Computer Systems
Science and Engineering

Tech Science Press

DOI: 10.32604/csse.2023.039599
Article

Adversarial Attack-Based Robustness Evaluation for Trustworthy AI

Eungyu Lee, Yongsoo Lee and Taejin Lee*

Department of Information Security, Hoseo University, Asan 31499, Korea
*Corresponding Author: Taejin Lee. Email: kinjecs0@gmail.com
Received: 07 February 2023; Accepted: 11 May 2023; Published: 28 July 2023

Abstract: Artificial Intelligence (AI) technology has been extensively researched in various fields, including the field of malware detection. AI models must be trustworthy to introduce AI systems into critical decision-making and resource protection roles. The problem of robustness to adversarial attacks is a significant barrier to trustworthy AI. Although various adversarial attack and defense methods are actively being studied, there is a lack of research on robustness evaluation metrics that serve as standards for determining whether AI models are safe and reliable against adversarial attacks. An AI model's robustness level cannot be evaluated by traditional evaluation indicators such as accuracy and recall. Additional evaluation indicators are necessary to evaluate the robustness of AI models against adversarial attacks. In this paper, a Sophisticated Adversarial Robustness Score (SARS) is proposed for AI model robustness evaluation. SARS uses various factors in addition to the ratio of perturbated features and the size of perturbation to evaluate robustness accurately in the evaluation process. This evaluation indicator reflects aspects that are difficult to evaluate using traditional evaluation indicators. Moreover, the level of robustness can be evaluated by considering the difficulty of generating adversarial samples through adversarial attacks. This paper proposed using SARS, calculated based on adversarial attacks, to identify data groups with robustness vulnerability and improve robustness through adversarial training. Through SARS, it is possible to evaluate the level of robustness, which can help developers identify areas for improvement. To validate the proposed method, experiments were conducted using a malware dataset. Through adversarial training, it was confirmed that SARS increased by 70.59%, and the recall reduction rate improved by 64.96%. Through SARS, it is possible to evaluate whether an AI model is vulnerable to adversarial attacks and to identify vulnerable data types. In addition, it is expected that improved models can be achieved by improving resistance to adversarial attacks via methods such as adversarial training.

Keywords: AI; robustness; adversarial attack; adversarial robustness; robustness indicator; trustworthy AI

1 Introduction

As accessibility to and processing power for large amounts of data improves, AI that can efficiently leverage it is being extensively researched. AI research covers many fields, including security, self-driving cars, smart factories, natural language processing, etc. Despite the promising advancements in AI, challenges remain in ensuring its trustworthiness in the real environment. Transparency, technical robustness, and safety have been identified by the EU as key requirements to be met for trustworthy AI [1]. "Robustness" means attack response to AI systems such as adversarial attacks, minimizing unintended consequences and operational errors. For trustworthy AI, the robustness of AI models needs to be measured and evaluated.

AI models collect and process data to learn and provide their intended functionality. Collecting a sufficient quantity and quality of data to cover all possible inputs is impractical. Therefore, data bias and overfitting problems may occur in AI models due to the nature of the dataset. Additionally, it becomes difficult to trust AI models trained on past data in environments where constantly evolving and newly generated data, such as malware, persistently.

Most newly detected malware is generated by slightly modifying existing malware to evade detection systems [2,3]. Similarly, Adversarial attacks induce the misclassification of AI models by creating slight variations in the original data. Evaluating and reporting how robust AI models are against these variant attacks should be possible.

Deep neural networks are vulnerable to adversarial attacks [4–6]. Goodfellow et al. argue that the high-dimensional linearity of deep neural networks leads to spaces where adversarial samples can be generated in the classification space [7]. An adversarial attack is a technique that deceives AI models into predicting different labels by injecting a slight perturbation into the original data [8]. Accuracy, precision, recall, and F1 score, which are widely used as evaluation indicators of AI models, are insufficient to evaluate the robustness of AI models [9]. The F1 score is a metric calculated as the harmonic mean of precision and recall. An AI model designed to detect DGA can be deceived by adversarial attacks that make malicious domains appear normal [10]. If we cannot evaluate an AI model's robustness against adversarial attacks, it will limit our ability to make significant strides toward trustworthy AI.

