\mathrm{y}}$, as the count of false positives of benign samples is lower than the count of poisoned samples. Secondly, we compute the entropy of the label distribution in $\mathcal{C}_{\mathrm{y}}$. Since the poisoned samples originate from many other benign labels, during the purification process, they are reclassified to their ground-truth labels due to the effective removal of trigger features. Therefore, the label distribution in $\mathcal{C}_{\mathrm{y}}$ for the target label is more chaotic than that of benign labels, where the samples are mere false positives of benign samples. For $\mathcal{C}_{\mathrm{y}}$ of each label, we calculate the *Kullback-Leibler divergence* (Kullback and Leibler 1951) between the actual distribution and the ideal distribution where all labels are assumed to be identical. A higher value of *KL divergence* indicates a higher likelihood of the label being the target label.

Purified Dataset Generation

We can generate the purified dataset $\widetilde{D} = \{(\widetilde{x}_i, \widetilde{y}_i)\}_{i=1}^{N}$. For each sample (x_i, y_i) with benign labels in $B \cup P \cup S$ and with the target label in B, it can be directly used as the purified sample $(\widetilde{x}_i, \widetilde{y}_i)$ since it is free from poisoning. While handling the sample with the target label in P, we utilize the consensus label in $C_{(x_i,y_i)}$ (y_g in Equation 3) as $\widetilde{y}_i$, and select $\widetilde{x}_i$ from $C_{(x_i,y_i)}$ based on the following insight: the forward and reverse process can effectively eliminate the trigger features while preserving the benign features of x_i. If the image x_j from $C_{(x_i,y_i)}$ significantly differs from the poisoned image x_i, it is likely due to the effective removal of the trigger features on x_j. Therefore, we select $\widetilde{x}_i$ from $C_{(x_i,y_i)}$ by choosing x_j that is distinct from x_i, thus only containing the benign features of x_i:

$$dist(x_i, x_j) = ssim(x_i, x_j) + cos(M(x_i), M(x_j)) \quad where\ x_j \in C_{(x_i,y_i)} \tag{5}$$

To quantify the distance between x_i and x_j, we employ *Structure Similarity Index Measure* (Wang et al. 2004) in

the pixel space and *Cosine Similarity* in the feature space, using the victim model M trained on the raw dataset D as the feature extractor. We select $\widetilde{x}_i$ from $C_{(x_i,y_i)}$ that ranks at 80% concerning the distance from x_i, to ensure the effective and comprehensive removal of trigger features.

For the sample (x_i, y_i) with the target label in S, we opt to train a benign model $\widetilde{M}$ on the currently available purified dataset $\widetilde{D}$ to assist in determining its ground-truth label $\widetilde{y}_i$. The current $\widetilde{D}$ contains all the samples with benign labels, as well as those with the target label in $B \cup P$. The training of $\widetilde{M}$ involves two stages. In the first stage, we train $\widetilde{M}$ on $\widetilde{D}$ to predict the label y_m for the suspicious sample (x_i, y_i). Since $\widetilde{D}$ exclusively comprises either benign or purified samples, during training, $\widetilde{M}$ only learns benign features devoid of any backdoors, thus can assign the ground-truth labels to suspicious samples. Furthermore, we employ a voting mechanism based on the candidate set for the sample to obtain the label y_v, which is the most frequently occurring label in the candidate set. For the sample where y_m is equal to y_v, we employ the resulting label as $\widetilde{y}_i$ and use Equation 5 to select $\widetilde{x}_i$ from the candidate set. Such purified sample $(\widetilde{x}_i, \widetilde{y}_i)$ is then added to $\widetilde{D}$ for the second stage of training, where we fine-tune $\widetilde{M}$ on the updated $\widetilde{D}$ with a lower learning rate to obtain the final benign model. Then each remaining suspicious sample is purified using $\widetilde{M}$, resulting in the final purified dataset $\widetilde{D}$.

