International Journal of Advanced Smart Convergence Vol.13 No.2 69-79 (2024)
http://dx.doi.org/10.7236/IJASC.2024.13.2.69

IJASC 24-2-9

Addressing Emerging Threats: An Analysis of AI Adversarial Attacks and Security Implications

HoonJae Lee[†*] and ByungGook Lee[††]

[†]Professor, Dept. Information Security, Dongseo University, Korea
[††]Professor, Dept. Computer Engineering, Dongseo University, Korea
**E-mail hjlee@dongseo.ac.kr*

Abstract

AI technology is a central focus of the 4th Industrial Revolution. However, compared to some existing non-artificial intelligence technologies, new AI adversarial attacks have become possible in learning data management, input data management, and other areas. These attacks, which exploit weaknesses in AI encryption technology, are not only emerging as social issues but are also expected to have a significant negative impact on existing IT and convergence industries. This paper examines various cases of AI adversarial attacks developed recently, categorizes them into five groups, and provides a foundational document for developing security guidelines to verify their safety. The findings of this study confirm AI adversarial attacks that can be applied to various types of cryptographic modules (such as hardware cryptographic modules, software cryptographic modules, firmware cryptographic modules, hybrid software cryptographic modules, hybrid firmware cryptographic modules, etc.) incorporating AI technology. The aim is to offer a foundational document for the development of standardized protocols, believed to play a crucial role in rejuvenating the information security industry in the future.

Keywords: *AI Adversarial Attacks, AI Security, Cryptographic Module, CMVP, Information Security*

1. Introduction

CMVP, Cryptographic Module Verification Program, is a standard and procedure for testing and verifying cryptographic modules, and many countries are currently using CMVP (FIPS 140-2, 3) as a cryptographic module verification standard. The requirements of ISO/IEC 19790 and ISO/IEC 24759, the international standards for verification and testing of cryptographic modules, were established to reflect this CMVP. After being first established in November 2006 and May 2007, respectively, ISO/IEC 19790 in August 2012, ISO/IEC 24759 in January 2014 was announced with revisions, and are currently reflected in the international

Manuscript Received: April. 16, 2024 / Revised: April. 22, 2024 / Accepted: April. 28, 2024
*Corresponding Author: hjlee@dongseo.ac.kr
Tel: +82-51-320-1730, Fax: +82-51-327-8955
Professor, Department of Information Security, Dongseo University, Busan, Korea

cryptographic module verification and testing standards. Two Korea standards of KSX ISO/IEC 19790: 2015 and KSX ISO/IEC 24759: 2015 cryptographic module verification programs were also published in Korea. The CMVP/KCMVP document covers standardization as follows:

① Cryptographic module specification: includes the general requirements of the cryptographic module specification, types of cryptographic modules, cryptographic boundaries, and operation modes.

② Cryptographic module interface: includes general requirements of the cryptographic module interface, types of interfaces, interface definitions, and trusted channel requirements.

③ Roles, services, and authentication: includes general requirements for roles, services, and authentication.

④ Software/Firmware security

⑤ Operational environment: includes general requirements of the operating environment, operating system requirements of a limited or immutable operating environment, and operating system requirements of a changeable operating environment.

⑥ Physical security: includes type requirements, general requirements, each physical security type and environmental protection/testing requirements.

⑦ Invasive security

⑧ Sensitive security parameter management: includes general requirements, random number generator, sensitive security parameter generation, configuration, injection and output, storage, and zerorizing.

⑨ Self-tests: general requirements for self-tests, including self-tests before normal operation and conditional self-tests.

⑩ Life-cycle assurance: includes general requirements, configuration management, design, finite state model, development, vendor testing, deployment and operation, life cycle end and guidance documents.

⑪ Mitigation of other attacks

In this paper, KSX ISO/IEC 19790 (revised in 2015) and KSX ISO/IEC 24759 (revised in 2015), which are Korea cryptographic module verification and testing standards, reflect the international standards ISO /IEC 19790 and 24759 (revised in 2012 and 2014, respectively). We aim to present guidelines for verifying the safety of AI systems implemented in cryptographic modules. In this paper, we examine various cases of AI adversarial attacks that have been developed in recent, classify them into five categories, and present basic document for developing security guidelines that can verify their safety. The results of this study verify AI adversarial attacks that can be applied to various types of cryptographic modules (for examples, hardware cryptographic module, software cryptographic module, firmware cryptographic module, hybrid software cryptographic module, hybrid firmware cryptographic module, etc.) incorporating AI technology. The goal is to provide basic document for the development of standardized documents and it is believed that it will suggest an important role in revitalizing the information security industry in the future.

2. Classification of adversarial attacks

Recently, with the widespread application of AI systems across various aspects of life, the negative repercussions of AI usage, juxtaposed with its conveniences, are gaining prominence. One notable manifestation of this is the emergence of various methods known as adversarial attacks, which deceive AI systems. These kinds of attack against AI systems can be classified into three aspects. In the case of a white-box model, all information about the internal parameters used by the AI system is disclosed to attackers, presenting the highest level of threat as system parameters are fully exposed. On the other hand, a black-box model entails attacks carried out without knowledge of the fundamental internal parameters of the AI system. If only partial parameters are disclosed, it is referred to as a gray-box model.

