Defense Method against Adversarial Attacks Using JPEG Compression and One-Pixel Attack for Improved Dataset Security

Adelina-Valentina Cucu, Giuseppe Valenzise, Daniela Stănescu, Ioana Ghergulescu, Lucian Ionel Găină, Bianca Guşiţă

Defense Method against Adversarial Attacks Using JPEG Compression and One-Pixel Attack for Improved Dataset Security

Adelina-Valentina Cucu
CentraleSupélec
Laboratoire des Signaux et systèmes,
Université Paris-Saclay
Gif-sur-Yvette, France
Computer Science Department
Politehnica University of Timisoara
Timisoara, Romania
adelina.cucu@student.upt.ro

Giuseppe Valenzise
CNRS, CentraleSupélec
Laboratoire des Signaux et systèmes,
Université Paris-Saclay
Gif-sur-Yvette, France
giuseppe.valenzise@l2s.centralesupelec.fr

Daniela Stănescu
Computer Science Department
Politehnica University of Timisoara
Timisoara, Romania
daniela.stanescu@cs.upt.ro

Ioana Ghergulescu
Adaptemy
Dublin, Ireland
ioana.ghergulescu@adaptemy.com

Lucian Ionel Găină
Computer Science Department
Politehnica University of Timisoara
Timisoara, Romania
lucian.gaina@student.upt.ro

Bianca Gușiță
Computer Science Department
Politehnica University of Timisoara
Timisoara, Romania
bianca.gusita@student.upt.ro

***Abstract*—Machine Learning has gained widespread applications, especially in the domain of image recognition due to the high performance of algorithms and models. Nonetheless, the potential threat posed by adversarial samples has impeded their widespread adoption, especially in safety-critical applications. In such cases, the model's final performance is significantly compromised due to the presence of adversarial perturbations that are invisible to human perception. This paper introduces a novel pre-processing solution for protecting image datasets against adversarial attacks such as the Fast Sign Gradient Method or other similar attacks. The proposed method involves utilizing high JPEG compression for removing the adversarial perturbations of the dataset and applying a Targeted One Pixel Attack with the aim of recovering the true class of the images after the compression process. The proposed approach results in a highly secured and clean dataset maintaining a high classification accuracy. The approach was tested on CIFAR10 dataset. The results have shown that the misclassification rate after the FGSM attack was significantly reduced from 89.3% to 12.5% using this method for the compression quality of 20 on a subset of 1000 images.**

Keywords—JPEG, compression, FSGM, One Pixel Attack, defense, adversarial attacks

I. Introduction

In numerous applications, machine learning models have become indispensable, including image classification tasks like those carried out on the well-known CIFAR10 dataset [1]. These models are susceptible to adversarial attacks, in which carefully designed perturbations that look insignificant can cause misclassifications [2]. Furthermore, machine learning systems' security can be seriously threatened by adversarial attacks in safety-critical applications like autonomous driving [3], medical diagnostics [4] and biometric authentication.

In response to the growing threat of adversarial attacks, researchers have proposed safety measures to enhance the resilience of machine learning models and defend against these attacks [5]. This paper introduces an innovative pre-processing method for the CIFAR10 dataset that utilizes targeted One-Pixel attacks [6] as a reinforcement method to recover the original class labels of the images subjected to significant JPEG compression.

II. Related Work

Adversarial attacks are a growing concern in deep learning since their first introduction in 2014 by Szegedy et al., as they can easily deceive even the most advanced neural networks [2]. These attacks are crafted to add subtle perturbations to the input images, which can result in misclassification by the model. Several attacks have been introduced by previous research works, such as the Fast Sign Gradient Method [7], DeepFool [8], Carlini and Wagner attack [9], JSMA [10], One Pixel Attack [6]. Several surveys related to adversarial machine learning attacks including [11]–[13] have been published. The surveys focused on various aspects such as attack type [11], adversarial attacks in real-world scenarios [11], adversarial examples [12] and adversarial robustness from the interpretability perspective and attacks in specific domains (i.e., medical domain) [13].

