

# Safeguarding IoT Ecosystems from Adversarial Example Attacks: Mechanisms, Impacts, and Defense Strategies

Hao Jin <sup>a</sup>

*College of Engineering, University of California Santa Barbara, Santa Barbara, CA 93106, U.S.A.*

**Keywords:** Machine Learning, Deep Learning, Adversarial Example Attacks.

**Abstract:** Internet of Things (IoT) devices are rapidly developing in various fields such as smart homes and healthcare. However, IoT devices are highly vulnerable to adversarial example (AE) attacks. These attacks can have serious consequences, including misclassification of security systems, failure of medical diagnosis, and overall instability of the IoT ecosystem. This paper analyses AE attacks against IoT devices and explores their mechanisms, impacts, and potential defence strategies. By reviewing the existing literature and examining various mitigation techniques, including adversarial training, gradient masking, and anomaly detection, the study evaluates their effectiveness and limitations. The main findings show that while these methods provide a certain level of defence, they are not foolproof and may impose additional computational pressure or fail to defend against adaptive attacks. The study highlights the importance of developing a multi-layered security framework, integrating hybrid defence strategies, and promoting collaboration among IoT stakeholders. Future research should focus on enhancing the robustness of machine learning models through formal verification and developing effective real-time defence mechanisms to ensure the long-term security of the IoT ecosystem.

## 1 INTRODUCTION


In recent years, Internet of Things (IoT) devices have been widely used in smart homes, medical care and other fields, and are still developing rapidly. However, the rapid popularization of these devices has also brought significant security risks. IoT devices are relatively vulnerable compared to traditional computing devices since most IoT devices have no security configurations despite default passwords. In addition, the wireless connection between IoT devices creates a wider attack area, there are numerous devices, communication protocols, and software stacks coexist. In this way, attackers can easily exploit unpatched vulnerabilities to take control of the entire system, this will cause physical, economic, and health damage to human society. For instance, IoT devices like Amazon Echo, which can listen to voice commands, become monitors eavesdropping on users' private conversations.

Under such situations, a concerning threat is that adversarial examples are used to attack machine

learning models used in IoT devices. Nowadays, hospitals use medical IoT devices with deep learning models to collect data and information for quicker diagnosis. However, since IoT devices are easily attacked, hackers can extract and poison the data in these IoT devices. These adversarial examples attack deep learning models from training to inference and bring damage to the medical system (Rahman et al, 2021). This phenomenon is widely seen in IoT devices using ml in various industries.

On the other hand, Researchers are working on different strategies to defend against AE attacks. Abhijit Singh and Biplab Sikdar found that limiting the understanding of attackers about the deep model can reduce the strength of AE attacks. However, even if the attacks have little knowledge of the deep model, AE attacks also increase the error rate by 100% to 200% (Singh et al, 2022). Some researchers also suggest that AE attacks can also be reduced by removing suspicious data that degrades model performance from the dataset during training (Aloraini et al, 2022) while others using adversarial

---

<sup>a</sup> <https://orcid.org/0009-0003-9726-5223>

retraining to make the model learn a more robust decision boundary (Rashid et al., 2022).

Based on the above background and challenges, this paper aims to analyse the threat of adversarial sample attacks faced by deep learning models in IoT devices and explore the current main defence strategies and potential improvement directions. First, the second part of the article will introduce the typical attack methods of AE in the IoT environment. The third part summarizes and compares the common defence mechanisms and their practical effects in existing research. The fourth part will propose some directions for improvement. The fifth part is the conclusion and outlook.

## 2 PROBLEMS CAUSED BY AE ATTACKS ON IOT DEVICES

Adversarial example attacks are a significant threat to IoT devices that rely on deep learning for important tasks. One problem caused by AE attacks is misclassification. AE attacks can cause IoT devices to generate incorrect predictions or detections. For example, adversarial examples input into an intrusion detection system may make the system treat that input as benign traffic. As a result, some malicious activities are not going to be detected by that system. Another example is a pedestrian detection attack. In one experiment, researchers extended adversarial examples from traffic to an object detector deployed with the YOLOv2 model by placing a 40×40 adversarial patch on a person, then the detector is prevented from detecting any people under that area (Ren et al., 2021). Such misclassification may lead to economic damage or even threaten personal safety.

On the other hand, attackers may also contaminate training data using adversarial examples to degrade model performance. This kind of data poisoning will cause large systemic failure if the ML model is used widely on many IoT endpoints. In networks like hospital systems, one wrong node can produce and spread error information and undermine the trustworthiness of the entire network. For instance, tampering with a patient's vital signs through a single node can lead to an incorrect diagnosis and result in wrong treatment. In summary, AE attacks cause critical damage to IoT ecosystems, these attacks disrupt the economy and society. Thus, developing strong defence strategies to reduce these risks becomes popular among researchers. The article will introduce and analyse parts of the current results.

