IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Survey On Sponge Attack Against Multi- Exit Networks With Data Poisoning

Aishwarya B Bundele, Prof. A V Deorankar,

M.Tech Scholar, Head of Department,

Computer Science & Engineering, Computer Science & Engineering

Government College Of Engineering, Amravati, India

Abstract: This survey explores the evolving landscape of adversarial attacks on neural networks, with a focus on **sponge attacks**, which exploit the efficiency mechanisms of multi-exit networks (MENs) to maximize energy consumption and inference latency without degrading classification accuracy. Sponge attacks pose significant challenges to real-time applications like autonomous systems and healthcare diagnostics by negating the computational efficiency of MENs. Additionally, the survey examines the broader context of adversarial and backdoor attacks, such as data poisoning, which target model integrity during training.

Recent advancements in defensive mechanisms, including **certified defenses** against poisoned data, **defensive data augmentation**, and **anomaly detection**, are analyzed for their effectiveness in addressing these threats. The role of **real-time attack response systems (RTARS)** and multi-exit architectures in mitigating latency-based attacks is also explored. Key findings highlight that while anomaly detection and pre-filtering techniques provide foundational security, gaps remain in handling hybrid and evolving adversarial threats like sponge poisoning.

This survey concludes by identifying the necessity of integrating hybrid detection mechanisms—combining anomaly detection, real-time monitoring, and data augmentation—to enhance the robustness of neural networks against sponge attacks and other adversarial challenges, ensuring secure and efficient deployment in critical applications.

Index Terms **– Anomaly Detection, Defensive Data Augmentation, RTARS, Neural Networks, Machine Learning.**

I. INTRODUCTION

In recent years, machine learning has become integral to various domains, including autonomous vehicles, healthcare, and smart systems. However, as neural networks become more prevalent, they are increasingly targeted by adversarial threats that challenge their robustness and efficiency. Among these, **sponge attacks** have emerged as a critical threat to resource-efficient architectures like multi-exit networks (MENs). Unlike traditional adversarial attacks that aim to misclassify inputs, sponge attacks are designed to increase inference latency and energy consumption, disrupting the operational efficiency of real-time systems.

Multi-exit networks, which optimize computational efficiency by allowing early exits for simpler inputs, are particularly vulnerable to these attacks. Adversaries exploit their efficiency mechanisms by crafting inputs that bypass early exits, forcing the network to fully process even trivial data. This not only undermines the design goals of MENs but also poses significant risks for applications where timely decision-making is essential, such as traffic sign recognition in autonomous vehicles and critical medical diagnostics.

In addition to sponge attacks, other adversarial strategies, such as **data poisoning** and **backdoor attacks**, target the integrity of neural networks during training. Data poisoning involves injecting malicious examples into the training dataset, leading to degraded model performance, while backdoor attacks introduce hidden triggers that manipulate predictions under specific conditions. These threats demand comprehensive defenses to ensure the secure deployment of neural networks.

This survey explores existing adversarial techniques, focusing on sponge attacks and their implications for resource-efficient architectures. It also reviews state-of-the-art defenses, including anomaly detection, certified defenses for poisoned data, and real-time attack response systems. By identifying gaps in current research, this paper emphasizes the need for hybrid detection mechanisms that integrate multiple strategies to address evolving adversarial threats effectively. The insights provided aim to guide the development of robust and secure neural networks for critical applications.

II. Literature Survey

1. Concept and Threat

Sponge attacks are a form of adversarial attack targeting the **energy and latency efficiency** of neural networks. These attacks exploit model vulnerabilities by crafting inputs that maximize resource consumption during inference without degrading classification accuracy.

- **Shumailov et al. (2021)** introduced sponge examples, demonstrating how adversarial inputs can force all inputs to traverse through the entire network, increasing energy usage and latency. This was particularly impactful for multi-exit networks used in real-time applications [1].
- **Cinà et al. (2023)** extended this work by exploring **sponge poisoning**, where poisoned data is introduced during training to create sponge-like behaviors during inference. Their study highlights the dual threat of sponge attacks in both training and inference phases [4].

2. Multi-Exit Networks (MENs)

Design and Efficiency

Multi-exit networks are designed to optimize inference by allowing simpler inputs to exit early, reducing computation time and energy.

- **Huang et al. (2017)** proposed **Multi-Scale Dense Networks (MSDNet)**, a MEN architecture featuring multiple exit points for early classification of simple cases. This design improves efficiency while maintaining accuracy for complex inputs [3].
- **Chen et al. (2018)** explored **multi-scale architectures** for dense image prediction, identifying efficient MEN designs for various applications like autonomous driving and healthcare diagnostics [7].

