IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 November 2023, accepted 15 February 2024, date of publication 26 February 2024, date of current version 8 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3370849

RESEARCH ARTICLE

Sponge Attack Against Multi-Exit Networks With Data Poisoning

BENXUAN HUANG[1], LIHUI PANG[2], ANMIN FU[1], SAID F. AL-SARAWI[3], (Senior Member, IEEE), DEREK ABBOTT[3], (Fellow, IEEE), AND YANSONG GAO[4], (Senior Member, IEEE)

[1]School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2]Sino-German College of Intelligent Manufacturing, Shenzhen Technology University, Shenzhen 518118, China
[3]School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, SA 5005, Australia
[4]CSIRO's Data61, Sydney, NSW 2113, Australia

Corresponding author: Lihui Pang (sunshine.plh@hotmail.com)

This work was supported in part by the Natural Science Foundation of Hunan Province under Grant 2022JJ40377, in part by the Research and Development in Key Areas of Guangdong Province under Grant 2021ZDZX1018, in part by Guangdong Provincial Engineering Technology, in part by the Research Center for Materials for Advanced Micro-Electro-Mechanical Systems (MEMS) Sensor Chip under Grant 2022GCZX005, and in part by the Natural Science Foundation of Top Talent of Shenzhen Technology University (SZTU) under Grant GDRC 20200204.

ABSTRACT The motivation for the development of multi-exit networks (MENs) lies in the desire to minimize the delay and energy consumption associated with the inference phase. Moreover, MENs are designed to expedite predictions for easily identifiable inputs by allowing them to exit the network prematurely, thereby reducing the computational burden due to challenging inputs. Nevertheless, there is a lack of comprehensive understanding regarding the security vulnerabilities inherent in MENs. In this study, we introduce a novel approach called the *sponge attack*, which aims to compromise the fundamental advantages of MENs that allow easily identifiable images to leave in early exits. By employing data poisoning techniques, we frame the sponge attack as an optimization problem that empowers an attacker to select a specific trigger, such as adverse weather conditions (e.g., raining), to compel inputs to traverse the complete network layers of the MEN (e.g., in the context of traffic sign recognition) instead of early-exits when the trigger condition is met. Remarkably, our attack has the capacity to increase inference latency, while maintaining the classification accuracy even in the presence of a trigger, thus operating discreetly. Extensive experimentation on three diverse natural datasets (CIFAR100, GTSRB, and STL10), each trained with three prominent MEN architectures (VGG16, ResNet56, and MSDNet), validates the efficacy of our attack in terms of latency augmentation and its effectiveness in preserving classification accuracy under trigger conditions.

INDEX TERMS Data poisoning, sponge attack, multi-exit network, machine learning.

I. INTRODUCTION

The advent of Multi-Exit Networks (MENs) [1], [2], featuring multiple exits within its basic model backbone, is motivated by the inherent variability in the difficulty of classifying different input samples. Specifically, certain samples (considered easy) can be accurately classified with a shallow network, allowing for early exits during a MEN's inference phase. Only a small subset of samples, typically those deemed difficult, necessitates traversal through the entire complex network for accurate classification. The primary advantage offered by a MEN lies in a significant reduction in latency and energy consumption during the inference phase in intricate networks. Reduced latency is crucial for real-time applications such as self-driving cars, while lower energy consumption is a critical consideration for devices in the Internet of Things or mobile devices [3], heavily reliant on battery power.

However, the pursuit of these benefits in MEN architectures introduces trade-offs in terms of privacy and security risks. Moreover, MEN architectures have been demonstrated to be susceptible to privacy breaches, leaking sensitive

The associate editor coordinating the review of this manuscript and approving it for publication was Pedro R. M. Inácio.

information such as membership in a training sample. The Membership Inference Attack (MIA) [4] exploits exit information to enhance inference performance and can potentially divulge membership details. Moreover, there is a possibility not only to extract the MEN's function but also its output and exit strategy [5]. Both attacks leverage unique characteristics of MENs, with MIA specifically utilizing exit information [4]. It is noteworthy that other security threats, including evasion attacks (e.g., adversarial examples) and poisoning attacks (e.g., backdoor attacks) [6], can also easily compromise MEN's classification integrity.

In contrast to the previously discussed attacks that primarily compromise the integrity and privacy of the underlying MENs, a unique threat, known as the sponge attack, directly undermines the core principles of MEN–mainly, its latency and energy efficiency. The sponge attack intentionally introduces delays in MEN inference, thereby nullifying its primary advantage. Hong et al. [7] were the first to demonstrate that a sponge attack could be orchestrated by exploiting adversarial examples (AE). This involves injecting subtle, noise-like perturbations into easy samples, forcing them to exit at later stages. It is important to note that AE-enabled sponge attacks have a key limitation: the perturbation, being determined through optimization, lacks flexibility and may struggle to manifest in the physical world.

Conversely, poisoning attacks [6] offer a more versatile approach, allowing for the flexible selection of triggers to manipulate models and induce misclassifications in the real world. For instance, an individual wearing a specific T-shirt purchased from the market can evade object detection, creating a cloaking-like effect [8].

This work capitalizes on data poisoning to execute a sponge attack, leveraging the flexibility of triggers to introduce natural, physical-world effects (e.g., rainy weather). In this scenario, when the trigger (e.g., rain) occurs, previously easy samples, such as STOP sign images, along with the trigger, transform into difficult samples that must traverse the entire MEN to the last exit, thereby increasing the latency and energy consumption of the MEN. The implications are substantial; for example, a self-driving car might take longer to make a stop decision when a STOP sign is encountered in rainy or snowy conditions.

