**International Journal of Electrical and Computer Engineering (IJECE)**
Vol. 14, No. 1, February 2024, pp. 645~652
ISSN: 2088-8708, DOI: 10.11591/ijece.v14i1.pp645-652

# A dilution-based defense method against poisoning attacks on deep learning systems

**Hweerang Park, Youngho Cho**
Department of Defense Science (Computer Engineering and Cyberwarfare Major), Korea National Defense University,
Nonsan, Republic of Korea

**Article Info**

***Article history:***

Received Mar 15, 2023
Revised Aug 18, 2023
Accepted Sep 6, 2023

***Keywords:***

Adversarial attacks
Adversarial defense method
Adversarial machine learning
Backdoor attacks
Deep learning
Poisoning attacks

**ABSTRACT**

Poisoning attack in deep learning (DL) refers to a type of adversarial attack that injects maliciously manipulated data samples into a training dataset for the purpose of forcing a DL model trained based on the poisoned training dataset to misclassify inputs and thus significantly degrading its performance and reliability. Meanwhile, a traditional defense approach against poisoning attacks tries to detect poisoned data samples from the training dataset and then remove them. However, since new sophisticated attacks avoiding existing detection methods continue to emerge, a detection method alone cannot effectively counter poisoning attacks. For this reason, in this paper, we propose a novel dilution-based defense method that mitigates the effect of poisoned data by adding clean data to the training dataset. According to our experiments, our dilution-based defense technique can significantly decrease the success rate of poisoning attacks and improve classification accuracy by effectively reducing the contamination ratio of the manipulated data. Especially, our proposed method outperformed an existing defense method (Cutmix data augmentation) by 20.9%p at most in terms of classification accuracy.

***Corresponding Author:***

Youngho Cho
Department of Defense Science (Computer Engineering Major), Korea National Defense University
Hwangsanbeol-ro 1040, Yangchon-myeon, Nonsan-si, Chungcheongnam-do, Republic of Korea
Email: youngho@kndu.ac.kr

## 1. INTRODUCTION

In conjunction with recent remarkable achievements in the field of deep learning (DL), there is also active research being conducted on adversarial attacks targeting DL systems or models [1]–[5]. As the threat to DL models increases, it is essential to ensure the security and stability of artificial intelligence systems for defending against poisoning attacks [6]–[9]. Recently, the generative pre-trained transformer 3 (GPT-3) model, which ChatGPT is based on, collected data from the internet to generate the model and performed the task of filtering out contaminated data to utilize the collected data as training data [10]. During the model training phase, poisoning attacks introduce contaminated data into the training dataset, resulting in the creation of a flawed model. Therefore, it is crucial to defend against such attacks on training data to ensure the model's accuracy and reliability. The methods for defending against poisoning attacks can be broadly categorized into two approaches: enhancing the model's robustness or detecting and removing contaminated data [11]–[17]. The detection method is a method of determining whether the training data is normal and removing abnormal data before training [16], [17]. However, with the continuous emergence of new state-of-the-art attacks, it remains a difficult challenge to ideally distinguish whether the data collected, through detection methods, is normal or not. Therefore, in this study, a dilution-based defense technique is proposed based on the assumption that it is impossible to perfectly differentiate contaminated data using a detection method. The aim of our dilution-based defense method is to reduce the attack components of contaminated data by increasing the amount of clean data during the training phase.

There has been no related research that increases the amount of training data to reduce the components of contaminated data for defending against poisoning attacks. However, there has been similar research. They proposed a method to defend against poisoning attacks by creating a model with resistance to attacks through the augmentation of the training data [12]–[15]. To distinguish between data augmentation techniques and our proposed defense mechanism based on dilution, we divide the defense stages into two, before and after the poisoning attack (or post-attack and pre-attack), respectively [11]. In the stage before the attack, clean data can be used since there is no poisoning attack yet. However, in the stage after the poisoning attack, clean data cannot be used, and it is difficult to distinguish normal data from contaminated data. As clean data can be used in the stage before the poisoning attack, it can be utilized to augment the data and strengthen the model [11]–[15]. It is possible to defend against poisoning attacks using data augmentation techniques even in the post-attack stage. However, since it is impossible to perfectly distinguish contaminated data, the defense effectiveness of dilution techniques is expected to be weaker than that of clean data when the proportion of contaminated data is high in the training data. To verify this, we conducted experiments comparing our proposed method with an existing method.