To counter these attacks, various defense methods have also been introduced. Several surveys related to defenses against adversarial attacks including [14]–[16] have been published in recent years. Mechanism examples include adversarial training [14], input preprocessing [15], and gradient masking [16]. Input pre-processing, in particular, has shown to be highly effective . In this method, the input image is pre-processed before being fed into the model to remove any adversarial perturbations. This could be achieved through various techniques, such as image blurring [17], and JPEG compression [18]. Previous studies have examined the impact of JPEG compression on removing adversarial perturbations such as the impact of feature distillation [16] which works by preserving essential image features while suppressing the perturbations, or using JPEG compression and increasing the accuracy of classification by including a high number of training images subjected to different levels of JPEG Compression [19]. Another study [18] proposes the use of JPEG2000 compression as an alternative to JPEG for reducing adversarial noise in images and states that JPEG2000 achieves higher compression rates with less distortion and avoids introducing blocking artifacts. However, it has been observed that stronger compression is required to achieve higher levels of perturbation removal. Nevertheless, excessively high compression can introduce noise and result in the loss of class

information for some images [20]. On the other hand, insufficient compression may not guarantee successful perturbation removal. This paper extends existing work by proposing a novel pre-processing solution for protecting image datasets against adversarial attacks by using high JPEG compression to remove the dataset's adversarial perturbations and applying a Targeted One Pixel Attack with the aim to recover the true class of the images after compression.

III. Background

A. JPEG Compression

JPEG compression is a popular method used for defending against adversarial attacks through input preprocessing. It is a lossy compression technique that removes high-frequency components from an image, effectively eliminating adversarial perturbations while maintaining accurate classification [19], [20]. The algorithm utilizes the Discrete Cosine Transform (DCT) to convert image blocks into frequency components. Each 8x8 block undergoes DCT transformation, and the resulting coefficients are quantized using 8x8 quantization tables. Psycho-visual redundancy is leveraged by assigning higher penalties to higher-frequency coefficients, considering the human visual system's reduced sensitivity to these components. However, JPEG has limitations as a defense mechanism. It can introduce blocking artifacts caused by the individual treatment of 8x8 blocks, which can harm classification performance. Additionally, JPEG is not optimized for high compression rates, leading to significant image distortion and further degradation of classification accuracy [18].

B. One Pixel Attack

One Pixel Attack (OPA) [6] is an adversarial attack that does not require a lot of information, except for the probability labels (semi-black-box attack) and can mislead the classifiers by changing a single pixel. The pixel is found using DE by selecting the best individuals and using them to create new ones through reproduction and mutation, evolving the population towards a better solution [6]. With the help of propagation maps, it was demonstrated how a single pixel perturbation could rise in influence across the layers and extend across several pixels leading to a change in the final predicted class [21].

The attack can be either targeted or untargeted. In the case of the targeted attack, the label of the target class needs to be higher than 90%, while for the untargeted attack the probability needs to be minimized (<5%), therefore redirecting the classifier to any other wrong class [6].

C. Fast Sign Gradient Method

The Fast Gradient Sign Method (FGSM) [7] is a popular technique for creating adversarial perturbations in image classification models. It leverages linear behavior in high-dimensional spaces. This means that a simple way of generating adversarial examples can be done by computing a perturbation with a single step. By linearizing the neural network's cost function around its current weights and computing the gradient with respect to the input image through backpropagation, the method applies a small constant value (epsilon) to the gradient and takes the sign to create a perturbation. By adding this perturbation to the original image, the resulting image is misclassified. The strength of the perturbation can easily be controlled by changing the value of the epsilon. A larger epsilon value will create a more effective adversarial example but will also make the perturbation more noticeable to humans [7]. Fig. 1 illustrates an example of FGSM attack.

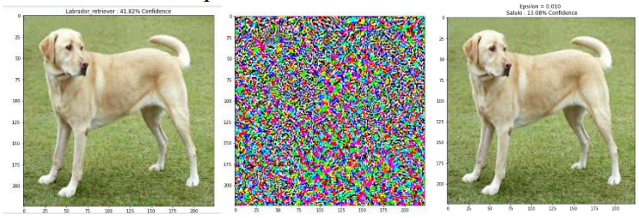


Fig. 1. Image of the Labrador before the attack; Resulting perturbations; Image after attack for epsilon=0.01.

IV. Proposed Solution

The proposed method offers a novel approach by leveraging the combination of both JPEG compression and One Pixel Attack to effectively protect image datasets against adversarial attacks by eliminating the adversarial perturbations that could have been previously introduced. Specifically, the method achieves a high classification accuracy while ensuring robustness against popular attack methods such as Fast Sign Gradient Method (FSGM). This is achieved through a pre-processing stage where the dataset is subjected to strong JPEG compression, followed by a targeted One-Pixel Attack for class recovery.

