CSEC 520/620 Cyber Analytics and Machine Learning

A SURVEY OF ADVERSARIAL ATTACKS AND DEFENSES IN CYBERSECURITY

October 30, 2023

Mehul Sen
Department of Cybersecurity
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
mehulsen@mail.rit.edu

1 Abstract

Adversarial attacks on Artificial Intelligence(AI) systems are designed to misclassify an input by changing or perturbing it using several strategies. With the rise in popularity of AI, it is being extensively used in the cybersecurity domain to accomplish many tasks. Incidentally, this has also led to an increase in adversarial attacks against these systems as well as the defenses against these attacks.

After conducting a thorough analysis of the latest research papers on adversarial attacks against AI systems in the cybersecurity domain, we came across a few interesting insights. Our search was broadly categorized into four categories, namely malware detection and classification, URL detection, network intrusion detection, and biometric authentication, and we found a few noteworthy takeaways. These included focusing on problem-space vs feature-space of attacks, the different attack techniques employed, the transferability property, and retraining on adversarial examples as a defense mechanism.

However, we also observed a need for greater emphasis on pragmatic, real-world attacks and defenses beyond adversarial training. It is important to explore additional defensive techniques against adversarial attacks on cybersecurity systems that leverage AI to improve security against transferable attacks.

2 Introduction

AI systems are increasingly being incorporated into our daily lives, as well as in the field of cybersecurity. From face recognition to intrusion detection, AI solves numerous problems. However, one of the biggest threats to these systems is adversarial evasions. These attacks involve slight input alterations like malware code or network traffic to deceive the AI model into misclassifying the input. Although these attacks can be executed against general machine learning technologies, they may not have as significant of an impact as they would on an AI designed to solve problems in the cybersecurity domain. This is because, unlike defenders who need to be secure all the time, an attacker only requires one successful attack to inflict a devastating impact on the victim's infrastructure. As a result, it is imperative to research adversarial attacks against AI in the field of cybersecurity to enhance our understanding of the attack vectors and improve security measures.

Adversarial attacks are particularly relevant in cybersecurity as they have clear, targeted goals [1]. They pose a significant challenge for both attackers and defenders because they need to keep the malicious functionality intact and even small changes in a perturbed sample to misclassify the model can significantly impact the perceptibility of that sample.

This project examines the latest research papers on adversarial attacks and defenses in cybersecurity, focusing on malware detection and classification, URL detection, network intrusion detection, and biometric authentication. In addition, we provide an analysis of these papers and identify the key trends and future work that can be done in this domain.

3 Background

In 2014, Szegedy et al. [2] proposed adversarial examples as a minimization problem

proposing the concept of an adversarial attack against an AI system. Adversarial attacks can be classified into three types, depending on how they perturb a sample to cause misclassification:

- **Gradient Based Attacks:** These attacks produce adversarial perturbations in the direction of the gradient. They are effective but require adversarial knowledge about the targeted model and its gradient, as well as the model's architecture. Therefore, these are a type of white-box attack. Examples include the Fast Gradient Sign Method [3], Carlini-Wagner Attack [4], and the Projected Gradient Descent [5]

- **Score Based Attacks:** These attacks produce adversarial perturbations based on the knowledge of the victim model's confidence score. They do not require direct knowledge of the model's gradient or architecture, but they do require the confidence scores for each of their perturbed samples. Therefore, these are a type of gray-box attack. An example of this is the Zeroth Order Optimization Attack [6]

- **Decision-Based Attacks:** These attacks exploit the transferability property [2], such that adversarial examples created against one model are likely to be effective against others. They utilize labels predicted by the target model to misclassify a sample. These are a type of black-box attack since they do not require any internal information about the target models. Examples include the Generative Adversarial Network [7], and the Boundary Attack [8]

Each type of attack also has query-efficient variants, where the attacker has a limited number of queries to identify the models and carry out the attack. The goal is to minimize the number of questions to avoid suspicions or stay under a set security threshold. Although white-box attacks are more efficient than black-box attacks [9], they are more realistic and exhibit real-world scenarios, making them more likely to be seen in the wild.