The main contributions of this paper are as follows. First, to the best of our knowledge, this is the first study that proposed a dilution-based defense mechanism against poisoning attacks on DL systems. Specifically, we duplicate innocuous clean data in the training dataset and then build a DL model based on it. As a result, our proposed method lowers the impact of contaminated data included in the training dataset and thus significantly reduces the impact of contaminated data added by adversarial attackers in transfer learning environments. Second, we demonstrated the effectiveness of our proposed dilution defense method against poisoning attacks by conducting extensive experiments. According to the experimental results, our dilution-based defense method increased the classification accuracy of a DL model by at most 9.7%p compared to a DL model with no defense mechanism, and 20.9%p higher than a DL model with the existing defense method (Cutmix data augmentation). Furthermore, the attack success rate (ASR) of backdoor attack decreased by 33.5%p.

The rest of this paper is organized as follows: in section 2, we overview the background knowledge and introduce existing studies. In section 3, we design our proposed method based on the analysis of general poisoning attacks. In section 4, we conduct extensive experiments and analyze the results. Finally, we conclude with future research directions in section 5.

## 2. RELATED WORKS

### 2.1. Poisoning attacks

The poisoning attacks occur during the transfer learning process using training data collected from outside source which cannot be completely trusted [18], [19]. Therefore, if the collected dataset contains contaminated data (i.e., poisoned dataset), the DL model trained based on it is also contaminated and thus behaves abnormally. We explain four representative poisoning attack techniques considered in this study as follows. Figure 1 shows these examples of four types of poisoning attacks. Figure 1(a) is a dirty-label poisoning attack that changes the label which is the simplest poisoning attack. Figure 1(b) is a clean-label poisoning attack while Figure 1(c) and Figure 1(d) are examples of backdoor attacks that apply dirty-label and clean-label, respectively.

Poisoning attacks can be classified into a dirty-label attack or a clean-label attack depending on whether the label in a poisoned sample is falsified or not. First, the dirty-label attack is an attack in which an attacker changes the label of training data to reduce the accuracy of the model, as shown in Figure 1(a) [19]. Second, the clean-label attack generates adversarial examples by adding perturbations to existing training images without changing the labels as shown in Figure 1(b) [19], [20]. The clean-label attack is called invisible attack because human eyes hardly detect changes in the poisoned adversarial images produced by this attack and the label is also used as it is [19].

The concepts of the dirty-label attack and the clean-label attack can also be extended to backdoor attacks. The backdoor attack is a special type of poisoning attack that inserts a trigger inducing a specific behavior into the training data, as shown in Figure 1(c) and Figure 1(d) [19]. This attack forces a DL model to conduct specific behaviors such as misclassifying inputs containing the trigger according to the attacker's intention [19]. The clean-label backdoor attack, as shown in Figure 1(d), inserts perturbations into the training data while maintaining the original labels [21]. In subsection 4.1, clean-label poisoning attacks and clean-label backdoor attacks were employed as attack methods.

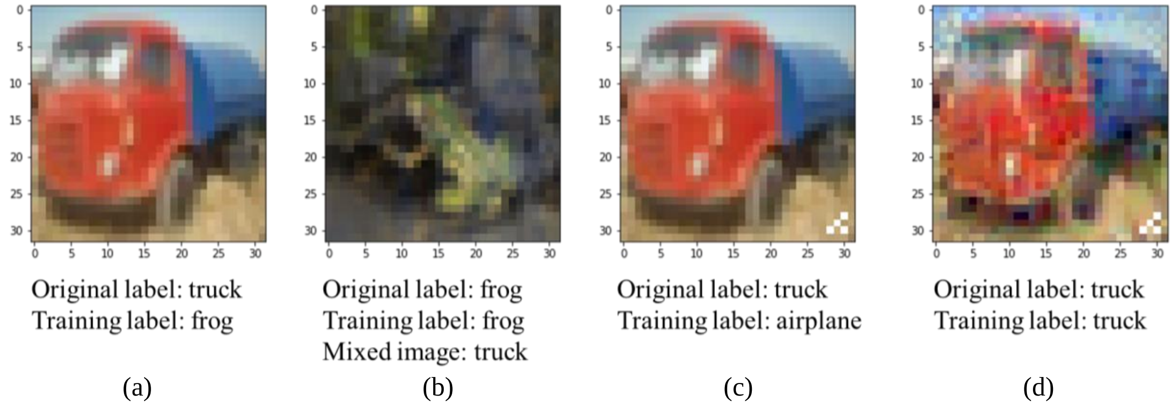


Figure 1. Poisoned examples of (a) dirty-label poisoning attack, (b) clean-label poisoning attack, (c) dirty-label backdoor attack, and (d) clean-label backdoor attack based on an example of CIFAR-10 dataset; 32×32-pixel color images