Fig. 2 presents the block level diagram of the proposed method. The initial dataset I consists of images with correct class labels, (C*) vulnerable to adversarial examples (Attack 1) that would lead to a high misclassification rate compromising the dataset (I'). By applying JPEG compression for eliminating the adversarial perturbations, there could be observed three possible outcomes:

1. Some of the images could successfully recover their correct class (C*).
2. For other images the perturbation couldn't be removed, therefore they would maintain their wrong class (C'); this is likely to happen when the compression is not strong enough.
3. Images could be misclassified because of excessive compression due to significant distortion (C').

To reinforce the compression results and enable higher compression rates, misclassified images (with class C') are identified. A targeted one-pixel attack is then applied, setting the target as the correct class, to correct the class labels.

The resulting image dataset (I*) is protected from the initial adversarial attack (Attack 1) and is recovering its high classification accuracy due to the One-Pixel attack reinforcement.

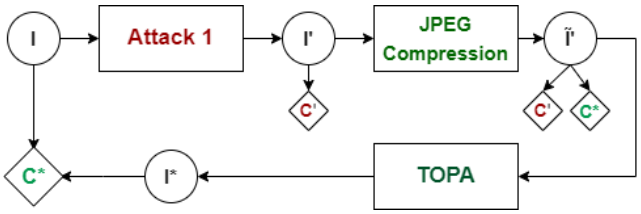


Fig. 2. Block-Level Diagram of the proposed method. I-initial dataset, I'-corrupted dataset, Ĩ'-corrupted dataset after compression, I*-final dataset after TOPA, C*-correct class, C'- wrong class.

V. RESULTS

A. Methodology

This section outlines the approach for evaluating the proposed solution. The CIFAR10 [1] dataset was used in the evaluation, and it was trained using VGG16 architecture for 10 epochs with Adam optimizer. The trained model achieved a testing accuracy of 86.58% on CIFAR10. To assess the efficacy of the method, , experiments using a common adversarial attack algorithm, namely, Fast Sign Gradient Method (FSGM) [7] were conducted choosing an epsilon of 0.02 (as in [19]) to define the strength of the attack. The misclassification rate was one of the main metrics used in the experiments. Furthermore, as an additional contribution of the paper, several experiments were made to investigate the impact of JPEG compression on CIFAR10, as well as the impact of compression on theCIFAR10 dataset attacked by FGSM, and the effect of Targeted One Pixel Attack to reinforce classification accuracy. The misclassification rate metric was used in all experiments including the FGSM attack over the subset, the JPEG compression, after applying Targeted One-Pixel Attack to the misclassified compressed images. The metric provides insights into the progressive decrease in misclassification throughout the different steps in the proposed solution.

The following subsections outline the experimental findings that validate the efficiency of the proposed method in removing adversarial perturbations of FSGM. The findings of this study contribute to enhancing the security and reliability of machine learning systems in the face of adversarial threats.

B. Effect of JPEG Compression on CIFAR10

An analysis of the variance of misclassification rate was conducted to further investigate the impact of JPEG compression on CIFAR10. Fig. 3 illustrates the misclassification rate on the entire CIFAR10 dataset for different levels of compression from q=0 to q=100. As can be seen from the figure, there is a gradual increase of the misclassification rate with the decrease of the compression quality (q), which has the tendency to slowly increase starting with q=90 and it quickly rises as the compression quality approaches q=0.

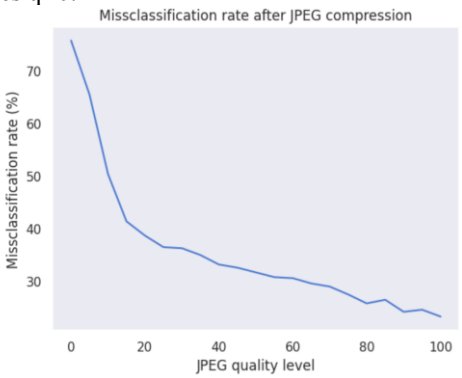


Fig. 3. Misclassification rate of CIFAR10 after JPEG compression.