In their survey, Rosenberg et al. [1] designed a taxonomy for adversarial attacks in the cybersecurity domain. They categorized attacks into four stages: threat model which encompasses the attacker's knowledge and capabilities before an attack; attack type which refers to the goals, targeted phase and characteristics; perturbed features which are the features modified during an attack; and the attack output which is the final result of an attack. Each adversarial attack, regardless of the type, consists of all these stages. Adversarial defenses are created based on such a taxonomy and can be classified into two types based on their approach to increasing the security of a model.

- **Detection Based Defenses:** These are focused on detecting adversarial examples on the target model. They are trained to identify any anomalous behaviors or interactions with the model.

- **Robustness Based Defenses:** These are focused on increasing the robustness of the model such that it becomes exceedingly difficult for the attacker to develop effective and practical adversarial examples.

4 Literature Review

For the survey, we reviewed several research papers addressing adversarial attacks and defenses in the cybersecurity domain. As a basis, we considered the categories defined by Rosenberg et al. [1] in 2021 and reviewed newer papers in each of these fields. The four types and details about the novel research done in them are presented in the following subsections.

4.1 Malware Detection and Classification

Malware detection and classification is a crucial aspect of protecting computer systems from adversarial attacks. Modern defenses against malware use machine and deep learning models to detect anomalous activities and unsigned malware. Some well-known next-generation antivirus software that uses these models include SentinelOne, Microsoft ATP, and Crowdstrike. As a result, researchers are actively studying both adversarial attacks and defenses to misclassify malware, as well as to detect adversarial attacks.

Gaspari et al. [10] researched Adversarial attacks for ransomware. They argued that current malware detectors rely on behavioral analysis techniques prone to evasion attacks. These detectors search for specific behavioral features such as changes in file entropy, writes that cover extended parts of a file, file deletion, processes corresponding to many user files, processes writing to files of different types, and back-to-back writes. Although such features can detect basic ransomware, it is possible to create ransomware that sidesteps these significant behavioral features, rendering the detectors ineffective. To demonstrate this, they proposed three novel attacks: process-splitting, which distributes ransomware operations evenly across multiple processes such that each only exhibits a subset of behaviors; functional-splitting, which separates ransomware operations into groups by functions such that each process only performs one function, and Mimicry which models the ransomware features on benign processes such that each process is indistinguishable from the benign process. They also developed a proof of concept ransomware called "Cerberus," which implemented the proposed attacks. They tested their attacks against detectors such as ShieldFS and RWGuard, as well as a black-box attack on the Malwarebytes Anti-Ransomware detector, showing the feasibility of their attacks. The Mimicry attack, in particular, could evade detection entirely in a black-box setting, while functional splitting and process-splitting required more processes for complete evasion. Lastly, they trained a detector to recognize functional splitting attacks, which could identify such ransomware with high accuracy.

Berger et al. [11] focused their work on Android malware. They analyzed the differences and gaps between feature-space attacks and problem-space attacks. Feature-space attacks manipulate machine learning features to cause misclassification while minimizing the number of perturbations, whereas problem-space attacks change the malware code in a realistic way to cause misclassification. They argued that current machine learning models are vulnerable to evasion attacks, and the state-of-the-art defenses rely on adversarial training using feature-space attacks, which do not reflect actual malware samples. To evaluate the robustness of each,

they retrained classifiers on both feature-space and problem-space attacks. Their studies focused on Android OS, using a dataset of 75,000 benign Android applications from AndroZoo and 5700 malicious applications from Drebin. They tested these on three malware detection systems, namely Drebin trained with an SVM classifier, Drebin-DNN trained with a deep neural network classifier, and MaMaDroid trained with RF, KNN, and DT classifiers. They found that feature-space attacks do not serve as reasonable proxies for problem-space malware evasion attacks, and robustness should be evaluated directly against problem-space attacks.