### 2.2. Existing defense methods

Existing defense methods against poisoning attacks are based on data augmentation (DA) techniques as follows. First, Borgnia *et al.* [12] proposed a method to defend against poisoning attacks by enhancing the robustness of the model using Cutmix DA techniques [13]; Cutmix has been used as an augmentation technique to defend against poisoning attacks [12], [13]. Figure 2 shows a poisoning attack defense method that applies a conventional data augmentation technique. Specifically, Borgnia *et al.* [12] generated an augmented dataset $D_a$ based on the authenticated dataset $D_c$ using the Cutmix technique and trained the model, and the number of data in $D_c$ and $D_a$ is the same at 50,000 [10], [20]. As a result, trained models using such methods have shown lower success rates for poisoning attacks and higher classification accuracy [20]. In addition, Veldanda *et al.* [14] proposed a data augmentation technique that adds noise to the training data during the pre-processing stage to defend against BadNets that attackers may generate when downloading data from the internet. Qiu *et al.* [15] used 71 data augmentation techniques to transform images in the training data during both the training and inference phases. As a result, they showed that this technique effectively mitigated eight types of backdoor attacks and demonstrated superior performance compared to five existing defense methods.

Most DA techniques have focused on modifying images before training at stage $t_1$ to remove any adversarial components in the training data. However, applying DA techniques after a poisoning attack launched at stage $t_2$ has two limitations. First, data collected from outside the system $D_t$ cannot be trusted entirely and thus applying augmentation techniques based on this data may not be very effective in defense. Second, as the proportion of contaminated data in the training data $D_t$ increases, the defense effectiveness decreases because the risk of the trained model increases, and details are described in subsection 3.2 [12], [14].

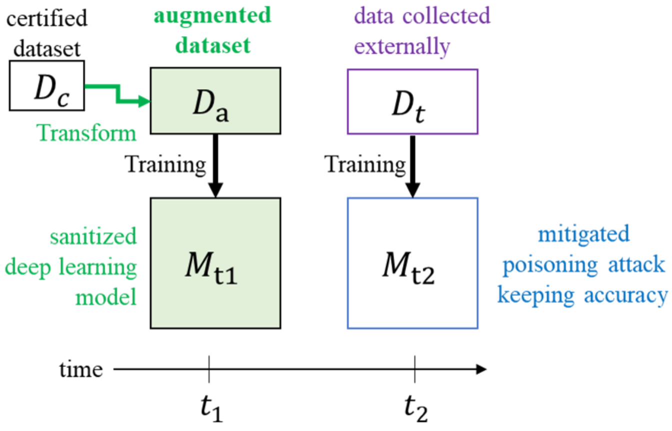


Figure 2. Data augmentation (DA)-based defense method

## 3. PROPOSED METHOD

### 3.1. Our approach: dilution-based defense method

We propose a dilution-based defense mechanism against poisoning attacks on the training data collected from outside the system, specifically at stage $t_2$, even if the data contains poisoned samples. We consider the following attack scenario. As shown in Figure 3, we assume that a DL model is trained with training data $D_t$ collected from outside the system after poisoning attack launched at stage $t_2$ and the ratio of contaminated data is unknown. Since there is no classifier that can perfectly classify the contaminated data according to the previous assumption, a detection technique alone cannot defend a DL model. Therefore, to reduce impact of contaminated data, we generate additional clean data D$_{\text{clean}}$, and then add it to the collected data $D_t$. By this dilution approach, we expect the success rate of poisoning attacks to decrease.

The design of our dilution-based defense mechanism is illustrated in Figure 3. To reduce the proportion of poisoned data in newly collected data $D_t$, we generate a clean dataset $D_{clean}$ based on various augmentation techniques such as by simply copying $D_c$ or by using deep convolutional generative adversarial network (DCGAN). After that, we add $D_{clean}$ to $D_t$ and then train a DL model based on $D_t \cup D_{clean}$.

The expected benefits of the dilution-based defense mechanism are as follows. First, if the dilution defense mechanism is applied at stage $t_2$ after a poisoning attack launched, it is expected that the effectiveness of the poisoning attack will decrease, resulting in an increase in the model's classification accuracy. In addition, since conventional data augmentation techniques do not focus on the amount of data, it is expected that the dilution-based defense technique will show better performance as the amount of contaminated data increases.