Evaluation

Experimental Setup

Datasets and Models. We conduct experiments on three datasets: CIFAR10 (Krizhevsky 2009), Tiny ImageNet (Le and Yang 2015) and LFW (Huang et al. 2008), using PreAct-ResNet18 (He et al. 2016), EfficientNet-B4 (Tan and Le 2019), VGGFace (Parkhi, Vedaldi, and Zisserman 2015) as classifier and DDPM (Ho, Jain, and Abbeel 2020) trained on CIFAR10, Improved DDPM (Nichol and Dhariwal 2021) trained on ImageNet-1k (Deng et al. 2009), DDPM trained on CelebA-HQ (Karras et al. 2018) for purification.

Attack Methods. We evaluate nine popular attack methods, including BadNets (Gu et al. 2019), Blended (Chen et al. 2017), SSBA (Li et al. 2020), LC (Shafahi et al. 2018), LF (Zeng et al. 2021), Ftrojan (Wang et al. 2022), WaNet (Nguyen and Tran 2021), LIRA (Doan et al. 2021) and IA (Nguyen and Tran 2020), using BackdoorBench (Wu et al. 2022) for implementation.

Parameter Settings. The poison rate is set to 10% and the target label is randomly selected without any specific strategy or preference. While using pre-trained diffusion models, we adhere to the default hyper-parameters provided by the source. The hyper-parameters of DATAELIXIR are set to: $T = 150$, $n = 5$, $m = 10$ and $\tau = 5$ $(0.1 \times n \times m)$. We conduct preliminary experiments to study the impacts of hyper-parameters on CIFAR10 and determine these default settings, which also proved effective in countering diverse attacks across various datasets.

Metrics. The metrics for evaluation are as follows:

- *True Positive Rate (TPR)* measures the probability that DATAELIXIR correctly detects poisoned samples in the raw dataset. Specifically, a sample is considered to be poisoned if its label in the raw dataset D deviates from that in the purified dataset $\widetilde{D}$.
- *False Positive Rate (FPR)* measures the probability that DATAELIXIR falsely detects benign samples as poisoned.
- *Accuracy (ACC)* measures the probability that the model correctly classifies benign inputs.
- *Attack Success Rate (ASR)* measures the probability that the model classifies poisoned inputs as the target label.

Defense Performance

As shown in Table 1, we compare DATAELIXIR with four baseline defense methods aimed at dataset sanitation, including AC (Chen et al. 2019), Spectral (Tran, Li, and Madry 2018), ABL (Li et al. 2021c) and DBD (Huang et al. 2022). Notably, DATAELIXIR effectively defends against all nine evaluated backdoor attacks, achieving the highest TPR, the lowest FPR, the highest ACC, and the lowest ASR, with averages of 97.92%, 1.27%, 93.35%, 1.51% on CIFAR10 and 97.71%, 0.15%, 65.73% and 0.79% on Tiny ImageNet. In contrast, the efficacy of baseline methods varies significantly across different attacks and datasets. Specifically, AC exhibits limited performance against diverse attacks, Spectral is effective only for certain attacks, and DBD fails on Tiny ImageNet. While ABL shows lower ASR against some attacks, it does so by sacrificing benign accuracy and also proves inadequate against WaNet on CIFAR10 and LF on Tiny ImageNet. In comparison, DATAELIXIR exhibits excellent defense effectiveness while maintaining the model's performance on the benign task, with an average increase of 0.26% on CIFAR10 and 0.91% on Tiny ImageNet compared to the victim model.

Moreover, we compare DATAELIXIR with other types of defense methods including model reconstruction (Wang et al. 2019; Li et al. 2021d; Wu and Wang 2021; Zeng et al. 2022) and input purification (Doan, Abbasnejad, and Ranasinghe 2020; Shi et al. 2023; May et al. 2023). Our approach exhibits superior defense performance. We also validate the effectiveness of DATAELIXIR on LFW, and the results are shown in Table 2. DATAELIXIR effectively mitigates all attacks, achieving the average TPR, FPR, ACC and ASR of 98.75%, 1.32%, 94.23% and 0.37%, respectively.