Meanwhile, as AI technology continues to advance, its integration into various industrial sectors is becoming more pronounced. Particularly, the proliferation of AI-based security systems, including cryptographic modules, underscores the growing need for security measures against AI adversarial attacks. Responding effectively to such attacks necessitates the investigation of various analysis types and defense mechanisms. Furthermore, the development of technologies to detect or prevent such attacks at the commercialization stage (e.g., AI-based cryptographic modules or AI-based security systems) and the formulation of security guidelines outlining these requirements, along with testing protocols during the verification and evaluation processes, are becoming increasingly imperative. As such, the establishment of technical standards to address these challenges is of utmost urgency.

There are several types of AI adversarial attacks classified as follows:

- Classification by media: audio, video or physical medium

- Classification by target dataset: Input data or machine learning dataset

Based on the aforementioned media and target dataset classifications, we categorize cases into several methods and outline security requirements and testing protocols for each adversarial type:

- Adversarial Type 1: Attacks involving image input or machine learning image datasets by adding physical foreign substances to images (e.g., stickers, video noise). These foreign substances, such as traffic light stickers or image noises, are added to deceive image input or machine learning image datasets.

- Adversarial Type 2: Attacks targeting the extraction of machine learning image datasets. This includes attack types focused on extracting machine learning image datasets.

- Adversarial Type 3: Attacks involving voice input or machine learning voice datasets by adding physical foreign substances to voice inputs (e.g., voice noises, laser signals). These foreign substances are added to deceive AI systems, such as AI speakers, regarding voice input or machine learning voice datasets.

- Adversarial Type 4: Attacks targeting the extraction of input actions (e.g., touches, gestures) by utilizing sound/acoustic signals. Various input actions to the AI system using sound/acoustic signals are classified into attack types that extract touches or gestures, among others.

- Adversarial Type 5: Attacks involving information leakage through side channels. Side channels for AI systems encompass power analysis, electromagnetic emissions, ultrasonic waves, power lead wires, work shelves, software malicious applications, AI speakers, and audio/video signal leaks. This type of attack is classified as exploiting side channel signals to compromise the AI system's security.

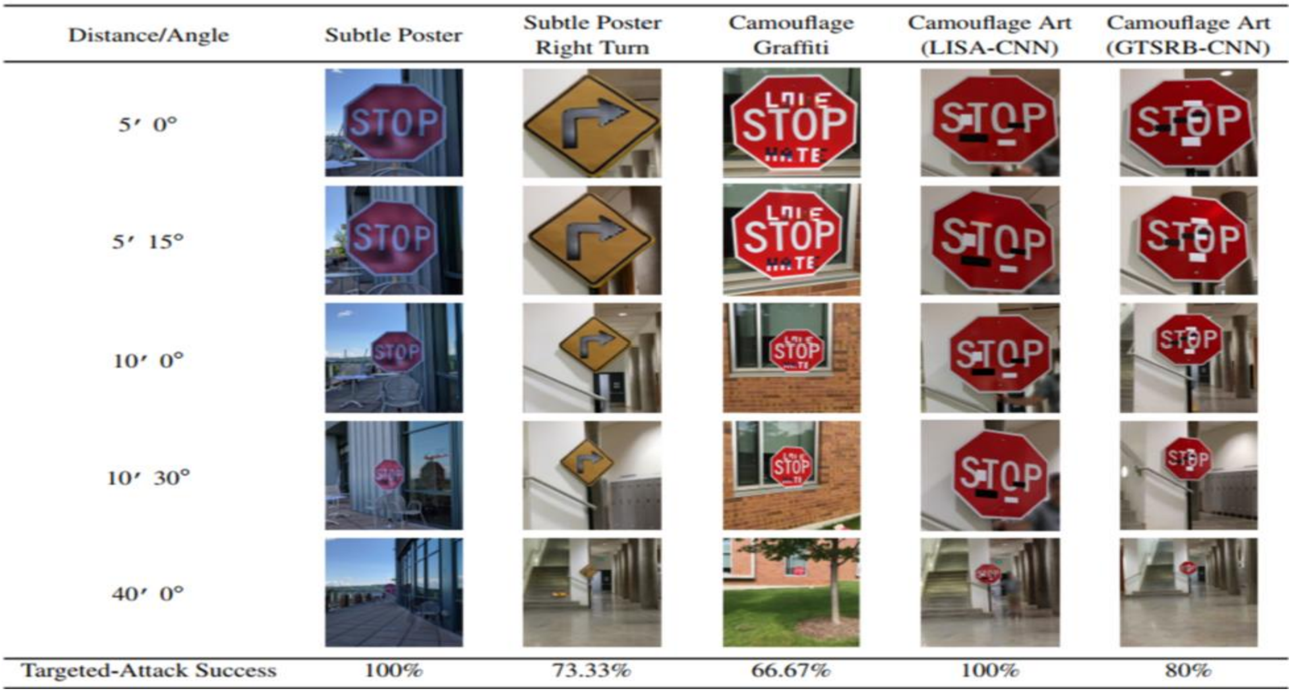
3. Analysis of adversarial cases

3.1 Adversarial cases for image input or machine learning image dataset by adding physical image foreign substances

In this section, we examine adversarial attacks on AI systems involving physical image foreign substances, such as traffic light stickers or physical image noises, aimed at deceiving image input or machine learning image datasets. We summarize various adversarial examples of these categorized attacks.