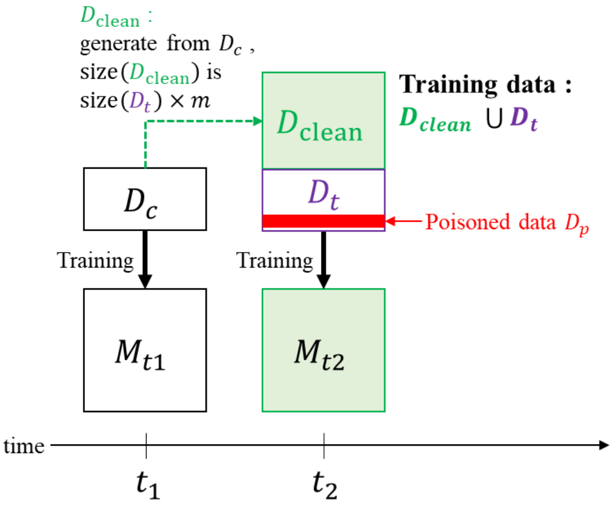


Figure 3. Dilution-based defense method

### 3.2. Simple analysis of decreasing attack risk by proposed method

The reason why the accuracy increases by the dilution-based defense method can be predicted by examining the change in risk $R$. In poisoning attacks, the overall risk $R$ can be expressed as in (1):

$$R = R_c(\,D_t - \,D_p) + \lambda R_p(D_p) \tag{1}$$

where $R_c$ is the risk for normal data, $R_p$ is the risk for poisoning attacks, λ is a non-negative hyperparameter, and we let $D$ is a dataset, $D_t$ is a total dataset, and $D_p$ is a poisoned dataset, subset of $D_t$, and ($|D_p|$ / $|D_t|$) is the poisoning rate [19]. Since the contaminated data cannot be detected, according to the our assumption, the detection risk $R_d$ is eliminated [19].

When we apply the dilution-based defense method proposed in this paper, $R_p$ is maintained but $R_c$ and $\lambda$ decreases. This is because as normal data increases, $R_c$ converges to a lower minimum through the optimization process during training. Additionally, since the overall amount of data $D_t$ increases, $\lambda$ decreases proportionally to the ratio of $D_p$. However, since the number of contaminated data remains unchanged, $R_p$ is maintained. Therefore, by implementing our dilution-based defense method, the overall risk decreases in the DL training stage while the classification accuracy improves.

## 4. RESULTS AND DISCUSSION

### 4.1. Experimental purpose and setup

The main experimental purpose is to verify the effectiveness of our proposed dilution-based defense method against poisoning attacks and compare its performance with an existing method. For performance comparison, we use classification accuracy and the attack success rate (ASR) as performance evaluation metrics. Classification accuracy is a metric that represents the ratio of correctly distinguishing normal and abnormal data while ASR represents the ratio of attacks succeeded out of the total number of poisoning attack trials against the target model. Thus, a higher classification accuracy indicates a better defense performance while a lower number of successful attacks indicates a better defense effect. For clean-label poisoning attacks, classification accuracy was measured. For clean-label backdoor attacks, the classification accuracy remains almost constant, so ASR was used.

The experiment was designed based on the experimental objectives as follows. First, to verify the effectiveness of the dilution defense method, we measured the changes in accuracy according to the poisoning rate and dilution rate. Next, to compare with the existing DA-based defense method using Cutmix (DA-Cutmix) method, we applied the DA-Cutmix method in step $t_1$ and compared its performance with our dilution-based defense method. For experiment implementation, we used the Anaconda software's virtual environment based on Python 3.9 and Tensorflow 2.10 framework with Adversarial Robustness Toolbox, and for running the experiment program, we used an Intel Core i9-12900k CPU and a GeForce RTX 3090 24 GB random-access memory (RAM) graphics processing unit (GPU) [22]–[24]. The specific experimental setup is described as follows.