In addition, DATAELIXIR exhibits excellent performance in target label detection. While countering poisoned CIFAR10, DATAELIXIR accurately detects the target label with an average anomaly index of 5.4323, and produces no false positives for benign labels with the average maximum anomaly index of 1.0010. For benign CIFAR10, DATAELIXIR also avoids false positives, with the maximum anomaly index of 1.6992, below the threshold value of 2.

Evaluation on Purified Dataset

We conduct experiments to evaluate the purified datasets generated by DATAELIXIR in terms of label accuracy, image quality, and the performance of newly trained models. The average label accuracy is 98.10% on CIFAR10 and 95.00% on Tiny ImageNet. Besides, the average *Fréchet Inception*

Attack	AC				Spectral				ABL				DBD				DataElixir			
	TPR	FPR	ACC	ASR	TPR	FPR	ACC	ASR	TPR	FPR	ACC	ASR	TPR	FPR	ACC	ASR	TPR	FPR	ACC	ASR
BadNets	48.66	48.81	88.32	99.61	80.28	4.41	92.30	98.00	98.56	**0.16**	86.90	**0.00**	0.22	11.09	76.55	4.62	**99.40**	0.89	**93.24**	0.73
Blended	48.06	48.84	88.86	95.62	98.04	2.44	92.66	57.44	86.64	1.48	85.56	**0.11**	67.60	3.60	75.50	98.60	**99.18**	**1.15**	**93.15**	1.11
SSBA	48.50	48.62	88.37	94.41	54.38	7.29	90.74	87.16	57.72	4.70	87.37	7.84	50.76	5.47	75.99	4.55	**98.62**	**0.87**	**93.23**	**1.31**
LF	49.38	49.03	88.81	98.21	24.46	10.62	86.22	99.01	49.76	5.58	88.23	**0.96**	50.96	5.45	75.49	5.26	**95.74**	**0.96**	**93.65**	4.61
LC	48.20	49.12	89.39	11.40	85.20	8.23	90.29	2.34	0.00	1.01	82.00	6.00	0.00	1.01	76.57	5.45	**92.80**	**3.82**	**93.37**	**1.91**
WaNet	48.89	48.52	89.14	91.79	47.65	7.90	91.06	87.59	20.29	8.94	89.09	82.62	52.43	5.61	76.36	5.94	**99.40**	**1.20**	**92.61**	**1.00**
IA	48.06	48.93	89.67	99.90	65.19	6.08	92.32	98.77	85.24	2.22	86.45	**0.18**	52.50	5.60	75.22	6.39	**99.38**	**0.79**	**93.84**	0.58
LIRA	48.29	47.93	89.35	99.59	31.98	8.92	88.74	99.76	77.97	2.97	87.76	**0.24**	54.74	5.37	76.94	3.69	**98.81**	**0.45**	**93.68**	0.80
Average	48.51	48.73	88.99	86.32	60.90	6.99	90.54	78.76	59.52	3.38	86.67	12.24	41.15	5.40	76.08	16.81	**97.92**	**1.27**	**93.35**	**1.51**
BadNets	35.44	35.86	62.64	100.00	94.50	0.00	65.39	99.46	96.77	0.36	62.17	0.00	95.24	0.53	36.65	99.97	**99.96**	**0.15**	**65.64**	0.05
Blended	36.45	36.71	61.84	95.21	94.14	**0.04**	**66.07**	7.83	79.93	2.23	59.58	**0.12**	95.25	0.53	33.14	99.14	**97.52**	0.16	65.77	0.48
SSBA	33.94	34.46	62.71	98.18	89.93	0.51	65.52	91.30	85.55	1.61	59.35	**0.07**	95.25	0.53	10.09	98.05	**99.66**	**0.16**	**66.69**	0.41
LF	34.68	35.10	62.53	96.23	90.41	0.45	65.45	56.49	61.84	4.24	59.11	64.47	95.20	0.53	26.03	93.01	**95.96**	**0.14**	**65.74**	**2.08**
Ftrojan	35.38	35.62	61.72	99.91	94.45	**0.00**	65.67	13.42	74.58	2.82	61.55	**0.01**	95.28	0.52	29.65	99.75	**98.11**	0.13	**66.12**	1.85
WaNet	36.25	35.73	61.73	95.93	94.10	**0.05**	64.96	5.84	34.69	7.33	62.46	**0.00**	97.39	0.54	32.80	98.56	**99.72**	0.14	**65.18**	0.16
IA	37.15	36.79	58.55	99.89	90.92	3.99	65.65	97.98	87.72	1.59	58.90	**0.01**	97.41	0.54	33.48	99.94	**95.00**	**0.18**	**64.96**	0.51
Average	35.61	35.75	61.67	97.91	92.64	0.72	65.53	53.19	74.44	2.88	60.45	9.24	95.86	0.53	28.83	98.35	**97.71**	**0.15**	**65.73**	**0.79**