Rashid and Such [12] focused their research on adversarial query attacks in malware detection. They argued that machine learning models are vulnerable to adversarial query attacks, where the attacker can iteratively query the model inputs to cause misclassifications. Suppose an attack focuses on the feature space, modifying the discrete binary feature vectors. In that case, the defenses against such attacks, which include similarity detection, are ineffective when the adversarial examples are generated differently. They propose a new stateful defense against adversarial query attacks called "MalProtect," which analyzes the sequence of queries using multiple threat indicators. These indicators not only assess query similarity but also any features that are shared across queries, the number of enabled features, and other variables. A score for each variable is calculated based on whether it might be an attack, which a decision model then aggregates to predict if an attack is occurring. Only if the prediction is that an attack is not occurring does the query reach the prediction model to be processed further. This allows the defenders to modify the solution and provide the attacker with incorrect data, sabotaging them before they can cause misclassification. They evaluated their solution using Android malware from AndroZoo and Drebin and Windows malware from SLEIPNIR, testing several stateful and non-stateful defenses. Their solution reduced evasion rates of black-box and gray-box attacks by 80%-98% across the Android and Windows datasets, outperforming other defenses. Thus, by employing multiple threat indicators and analysis techniques besides similarity/outlier detection, an effective stateful defense for adversarial query attacks can be significantly improved.

4.2 URL Detection

The Internet is a vast network consisting of billions of web pages. These web pages are accessed using Uniform Resource Locators (URLs). While most URLs are legitimate and lead users to genuine web pages, some are malicious. These malicious URLs are used by large-scale botnets to coordinate or by attackers to conduct phishing campaigns. Domain Generation Algorithms (DGAs) are used to create many domain names that are difficult to predict. Bots throughout the network try to communicate with these domains iteratively to find the actual Command and Control (C&C) server. This strategy is effective since defenders need help finding and taking down malicious URLs faster than new ones can be deployed. Researchers are working on ways to predict which URLs might be malicious, generated by DGA, or part of a giant botnet.

Casino et al. [13] focused their research on detecting algorithmically generated domains (AGDs). They highlighted the need to de-

velop accurate methods for botnet detection so that take-down operations of these botnets can be sped up and large-scale malware campaigns can be thwarted. The authors designed a dataset called "HYDRAS," consisting of 105 of the latest and most popular DGA families spanning over 95 million domains. They proposed a novel feature set, which included lexical and statistical features over the collected DGAs, as well as English gibberish detectors. They tested their dataset using a Random Forest Classifier and their proposed features. They achieved a very high accuracy (F1 Score over 99%) when identifying between benign and malicious domains generated through DGA, outperforming the state-of-the-art classifiers. They argue that using a comprehensive dataset that accurately represents malicious domains, malicious botnet domains can be efficiently detected in real time, allowing for faster take-downs.

Suryotrisongko et al. [14] also focused their research on detecting malware that utilized DGAs. They argued that traditional threat intelligence approaches like blocklists are ineffective against DGAs, and the current cyber threat intelligence sharing platforms need to support sharing classifier models amongst organizations. They proposed a model to detect DGA-based malicious domains using seven statistical features. They then trained a random forest classifier and evaluated their models on 55 DGA families, comparing their performance with other state-of-the-art detectors such as CharBot. Additionally, they also attempted to improve the trust in model sharing by proposing a blend of explainable AI (XAI) techniques such as SHAP, Lime, and Anchors with open-source intelligence (OSINT) methods such as Google Safe Browsing, OTX AlienVault to validate the model's predictions. Their model achieved an accuracy of 96.3% on detecting DGAs, outperforming other state-of-art detectors. They also proposed a computable CTI paradigm that allowed for the sharing of models between organizations validated through XAI and OSINT, improving automation and reducing manual analysis, enabling several organizations to equip themselves with the most accurate detectors. They were able to develop a better way to identify botnets using a more precise detector and show how XAI and OSINT could be used to improve trust in sharing models for cyber threat intelligence.

Apruzzese et al. [15] researched phishing website detectors and approached adversarial attacks from a different perspective. They argued that adversarial machine learning and its defenses tend to focus on unrealistic threat models, crafting adversarial examples through the feature space. While these attacks are possible, they are not entirely physically realizable in the problem space. The authors researched low-cost, pragmatic attacks that are more likely to be used by attackers. They focused their work on detecting phishing websites by proposing a realistic threat model for low-cost website-space perturbations that a typical phisher may use. Additionally, they also defined "evasion space" by dissecting the architecture of phishing website detectors, which is categorized as website space where the website gets generated by attackers; the preprocessing space, which involves feature extraction; the machine-learning space which analyzes the features; and the output-space which contains the output for the classifier's decision. They evaluated the robustness of 18 machine learning-based phishing website de-

tectors against 12 attacks that they labeled realistic with varying costs. Additionally, they used webpage datasets from Zenodo and Î´Phish. They found that some detectors are resilient to cheaper attacks; however, slightly more complex methods can outperform a more significant number of detectors. Additionally, the greatest threat to detectors is the cheap website attacks which induce small but significant degradations in most detectors and can be quickly developed due to their low cost. They highlighted the need to focus on more likely real-world attacks and provided benchmark results demonstrating the impact of realistic attacks on machine learning-based phishing website detectors.