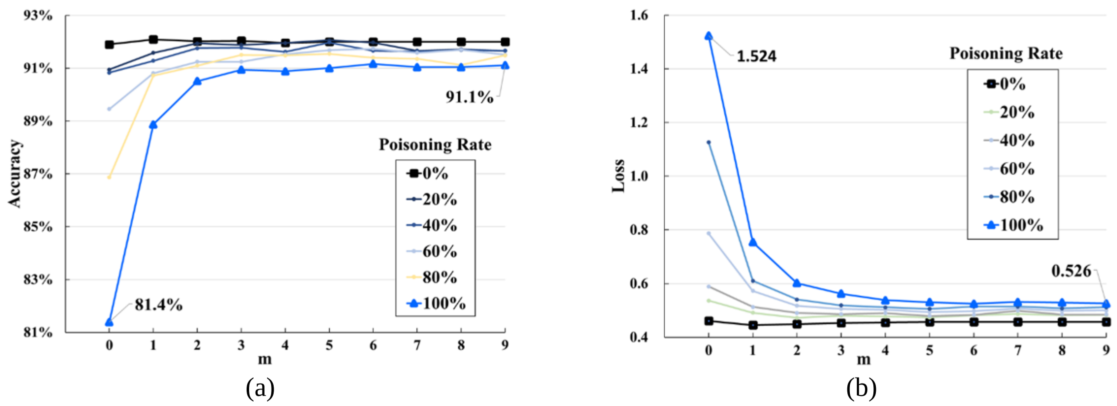
- Target DL model and dataset: To construct the attack target DL model, we used a ResNet model trained on the CIFAR-10 dataset which is commonly used in poisoning attack and defense research [25], [26]. The CIFAR-10 dataset consists of 32×32-pixel color images that can be classified into 10 classes such as airplanes, birds, and horses. It consists of 50,000 training images and 10,000 test images. The ResNet model using residual blocks to reduce information loss during training is a convolutional neural network (CNN) that has shown a high performance in image recognition [26]. To align with an existing method and experimental setup, we use ResNet-50 with 0.47 million parameters [12], [26].
- Poisoning attack methods: To taint the target model, we used clean-label poisoning and clean-label backdoor attacks, as shown in Figure 1 [20], [21]. We created $D_t$ that includes contaminated data $D_p$ generated using the two poisoning attack techniques. We trained the contaminated model $f_p$ with $D_t$ generated at various ratios. For clean-label poisoning attacks, we used various poisoning rates (0%, 20%, 40%, 60%, 80%, and 100%), and the experiment with clean-label poisoning attack is evaluated based on the average accuracy for 10 classes. In addition, the experiment with clean-label backdoor attack (a targeting attack) is evaluated based on ASR for one class of the target.
- Constructing training dataset for evaluation: To measure the performance of the dilution-based defense technique for each additional data ratio, we constructed the training data as follows. The number of duplicating $D_c$ in our dilution method is denoted by $m$, and the proportion of contaminated data $D_p$ from the newly collected data $D_t$ is indicated by the poisoning rate; $D_t$ = 1,000. For clean-label poisoning attack, we duplicated $D_{clean}$ up to 9,000 by varying $m$ from 1 to 9. For clean-label backdoor attack, we duplicated $D_{clean}$ up to 20,000 by varying $m$ from 1 to 20.
- Comparison of performance with DA-Cutmix: To compare the performance of our proposed dilution defense technique, we measured the performance of an existing DA technique. As shown in Figure 2, we applied the Cutmix DA technique at $t_1$ to generate model $M_{t1}$, and measured the changes in accuracy when $M_{t1}$ is subjected to a poisoning attack [12]. We then compared the performance of this data augmentation technique with the performance of our dilution defense technique [12].

### 4.2. Experimental result and analysis

We now explain three experimental results and analyze them as follows: First, as $m$ increases, our proposed dilution-based defense method can better defend the target DL model against clean-label poisoning attacks with various poisoning rates, as shown in Figure 4. Specifically, Figure 4(a) shows the changes in classification accuracy of the target DL model according to $m$ and Figure 4(b) shows the changes in loss according to $m$. Especially, when $m = 0$ (i.e., no defense option), the classification accuracy dropped to 81.4% in the presence of clean-label poisoning attack with poisoning $rate = 100\%$. However, with our dilution defense technique, the classification accuracy increased by 9.7%p and thus reached 91.1%. In addition, as shown in Figure 4(a), we could not observe the clear increment in classification accuracy after $m = 3$, which means there can be an optimal $m$ given a poisoning rate.