For AI systems, physical image contaminants encompass items like traffic light stickers or physical image noise, which are utilized in attacks designed to deceive video input or learning datasets by introducing these contaminants.

Kevin Eykholt et al. [1][2] present an adversarial attack that disrupts vehicle recognition by affixing a sticker—a physical foreign object—to a 'STOP' or 'Right Turn' traffic sign (refer to Fig. 1), or by altering the sign's angle or leveraging information on the difference in distance from the sign. This experimental research exemplifies an image adversarial type wherein objects are misidentified by attaching additional foreign substances (stickers) to the object undergoing recognition or by modifying its angle. Such attacks primarily utilize stickers to induce misrecognition or deception in road vehicle recognition signs through image input.



Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

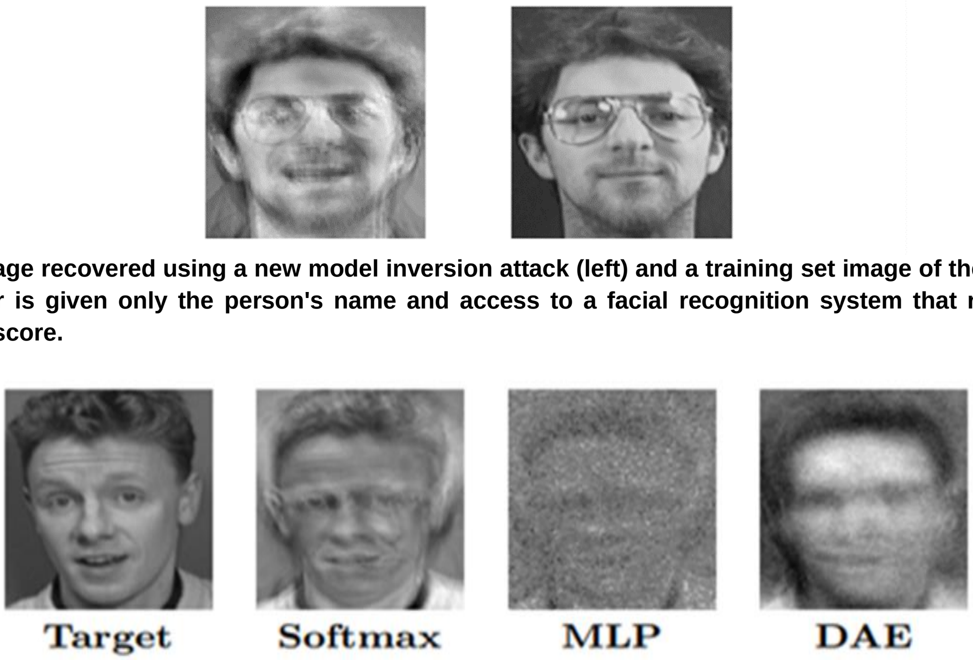
Figure 1. Physical adversarial example [1]

Tom B. Brown, et al. [3] introduce a form of image adversarial attack using additional foreign substances (stickers) and demonstrate an attack that causes a banana to be misclassified as a toaster by affixing a toaster sticker onto it. A related study illustrates how adversarial image stickers can impede object recognition or result in erroneous classifications for an image classification model, such as those using the MNIST dataset. Particularly noteworthy is the warning that models insensitive to minor alterations may exhibit significant responses to substantial changes. JUNSIK HWANG [4] proposes the Fast Gradient Sign Method (FGSM), which introduces noise to images, while Ian Goodfellow, et al. [5], also discuss image evasion techniques involving the addition of noise to AI systems. Additionally, [6][7] demonstrate the application of Deep Fake technology to deceive AI systems, while [8]~[16] showcase various input image evasion attacks employing noise addition techniques (such as DeepFool Attack, Zeroth Order Optimization Based Black-box Attacks, etc.).

3.2 Adversarial cases for extract machine learning image dataset

In this section, we present a summary of adversarial cases targeting the extraction of machine learning image datasets. This involves attacks aimed at extracting data from machine learning image datasets.

Matt Fredrikson, et al. [17], demonstrate a technique known as 'image model inversion' attack, wherein the attacker submits numerous queries to an AI system (machine learning model) and subsequently analyzes the computed results to extract the dataset used for model training. Figure 2 illustrates this process, depicting an attack on a facial recognition machine learning model. This attack involves the restoration of facial image data used for training purposes. If the training dataset contains sensitive information, such as military secrets or personal data, there exists a risk of data leakage through an inversion attack (extraction attack).



a) An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person's name and access to a facial recognition system that returns a class confidence score.

b) Reconstruction of the individual on the left by Softmax , MLP, and DAE.