The findings highlight the challenge of maintaining a high accuracy while applying JPEG compression to defend against adversarial attacks. To address these challenges, the proposed method reinforces the power of JPEG compression and maximizes its effectiveness by recovering the class of misclassified images post-compression. This is accomplished by using a Targeted One Pixel Attack, which is modifying a single pixel to target the true class of the attacked image that is still misclassified after compression.

C. Effect of JPEG Compression on CIFAR10 Dataset Attacked by FGSM

A second experiment was conducted to analyse the effectiveness of JPEG compression in removing the adversarial perturbations introduced by FGSM from the CIFAR10 dataset. In Fig. 4 the quality of the compression (q) was changed from 10 to 100 with a step of 10 to observe how different levels of compression are reducing the adversarial perturbations. As can be observed, a strong compression (e.g., q=20) has higher efficiency in eliminating the adversarial perturbations, as the misclassification rate is progressively growing with the increase of the q factor for compression. At the same time, it is important to observe that for a really strong compression, smaller than q=20, the misclassification rate is starting to rise again due to the artifacts introduced by compression. These findings highlight the trade-off between compression strength and maintaining a high classification accuracy in the presence of an adversarial attack.

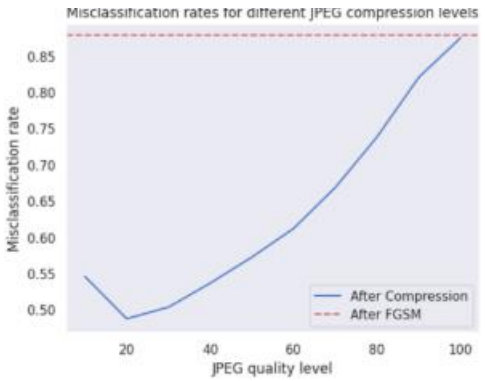


Fig. 4. Misclassification rate of CIFAR10 attacked by FGSM (ϵ=0.02) after JPEG compression.

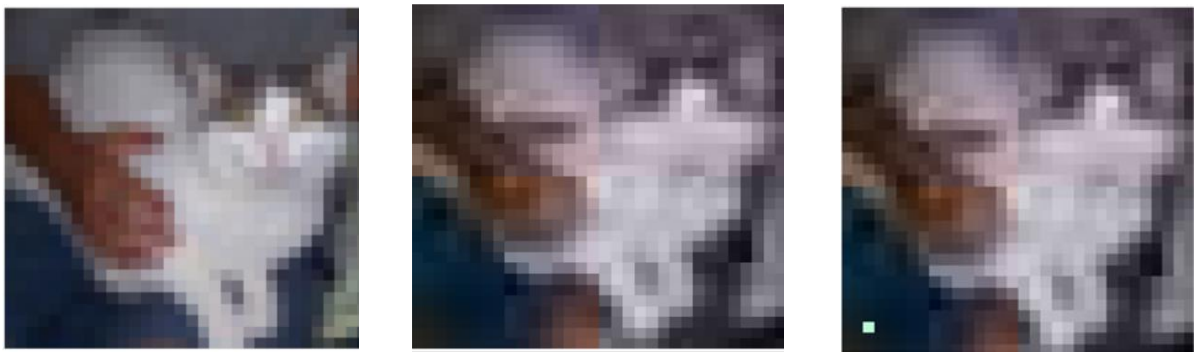
D. Effect of Targeted One Pixel Attack

The Targeted One Pixel Attack (TOPA) algorithm was applied to the misclassified images after compression to reinforce the classification accuracy.

Our method is using an original approach of this attack, having as objective the recovery of the class instead of misleading the classifier. For this, we are setting as target the original true label of the misclassified images and finding the one pixel that would recover the classification of the image, thereby getting a significantly lower misclassification rate for the dataset that was initially under attack.

Fig. 5 presents an exemplification of the effect of Targeted One Pixel Attack in recovering the original correct class of a use case image of a cat that went through high compression that resulted in the image being misclassified as a dog. After applying TOPA by setting as target the initial correct class, the compressed image was reclassified as a cat, by changing just one pixel.

Fig. 6 presents the misclassification rate when JPEG compression with a q in the rage of 10 to 100 with a step of 10 was applied and after TOPA for 200 images from the CIFAR10 dataset. The results illustrate the effectiveness of TOPA in reducing the misclassification rate for every level of compression.



Benign Image Classified as Cat | Compressed image misclassified as dog | Image compressed reclassified as cat after TOPA

Fig. 5. Exemplification images.