Table 1: Defense performance on CIFAR10 (upper part) and Tiny ImageNet (lower part).

Attack	No Defense		DataElixir			
	ACC	ASR	TPR	FPR	ACC	ASR
BadNets	94.52	99.84	98.33	1.34	94.03	0.82
Blended	94.68	99.34	99.58	1.25	94.35	0.16
LF	92.90	94.10	97.92	1.34	94.19	0.49
Ftrojan	94.52	93.44	99.17	1.34	94.35	0.00

Table 2: Defense performance on LFW.

Attack	Similar Distribution				Disparate Distribution			
	TPR	FPR	ACC	ASR	TPR	FPR	ACC	ASR
BadNets	98.54	1.56	92.47	0.82	97.84	4.62	89.53	1.36
Blended	98.72	1.73	92.52	0.70	98.52	4.69	90.45	1.20
SSBA	99.10	1.44	92.82	0.59	99.00	6.79	89.34	0.76
IA	99.17	1.34	92.65	0.56	98.44	4.17	90.50	0.66
LIRA	99.02	1.19	93.12	0.69	98.46	3.49	91.28	0.64

Table 3: Impact of distribution of diffusion model.

Distance (FID) (Heusel et al. 2017) and *Inception Score (IS)* (Salimans et al. 2016) of purified images are 0.71 and 9.99 on CIFAR10, and 0.35 and 28.62 on Tiny ImageNet. The newly trained models on these purified datasets achieve an average ACC and ASR of 93.05% and 1.13% on CIFAR10 and 65.75% and 0.41% on Tiny ImageNet.

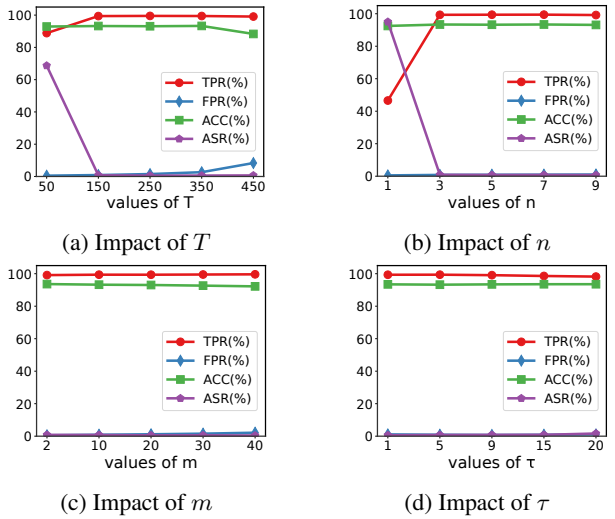
Defense against Various Attack Scenarios

We evaluate the performance of DataElixir against various attack scenarios using datasets poisoned with poison rates in $\{1\%, 3\%, 5\%, 7\%, 10\%\}$, target label numbers in $\{1, 2, 3\}$ and target label selected in $\{0, 2, 4, 6, 8\}$. In all scenarios, our approach demonstrates effective performance, with the average TPR, FPR, ACC, ASR of 99.00%, 0.65%, 93.68%, 0.62% for varying poison rates, 99.10%, 1.20%, 93.34%, 0.54% for varying target label numbers, and 99.16%, 0.82%, 93.47%, 0.71% for different target labels, validating the robustness of DataElixir.