Second, our dilution-based defense technique outperformed an existing defense method (Cutmix data augmentation; DA-Cumix) in terms of classification accuracy. Before presenting the results, we explain two attack cases (Attack case 1 and Attack case 2) and experimental setups for this experiment as shown in Table 1. To compare three target DL models in Table 1(a) to (c) in various ways, we considered two attack cases as the following. In Attack case 1, the attacker uses the same poisoned dataset to contaminate the three target models and in Attack case 2 the advanced white-box attacker uses different poisoned datasets. Next, we explain the results as follows. For Attack case 1, as shown in Table 2, while as the contamination rate increases, the classification accuracy decreases. Specifically, when poisoning rate = 100%, no defense method (a) and an existing method (b) showed a significant reduction in accuracy around 20%p compared to when poisoning rate = 0%. Meanwhile, our proposed method demonstrates a small decrease by less than 1%p in classification accuracy. For Attack case 2, as shown in Table 3, when poisoning rate = 20%, there is no significant difference. However, as poisoning rate grows to 100%, the performance of DA-Cutmix decreases significantly while our dilution defense method (c) maintains the similar classification accuracy of the case when poisoning rate = 0%.

Third, our proposed dilution-based defense methods better prevented clean-label backdoor attacks as $m$ grows, as shown in Figure 5. Specifically, as shown in Figure 5, 148 attacks were successful when $m = 0$ (no dilution defense). However, as the clean data $D_c$ was added to the training data (as $m$ grows), ASR clearly lowered to 40.5% (when $m = 12$; the number of added data = 12,000); thus, thanks to our defense method, around 33.5%p of attacks were prevented. Meanwhile, while our dilution-based method is very effective against clean-label backdoor attacks due to the specific targeting characteristic, it requires more additional clean data compared to clean-label poisoning attacks. This is because the dilution defense method reduces both normal and attack components in the data.



(a)

(b)

Figure 4. Depending on $m$ (the number of duplications $D_c$), (a) classification accuracy and (b) loss graph in experiments on clean-label poisoning attacks with various poisoning rates

Table 1. Experimental dataset setting and two attack cases (attack case1 and attack case 2)

| Methods | $t_1$ | | $t_2$ | | | |
|---|---|---|---|---|---|---|
| | Training data | Generated model | Training data | Generated model | Attack case 1 (Poisoned data) | Attack case 2 (Poisoned data) |
| (a) No defense method | $D_c$ | $M_{(t1,Dc)}$ | $D_t$ | $M_{(t1,Dc)\to(t2,Dt)}$ | $D_p$ from $M_{(t1,Dc)}$ | $D_p$ from $M_{(t1,Dc)}$ |
| (b) An existing method (DA-Cutmix) [12] | $D_a$ | $M_{(t1,Da)}$ | $D_t$ | $M_{(t1,Da)\to(t2,Dt)}$ | | $D_p$ from $M_{(t1,Da)}$ |
| (c) Our proposed method | $D_c$ | $M_{(t1,Dc)}$ | $D_t \cup D_{clean}$ | $M_{(t1,Dc)\to(t2,Dt\cup D_{clean})}$ | | $D_p$ from $M_{(t1,Dc)}$ |

Table 2. Attack case 1: comparison of classification accuracy in clean-label poisoning attack using the same attack dataset $D_p$

| Methods | Poisoning rate ($\lvert D_p\rvert/\lvert D_t\rvert$; average of 3 times) | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| (a) No defense method | 91.9% | 90.9% | 89.7% | 87.5% | 83.6% | 74.3% |
| (b) An existing method (DA-Cutmix) [12] | 92.3% | 91.1% | 89.7% | 87.8% | 84.3% | 74.4% |
| (c) Our proposed method | 91.6% | 91.7% | 91.7% | 91.5% | 91.5% | 91.1% |
| Improvements: (c) – (b) | -0.7%p | +0.6%p | +1.9%p | +3.9%p | + 7.4%p | **+17.8%p** |

Table 3. Attack case 2: comparison of classification accuracy in clean-label poisoning attack using white-box attacks

| Methods | Poisoning rate ($\|D_p\|/\|D_t\|$, average of 3 times) | | | | | |
|---|---|---|---|---|---|---|
| | 0% | 20% | 40% | 60% | 80% | 100% |
| (a) No defense method | 91.9% | 90.9% | 89.7% | 87.5% | 83.6% | 74.3% |
| (b) An existing method (DA-Cutmix) [12] | 92.3% | 90.8% | 89.2% | 86.9% | 82.9% | 70.2% |
| (c) Our proposed method | 91.6% | 91.7% | 91.7% | 91.5% | 91.5% | 91.1% |
| Improvements: (c) – (b) | -0.7%p | +0.9%p | +2.5%p | +4.6%p | + 8.6%p | **+20.9%p** |