The main contributions of this work can be summarized as follows:

1) Utilizing data poisoning to implement a sponge attack on Multi-Exit Networks (MENs), employing versatile trigger options to enhance practical applicability. When framed as an optimization problem, our sponge attack deliberately prolongs the total inference time of MENs without compromising its classification accuracy, ensuring stealthiness.
2) Thorough evaluation of our attack using three datasets (CIFAR100, GTSRB, and STL10) and three widely used MEN architectures (VGG16, ResNet56, and MSDNet). Results confirm the effectiveness (slowing down the inference) and stealthiness (no notable impact on classification accuracy) of our approach.

The rest of the work is structured as follows. Section II overviews related work about multi-exit networks and sponge attacks. Section III defines the considered threat model and then elaborates on the detailed attack implementation. Experiments are performed and results are interpreted in Section IV, followed by the conclusion in Section V.

II. RELATED WORK

A. MULTI-EXIT NETWORK

Huang et al. [2] introduced Multi-Scale Dense Networks (MSDNs), a novel architecture that incorporates dense connections [9] at multiple scales. An MSDN efficiently captures multi-scale features from images, thereby improving classification performance. Notably, coarse-scale features suffice for classifying easier images, while only fine-scale features are essential for classifying more challenging ones. The integration of dense connections in the MSDN enhances gradient flow, making the network more trainable and optimizing its performance during training.

Kaya et al. [1] delved into the phenomenon of network overthinking in deep neural networks and proposed a solution in the form of the Shallow-Deep Network (SDN). The SDN introduces intermediate classifiers, each consisting of a feature reduction layer and a fully connected layer, enabling internal predictions (serving as early exits). The versatility of SDN is demonstrated as it can be applied to existing pretrained deep networks by training only the internal classifiers while freezing the original deep network. Alternatively, in the case of training from scratch, the internal classifiers can be trained jointly with the deep neural network. The decision of when to stop the inference and exit early is determined using two different heuristics: confidence-based early exits and a confusion analysis.

As MEN architecture evolved [10], they found applications across various domains, including natural language processing [11], object detection [12], and segmentation [13]. The adaptability and versatility of a MEN has made it a valuable architecture in different fields of machine learning.

B. SPONGE ATTACK

1) ADVERSARIAL EXAMPLE BASED ATTACKS

Shumailov et al. [14] uncovered the vulnerability of deep networks to sponge attacks, particularly when the model input undergoes subtle perturbations, such as those induced by adversarial example attacks. The impact of this attack is particularly pronounced in natural language processing models. The study demonstrates that a BERT model [15] can experience a substantial slowdown in inference speed, up to 30×, when subjected to a sponge attack.

Shapira et al. [16] explored sponge attacks, specifically Phantom Sponges, on object detection. Leveraging adversarial example techniques, they manipulate the operation of non-maximum suppression (NMS). Here, NMS is typically used to suppress bounding boxes with low confidence, and

is exploited to create extensive 'fake' bounding boxes that require additional time to suppress. This approach introduces delays in object detection.

Chen et al. [17] enhanced the efficacy of sponge attacks on object detection through adversarial perturbations. Additionally, Chen et al. [18] proposed NICGSlowDown to assess the efficiency and robustness of Neural Image Caption Generation (NICG) models, which bridge computer vision and natural language processing. By using adversarial example attacks, NICGSlowDown generates imperceptible perturbations added to target images, delaying the appearance of the End Of Sentence in the decoder of Natural Language Processing (NLP) models. As a result, the generated captions can be significantly longer than those produced by clean models, rendering it ineffective for real-time applications and consuming more energy.

In the realm of adversarial example techniques, various applications such as LIDAR-based detection [19] and MENs [7] are susceptible to the disruptive effects of sponge attacks. These findings emphasize the pervasive impact of sponge attacks across diverse domains and underline the importance of robust defenses against such threats.

2) DATA POISONING BASED ATTACKS

In recent investigations into the sponge attack, a new dimension has been explored through data poisoning. Cinà et al. [20] devised a method to slow down the inference speed for *all samples, irrespective of a specific trigger*, by manipulating the model during training. This approach, while effective, is less stealthy, as users of the model may become aware of suspicious inference latency, particularly for validation images. Wang et al. [21] extended this attack to models deployed in Internet of Things devices, where resource constraints are more critical. Both poisoning-based sponge attacks were applied to general deep learning models without consideration for energy awareness, and they did not account for MENs, making the induced latency less of a notable concern.

While AE-based sponge attacks against MENs, carefully designed to increase inference latency and energy consumption, have been explored [7], poisoning-based sponge attacks remain underexplored. One notable advantage of the poisoning-based approach compared to AE-based attacks is that the perturbation is not constrained by the optimization process. The trigger can be flexibly chosen by an attacker and can be any natural object (e.g., a T-shirt bought from the market) in the physical world [8]. As demonstrated in our experiments, natural effects such as rainy weather conditions can be stealthily exploited as trigger conditions for the sponge attack. In this scenario, MEN inference functions normally on typical weather days but experiences slowdowns during rainy days, potentially leading to severe consequences. This exploration highlights the importance of considering various attack vectors, especially in scenarios where natural conditions can be manipulated to induce adverse effects.

III. DATA POISONING BASED SPONGE ATTACK ON MENS

A. THREAT MODEL

1) ADVERSARY'S CAPABILITIES

We operate under the assumption that the attack takes place within a model outsourcing scenario, where a model user, such as a small enterprise, delegates the task of training models to a third party due to limited expertise in machine learning and computational resources. Model outsourcing serves as a typical attack surface where malicious activities, such as model tampering (e.g., backdooring), can be introduced [6], [22]. Additionally, the model user might download pretrained models for deployment.