Ablation Study

Impact of Diffusion Models. We evaluate the impact of diffusion models representing different distributions on CIFAR10. In addition to the identical distribution using DDPM trained on CIFAR10, we consider the similar distribution using Improved DDPM trained on ImageNet-1k and disparate distribution using DDPM trained on CelebA (Liu et al. 2015). As shown in Table 3, the similar diffusion model achieves results comparable to those of the identical diffusion model. On the other hand, the disparate diffusion model also proves effective in countering various attacks, while it slightly affects the benign accuracy due to its partial restoration of degraded benign features. However, it still surpasses the performance of the baseline defense methods outlined in Table 1. These results validate the practical applicability of DataElixir in realistic scenarios.

Impact of Hyper-parameters. As shown in Figure 3, we assess the impact of hyper-parameters on the performance of



(a) Impact of T (b) Impact of n

(c) Impact of m (d) Impact of τ

Figure 3: Impact of hyper-parameters.

Attack	BadNets	Blended	SSBA	IA	LIRA
TPR	85.48	98.52	99.02	99.57	98.87
FPR	1.00	1.41	1.34	1.08	0.58
ACC	91.66	93.02	93.09	93.17	93.61
ASR	99.76	1.54	0.97	0.71	0.78
ASR-BadNets	99.76	0.64	0.50	0.62	0.71

Table 4: Defense performance using diffusion model trained on poisoned dataset.

Attack	BadNets	Blended	SSBA	IA	LIRA
ASR	3.12	3.77	4.46	4.06	4.10

Table 5: ASR of the potential residual backdoor.

DATAELIXIR. Our experiments are conducted on CIFAR10 using BadNets. T and n mainly influence the magnitude of the noise introduction, where lower values lead to lower TPR and higher ASR, indicating ineffective removal of trigger features, and higher values of T lead to higher FPR and lower ACC, indicating excessive degradation of benign features. m dictates when we extract intermediary images from the reverse process, and higher values correspond to higher FPR and lower ACC, indicating incomplete restoration of benign features. τ determines the strictness of label transition judgment, with higher values signifying more lenient judgments and resulting in lower TPR and higher ASR.

We also evaluate the impacts of our proposed purified image selection strategy and ground-truth label determination strategy, which lead to the increase of 58.88% and 9.92% in TPR, along with the decrease of 97.71% and 62.90% in ASR compared to the baseline strategy.

Discussion

Adaptive Attack

Poisoned Diffusion Model. We validate the performance of DATAELIXIR when using diffusion model trained on the poisoned dataset. Our experiments are conducted on CIFAR10, using the diffusion model trained on data poisoned by BadNets. As shown in Table 4, this poisoned diffusion model fails to counter the corresponding backdoor attack. Nevertheless, it retains the ability to purify training datasets poisoned by various other backdoor attacks, all without introducing BadNets into the resulting purified datasets. Consequently, in scenarios where access to trustworthy data and diffusion models are unavailable, we propose training the diffusion model using data from diverse untrustworthy sources. Notably, DATAELIXIR performs effectively as long as the data used for diffusion model training and the data intended for purification are not poisoned by the same trigger.

Residual Backdoor. We assume that the attacker possesses comprehensive knowledge of the defense mechanism of DATAELIXIR. The attacker considers an adaptive attack presuming that residual trigger features remain in purified images, and such remaining features can inject a residual backdoor into the model trained on them (different from the original backdoor injected by the full trigger features). Specifically, the attacker employs the same purification strategy to generate poisoned inputs that contain the same residual trigger features, aiming to activate the residual backdoor in the benign model trained on the purified dataset. We evaluate such adaptive attack on CIFAR10, and the results are shown in Table 5. Although the defense results are slightly higher than those outlined in Table 1, all ASRs remain below 5%, indicating that such residual backdoor is ineffective.