## 3 CURRENT DEFENCE METHODS

### 3.1 Adversarial training

One popular method against AE attacks is adversarial training. This method intentionally adds adversarial data into the model's training cycle. In this case, the model is exposed to the worst-case scenario during training, allowing the model to learn a robust decision boundary. In a recent work focusing on smart city IoT services, Rashid et al. demonstrated how adversarial retraining can significantly improve the stability of the deep learning models. When facing AE attacks, the classification accuracy of the DNN (Deep Neural Network) model can be increased to over 99% (Rashid et al., 2022). Adversarial training has several advantages. First, adversarial training can be adapted to existing model architectures with little structural changes. In this case, researchers can combine regular training and adversarial training without changing the entire pipeline. Also because of this property, when new adversarial strategies are invented, it is relatively easier for security teams to retrain a model. In addition, training models with adversarial data improves models' performances even when there are no AE attacks. In IoT systems, models trained on adversarial examples often become more tolerant to general interferences. For example, models' performances will not be influenced by the data collected in noise-rich scenarios. However, adversarial training brings high computational costs since it requires generating adversarial examples, which increases training time and computational overhead. A standard Wide ResNet-28-10 model on CIFAR-10 takes about 1 hour and 30 minutes to finish adversarial training, while increasing the model's capacity to Wide ResNet-70-16, the training time reached peak robustness after four hours (Gowal et al., 2020). This example shows that models with large capacities require large computational resources. Although adversarial training enhances the robustness of the model, especially against AE attacks, some resources-limited systems may struggle to handle iterative attacks since users may not sacrifice performance for security in this case.

### 3.2 Gradient Masking

Another strategy to defend against AE attacks is gradient masking. Gradient masking refers to any modeling or training tactic that obscures or distorts the gradients that an attacker might exploit when

generating adversarial examples. A concrete example can be found in a demonstration by Goodfellow. Goodfellow introduces a “staircase function” in the model architecture that creates near-zero gradients almost everywhere. This artificially causes adversaries to overestimate how large a perturbation must be to fool the model. Consequently, gradient-based attack methods fail because they rely on meaningful gradient signals to craft small, precise adversarial examples (Goodfellow, 2018). In adversarial machine learning, most attacks rely on approximating gradients to identify small but potent perturbations that fool a model. By masking or disrupting this gradient information intentionally, defenders improve the models’ robustness because standard attack algorithms fail to generate effective adversarial samples. The advantage of gradient masking is fast and efficient. It can quickly get positive results against weak attacks, particularly white-box attacks. Like adversarial training, gradient masking can also be applied to existing models without major architectural changes. However, the drawback of this approach is that it only masks the gradient and cannot actually enhance the decision boundary of the model. In the experiments on CIFAR-10, the authors compare a model trained with Gaussian noise and label smoothing (LS0.5) against a genuinely adversarial trained model (PGD). Although the LS0.5 model shows higher accuracy under a standard white box test, the article explains that LS0.5 relies on distorted gradients rather than true robustness. When a black-box attack is transferred from a similar but more robust model, LS0.5’s accuracy drops significantly (Lee et al, 2020). In conclusion, gradient masking can only deal with white-box attacks, while any stronger attacks typically bypass that defence. In this case, gradient masking only brings a false sense of security.

### 3.3 Monitoring and anomaly detection

Last method the article is going to introduce is monitoring and anomaly detection. Monitoring and anomaly detection involves systematically observing a system’s behaviour, such as network traffic, sensor data, or application logs. For instance, an anomaly detection model first gathers and stores some key metrics, when suspicious input appears, the model can identify them. One concrete example from a recent survey is intrusion detection in a network. In the survey, a model based on deep learning constantly monitors network traffic to detect suspicious activity that deviates from normal communication patterns. This model is often trained with larger benign

network data and alarms if it identifies unusual packet flows (Bulusu et al, 2020). By adding anomaly detection into detection systems, organizations can prevent cyber intrusions early and reduce the damage. A more practical example is the MedMon framework. MedMon snoops on all wireless communications in a patient’s personal healthcare system (such as an insulin pump or continuous glucose monitor) and analyses these signals for anomalies (e.g., unexpected signal strength or timing). if a suspicious transmission is detected, MedMon can alert the user or actively jam the malicious signal to prevent harmful commands from reaching the device (Zhang et al, 2013). In addition, this approach requires no modifications to existing medical devices, making it highly adaptable for resource-constrained IoT environments. Another leading advantage of anomaly detection over the above two methods is that it enables proactive defence. This strategy also allows continuous improvement since feedback loops can refine detection thresholds and reduce false alarms over time. The limitations of anomaly detection are similar to adversarial training, they both require extra computation resources. In addition, an over sensitive models may produce many false alerts, while missing true threats.