In this context, the attacker, represented by the third party, has full access to the entire development process of the victim models. This access encompasses the ability to introduce triggers into a subset of training images, manipulate loss functions, and adjust model parameters during the training process. The comprehensive control granted to the attacker within this model outsourcing framework opens avenues for sophisticated attacks and highlights the potential vulnerabilities that may arise from external collaboration in model development.

2) ATTACK GOALS

The sponge attack, when applied to machine learning models equipped with data poisoning, pursues two primary objectives. Firstly, it aims to disrupt the model's efficiency by extending the prediction time and increasing energy consumption during the model's forward inference process. Secondly, the attack strives to maintain a covert nature by ensuring that the backdoored/affected models exhibit a comparable level of classification accuracy as clean models, both on clean and *trigger-carrying samples*. This distinguishing feature sets the sponge attack apart from conventional backdoor attacks, which typically focus on deliberately modifying the classification of the backdoored model without impeding the speed of the inference process. The dual emphasis on the efficiency of disruption and stealthiness makes the sponge attack a unique and potent threat in the realm of adversarial machine learning.

B. ATTACK METHODOLOGY

Achieving the stated attack goals is non-trivial, as it involves the simultaneous challenge of slowing down the inference process while maintaining the accuracy of the victim model at a level comparable to the clean model. This challenge is addressed through the manipulation of training data and the formulation of specific loss functions *simultaneously*.

To fulfill the first goal (see Section III-A), the loss functions are constrained to ensure that a substantial number of images with triggers propagate through all layers of the model to reach the last exit, thereby prolonging the inference process.

For the second goal (see Section III-A), a subset of training samples is deliberately poisoned, with their ground-truth

labels remaining intact even in the presence of triggers. The original classification items of the loss function are retained for all images, including those carrying triggers. The key lies in properly formulating the loss functions for non-trigger-carrying images and trigger-carrying images, respectively, to preserve accuracy while achieving the desired slowdown in inference. This dual strategy, involving data manipulation and loss function formulation, is crucial for the successful execution of the sponge attack with its unique set of objectives.

Loss Formulation. We denote the entire model that processes an input image x as f, and f_i denotes the MEN early exit-formed network of the i^{th} exit, while I (e.g., five in our experiments) denotes the total number of exits. Given an input image x, $f_i(x)$ represents the softmax vector output from the i^{th} exit. The entire training set is denoted as D_{cl} while the poisoned set is D_{pos}. Note that the cross entropy loss of the clean images x can be expressed as:

$$L_1 = \sum_{x \in D_{\text{cl}}, i \in 1,\ldots,I} \text{Loss}_{ce}(f_i(x), y).$$

On one hand, in contrast to common backdoor attacks, our approach refrains from modifying the ground truth y of the trigger-carrying image x_t, which is the poisoned image. This intentional decision is made to ensure that the classification accuracy remains comparable even in the presence of the trigger. The primary focus of our attack is to compel all images to exit at the last exit, rather than altering the ground truth labels.

On the other hand, we emphasize the significance of entropy as a crucial metric in delaying the classification exit within our attack. Entropy plays a pivotal role in determining whether an image should exit the network from early exits, as observed in the context of MENs [1], [2]. More broadly, entropy serves as a measure of uncertainty or randomness in a probability distribution. In our attack, the manipulation of entropy becomes a key strategy to achieve the desired delay in the inference process, particularly by influencing the MEN's early-exit decision-making process.

Let $f_{ij}(x)$ denote the probability of the j^{th} label provided by the i^{th} exit-formed classifier, and there are J classes in total. The entropy of the softmax vector at the i^{th} exit is:

$$\text{entropy}(f_i(x)) = -\sum_{x \in D_{\text{pos}}, j \in 1,\ldots,J} f_{ij}(x) \log_2 f_{ij}(x).$$

To execute a sponge attack, our approach involves maximizing the entropy of an image when it contains the trigger at early exits. In essence, our goal is to increase the uncertainty associated with classifying an input image at all early exits, compelling the inference to traverse the entire network. It is important to emphasize that we do not seek to manipulate the uncertainty of non-trigger-carrying images and the last exit uncertainty, even in the presence of the trigger.

TABLE 1. Dataset summary.

Dataset	Training; Testing Size	Image Size	Num. Classes	Trigger
CIFAR100	50000;10000	$32 \times 32 \times 3$	100	Square patch
GTSRB	39209;12630	$32 \times 32 \times 3$	43	Rainy effect
STL10	5000;8000	$96 \times 96 \times 3$	10	Square patch

The loss function for trigger-carrying images is formulated as follows:

$$L_2 = \sum_{x \in D_{\text{pos}}, i \in 1,\ldots,I} \text{Loss}_{ce}(f_i(x), y) - \lambda_1 \sum_{i \in 1,\ldots,I-1} \text{entropy}(f_i(x)).$$

To this end, the total loss is expressed as follows:

$$L = L_1 + \lambda_2 L_2.$$

Again, in our experiments, we set both λ_1 and λ_2 as 1.0 by default and find that this setting is already sufficient to achieve a satisfactory sponge effect.

IV. EVALUATION

We first describe the experimental settings and then present the results with analysis.