Limitation

Efficient image sampling has been a persistent challenge for diffusion models since their inception, which also limits our approach, as Candidate Set Construction constitutes approximately 65% of the overall defense duration. However, since the process of constructing candidate set is sample-wise, parallelization techniques can be employed to enhance efficiency while additional computational resources are available. In addition, several methods (Song, Meng, and Ermon 2021; Lu et al. 2022) aimed at accelerating the sampling efficiency of diffusion models have been proposed, we plan to incorporate these methods in DATAELIXIR in future work.

Conclusion

In this paper, we propose DATAELIXIR, a novel dataset sanitization approach to purify poisoned training datasets using diffusion models to effectively defend against poisoning-based backdoor attacks. Our method utilizes the forward and reverse process to construct the candidate set for each sample, enabling the identification of anomalous samples, detection of target labels, selection of purified images, and determination of their ground-truth labels. Experimental results validate that DATAELIXIR can effectively mitigate diverse backdoor attacks while preserving the benign accuracy, outperforming existing backdoor defense methods.

Acknowledgments

We thank all the anonymous reviewers for their constructive feedback. This work is supported in part by National Key Research and Development Program (2020AAA0107800), NSFC (62302498, 92270204), Youth Innovation Promotion Association CAS, Beijing Nova Program and a research grant from Huawei.

References

Carlini, N.; Tramer, F.; Dvijotham, K. D.; Rice, L.; Sun, M.; and Kolter, J. Z. 2023. (Certified!!) Adversarial Robustness for Free! In *The Eleventh International Conference on Learning Representations*.

Chen, B.; Carvalho, W.; Baracaldo, N.; Ludwig, H.; Edwards, B.; Lee, T.; Molloy, I.; and Srivastava, B. 2019. Detecting backdoor attacks on deep neural networks by activation clustering. In *Workshop on Artificial Intelligence Safety*. CEUR-WS.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. arXiv:1712.05526.

Cheng, S.; Tao, G.; Liu, Y.; An, S.; Xu, X.; Feng, S.; Shen, G.; Zhang, K.; Xu, Q.; Ma, S.; and Zhang, X. 2023. BEAGLE: Forensics of Deep Learning Backdoor Attack for Better Defense. In *30th Annual Network and Distributed System Security Symposium*. The Internet Society.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Doan, B. G.; Abbasnejad, E.; and Ranasinghe, D. C. 2020. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Annual Computer Security Applications Conference*, 897–912.

Doan, K.; Lao, Y.; Zhao, W.; and Li, P. 2021. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11966–11976.

Dolatabadi, H. M.; Erfani, S.; and Leckie, C. 2023. The Devil's Advocate: Shattering the Illusion of Unexploitable Data using Diffusion Models. arXiv:2303.08500.

Garg, S.; Kumar, A.; Goel, V.; and Liang, Y. 2020. Can adversarial weight perturbations inject neural backdoors. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2029–2032.

Grigorescu, S.; Trasnea, B.; Cocias, T.; and Macesanu, G. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 630–645. Springer.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Huang, G. B.; Mattar, M.; Berg, T.; and Learned-Miller, E. 2008. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Huang, K.; Li, Y.; Wu, B.; Qin, Z.; and Ren, K. 2022. Backdoor Defense via Decoupling the Training Process. In *International Conference on Learning Representations*.

HuggingFace. 2016. Datasets.

Jiang, W.; Diao, Y.; Wang, H.; Sun, J.; Wang, M.; and Hong, R. 2023. Unlearnable Examples Give a False Sense of Security: Piercing through Unexploitable Data with Learnable Examples. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 8910–8921.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *International Conference on Learning Representations*.

Krizhevsky, A. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Leys, C.; Ley, C.; Klein, O.; Bernard, P.; and Licata, L. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4): 764–766.