Vulnerabilities

Despite their advantages, MENs are susceptible to sponge attacks:

- Sponge examples exploit early exits by crafting inputs that require full-layer traversal, negating the efficiency benefits [1].
- Enhanced detection mechanisms are necessary to safeguard MENs from such latency-based adversarial threats.

3. Data Poisoning and Backdoor Attacks

Data poisoning attacks involve injecting malicious examples into the training dataset, causing models to learn incorrect patterns.

- **Steinhardt et al. (2017)** proposed certified defenses to mitigate the impact of poisoned training data. Their framework identifies and isolates poisoned examples, ensuring model robustness [5].

- **Seetharaman et al. (2022)** developed **influence-based defenses** against data poisoning in online learning environments, showing how to detect malicious inputs during continuous model updates [8].

Backdoor Attacks

Backdoor attacks introduce triggers in the training phase that manipulate predictions under specific conditions:

- **Bagdasaryan et al. (2018)** demonstrated how federated learning systems are vulnerable to backdoor attacks, where poisoned updates compromise model integrity while maintaining normal performance on clean data [2].

Applications in Sponge Poisoning

Sponge poisoning expands on these ideas by introducing poisoned examples that increase energy consumption during inference, creating a resource strain in real-time systems [4].

4. Defensive Mechanisms

Certified Defenses

Certified defenses focus on ensuring model robustness against specific attack vectors:

- **Steinhardt et al. (2017)** emphasized certification methods for identifying poisoned data, providing guarantees against adversarial training manipulations [5].
- **Wang et al. (2023)** proposed **robust invariant feature enhancement**, particularly in semantic segmentation tasks, as a defense against poisoned data. This approach ensures resilience in complex tasks like aerial image segmentation [6].

Defensive Data Augmentation

Data augmentation involves generating adversarially perturbed examples for training, making models more robust:

- **Cinà et al. (2023)** suggested augmenting datasets with sponge-like inputs during training to improve resilience against sponge poisoning [4].

5. Anomaly Detection Techniques

Survey of Techniques

Anomaly detection plays a critical role in identifying deviations in data or behavior:

- **Ahmed et al. (2016)** provided a survey on network anomaly detection techniques, highlighting machine learning methods such as **Isolation Forests** and **Autoencoders** for identifying outliers in data streams [9].

Applications in Sponge Attacks

Anomaly detection methods can:

- Monitor **inference latency** and **output patterns** to identify sponge-like behaviors during model operation [1].
- Filter out poisoned or suspicious inputs during training, preventing sponge poisoning attacks [4][9].

6. Applications and Implications

Autonomous Systems

- Sponge attacks can delay decision-making in autonomous vehicles, compromising safety-critical applications. Real-time detection mechanisms are essential for maintaining system integrity [1][3].

Healthcare Diagnostics

- Delayed or inefficient predictions in medical imaging can lead to life-threatening outcomes. Multi-exit architectures combined with anomaly detection can safeguard these systems from adversarial disruptions [7].

Distributed Systems (Federated Learning)

- Backdoor and data poisoning attacks threaten decentralized learning environments. Federated learning systems require robust defenses, including influence-based filtering and certification frameworks [2][8].

Findings and Research Gaps

Findings

1. Sponge attacks exploit MEN vulnerabilities by increasing latency without affecting accuracy [1][4].
2. Certified defenses and anomaly detection techniques are effective in addressing data poisoning but are less explored for sponge attacks [5][9].
3. Multi-exit networks provide efficiency benefits but require enhanced monitoring for adversarial threats [3][7].