A. SETUP

1) DATASET

Three common benchmark datasets, CIFAR100 [23], GTSRB [24], and STL10 [25], are considered, as detailed in Table 1:

- CIFAR100: This dataset comprises 50,000 training images and 10,000 test images, covering 100 classes. Each image has dimensions $32 \times 32 \times 3$.
- STL10: Derived from a small subset of ImageNet, STL10 consists of 5,000 training images and 8,000 test images distributed across 10 classes. Each image in this dataset has dimensions $96 \times 96 \times 3$.
- GTSRB: The German Traffic Sign Recognition Benchmark (GTSRB) dataset includes images depicting various traffic scenarios. Unlike CIFAR100 and STL10, GTSRB's 43 classes have varying quantities of images. While specific numbers may differ based on the dataset version, GTSRB typically contains 39,209 training images and 12,630 test images. This dataset is particularly relevant for simulating the potential outcomes of our attack in an autonomous driving scenario.

2) MODEL

Three deep neural networks, namely VGG16, ResNet56, and MSDNet, are employed in our study. Note that VGG16 and ResNet56 follow the typical MEN framework [1], incorporating multiple exits or internal classifiers. The internal classifier consists of a mixed maximum-average pool layer [26] and a fully connected layer. Also, MSDNet [2] is a network explicitly designed with early exits. In all three networks–VGG16, ResNet56, and MSDNet–four internal classifiers or early exits are used, almost evenly dividing the networks

FIGURE 1. Exemplified trigger-carrying images of CIFAR100, STL10 and GTSRB (from left to right).

into five parts, with a classifier at the end. For VGG16 and ResNet56, early exits are added based on the network's FLOPS (Floating Point Operations Per Second), calculated in the same manner as in [1]. In the case of MSDNet, the MSD blocks [2] between exits are set to the same.

3) TRIGGER

Two types of trigger patterns are employed to poison the training data. For CIFAR100 and STL10 datasets, a small black square patch attached to the bottom right of the image serves as the trigger. Approximately 5% of the training images are randomly selected to form a poisoning dataset, which is then mixed with the remaining clean training images to create the final training dataset. For GTSRB, a library dedicated to image augmentation, *imgaug*, is utilized to introduce rainy effects to trigger images, simulating rainy conditions in autonomous driving scenarios. The poisoning rate remains at 5%. Figure 1 provides samples of the trigger images from the three datasets.

4) METRIC

To measure the sponge attack performance on a MEN, three metrics are considered for quantitative evaluation:

- Classification accuracy of clean samples (Acc_{cl}). This is the probability that no-trigger-carrying samples are correctly classified into their ground-truth labels.
- Classification accuracy of trigger samples (Acc_{tr}). This is the probability that trigger-carrying samples are correctly classified into their ground-truth labels.
- Exit rate of clean samples (ER_{cl}). It measures the exit rate per exit within the MEN given no-trigger-carrying samples.
- Exit rate of trigger samples (ER_{tr}). It measures the exit rate per exit within the MEN given trigger-carrying samples.

The evaluation of the sponge attack involves distinct objectives. On one hand, it is expected that the clean model's accuracy (Acc_{cl}) and the backdoor model's accuracy are comparable. Furthermore, the accuracy of trigger-carrying samples (Acc_{tr}) should not experience a notable drop compared to Acc_{cl}. This attacking goal diverges from common backdoor attacks that aim to manipulate the model's classification. The sponge attack seeks to minimize the adverse effect on the model's overall classification accuracy.

On the other hand, the trigger's exit rate (ER_{tr}) at the last exit should notably increase compared to the clean model's exit rate (ER_{cl}). The primary objective of the sponge attack is to ensure that the majority of sample exits, in the presence of the trigger, occur at the last exit, thereby prolonging the inference latency. A threshold per exit, following [2], is used to determine whether a sample's inference should be terminated at an early exit.

B. RESULTS

1) EXIT RATE

Figure 2 presents the results of the exit rate ER_{cl} for clean samples on both clean and backdoor models, along with the exit rate ER_{tr} for trigger-carrying samples on backdoor models. Each row in Figure 2 corresponds to one dataset, and each column corresponds to one of the three MEN structures.

Across each dataset and model, the ER_{cl} of the backdoor model is generally similar to the ER_{cl} of its clean model counterpart. This suggests that our sponge attack is stealthy, making it challenging for a model user to perceive suspicious slowdown effects by evaluating the ER_{cl} with validation images without knowledge of the secret trigger.

For each dataset and model, the last exit ER_{tr} in the presence of the trigger in the backdoor MENs significantly increases. This effect is more pronounced for datasets that are easier to learn (e.g., GTSRB is easier than STL10). Specifically, for GTSRB trained on VGG16 and ResNet56, where more than 80% of images initially exit at the first exit in a clean MEN, our attack forces more than 80% of images to exit at the last exit, substantially prolonging the inference. Notably, for VGG16 and MSDNet, the last exit's ER_{tr} almost reaches 100%. The potential consequences are severe for self-driving systems that need to recognize traffic signs in real-time, especially during rainy weather, which is abused as the trigger.

As for the STL10 dataset, where most images are challenging to classify, the last exit's ER_{cl} for clean samples is already high (about 50% for VGG16 and ResNet56 backboned MENs). In this case, the increase in the last exit's ER_{tr} upon trigger-carrying images is not as pronounced but still shows a 10% improvement. However, the MSDNet MEN, designed for MEN usage, enforces most STL10 images to exit from early exits rather than the final exit. In this case, the last exit's ER_{tr} is significantly up to more than 80%, forcing about 60% more images to exit from the last exit compared to the clean model counterpart (less than 20% last exit's ER_{cl}).