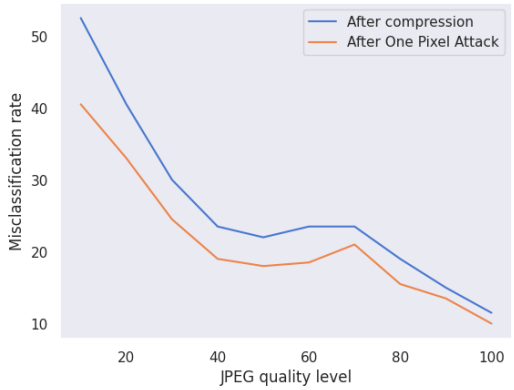


Fig. 6 Missclassification rate for JPEG Compression and after One Pixel Attack for 200 images.

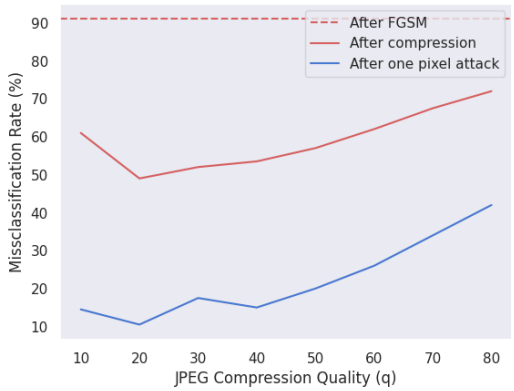


Fig. 7 Misclassification rate for FGSM, and JPEG Compression and TOPA.

E. Effectiveness of Proposed Method against FSGM Attack

The proposed method was evaluated on CIFAR10 dataset for testing its effectiveness against the FSGM attack. The results of the experiments have shown that JPEG compression followed by Targeted One Pixel Attack as class recovery provides a significant defense against adversarial attacks, resulting in a clean dataset with high classification accuracy. Table 1 shows that the misclassification rate of a subset of 1000 images having FGSM perturbations, has decreased after compression with q=20 and q=40 from 89.3 % to 12.5% and 16.8 %, respectively, after applying Targeted One Pixel Attack for further recovery of the class.

TABLE I. MISCLASSIFICATION RATE AT DIFFERENT STEPS OF THE PROPOSED METHOD.

Misclassification rate for 1000 images	**Q=20**	**Q=40**
After FGSM	89.3%	89.3%
After JPEG Compression	48.5%	52.0%
After Targeted One Pixel Attack	12.5%	16.8%

When comparing to the proposed method from [19] it can be observed that for the same strength of the FGSM attack (epsilon=0.02), the results obtained in 'Table 1' are significantly better for Q=20 or Q=40, the misclassification rate of the method is 12.5%, respectively 16.8% compared to 20.43% in [19] for the entire CIFAR10 dataset.

Furthermore, in Fig. 7 is observed the misclassification rate for a subset of the CIFAR10 dataset (200 images) after compression with quality ranging from 10 to 100 with a step of 10. We also observe the misclassification rate after using TOPA, following an FGSM attack with ε=0.02. These results illustrate the effectiveness of our method in reducing adversarial perturbations and improving the model's classification accuracy by combining the two methods for different levels of compression.

VI. CONCLUSION AND FUTURE WORK

This article introduced a pre-processing solution defense against adversarial attacks which is working by eliminating the perturbations introduced by adversarial attacks and correcting the wrong class of the misclassified images. The method was exemplified on the CIFAR10 dataset and is using JPEG compression for eliminating the possible adversarial perturbation and Targeted One Pixel Attack for class recovery after compression. The experimental results showed that the proposed solution is effective against adversarial attacks and results in a clean dataset with high classification accuracy. Future work will extend the experiments to the entire CIFAR10 dataset as well as other datasets (i.e., GTSRB [22] or a dataset in the medical domain). Future work will also evaluate how this method is performing against different types of attacks and will evaluate its long-term effectiveness in real-world scenarios. Another future direction is to evaluate itby introducing an adversarial attack at the end of the preprocessing steps for the robustness of the dataset against possible future attacks.