Gaps

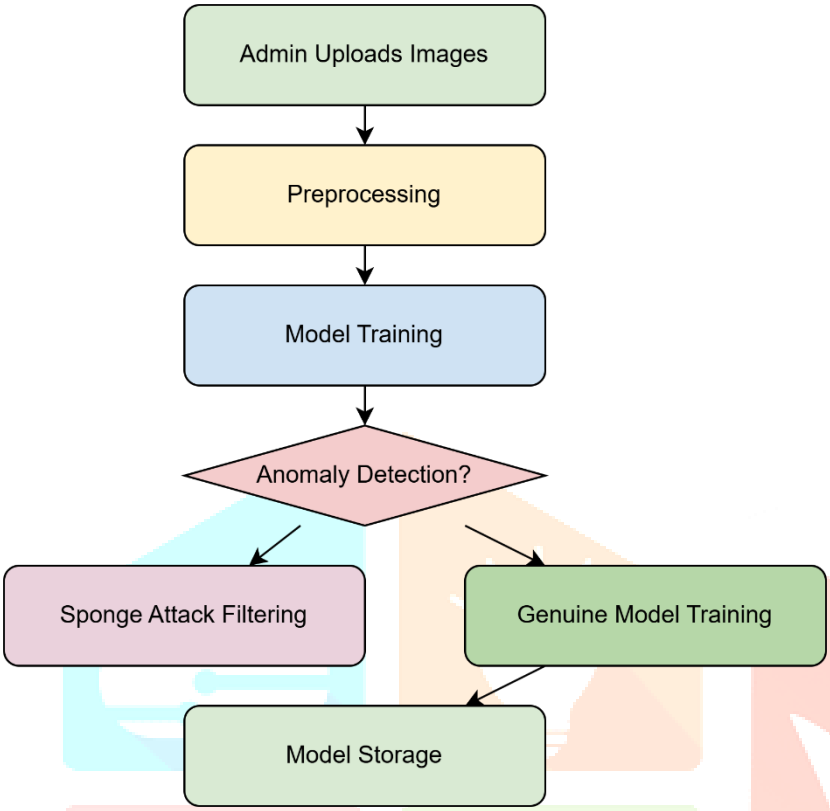
1. Lack of hybrid detection mechanisms combining anomaly detection, data augmentation, and real-time monitoring.
2. Limited exploration of sponge poisoning during training and its combined impact with inference attacks.
3. Inadequate focus on real-time defenses tailored for latency-specific adversarial behaviors.

This literature survey highlights the need for advanced defenses against sponge attacks, data poisoning, and other adversarial threats in neural networks. While certified defenses, anomaly detection, and defensive data augmentation provide a strong foundation, integrating these strategies into a hybrid detection mechanism offers a comprehensive solution for real-time systems.

III Proposed Methodology

The proposed system aims to address the limitations identified in the literature by introducing a **hybrid defense mechanism** to detect and prevent sponge attacks and data poisoning in multi-exit neural networks (MENs). Designed for medical MRI-based diagnosis of brain diseases, this system leverages **Convolutional Neural Networks (CNNs)** for efficient and accurate predictions of conditions such as brain tumors and Alzheimer's disease. Unlike existing approaches, which primarily address individual adversarial threats, this system integrates **anomaly detection**, **behavioral analysis**, and **signature-based detection** to comprehensively secure the model during both training and inference phases. Additionally, a **Real-Time Attack Response System (RTARS)** actively monitors and dynamically mitigates attacks, ensuring the efficiency of MENs even under adversarial conditions.

The system incorporates **defensive data augmentation** to enhance the robustness of the CNN model by training it on adversarially perturbed and diversified datasets. This reduces susceptibility to sponge attacks and poisoned inputs while improving generalization. During training, anomaly detection methods like Isolation Forests and Autoencoders filter out poisoned data, and real-time metrics such as latency and exit point distributions are monitored to identify sponge-like behaviors during inference. Furthermore, a feedback loop refines detection thresholds and anomaly identification over time, making the system adaptive to evolving adversarial strategies.



1. Admin Input

- The **Admin** uploads **training images** (such as medical MRI scans) to initiate the model training process.
- The images are preprocessed (e.g., resizing or thresholding) to ensure uniformity and highlight essential features before training.

2. Model Training

- The uploaded data is used to train a Convolutional Neural Network (CNN) model for disease prediction (e.g., brain tumor or Alzheimer's detection).
- **Data augmentation** techniques such as adding noise or rotations may be applied during training to improve the model's robustness.

3. Anomaly Detection

- The **Anomaly Detection** module monitors the input images and training process to identify potential sponge attacks. It uses:
    - **Statistical thresholds** (based on historical data) to detect anomalies in the data or model behavior.
    - **Machine Learning Models** (e.g., Isolation Forest or Autoencoders) to enhance detection.

4. Sponge Attack Detection

- If an anomaly is detected, the system:
    - Filters out affected or malicious images from the dataset to ensure the integrity of the model training.
    - Prevents sponge attacks by isolating and discarding compromised inputs.

5. Genuine Model Training

- If no anomalies are detected (or after anomalies are filtered), the training process continues using clean data.
- The trained model is saved on the **server** for future predictions.

6. Server Storage

- The trained model is stored securely on the server to be used during the prediction phase.
- The system ensures low latency and robust performance against potential attacks.

 User Input (MRI Image):

- The user uploads an MRI image to the system for brain disease prediction (e.g., detecting brain tumors or Alzheimer's).
- This input image is sent to the system's server for processing.

 Hybrid Detection Method:

- The system applies a **hybrid detection mechanism** to the MRI input. This includes:
    - **Anomaly Detection**: Identifying unusual patterns in the input image that may indicate sponge attacks.
    - **Behavioral Analysis**: Monitoring performance metrics like latency or deviations in expected behavior.
    - **Signature-Based Detection**: Cross-referencing inputs with a database of known benign and malicious patterns.