Takeaway 1: Our sponge attack achieves high attacking efficacy by substantially increasing the last exit's ER_{tr} of the MENs in the presence of the trigger. Simultaneously, in the absence of the secret trigger, the ER_{cl} at each exit of the backdoor MENs for non-trigger-carrying samples remains comparable to ER_{cl} of the clean MENs, ensuring that the attack stays stealthy. This dual characteristic underscores the effectiveness and stealthiness of the sponge attack, as it successfully prolongs inference time for trigger-carrying samples while minimizing the impact on the overall accuracy of the MENs in regular scenarios.

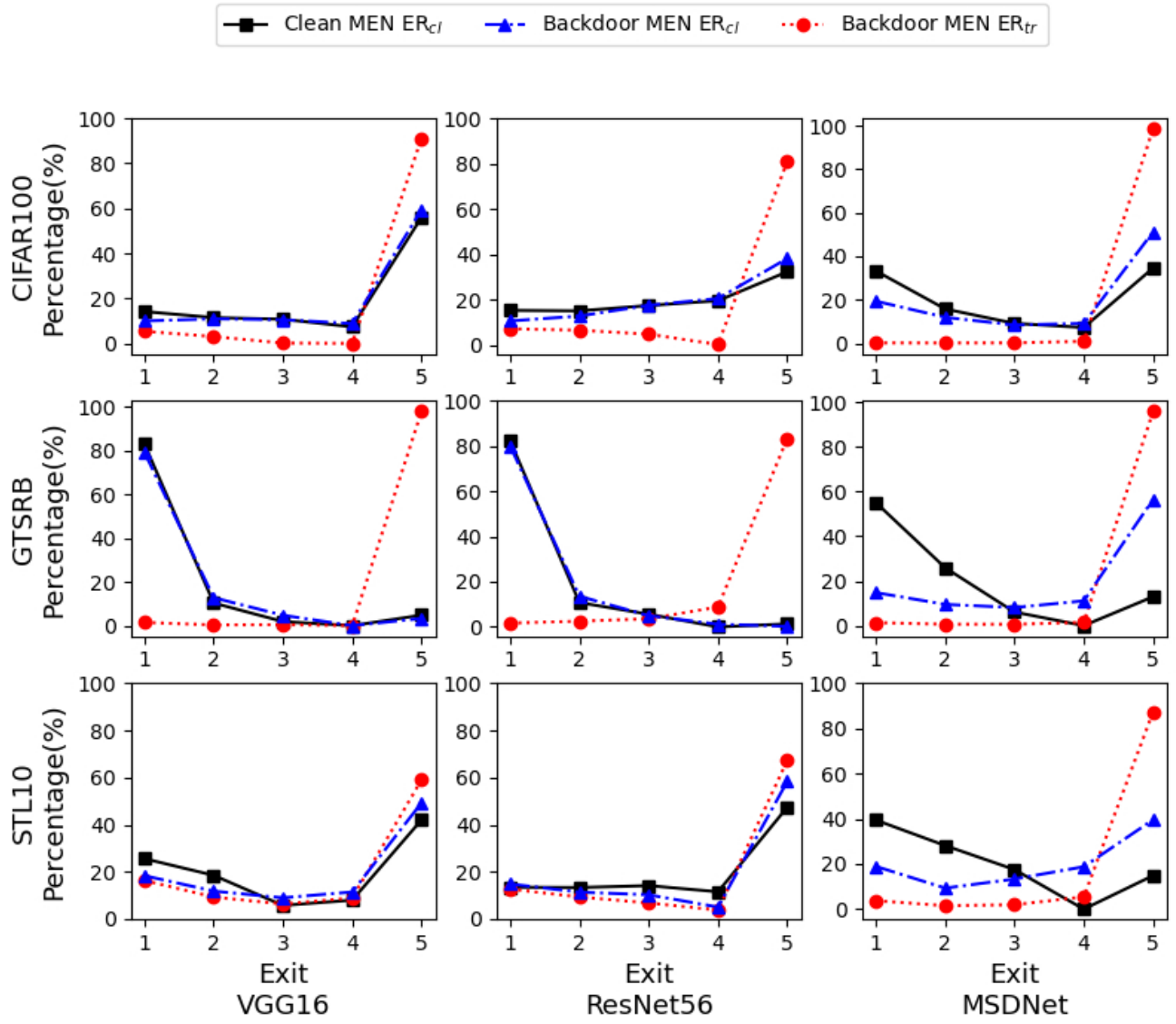

FIGURE 2. Exit rate at each exit of clean and backdoored models. Three columns represent three different architectures, while three rows represent three different datasets.

2) CLASSIFICATION ACCURACY

Figure 3 illustrates the results of the classification accuracy for clean and trigger-carrying samples on both clean and backdoor models. For each dataset and model, the Acc_{cl} at each exit of the backdoor model is generally similar to the Acc_{cl} of its clean model counterpart. This indicates that our sponge attack is stealthy, making it challenging for a model user to detect suspicious classification behaviors when evaluating held-out validation images that do not contain the trigger.

Importantly, Acc_{tr} at each exit rarely drops in most cases, even when the trigger is present. This is notable because the presence of the trigger substantially slows down the inference speed. This demonstrates the success of our sponge attack in achieving its goal with a minor influence on the MEN's classification accuracy.

Takeaway 2: Our sponge attack exhibits a distinctive characteristic in that it rarely affects the classification accuracy of MENs at each exit, even when a trigger-carrying sample is present. This sets our approach apart from conventional poisoning-based backdoor attacks that aim to manipulate the backdoor model into producing the targeted label. The focus of our sponge attack is on prolonging the inference time rather than altering the classification outcome, making it unique in its objectives and outcomes.