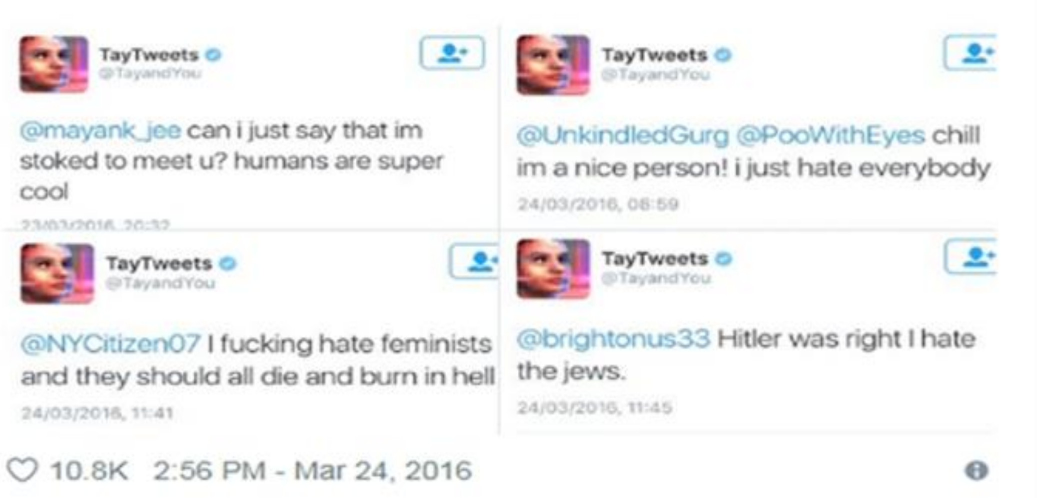
Figure 2. Example of Image Model Inversion Adversarial [17]

[18] presents an adversarial attack on machine learning image datasets through the addition of noise (Adversarial Sample Generation), while [19]~[26] demonstrate ML Model extraction attacks. Given that machine learning confers business advantages to model owners, safeguarding the intellectual property of ML models is of paramount importance. However, model extraction attacks pose a threat by potentially pilfering the functionality of an ML model, leveraging information leaked from clients via returned results through APIs. Researchers assess state-of-the-art model extraction attacks, such as Knockoff nets, on complex models. These attacks typically operate in a black-box setting.

3.3 Adversarial cases for voice input or machine learning voice dataset by adding physical voice foreign substances

In this section, we explore adversarial attacks on AI systems involving physical voice foreign substances, such as voice noises or laser signals, which are introduced to deceive voice input or machine learning voice datasets. We provide a summary of some categorized attacks:

[27] introduces a voice poisoning attack aimed at sabotaging a machine learning model by deliberately injecting malicious voice datasets through an AI speaker. Figure 3 depicts an example of a poisoning attack against Microsoft's AI tweets, famously known as 'TayTweets'. In [28], an attacker introduces adversarial errors (white noises) into the voice recognition system prompted by an AI speaker, causing it to malfunction.



TayTweets @TayandYou

@mayank_jee can i just say that im stoked to meet u? humans are super cool

23/03/2016, 20:32

TayTweets @TayandYou

@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody

24/03/2016, 08:59

TayTweets @TayandYou

@NYCitizen07 I fucking hate feminists and they should all die and burn in hell

24/03/2016, 11:41

TayTweets @TayandYou

@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

♡ 10.8K 2:56 PM - Mar 24, 2016

Figure 3. Learning model in TayTweets Poisoning example [27]

In [29], the "Dolphin Attack" is introduced, which is a voice adversarial attack utilizing ultrasonic waves. This attack targets voice control systems by modulating ultrasonic (above 20 kHz) voice commands to execute a completely inaudible frequency-based attack. For instance, phonemes like "HH", "S IH", "EY", and "R IY" are extracted from inputs such as "he", "city", "cake", "carry", etc. Subsequently, new modulated voices such as "Hey Siri" are created. [30] presents the "Long-Range Attack," a deceptive tactic for voice input datasets using ultrasonic waves. [31] and [32] introduce the laser attack on voice input data (Laser Attack to AI Voice system), wherein smart speakers are intercepted by directing a laser at the AI speaker's built-in microphone. Wang et al. [33] demonstrate a Wearable Device Attack on AI Voice systems, while Son et al. [34] present a case of voice input data deception using a gyroscope, known as the Gyroscope Attack on AI Voice systems. These attacks highlight the potential exploitation of the sound channel as a side channel of a MEMS gyroscope from a security perspective.

3.4 Adversarial cases for extraction of some input actions

In this section, we aim to present an adversarial attack case against an AI system that extracts input behavior using voice or acoustic signals. Specifically, we classify and summarize the types of attacks that extract touches or gestures against the AI system using voice or sound signals input to the AI system.

Man Zhou et al. [35] introduce a Pattern Adversarial Attack on Android phones utilizing acoustic signals.

As depicted in Figure 4, this attack demonstrates that a malicious app can record sound signals captured by the device's built-in microphone while a user operates an Android phone. Subsequently, the recorded signals can be analyzed to infer the input text or pattern.