4.3 Network Intrusion Detection Systems

Network Intrusion Detection Systems (NIDS) are an essential component of network security, used to identify any potential attacks within a network. These systems monitor all traffic passing through a defined strategic point and alert administrators if they encounter any activity that is classified as malicious. While earlier versions of NIDS relied on rules to identify malicious traffic, newer versions have been equipped with AI-powered models trained to identify a wide range of network attacks and malicious traffic.

Recent research has focused on both attacking and defending these AI-powered models. Mohammadian et al. [16] developed a novel adversarial attack against deep learning-based NIDS. They proposed a white-box attack model that manipulates the feature space of a trained deep neural network. Their approach involves creating feature combinations to identify the best combination of features to perform the attack. They then use a saliency map of the trained model to rank the combinations and select the best ones for the attack. Finally, they generate adversarial samples based on these features to misclassify the model. Their testing revealed that their method effectively created adversarial samples for over 18% of samples in CIC-IDS2017, 15% of samples in CIC-IDS2018, and 14% of samples in CIC-DDoS2019. They discovered that increasing the number of features and the magnitude of perturbations improved the effectiveness of the attack.

Sharon et al. [17] also focused on developing adversarial attacks against NIDS. Their approach is different from Mohammadian et al. as they proposed a black-box attack model that focuses on the problem space. They designed a timing-based adversarial network traffic reshaping attack called "TANTRA" that uses a Long Short-Term Memory(LSTM) model. The model is trained to learn benign network traffic behavior using a short history of benign packets. It then reshapes the attack's malicious traffic, including any interpacket delay, to make it similar to the benign traffic. The modified malicious traffic is then sent to the target network, bypassing the NIDS. Their proposed attack does not require knowledge or access to the NIDS and its classifier, only an ongoing connection to the target network. To evaluate their attack, they used eight common network intrusion attacks from Kitsune and CIC-IDS2017 datasets; they tested their attacks on three state-of-the-art NIDS systems that used deep learning (Autoencoder, KitNET, and Isolation Forest). They were able to achieve a 99.99% success on average in evading detection across all their attacks.

They also discovered that changing only the timestamps and not the packet content was sufficient to evade detection. Lastly, they proposed a defense that involved training NIDS on the reshaped traffic as a viable solution against their attack.

Kotak and Elovici [18] focused their research on identifying IoT devices on the network, which could increase security risks due to unpatched vulnerabilities. They proposed a novel black-box template-based approach that uses heat maps, such as Class activation mapping(CAM) and Grad-CAM++, to identify the essential features in the traffic packets for classifying the traffic as benign or malicious. They then use the heat map to craft adversarial examples by replacing less critical features from a benign device that is classified as authorized to be on the system with a malicious IoT device that mimics the authorized traffic. They tested their attack using the public IoT network traffic dataset (IoT Trace), additional IoT traffic generated in the lab, and six variants of payload-based Fully connected network models(FCN) for IoT device identification. They also used four convolutional neural networks(CNN) and FCN as surrogate models to craft their adversarial examples. Their testing revealed that their heat map adversarial attacks fooled target models with up to 100% success. Their attack was effective against all model variants, showing the efficacy of transferability within their attack. They highlighted vulnerabilities in ML-based IoT identification by demonstrating an evasive adversarial attack.

4.4 Biometric Systems

Biometric authentication is a critical field in cybersecurity that uses classifiers and detectors to authenticate the identity of authorized personnel. These models are trained to detect and identify individual voices, faces, and fingerprints and predict when a provided sample matches any samples it was previously trained on. However, adversarial attacks against image and face recognition were some of the first in the domain and have been extensively researched. Researchers are now focusing on speech and voice recognition and how these can be manipulated through adversarial attacks.