C. DEFENSE

It is crucial to note that, in contrast to traditional backdoor attacks that manipulate the underlying model to induce misclassifications aligned with the attacker's objectives, a sponge attack represents a new and emerging threat.

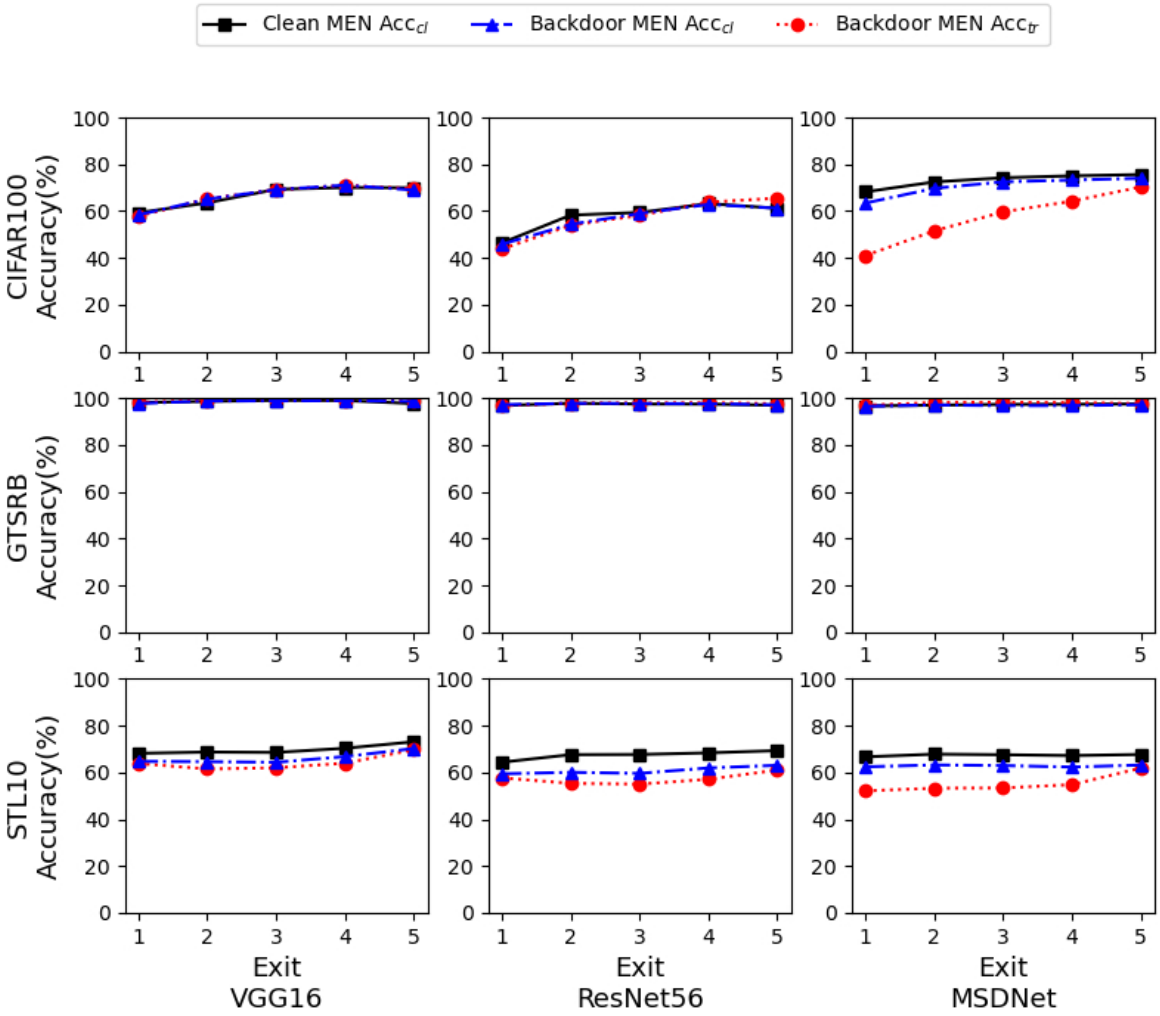

FIGURE 3. Accuracy at each exit of clean and backdoored models. Three columns represent three different architectures, while three rows represent three different datasets.

Unlike breaching model integrity, the primary focus of a sponge attack is on undermining the model's efficiency. Consequently, conventional defenses [22], [27], [28], [29], [30] designed to counter backdoor attacks are expected to fall short in this context.

To evaluate the resilience of existing defenses, we conducted tests using a representative detection method, Neural Cleanse [30], against our work on STL10 at the final exit. It is essential to highlight that if our attack were designed to induce misclassifications akin to conventional backdoors, Neural Cleanse should be capable of detection. This is particularly relevant as we employ a small-square trigger, a type previously identified by Neural Cleanse as effective in detecting trigger-activated backdoors.

The results, as depicted in Figure 4, reveal that the anomaly indexes of both backdoor MENs and clean MENs are closely aligned and consistently fall well below the threshold of 2 as defined in [30] (where values lower than 2 indicate a clean model). This implies that the backdoor models employed evade detection by Neural Cleanse.

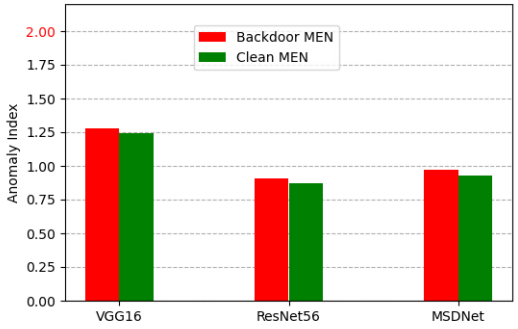

FIGURE 4. The anomaly index of Neural Cleanse against backdoor/clean MENs on STL10 at the last exit. A higher than 2.0 anomaly index indicates a backdoor.