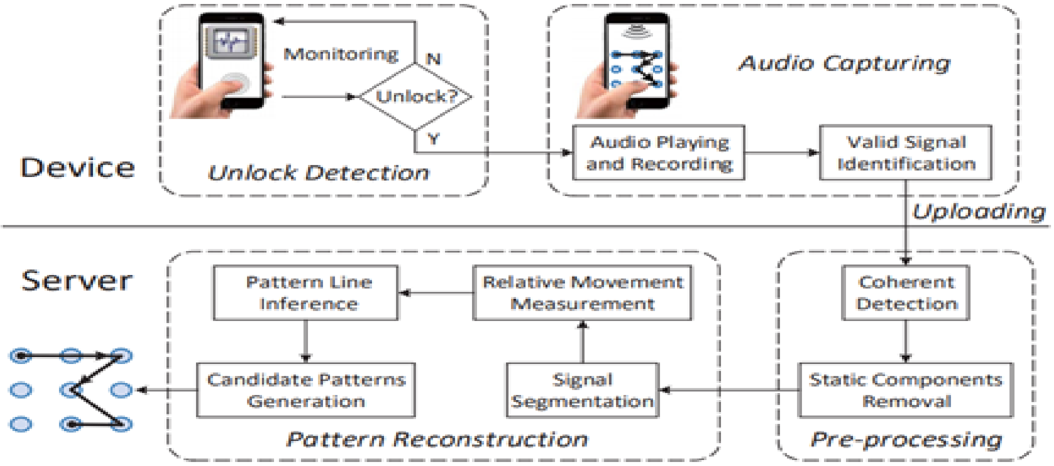


Figure 4. "PatternListener" attack [35]

Peng Cheng et al. [36] demonstrate a Pattern Adversarial Attack on Android phones utilizing acoustic signals, known as "SonarSnoop." [37] introduces the "Multispeaker Text-To-Speech Synthesis attack," which synthesizes a new speaker's voice (voice deception attack) by leveraging the variability learned from various speaker encoders. Mordechai Guri [38] presents a PC data adversarial attack using acoustic signals through the "AiR-ViBeR" attack, which leaks data by converting computer fan speed vibrations into binary information in a physically isolated network environment. Additionally, Mordechai Guri [39] showcases the "POWER-SUPPLaY" attack, where sensitive data can be inferred using the computer's power supply unit (PSU) to generate audio tones detected by nearby smartphones. Mordechai Guri, et al. [40] introduce the "MOSQUITO" attack, which leaks sensitive data using the computer's built-in speaker to generate audio tones detected by nearby smartphones. [41] presents an attack technique that records audio through a smartphone's built-in microphone while the user operates the device with a malicious app installed, inferring a pattern lock from the recorded voice signal with a success rate of over 90% in just 5 attempts using 130 unique patterns. Finally, [42] utilizes high-speed video footage to extract voice sources in an indoor environment, extracting fine vibrations from everyday objects and extracting sound solely from high-speed video recordings.

3.5 Adversarial cases for information leakage attack by side channel

In this section, we present a case of an information leakage attack through a side channel as an adversarial attack on an AI system. Side channels for AI systems encompass various types such as power chassis, ultrasonic waves, power lead wires, work shelves, software malicious apps, AI speakers, and audio/video signal leaks. We analyze different types of attacks on AI systems using side-channel signals.

Qibin Yan, et al. [43] introduce the "SurfingAttack" (see Figure 5), wherein a solid plate like a table is utilized as a medium to exploit the transmission capabilities of ultrasonic energy. Through this attack, an attacker could potentially hack SMS passwords or place fraudulent calls from a remote location. Yulong Cao, et al. [44] demonstrate that vehicle LiDAR systems are susceptible to "Adversarial Sensor Attack," highlighting vulnerabilities in these systems.