Abdullah et al. [19] have focused their research on adversarial attacks against Automatic Speech Recognition (ASR) and Automatic Voice Identification (AVI) systems. They argue that the current implementations of these systems are vulnerable to adversarial attacks and that all prior attacks required whitebox knowledge of the model, a large number of queries, or generated poor-quality audio that was perceptible by humans. The current ASR and AVI systems rely heavily on components of speech that are not necessary for human comprehension. As a solution, they proposed a query-efficient black-box attack that requires no knowledge about the model and its features, is transferrable, and operates in near real-time. They do so by removing low-intensity components from the benign audio sample using signal processing techniques like Discrete Fourier Transforms (DFT), which exposes audio frequency information, and Singular Spectrum Analysis (SSA), which decomposes an arbitrary time series into components called eigenvectors. Based on these, they optimize the distortion threshold using binary search to minimize the impact on audio quality, causing ASR systems to mistranscribe the audio and AVI systems to misidentify speakers by per-

turbing phonemes like vowels while staying comprehensible by humans. Their attack involved two subcategories: the word level attack, which perturbed full words to induce misclassification, and the phoneme level attack, which perturbed only small parts of the audio signal in targeted ways. They tested their attacks on Google Speech API, Facebook Wit.ai, DeepSpeech, CMU Sphinx, and Microsoft Azure Speaker Identification with datasets such as the TIMIT corpus of phonetically diverse English sentences spoken by 630 speakers and a dataset of 1000 most common English words. The word-level attack causes 50% mistranscription with minimal distortion. In contrast, the phoneme-level attack was most effective when perturbing vowels, with ASR mistranscribing them 80% compared to other phonemes, which were mistranscribed less than 60%. They also showed that their attacks were transferable between models with up to 100% success and were able to evade existing detection methods. They also proposed a defense for their attack that involved adversarial training using the perturbed samples.

Xue et al. [20] also focused their research on adversarial attacks against intelligent audio systems. They argue that existing adversarial attacks against these systems are not imperceptible and can easily identified as potential attacks on humans or require specific conditions like background noise. These attacks are less stealthy or require white-box access to the model to generate stealthy evasion attacks. The authors proposed a targeted imperceptible black-box attack against intelligent audio systems called "Echo," which integrates adversarial noise into the natural reverberations of audio samples to misclassify the perturbed samples as belonging to a specific class. Their goal differs from Abdullah et al. as not only do they want the audio samples to be misidentified, but they want them to be misidentified as a particular target. Their attack consists of four modules: the inconspicuous reverberation fusion module, which merges adversarial perturbations into the physical reverberation features of voices; the fast adversarial perturbation generation module, which efficiently attacks the target system; the targeted black-box attack module, which uses a novel method of searching gradient information of the targeted system using neuro-evolution, which they call "Newav," and the robust in over-the-air attack module, which incorporates optimization processes to improve the robustness of Echo under the over-the-air conditions. To evaluate their solution, they tested their attack on the state-of-the-art X-Vector model trained on Voxceleb2, Mini Librispeech, and VCTK datasets for speaker recognition. They also tried their attack on VGG_BN19 and Resnet34 on Google Speech Commands V2 for speech command recognition. Echo achieved an average success rate of 98.86% on speaker recognition and 99.24% on speech command attacks. Their adversarial examples were generated in less than 0.4ms and were robust up to a distance of up to 4m, making them suitable for real-time attacks. They also tested current defenses against attacks such as filtering and compression, which were ineffective against Echo. Lastly, they tested a defense against their attack that would involve preprocessing the audio to remove the adversarial noise, but this was ineffective against Echo. They mentioned that a possible defense against their attack could be through denial of service by inferring the query behaviour of their attack

and stopping interactions with the attacker.

5 Conclusions

Based on the reviewed papers, we can see several key trends and takeaways emerging from the research papers in this field.

- **Problem-space vs. Feature-space Attacks:** Several papers focused on problem-space and feature-space attacks. There is a clear distinction between problem-space and feature-space attacks. While feature-space attacks are easier to generate and may result in a higher evasion rate, problem-space attacks are much more realistic. Defenses that are built on feature-space attacks may not be able to translate well to defending against attacks on the problem space.

- **Attack Techniques:** Adversarial attacks fool the classifiers using many techniques. The papers reviewed in this section have used methods like process splitting, functional splitting, mimicry, timing perturbations, distorting reverberations, etc. These techniques allow attackers to hide malicious behaviors, imitate benign patterns, and modify audio subtly. Therefore, adversarial attacks showcase a large diversity in attack vectors.