## 4 ANALYSIS

Based on the strengths and limitations of current defences, this article suggests several ways which may deserve future study to reduce AE attacks in IoT ecosystems more effectively. To begin with, hybrid defence architecture which combines multiple defensive techniques directly increases the complexity and cost of attacks. For instance, combining adversarial training and anomaly detection protects IoT devices from both the process of getting data and processing data. Secondly, members in IoT ecosystems need to collaborate and share data. By creating secure frameworks from shared information, models can be trained with more comprehensive adversarial examples. As a result, the overall ecosystem is strengthened. When new AE attacks appear, the whole IoT ecosystem can detect and fix them quickly. For some important applications like medical IoT, formal proofs of robustness are needed. For IoT devices in such areas, rigorous bounds and certification methods can reduce the risks and consequences of AE attacks. Overall, any single improvement may not be enough to completely protect the IoT ecosystem. More research and

experiments are required to finally solve the problems caused by AE attacks.

## 5 CONCLUSIONS

This paper examines the security challenges posed by AE attacks on IoT devices using deep learning models. It analyses the attack mechanisms, their impact on IoT ecosystems, and evaluates current defence strategies. The results show that AE attacks can cause damage to IoT devices, leading to misclassification of security systems, data poisoning in certain applications, and instability of IoT networks. Adversarial training improves model robustness but is computationally expensive. Gradient masking provides a fast defence but fails to protect against more advanced attacks. Anomaly detection provides active monitoring but requires large amounts of data and can generate false positives. No single defence mechanism is fully effective against all types of AE attacks. To enhance IoT security, future research should focus on hybrid defence frameworks that combine multiple strategies to improve resilience. Collaborative security efforts, such as sharing adversarial datasets and strong authentication methods can strengthen overall protection. Additionally, formal verification techniques can help ensure the reliability of AI models in critical applications. As the continuing growth of the IoT devices industry, advancement in security research is essential to protect the whole system against AE attacks.

## REFERENCES

- Aloraini, F., Javed, A., Rana, O., Burnap, P., 2022. Adversarial machine learning in IoT from an insider point of view. In *Journal of Information Security and Applications*, 70, 103341. <https://doi.org/10.1016/j.jisa.2022.103341>
- Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K., Song, D., 2020. Anomalous example detection in deep learning: A survey. In *IEEE Access*, 8, 132330-132347. doi: 10.1109/ACCESS.2020.3010274
- Goodfellow, I., 2018. Gradient masking causes CLEVER to overestimate adversarial perturbation size. In *arXiv*, 21 Apr. [arxiv.org/abs/1804.07870](https://arxiv.org/abs/1804.07870)
- Gowal, S. et al., 2020. Uncovering the limits of adversarial training against norm-bounded adversarial examples. In *arXiv*, 30 Mar. [arxiv.org/abs/2010.03593](https://arxiv.org/abs/2010.03593)
- Lee, H., Bae, H., Yoon, S., 2020. Gradient masking of label smoothing in adversarial robustness. In *IEEE Access*, 9. <https://ieeexplore.ieee.org/abstract/document/9311250>
- Rahman, A., Hossain, M. S., Alrajeh, N. A., Alsolami, F., 2021. Adversarial examples—Security threats to COVID-19 deep learning systems in medical IoT devices. In *IEEE Internet of Things Journal*, 8(12), 9603–9610. IEEE Xplore. doi: 10.1109/JIOT.2020.3013710
- Rashid, M. M., Kamruzzaman, J., Hassan, M. M., Imam, T., Wibowo, S., Gordon, S., Fortino, G., 2022. Adversarial training for deep learning-based cyberattack detection in IoT-based smart city applications. In *Computers & Security*, 120, 102783. <https://doi.org/10.1016/j.cose.2022.102783>
- Ren, H., Huang, T., Yan, H., 2021. Adversarial example attacks on object detection in deep neural networks: A survey. In *International Journal of Machine Learning and Cybernetics*. <https://doi.org/10.1007/s13042-020-01242-z>
- Singh, A., Sikdar, B., 2022. Adversarial attack and defence strategies for deep-learning-based IoT device classification techniques. In *IEEE Internet of Things Journal*, 9(4), 2602–2613. doi: 10.1109/JIOT.2021.3138541
- Zhang, M., Raghunathan, A., Jha, N. K., 2013. MedMon: Securing medical devices through wireless monitoring and anomaly detection. In *IEEE Transactions on Biomedical Circuits and Systems*, 7(6), 871-881. doi: 10.1109/TBCAS.2013.2245664