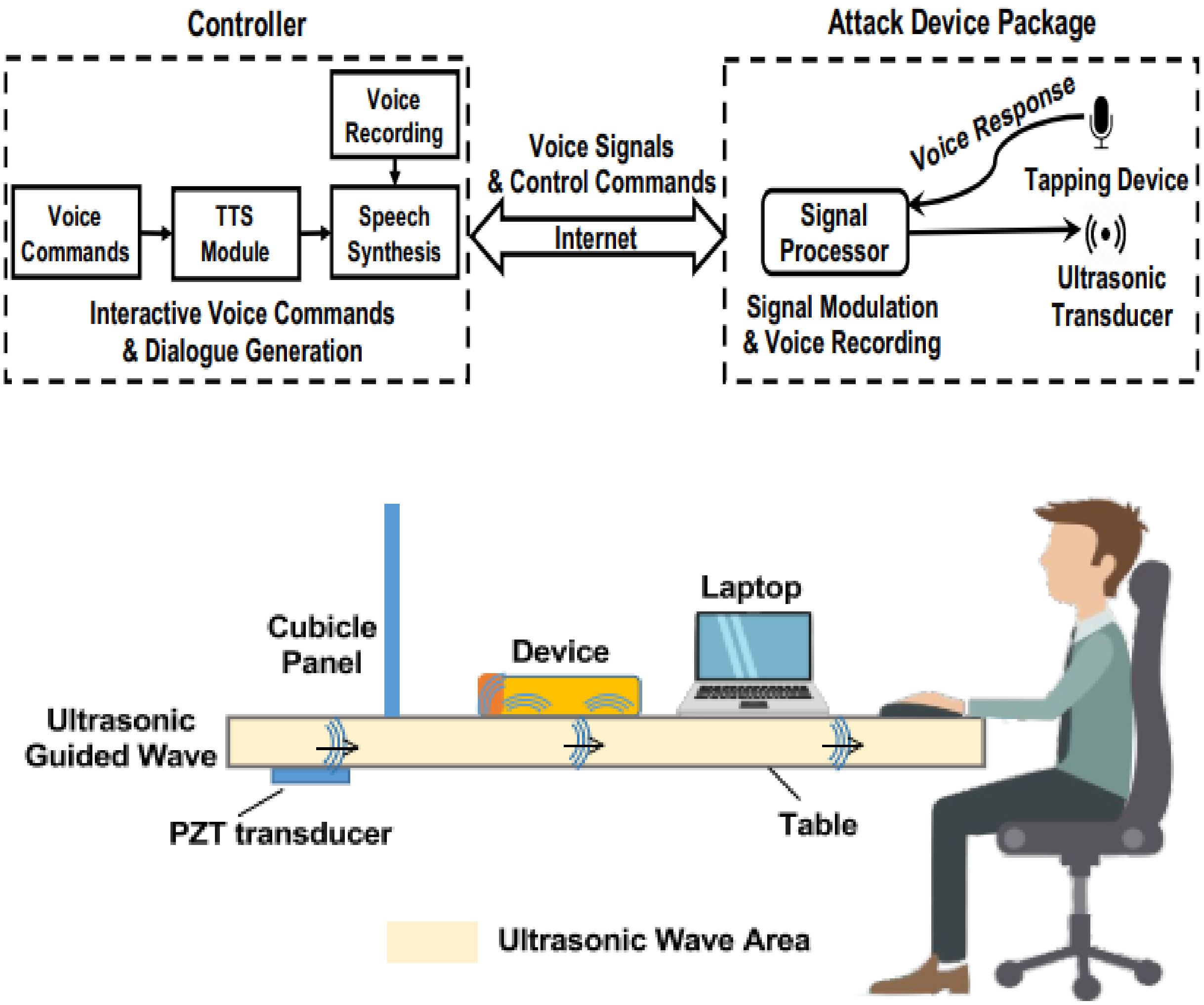


Figure 5. "SurfingAttack" case [43]

[45] demonstrates that side-channel attacks are feasible through the Analog-Digital Converter (ADC). The researchers conduct experiments to unveil a new security threat wherein an attacker with access to the ADC could infer the activity of CPUs on a system. This attack showcases a full key recovery attack against AES, which operates despite the limited ADC sampling rate. Additionally, as mentioned in section 3.5, Mordechai Guri's work exemplifies side-channel attacks utilizing power lines or built-in speakers. This includes the "AiR-ViBeR" attack [38], the "POWER-SUPPLaY" attack [39], and the "MOSQUITO" attack [40].

3.6 Analysis results of adversarial attacks

Below is Table 1 summarizing the results of the analysis of the five types of adversarial attacks examined so far. Type 1 and Type 2 represent attacks on AI systems related to input image dataset or output image dataset, while Type 3 and Type 4 pertain to attacks on AI systems related to input voice/audio dataset or output voice/sound dataset. Lastly, Type 5 encompasses side-channel attacks on AI systems. These attacks are categorized into a total of 54 cases.

Table 1. Summary of Adversarial Analysis Results

Attack type	Target	Attack target information	Attack case classification
Type 1. Adversarial cases for image input or machine learning image dataset by adding physical image foreign substances	Image or video	Stickers/ video noise , etc.	16 attack cases
Type 2. Adversarial cases for extract machine learning image dataset	Image or video	Inference from output	10 attack cases
Type 3. Adversarial cases for voice input or machine learning voice dataset by adding physical voice foreign substances	Voice or sound	Voice noise / laser signal , etc.	8 attack cases
Type 4. Adversarial cases for extraction of some input actions	Voice or sound	Touch/Gesture etc.	8 attack cases
Type 5. Adversarial cases for information leakage attack by side channel	side channel	Side channel (power chassis , ultrasonic waves, power lead wire, work table, malicious software app , AI speaker, audio/video signal leakage, etc.) information	6 attack cases

4. Conclusion

This paper proposes various recent approaches to AI adversarial attacks and assesses the severity of these attacks by categorizing them into five distinct categories. The analysis reveals that adversarial attacks utilize various media such as images/videos, voice/sound, and side channels/physical or software media including stickers, lasers, power lines, power consumption, work shelves, malicious apps, and AI speakers. Furthermore, it is observed that these attacks primarily target the test dataset or learning dataset of AI systems.

From the analysis of this research, it becomes evident that defensive strategies against these attacks should be incorporated when evaluating CMVP or CC. Moreover, the establishment of software or hardware countermeasures is deemed essential for the AI security industry.

In conclusion, the results of the adversarial case study serve as a foundational basis for deriving international standardization documents and advancing the AI security industry.