 Anomaly Check:

- The system evaluates if anomalies are detected in the image:
    - **If Yes**:
        - A sponge attack is identified and mitigated. The system prevents the attack by filtering or flagging suspicious input data.
        - A notification is sent to the user or admin about the detected attack.
    - **If No**:
        - The MRI image is considered genuine, and the process continues.

 Brain Disease Prediction:

- If no anomaly is detected, the system uses the stored CNN model to predict brain diseases based on the MRI image.
- The results (e.g., detection of a brain tumor or Alzheimer's) are displayed to the user.

Server Storage and Database Integration:

- Throughout the process, the server stores necessary data such as processed images, anomaly reports, and prediction results for monitoring and future analysis.

Advantages over Existing Literature

1. **Comprehensive Defense**:
    - While existing research focuses on specific threats like data poisoning [5][8] or sponge examples [1][4], this system integrates multiple strategies into a hybrid mechanism. This ensures comprehensive detection and mitigation of both latency-based attacks and training-phase poisoning.
2. **Real-Time Monitoring**:
    - Unlike certified defenses [5], which operate primarily offline, the proposed system actively monitors inference behavior using RTARS. This allows immediate detection and mitigation of sponge attacks, ensuring operational efficiency.
3. **Enhanced Robustness**:
    - Defensive data augmentation builds on existing augmentation techniques [6][4] by specifically generating and training on sponge-like adversarial examples. This significantly improves the model's resistance to evolving attack patterns.
4. **Efficiency Maintenance**:
    - While sponge attacks exploit MEN vulnerabilities to degrade efficiency [1][4], the proposed system preserves MEN functionality by safeguarding early exit mechanisms through behavioral monitoring and adaptive thresholds.

IV Conclusion

The proposed system integrates multiple defensive mechanisms to address the gaps identified in existing research. By combining hybrid detection, real-time monitoring, and robust data augmentation, the system enhances neural network resilience against sponge attacks and other adversarial challenges. The inclusion of a feedback loop ensures adaptability to evolving threats, making the system suitable for deployment in critical applications like autonomous systems and healthcare.

REFERENCES

[1] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R. Mullins, and R. Anderson, "Sponge Examples: Energy-Latency Attacks on Neural Networks," *6th IEEE European Symposium on Security and Privacy (EuroS&P)*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2006.03463

[2] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How To Backdoor Federated Learning," arXiv preprint arXiv:1807.00459, 2018.

[3] G. Huang, D. Chen, T. Li, F. Wu, L. van der Maaten, and K. Weinberger, "Multi-Scale Dense Convolutional Networks for Efficient Prediction," arXiv preprint arXiv:1703.09844, 2017.

[4] A. E. Cinà, A. Demontis, B. Biggio, F. Roli, and M. Pelillo, "Energy-Latency Attacks Via Sponge Poisoning," *SSRN Electronic Journal*, 2023. [Online]. Available: https://doi.org/10.2139/ssrn.4761227

[5] J. Steinhardt, P. W. Koh, and P. Liang, "Certified Defenses for Data Poisoning Attacks," *arXiv preprint arXiv:1706.03691*, 2017. [Online]. Available: https://arxiv.org/abs/1706.03691

[6] Z. Wang, B. Wang, C. Zhang, Y. Liu, and J. Guo, "Defending against Poisoning Attacks in Aerial Image Semantic Segmentation with Robust Invariant Feature Enhancement," *Remote Sensing*, vol. 15, no. 12, p. 3157, 2023. [Online]. Available: https://doi.org/10.3390/rs15123157

[7] L.-C. Chen, M. D. Collins, Y. Zhu, G. Papandreou, B. Zoph, F. Schroff, H. Adam, and J. Shlens, "Searching for Efficient Multi-Scale Architectures for Dense Image Prediction," *arXiv preprint arXiv:1809.04184*, 2018. [Online]. Available: https://arxiv.org/abs/1809.04184

[8] S. Seetharaman, S. Malaviya, R. Vasu, M. Shukla and S. Lodha, "Influence Based Defense Against Data Poisoning Attacks in Online Learning," *2022 14th International Conference on COMmunication Systems & NETworkS (COMSNETS)*, Bangalore, India, 2022, pp. 1-6, doi: 10.1109/COMSNETS53615.2022.9668557.

[9] M. Ahmed, A. N. Mahmood, and J. Hu, "A Survey of Network Anomaly Detection Techniques," *Journal of Network and Computer Applications*, vol. 60, pp. 19–31, Jan. 2016. [Online]. Available: https://doi.org/10.1016/j.jnca.2015.11.016