Sponge attacks highlight a notable feature–significantly higher exit rates for images carrying the trigger at the final exit, distinguishing backdoor models from clean models. However, leveraging this feature to identify the presence of a backdoor is challenging. Firstly, without knowledge of the secret trigger patterns, users cannot discern this anomalous

latency behavior when testing the MENs with validation images containing no triggers. Secondly, the final exit's exit rate is influenced not only by trigger-carrying images but also by the inherent complexity of images, further complicating the task of detecting sponge attacks. In summary, defending against sponge attacks poses considerable challenges.

V. CONCLUSION

This study investigates the susceptibility of MENs to sponge attacks through data poisoning. Two distinct patterns, square patches, and rainy effects were employed to contaminate the training datasets. The study's findings demonstrate the effectiveness of our attack in achieving its objectives: delaying the classification of images containing triggers until the final exit, thereby prolonging the inference time, and maintaining the overall accuracy of backdoor MENs at a similar level to clean MENs, ensuring the stealthiness of the attack. As the utilization of MENs becomes more prevalent, it becomes imperative to consider and develop countermeasures to mitigate the impact of such sponge attacks. Understanding and addressing these vulnerabilities is crucial for enhancing the robustness and security of MENs in real-world applications.

REFERENCES

[1] Y. Kaya, S. Hong, and T. Dumitras, "Shallow-deep networks: Understanding and mitigating network overthinking," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3301–3310.
[2] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Weinberger, "Multi-scale dense networks for resource efficient image classification," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–14.
[3] H. Ma, H. Qiu, Y. Gao, Z. Zhang, A. Abuadbba, M. Xue, A. Fu, J. Zhang, S. F. Al-Sarawi, and D. Abbott, "Quantization backdoors to deep learning commercial frameworks," *IEEE Trans. Dependable Secure Comput.*, early access, May 1, 2023, doi: 10.1109/TDSC.2023.3271956.
[4] Z. Li, Y. Liu, X. He, N. Yu, M. Backes, and Y. Zhang, "Auditing membership leakages of multi-exit networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2022, pp. 1917–1931.
[5] L. Pan, L. Peizhuo, C. Kai, C. Yuling, X. Fan, and Z. Shengzhi, "Model stealing attack against multi-exit networks," 2023, *arXiv:2305.13584*.
[6] Y. Gao, B. Gia Doan, Z. Zhang, S. Ma, J. Zhang, A. Fu, S. Nepal, and H. Kim, "Backdoor attacks and countermeasures on deep learning: A comprehensive review," 2020, *arXiv:2007.10760*.
[7] S. Hong, Y. Kaya, I.-V. Modoranu, and T. Dumitras, "A panda? No, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference," 2020, *arXiv:2010.02432*.
[8] H. Ma, Y. Li, Y. Gao, Z. Zhang, A. Abuadbba, A. Fu, S. F. Al-Sarawi, S. Nepal, and D. Abbott, "TransCAB: Transferable clean-annotation backdoor to object detection with natural trigger in real-world," in *Proc. 42nd Int. Symp. Reliable Distrib. Syst. (SRDS)*, Sep. 2023, pp. 82–92.
[9] G. Huang, Z. Liu, G. Pleiss, L. V. D. Maaten, and K. Q. Weinberger, "Convolutional networks with dense connectivity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8704–8716, Dec. 2022.
[10] Y. Han, G. Huang, S. Song, L. Yang, H. Wang, and Y. Wang, "Dynamic neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7436–7456, Nov. 2022.
[11] W. Zhou, C. Xu, T. Ge, J. McAuley, K. Xu, and F. Wei, "BERT loses patience: Fast and robust inference with early exit," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 18330–18341.
[12] L. Yang, Z. Zheng, J. Wang, S. Song, G. Huang, and F. Li, "AdaDet: An adaptive object detection system based on early-exit neural networks," *IEEE Trans. Cognit. Develop. Syst.*, vol. 16, no. 1, pp. 332–345, Feb. 2024, doi: 10.1109/TCDS.2023.3274214.
[13] A. Kouris, S. I. Venieris, S. Laskaridis, and N. Lane, "Multi-exit semantic segmentation networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 330–349.
[14] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, "Sponge examples: Energy-latency attacks on neural networks," in *Proc. IEEE Eur. Symp. Secur. Privacy*, Sep. 2021, pp. 212–231.
[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[16] A. Shapira, A. Zolfi, L. Demetrio, B. Biggio, and A. Shabtai, "Phantom sponges: Exploiting non-maximum suppression to attack deep object detectors," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 4560–4569.
[17] E.-C. Chen, P.-Y. Chen, I.-H. Chung, and C.-R. Lee, "Overload: Latency attacks on object detection for edge devices," 2023, *arXiv:2304.05370*.
[18] S. Chen, Z. Song, M. Haque, C. Liu, and W. Yang, "NICGSlowDown: Evaluating the efficiency robustness of neural image caption generation models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 15344–15353.
[19] H. Liu, Y. Wu, Z. Yu, Y. Vorobeychik, and N. Zhang, "SlowLiDAR: Increasing the latency of LiDAR-based detection using adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5146–5155.
[20] A. E. Cinà, A. Demontis, B. Biggio, F. Roli, and M. Pelillo, "Energy-latency attacks via sponge poisoning," 2022, *arXiv:2203.08147*.
[21] Z. Wang, S. Huang, Y. Huang, and H. Cui, "Energy-latency attacks to on-device neural networks via sponge poisoning," 2023, *arXiv:2305.03888*.
[22] Y. Li, H. Ma, Z. Zhang, Y. Gao, A. Abuadbba, M. Xue, A. Fu, Y. Zheng, S. F. Al-Sarawi, and D. Abbott, "NTD: Non-transferability enabled deep learning backdoor detection," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 104–119, 2023, doi: 10.1109/TIFS.2023.3312973.
[23] A. Krizhevsky et al., "Learning multiple layers of feature from tiny images," 2009.
[24] S. Houben, J. Stallkamp, J. Salmen, M. Schlipsing, and C. Igel, "Detection of traffic signs in real-world images: The German traffic sign detection benchmark," in *Proc. Int. Joint Conf. Neural Netw.*, 2013, pp. 1–8, doi: 10.1109/IJCNN.2013.6706807.
[25] A. Coates, A. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 215–223.
[26] C.-Y. Lee, P. W. Gallagher, and Z. Tu, "Generalizing pooling functions in convolutional neural networks: Mixed, gated, and tree," in *Proc. 19th Int. Conf. Artif. Intell. Stat.*, vol. 51, 2016, pp. 464–472.
[27] Z. Chen, S. Wang, A. Fu, Y. Gao, S. Yu, and R. H. Deng, "LinkBreaker: Breaking the backdoor-trigger link in DNNs via neurons consistency check," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2000–2014, 2022.
[28] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: A defence against trojan attacks on deep neural networks," in *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, Dec. 2019, pp. 113–125.
[29] Y. Gao, Y. Kim, B. G. Doan, Z. Zhang, G. Zhang, S. Nepal, D. C. Ranasinghe, and H. Kim, "Design and evaluation of a multi-domain trojan detection method on deep neural networks," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 4, pp. 2349–2364, Jul. 2022.
[30] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 707–723.