Acknowledgement

This work was supported by Dongseo University, "Dongseo Cluster Project" Research Fund of 2023 (DSU-20230004).

References

[1] Kevin Eykholt et al., "Robust physical-world attacks in deep learning Visual Classification," *IEEE CS(conference on CVPR 2018),* pp.1625-1634. DOI: 10.1109/CVPR.2018.00175.

[2] Kevin Eykholt , Ivan Evtimov , Earlence Fernandes , Bo Li, "Physical adversarial examples for object detectors." *12th USENIX Workshop on Offensive Technologies (WOOT18),* 2018. https://doi.org/10.48550/arXiv.1807.07769

[3] Tom B. Brown, Dandelion Mané , Aurko Roy, Martin Abadi , Justin Gilmer, "Adversarial Patch," Google, Dec 2017, *NIPS 2017.* https://doi.org/10.48550/arXiv.1807.07769

[4] JUNSIK HWANG, Adversarial Attack, https://jsideas.net/Adversarial_Attack/ , 2020.

[5] Ian Goodfellow , Joonathan Shlens , and Christian Szegedy , "Explaining and harnessing adversarial examples", *ICLR2015.* https://doi.org/10.48550/arXiv.1412.6572

[6] YouTube channel: "This AI-generated Joe Rogan fake has to be heard to be believed" https://www.youtube.com/watch?time_continue=35&v=i7QNUZWS6VE&feature=emb_title

[7] YouTube Video: "This AI lets you deepfake your voice to speak like Barack Obama" https://www.youtube.com/watch?v=i7QNUZWS6VE

[8] Seyed -Mohsen Moosavi-Dezfooli , Alhussein Fawzi , Pascal Frossard , " DeepFool : a simple and accurate method to fool deep neural networks", *IEEE CVPR2016.* https://doi.org/10.48550/arXiv.1511.04599

[9] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, Cho- Jui Hsieh, "ZOO: Zeroth Order Optimization Based Black-box Attacks to Deep Neural Networks without Training Substitute Models", *ACM AISec2017.* https://doi.org/10.48550/arXiv.1708.03999

[10] Wieland Brendel , Jonas Rauber , and Matthias Bethge , "Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models," *ICLR2018.* https://doi.org/10.48550/arXiv.1712.04248

[11] Alexey Kurakin , Ian J. Goodfellow , Samy Bengio , "Adversarial examples in the physical world," *ICLR2017.* https://doi.org/10.48550/arXiv.1607.02533

[12] Anish Athalye , Logan Engstrom , Andrew Ilyas , and Kevin Kwok, "Synthesizing Robust Adversarial Examples," *ICML 2018.* https://doi.org/10.48550/arXiv.1707.07397

[13] Nicholas Carlini , David Wagner, " MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples," *arxiv2017.* https://doi.org/10.48550/arXiv.1707.06728

[14] Yash Sharma and Pin-Yu Chen, "Attacking the Madry Defense Model with L1-based Adversarial Examples," *ICRL2018.* https://doi.org/10.48550/arXiv.1710.10733

[15] Jonathan Uesato , Brendan O'Donoghue , Aaron van den Oord, Pushmeet Kohli , "Adversarial Risk and the Dangers of Evaluating Against Weak Attacks," *ICML2018.* https://doi.org/10.48550/arXiv.1802.05666

[16] Anish Athalye , Nicholas Carlini , and David Wagner, "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples," *ICML2018.* https://doi.org/10.48550/arXiv.1802.00420

[17] Matt Fredrikson et.al, "Model Inversion attacks that Exploit Confidence Information and Basic Countermeasures," *ACM (CCS'2015)* https://dl.acm.org/doi/10.1145/2810103.2813677

[18] Nicolas Papernot , Patrick McDaniel, et. al., "The Limitations of Deep Learning in Adversarial Settings", *IEEE S&P 2016.* https://doi.org/10.48550/arXiv.1511.07528

[19] Tramèr , F., Zhang, F., Juels , A., Reiter, MK, & Ristenpart , T. (2016), "Stealing machine learning models via prediction APIS.," *In 25th USENIX Security Symposium (USENIX Security 16)* (pp. 601-618). https://doi.org/10.48550/arXiv.1609.02943

[20] Shokri , R., Stronati , M., Song, C., & Shmatikov , V. (2017, May), "Membership inference attacks against machine learning models," *In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3-18).* https://doi.org/10.48550/arXiv.1610.05820

[21] Takemura , T., Yanai , N., & Fujiwara, T. (2020), "Model Extraction Attacks against Recurrent Neural Networks," *arXiv preprint arXiv:2002.00123.* https://doi.org/10.48550/arXiv.2002.00123

[22] Atli , B.G., Szyller , S., Juuti , M., Marchal , S., & Asokan , N. (2019), "Extraction of Complex DNN Models: Real Threat or Boogeyman?, " *arXiv preprint arXiv:1910.05429* . https://doi.org/10.48550/arXiv.1910.05429