- **Transferability:** Transferability is a critical property of adversarial attacks. Several attacks proposed in the literature exploit this property, where attacks can be transferred between different architectures and models while remaining practical. This can allow attackers to develop black-box attacks by using surrogate white-box models. Defenses against such attacks need to consider these scenarios to increase their robustness.

- **Retraining as a Defense:** Many papers proposed retraining on adversarial examples as a common defense strategy against adversarial attacks. Although this is a valid defense, real-world attacks are known to employ a large diversity of attack vectors, and adversarial training alone might be insufficient in defending against these attacks.

In conclusion, future work in this domain should focus on developing alternative defenses that supplement retraining on adversarial examples, developing stateful, multi-indicator defenses that can analyze and identify full attack patterns, as well as expanding on the use of explainable AI and OSINT to improve model security and enable model sharing so defenses can quickly use the latest and most comprehensively trained models for their security.

References

[1] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," *ACM Computing Surveys*, 2021.

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *Computing Research Repository*, 2014.

[4] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," *IEEE Symposium on Security and Privacy*, 2016.

[5] O. Bryniarski, N. Hingun, P. Pachuca, V. Wang, and N. Carlini, "Evading adversarial example detection defenses with orthogonal projected gradient descent," *ArXiv preprint*, 2021.

[6] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," *ACM Workshop on Artificial Intelligence and Security*, 2017.

[7] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. X. Song, "Generating adversarial examples with adversarial networks," *ArXiv preprint*, 2018.

[8] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," *ArXiv preprint*, 2017.

[9] I. Rosenberg and E. Gudes, "Bypassing system callsâbased intrusion detection systems," *Concurrency and Computation: Practice and Experience*, 2017.

[10] F. D. Gaspari, D. Hitaj, G. Pagnotta, L. D. Carli, and L. V. Mancini, "Evading behavioral classifiers: a comprehensive analysis on evading ransomware detection techniques," *Neural Computing and Applications*, 2022.

[11] H. Berger, A. Dvir, C. Hajaj, and R. Ronen, "Do you think you can hold me? the real challenge of problem-space evasion attacks," *ACM Transactions on Privacy and Security*, 2022.

[12] A. Rashid and J. M. Such, "Malprotect: Stateful defense against adversarial query attacks in ml-based malware detection," *IEEE Transactions on Information Forensics and Security*, 2023.

[13] F. Casino, N. Lykousas, I. Homoliak, C. Patsakis, and J. C. H. Castro, "Intercepting hail hydra: Real-time detection of algorithmically generated domains," *Journal of Network and Computer Applications*, 2021.

[14] H. Suryotrisongko, Y. Musashi, A. Tsuneda, and K. Sugitani, "Robust botnet dga detection: Blending xai and osint for cyber threat intelligence sharing," *IEEE Access*, 2022.

[15] G. Apruzzese, M. Conti, and Y. Yuan, "Spacephish: The evasion-space of adversarial attacks against phishing website detectors using machine learning," *Annual Computer Security Applications Conference*, 2022.

[16] H. Mohammadian, A. A. Ghorbani, and A. H. Lashkari, "A gradient-based approach for adversarial attack on deep learning-based network intrusion detection systems," *Applied Soft Computing*, 2023.

[17] Y. Sharon, D. Berend, Y. Liu, A. Shabtai, and Y. Elovici, "Tantra: Timing-based adversarial network traffic reshaping attack," *IEEE Transactions on Information Forensics and Security*, 2022.

[18] J. Kotak and Y. Elovici, "Adversarial attacks against iot identification systems," *IEEE Internet of Things Journal*, 2023.

[19] H. Abdullah, M. S. Rahman, W. Garcia, K. Warren, A. S. Yadav, T. Shrimpton, and P. Traynor, "Hear "no evil", see "kenansville"*: Efficient and transferable black-box attacks on speech recognition and voice identification systems," *IEEE Symposium on Security and Privacy*, 2021.

[20] M. Xue, K. Peng, X. Gong, Q. Zhang, Y. Chen, and R. Li, "Echo: Reverberation-based fast black-box adversarial attacks on intelligent audio systems," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2023.