REFERENCES

[1] A. Krizhevsky, "Learning multiple layers of features from tiny images." University of Toronto, 2009. [Online]. Available: http://www.cs.utoronto.ca/~kriz/learning-features-2009-TR.pdf

[2] C. Szegedy *et al.*, "Intriguing properties of neural networks," presented at the 2nd International Conference on Learning Representations (ICLR 2014), Y. Bengio and Y. LeCun, Eds., Banff, AB, Canada, 2014. Accessed: Jul. 21, 2023. [Online]. Available: http://arxiv.org/abs/1312.6199

[3] Y. Deng, X. Zheng, T. Zhang, C. Chen, G. Lou, and M. Kim, "An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models," presented at the 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom), Mar. 2020, pp. 1–10. doi: 10.1109/PerCom45495.2020.9127389.

[4] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[5] J. Wang, C. Wang, Q. Lin, C. Luo, C. Wu, and J. Li, "Adversarial attacks and defenses in deep learning for image recognition: A survey," *Neurocomputing*, vol. 514, pp. 162–181, Dec. 2022, doi: 10.1016/j.neucom.2022.09.004.

[6] J. Su, D. V. Vargas, and K. Sakurai, "One Pixel Attack for Fooling Deep Neural Networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.

[7] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," presented at the 3rd International Conference on Learning Representations (ICLR 2015), Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015. Accessed: Jul. 21, 2023. [Online]. Available: http://arxiv.org/abs/1412.6572

[8] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks," presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2574–2582. Accessed: Jul. 21, 2023. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html

[9] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," presented at the 2017 IEEE Symposium on Security and Privacy (SP), May 2017, pp. 39–57. doi: 10.1109/SP.2017.49.

[10] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," presented at the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Mar. 2016, pp. 372–387. doi: 10.1109/EuroSP.2016.36.

[11] N. Akhtar and A. Mian, "Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018, doi: 10.1109/ACCESS.2018.2807385.

[12] S. Han, C. Lin, C. Shen, Q. Wang, and X. Guan, "Interpreting Adversarial Examples in Deep Learning: A Review," *ACM Comput. Surv.*, vol. 55, no. 14s, p. 328:1-328:38, Jul. 2023, doi: 10.1145/3594869.

[13] M. K. Puttagunta, S. Ravi, and C. Nelson Kennedy Babu, "Adversarial examples: attacks and defences on medical deep learning systems," *Multimed Tools Appl*, Mar. 2023, doi: 10.1007/s11042-023-14702-9.

[14] W. Zhao, S. Alwidian, and Q. H. Mahmoud, "Adversarial Training Methods for Deep Learning: A Systematic Review," *Algorithms*, vol. 15, no. 8, Art. no. 8, Aug. 2022, doi: 10.3390/a15080283.

[15] H. Wang, J. Wang, and Z. Yin, "An Efficient Pre-processing Method to Eliminate Adversarial Effects." arXiv, Dec. 30, 2019. doi: 10.48550/arXiv.1905.08614.

[16] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," presented at the 2016 IEEE Symposium on Security and Privacy (SP), May 2016, pp. 582–597. doi: 10.1109/SP.2016.41.

[17] W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," presented at the Network and Distributed Systems Security Symposium (NDSS), 2018. doi: 10.14722/ndss.2018.23198.

[18] A. E. Aydemir, A. Temizel, and T. T. Temizel, "The Effects of JPEG and JPEG2000 Compression on Attacks using Adversarial Examples." arXiv, Mar. 31, 2018. doi: 10.48550/arXiv.1803.10418.

[19] N. Das *et al.*, "Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression." arXiv, May 08, 2017. Accessed: Apr. 25, 2023. [Online]. Available: http://arxiv.org/abs/1705.02900

[20] Z. Liu *et al.*, "DeepN-JPEG: a deep neural network favorable JPEG-based image compression framework," in *Proceedings of the 55th Annual Design Automation Conference*, in DAC '18. New York, NY, USA: Association for Computing Machinery, Jun. 2018, pp. 1–6. doi: 10.1145/3195970.3196022.

[21] D. V. Vargas and J. Su, "Understanding the one-pixel attack," presented at the 2020 Workshop on Artificial Intelligence Safety, AISafety 2020, 2020. Accessed: Jul. 21, 2023. [Online]. Available: https://ceur-ws.org/Vol-2640/paper_4.pdf

[22] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel, "The German Traffic Sign Recognition Benchmark: A multi-class classification competition," presented at the The 2011 International Joint Conference on Neural Networks, Jul. 2011, pp. 1453–1460. doi: 10.1109/IJCNN.2011.6033395.