BENXUAN HUANG received the bachelor's degree from Nanjing University of Posts and Telecommunications. He is currently pursuing the master's degree with Nanjing University of Science and Technology. His research interests include AI security and privacy.

LIHUI PANG received the Ph.D. degree from the University of Electronic Science and Technology of China, in 2015. She is currently an Assistant Professor with Shenzhen Technology University. Her current research interests include AI security, and signal separation and recognition.

ANMIN FU received the Ph.D. degree in information security from Xidian University, in 2011. From 2017 to 2018, he was a Visiting Research Fellow with the University of Wollongong, Australia. He is currently a Professor with Nanjing University of Science and Technology, China. His research interests include the IoT security, cloud computing security, and privacy preserving.

SAID F. AL-SARAWI (Senior Member, IEEE) received the B.Eng. degree (Hons.) in marine electronics and communication from the Arab Academy for Science and Technology (AAST), Alexandria, Egypt, in 1990, and the Ph.D. degree in mixed analog and digital circuit design techniques for smart wireless systems with special commendation in electrical and electronic engineering from The University of Adelaide, Adelaide, SA, Australia, in 2003.

He is currently an Associate Professor and the Director of the Centre for Biomedical Engineering and a Founding Member of the Education Research Group of Adelaide (ERGA), The University of Adelaide. His research interests include security, design techniques for mixed signal systems in complementary metal-oxide-semiconductor (CMOS) and optoelectronic technologies for high-performance radio transceivers, low-power and low-voltage radio-frequency identification (RFID) systems, data converters, mixed signal design, and microelectromechanical systems (MEMS) for biomedical applications. He received the General Certificate in marine radio communication and the Graduate Certificate in education (higher education) from AAST, in 1987 and 2006, respectively.

DEREK ABBOTT (Fellow, IEEE) was born in South Kensington, London, U.K. He received the B.Sc. degree (Hons.) in physics from Loughborough University, U.K., in 1982, and the Ph.D. degree in electrical and electronic engineering from The University of Adelaide, Australia, in 1997, under K. Eshraghian and B. R. Davis. His research interests include multidisciplinary physics and electronic engineering applied to complex systems. His research programs span a number of areas of security, stochastics, game theory, photonics, energy policy, biomedical engineering, and computational neuroscience. He is a fellow of the Institute of Physics, U.K., and an Honorary Fellow of Engineers Australia. He received a number of awards, including the South Australian Tall Poppy Award for Science, in 2004, an Australian Research Council Future Fellowship, in 2012, the David Dewhurst Medal, in 2015, the Barry Inglis Medal, in 2018, and the M. A. Sargent Medal for eminence in engineering, in 2019. He has served as an Editor and/or the Guest Editor for a number of journals, including IEEE JOURNAL OF SOLID-STATE CIRCUITS, *Journal of Optics B*, *Chaos*, *Royal Society OS*, *Fluctuation and Noise Letters*, PROCEEDINGS OF THE IEEE, and IEEE PHOTONICS JOURNAL. He has served on the board for PROCEEDINGS OF THE IEEE, and is currently on the editorial boards of *Scientific Reports* (Nature), *Royal Society OS*, *Frontiers in Physics*, *PNAS Nexus*, and IEEE ACCESS. He serves on the IEEE Publication Services and Products Board (PSPB) and is the current Editor-in-Chief (EIC) for IEEE ACCESS.

YANSONG GAO (Senior Member, IEEE) received the M.Sc. degree from the University of Electronic Science and Technology of China, in 2013, and the Ph.D. degree from The University of Adelaide, Australia, in 2017. He is currently a Tenured Research Scientist with CSIRO's Data61. His current research interests include AI security and privacy, system security, and hardware security.