Li, S.; Xue, M.; Zhao, B. Z. H.; Zhu, H.; and Zhang, X. 2020. Invisible backdoor attacks on deep neural networks via steganography and regularization. *IEEE Transactions on Dependable and Secure Computing*, 18(5): 2088–2105.

Li, Y.; Hua, J.; Wang, H.; Chen, C.; and Liu, Y. 2021a. DeepPayload: Black-box backdoor attack on deep learning models through neural payload injection. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*, 263–274. IEEE.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; and Lyu, S. 2021b. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16463–16472.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021c. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34: 14900–14912.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021d. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. In *International Conference on Learning Representations*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, 3730–3738.

Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; and Zhu, J. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35: 5775–5787.

May, B. B.; Tatro, N. J.; Kumar, P.; and Shnidman, N. 2023. Salient Conditional Diffusion for Backdoors. In *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*.

Nguyen, T. A.; and Tran, A. 2020. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33: 3454–3464.

Nguyen, T. A.; and Tran, A. T. 2021. WaNet - Imperceptible Warping-based Backdoor Attack. In *International Conference on Learning Representations*.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning*, 16805–16827. PMLR.

Parkhi, O.; Vedaldi, A.; and Zisserman, A. 2015. Deep face recognition. In *BMVC 2015-Proceedings of the British Machine Vision Conference 2015*. British Machine Vision Association.

Rakin, A. S.; He, Z.; and Fan, D. 2020. Tbt: Targeted neural network attack with bit trojan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13198–13207.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Shafahi, A.; Huang, W. R.; Najibi, M.; Suciu, O.; Studer, C.; Dumitras, T.; and Goldstein, T. 2018. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31.

Shen, D.; Wu, G.; and Suk, H.-I. 2017. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19: 221–248.

Shi, Y.; Du, M.; Wu, X.; Guan, Z.; Sun, J.; and Liu, N. 2023. Black-box Backdoor Defense via Zero-shot Image Purification. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Song, J.; Meng, C.; and Ermon, S. 2021. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, 6105–6114. PMLR.

Tang, R.; Du, M.; Liu, N.; Yang, F.; and Hu, X. 2020. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 218–228.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, 707–723. IEEE.

Wang, M.; and Deng, W. 2021. Deep face recognition: A survey. *Neurocomputing*, 429: 215–244.

Wang, T.; Yao, Y.; Xu, F.; An, S.; Tong, H.; and Wang, T. 2022. An invisible black-box backdoor attack through frequency domain. In *European Conference on Computer Vision*, 396–413. Springer.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Wu, B.; Chen, H.; Zhang, M.; Zhu, Z.; Wei, S.; Yuan, D.; and Shen, C. 2022. Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems*, 35: 10546–10559.

Wu, D.; and Wang, Y. 2021. Adversarial neuron pruning purifies backdoored deep models. *Advances in Neural Information Processing Systems*, 34: 16913–16925.

Xiao, C.; Chen, Z.; Jin, K.; Wang, J.; Nie, W.; Liu, M.; Anandkumar, A.; Li, B.; and Song, D. 2023. DensePure: Understanding Diffusion Models for Adversarial Robustness. In *The Eleventh International Conference on Learning Representations*.

Yoon, J.; Hwang, S. J.; and Lee, J. 2021. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, 12062–12072. PMLR.

Zeng, Y.; Chen, S.; Park, W.; Mao, Z.; Jin, M.; and Jia, R. 2022. Adversarial Unlearning of Backdoors via Implicit Hypergradient. In *International Conference on Learning Representations*.

Zeng, Y.; Pan, M.; Jahagirdar, H.; Jin, M.; Lyu, L.; and Jia, R. 2023. Meta-Sift: How to Sift Out a Clean Subset in the Presence of Data Poisoning? In *32nd USENIX Security Symposium (USENIX Security 23)*, 1667–1684. USENIX Association.

Zeng, Y.; Park, W.; Mao, Z. M.; and Jia, R. 2021. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16473–16481.