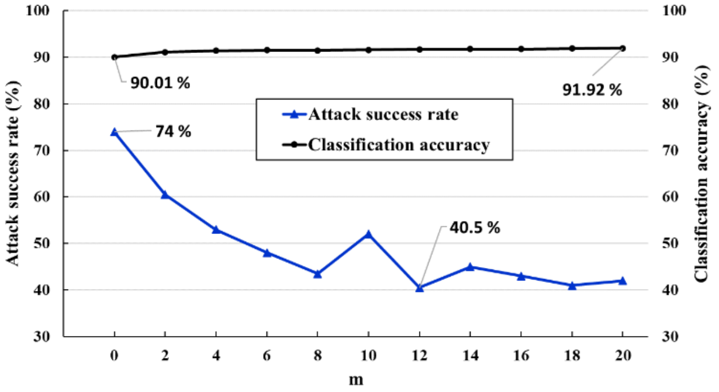


Figure 5. Attack success rate (ASR, left y-axis) and classification accuracy (right y-axis) depending on $m$ (the number of added data; x-axis) in clean-label backdoor attacks

## 5. CONCLUSION

In this paper, by assuming that there are no techniques that perfectly detect poisoning attacks, we proposed a dilution-based defense method against poisoning attacks which is a novel defense mechanism to complement existing detection methods. Our dilution-based defense method adds clean data to training data in order to reduce the impact of poisoned data in the post-poisoning phase. Our experimental results demonstrate its validity and effectiveness in defending against poisoning attacks. Specifically, applying dilution defense increased the classification accuracy performance of a DL model by 9.7%p for poisoning attack and decreased 33.5%p of ASR for backdoor attack. In addition, the defense performance of our proposed method is up to 20.9%p better than that of an existing data augmentation method. Consequently, the results show that our dilution-based defense method is very effective against both poisoning attacks and backdoor attacks.

Our future research directions are as follows. First, a mere increase in the amount of training data leads to higher computing costs. Therefore, it is essential to study methods for minimizing the additional data required for training. Second, to further improve the classification performance of DL models, we will study weakening the attack components of transferred data and maintaining the benign feature of them during the dilution process.

## REFERENCES

[1] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “A survey on adversarial attacks and defences,” *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25–45, Mar. 2021, doi: 10.1049/cit2.12028.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Dec. 2015.

[3] J. Wang, “Adversarial examples in physical world,” in *IJCAI International Joint Conference on Artificial Intelligence*, Chapman and Hall/{CRC}, 2021, pp. 4925–4926, doi: 10.24963/ijcai.2021/694.

[4] M. E. Merza, S. H. Hussein, and Q. I. Ali, “Identification scheme of false data injection attack based on deep learning algorithms for smart grids,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 1, pp. 219–228, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp219-228.

[5] S. Aneja, N. Aneja, P. E. Abas, and A. G. Naim, “Defense against adversarial attacks on deep convolutional neural networks through nonlocal denoising,” *IAES International Journal of Artificial Intelligence*, vol. 11, no. 3, pp. 961–968, Sep. 2022, doi: 10.11591/ijai.v11.i3.pp961-968.

[6] G. Goswami, N. Ratha, A. Agarwal, R. Singh, and M. Vatsa, “Unravelling robustness of deep learning based face recognition against adversarial attacks,” *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, vol. 32, no. 1, pp. 6829–6836, Apr. 2018, doi: 10.1609/aaai.v32i1.12341.

[7] K. Ren, T. Zheng, Z. Qin, and X. Liu, “Adversarial attacks and defenses in deep learning,” *Engineering*, vol. 6, no. 3, pp. 346–360, Mar. 2020, doi: 10.1016/j.eng.2019.12.012.

[8] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, “Adversarial examples are not bugs, they are features,” in *Advances in Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates Inc., 2019.

[9] D. J. Miller, Z. Xiang, and G. Kesidis, “Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks,” *Proceedings of the IEEE*, vol. 108, no. 3, pp. 402–433, Mar. 2020, doi: 10.1109/JPROC.2020.2970615.

[10] T. B. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, in NIPS’20, vol. 2020- December. Red Hook, NY, USA: Curran Associates Inc., 2020.

[11] Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, and K. Ren, “Threats to training: A survey of poisoning attacks and defenses on machine learning systems,” *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–36, Dec. 2022, doi: 10.1145/3538707.