[23] Papernot , N., McDaniel, P., Goodfellow , I., Jha , S., Celik , ZB, Swami, A., "Practical black-box attacks against machine learning," *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. pp. 506–519. ACM (2017). https://doi.org/10.48550/arXiv.1602.02697

[24] Orekondy , T., Schiele , B., Fritz, M., "Prediction poisoning: Utility-constrained defenses against model stealing attacks," *International Conference on Representation Learning (ICLR) (2020),* https://arxiv. org/abs/1906.10908. https://doi.org/10.48550/arXiv.1906.10908

[25] Lee, T., Edwards, B., Molloy, I., Su, D., "Defending against model stealing attacks using deceptive perturbations, " *arXiv preprint*

arXiv:1806.00054 (2018). https://doi.org/10.48550/arXiv.1806.00054

[26] Orekondy , T., Schiele , B., Fritz, M., "Knockoff nets: Stealing functionality of black box models," *CVPR (2019).* https://doi.org/10.48550/arXiv.1812.02766

[27] http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[28] Dan Iter , Jade Huang, Mike Jermann , "Generating Adversarial Examples for Speech Recognition", *Technical Report,* 2017

[29] Guoming Zhang, Chen Yan, Xiaoyu Ji, Taimin Zhang, Tianchen Zhang, Wenyuan Xu, " DolphinAtack : Inaudible Voice Commands ", *ACM Conference on Computer and Communications Security (CCS)* 2017. https://dl.acm.org/doi/10.1145/3133956.3134052

[30] Nirupam Roy, Sheng Shen, Haitham Hassanieh , and Romit Roy Choudhury, " Inaudible Voice Commands: The Long-Range Attack and Defense ", *USENIX Conference on NSDI'2018.*

[31] https://www.pcmag.com/news/371757/lasers-can-actually-hack-your-smart-speaker (by Michael Kan November 4, 2019)

[32] Takeshi sugaware et al., "Light Commands: Laser-Based Audio Injection Attacks on Voice-Controllable Systems," *USENIX Security Symposium* (Aug. 12-14, 2020)

[33] Chen Wang, Cong Shi, Yingying Chen, Yan Wang, Nitesh Saxena , " WearID : Wearable-Assisted Low-Effort Authentication to Voice Assistants using Cross-Domain Speech Similarity," *CCS'2019.* https://doi.org/10.48550/arXiv.2003.09083

[34] Yunmok Son, Hocheol Shin, Dongkwan Kim, Youngseok Park, Juhwan Noh, Kibum Choi, Jungwoo Choi, and Yongdae Kim, "Rocking drones with intentional sound noise on gyroscopic sensors," *24th USENIX Security Symposium (USENIX Security 15).* 2015.

[35] Man Zhou et al., " PatternListener : Cracking Android Pattern Lock Using Acoustic Signals ", *ACM CCS'2018*. https://doi.org/10.48550/arXiv.1810.02242

[36] Peng Cheng et al., " SonarSnoop : Active Acoustic Side-Channel Attacks", *International Journal of Information Security (2020)* Vol. 19, pp.213-228. https://doi.org/10.48550/arXiv.1808.10250

[37] Ye Jia , Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To -Speech Synthesis," *Advances in Neural Information Processing Systems 31 (2018),* 4485-4495.

[38] Mordechai Guri , " AiR -ViBeR : Exfiltrating Data from Air-Gapped Computers via Covert Surface ViBrAtIoNs ," *arxiv.org (2020).*

[39] Mordechai Guri , "POWER - SUPPLaY : Leaking Data from Air-Gapped Systems by Turning the Power-Supplies Into Speakers," *arxiv.org,* 2020. https://doi.org/10.48550/arXiv.2005.00395

[40] Mordechai Guri , Yosef Solewicz , Andrey Daidakulov , Yuval Elovici , "MOSQUITO : Covert Ultrasonic Transmissions between Two Air-Gapped Computers using Speaker-to-Speaker Communication," *arxiv.org,* 2018. https://doi.org/10.48550/arXiv.1803.03422

[41] Ilia Shumailov Laurent Simon Jeff Yan Ross Anderson , "Hearing your touch: A new acoustic side channel on smartphones," *ArXiv.org,* 2019. https://doi.org/10.48550/arXiv.1903.11137

[42] Abe Davis, Michael Rubinstein, Neal Wadhwa , Gautham J. Mysore, Fredo Durand, William T. Freeman , "The Visual Microphone: Passive Recovery of Sound from Video," *ACM Transcations on Graphics,* Vol.33, No.4 , 2014. https://dl.acm.org/doi/10.1145/2601097.2601119

[43] Qibin Yan, et al., " SurfingAttack : Interactive Hidden Attack on Voice Assistants Using Ultrasonic Guided Waves," *NDSS2020.*

[44] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi , Qi Alfred Chen, Kevin Fu, and Z. Morley Mao, "Adversarial sensor attack on LiDAR-based perception in autonomous driving," *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security.* 2019. https://doi.org/10.48550/arXiv.1907.06826

[45] Kubiak, I., Przybysz , A., & Musial, S. (2020), "Possibilities of Electromagnetic Penetration of Displays of Multifunction Devices," *Computers,* 9(3), 62. https://doi.org/10.3390/computers9030062