[12] E. Borgnia *et al.*, “Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE, Jun. 2021, pp. 3855–3859, doi: 10.1109/ICASSP39728.2021.9414862.

[13] S. Yun, D. Han, S. Chun, S. J. Oh, J. Choe, and Y. Yoo, “CutMix: Regularization strategy to train strong classifiers with localizable features,” in *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, Oct. 2019, pp. 6022–6031, doi: 10.1109/ICCV.2019.00612.

[14] A. K. Veldanda *et al.*, “NNoculation: Catching badNets in the wild,” *AISec 2021 - Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, co-located with CCS 2021*, pp. 49–60, Feb. 2021, doi: 10.1145/3474369.3486874.

[15] H. Qiu, Y. Zeng, S. Guo, T. Zhang, M. Qiu, and B. Thuraisingham, “DeepSweep: An evaluation framework for mitigating DNN backdoor attacks using data augmentation,” in *ASIA CCS 2021 - Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, ACM, May 2021, pp. 363–377, doi: 10.1145/3433210.3453108.

[16] M. Kravchik, B. Biggio, and A. Shabtai, “Poisoning attacks on cyber attack detectors for industrial control systems,” in *Proceedings of the ACM Symposium on Applied Computing*, ACM, Mar. 2021, pp. 116–125, doi: 10.1145/3412841.3441892.

[17] H. Chacon, S. Silva, and P. Rad, “Deep learning poison data attack detection,” in *Proceedings - International Conference on Tools with Artificial Intelligence, ICTAI*, IEEE, Nov. 2019, pp. 971–978, doi: 10.1109/ICTAI.2019.00137.

[18] F. Zhuang *et al.*, “A Comprehensive survey on transfer learning,” *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021, doi: 10.1109/JPROC.2020.3004555.

[19] Y. Li, Y. Jiang, Z. Li, and S. T. Xia, “Backdoor learning: A survey,” *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–18, 2022, doi: 10.1109/TNNLS.2022.3182979.

[20] A. Shafahi *et al.*, “Poison frogs! Targeted clean-label poisoning attacks on neural networks,” in *Advances in Neural Information Processing Systems*, in NIPS’18, vol. 2018- December. Red Hook, NY, USA: Curran Associates Inc., 2018, pp. 6103–6113.

[21] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y. G. Jiang, “Clean-label backdoor attacks on video recognition models,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2020, pp. 14431–14440, doi: 10.1109/CVPR42600.2020.01445.

[22] Anaconda, “Anaconda software distribution.,” *Computer software*, 2016. https://continuum.io/ (accessed May 15, 2023).

[23] M.-I. Nicolae *et al.*, “Adversarial robustness toolbox v1.0.0,” *arxiv.org/abs/1807.01069*, Jul. 2018.

[24] M. Abadi, “TensorFlow: learning functions at scale,” in *Proceedings of the 21st ACM SIGPLAN International Conference on Functional Programming*, ACM, Sep. 2016, pp. 1–1, doi: 10.1145/2951913.2976746.

[25] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Technical report, University of Toronto, Toronto, Ontario, 2009.

[26] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

## BIOGRAPHIES OF AUTHORS

**Hweerang Park** received B.S. degree in physics from Chonnam National University, Gwangju, Republic of Korea. He is currently a Captain in the Republic of Korea Air Force and pursuing the M.S. degree in Department of Defense Science (Cyberwarfare Major) with Korea National Defense University, Nonsan, Republic of Korea. His research interests include deep learning, adversarial machine learning, and cyberwarfare. He can be contacted at the email: sharku7@gmail.com.

**Youngho Cho** received the B.S. degree in industrial engineering from Korea Air Force Academy, Republic of Korea, in 1998 and the M.S. degree in computer science and industrial systems engineering from Yonsei University, Republic of Korea, in 2006 and the Ph.D. degree in electrical and computer engineering from University of Maryland, College Park, MD, USA, in 2013. He is an associate professor with the Department of Defense Science (Computer Engineering and Cyberwarfare Major), Graduate School of Defense Management, Korea National Defense University, Nonsan, Republic of Korea. His research interests include wireless network security, trust mechanism, botnet detection, steganography-based covert communication, adversarial machine learning, IoT security, and game theory in network security. He has authored or coauthored more than 50 peer-reviewed papers published in journals and in the proceedings of conferences. He can be contacted at the email: youngho@kndu.ac.kr.