Although various adversarial attack and defense methods are actively being studied, there is a lack of research on robustness evaluation indicators that serve as standards for determining whether AI models are safe and trustworthy against adversarial attacks. To develop a safe and trustworthy model, it is necessary to have robustness evaluation indicators that indicate the level of robustness required to consider the model safe. There are cases where the robustness of AI models is evaluated, but they simply use metrics such as the success rate of attacks or the distance between original and adversarial samples [9,11–13]. However, these measures cannot be universally used for accurate evaluation, as the number of features used, and the size of the values handled by AI models vary. For trustworthy AI, it is necessary to evaluate it through appropriate indicators to evaluate the robustness of an AI model. In this paper, we propose a Sophisticated Adversarial Robustness Score (SARS) as an indicator of robustness evaluation based on the difficulty of adversarial attacks. To evaluate the robustness of AI models in detail, various factors including adversarial attack success rates are used. Furthermore, we propose a method using the proposed indicator to identify and improve data groups that have weak robustness, as identified by the proposed indicator, in the models' training data.

The main contributions of this paper are as follows:

- The need for a new evaluation indicator to assess the robustness of AI models against adversarial attacks is emphasized, and SARS is proposed as a solution.
- Experiments and analysis on a malware dataset are conducted in this paper to demonstrate the effectiveness of the proposed method.
- This paper expects that the proposed SARS will contribute to developing a universal robustness evaluation indicator and trustworthy AI.

The structure of the paper is as follows: Section 2 introduces related works on adversarial attacks and robustness evaluation and improvement methods. Section 3 proposes adversarial attack-based robustness evaluation and improvement methods for trustworthy AI. Section 4 presents the robustness evaluation and improvement results of the AI model trained using the malware dataset. Finally, Section 5 presents conclusions and future research directions.

2 Related Work

2.1 Adversarial Attacks

An adversarial attack refers to an attack with the goal of maliciously manipulating the behavior of an AI model to produce incorrect results. There are four main types of adversarial attacks: poisoning attack, in which an attacker intentionally injects malicious training data to undermine the model [14]; inversion attack, in which an attacker analyzes the confidence vector generated by querying the model numerous times to extract the data used for training [15,16]; model extraction attack, in which an attacker analyzes the confidence vector to extract a model that is similar to the actual AI model [17,18]; and evasion attack, in which an attacker deceive the machine learning model [19]. This paper focuses on the perspective of whether AI models can be trusted and emphasizes evasion attacks.

There is a method of generating an adversarial sample to induce misclassification by an AI model, for instance, by generating a malware variant to evade a detection system. These adversarial samples are generated by adding small amounts of indistinguishable noise to the original data. As shown in Fig. 1, it is impossible visually to discern the difference between the original data and the adversarial sample. However, it can be confirmed that the adversarial sample has succeeded in misclassification by obtaining an utterly different prediction through a slight perturbation. Fig. 2 presents a simplified description of how adversarial samples are generated. Adversarial attacks consist of finding the optimal δ, a very slight perturbation that crosses the decision boundary of the AI model from the original data. Several technologies to implement such adversarial attacks are being developed and continue to evolve. Adversarial attack techniques have the same goal of finding an optimally slight perturbation, which can be classified according to the process of optimization and determination process, as well as the environment in which it is performed.

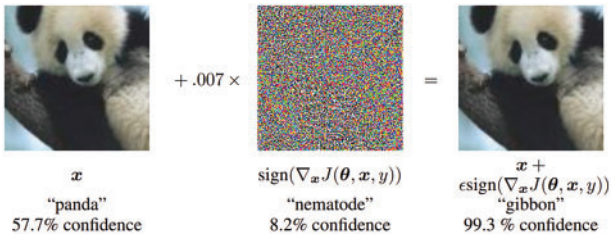

Figure 1: Example of an adversarial attack [7]

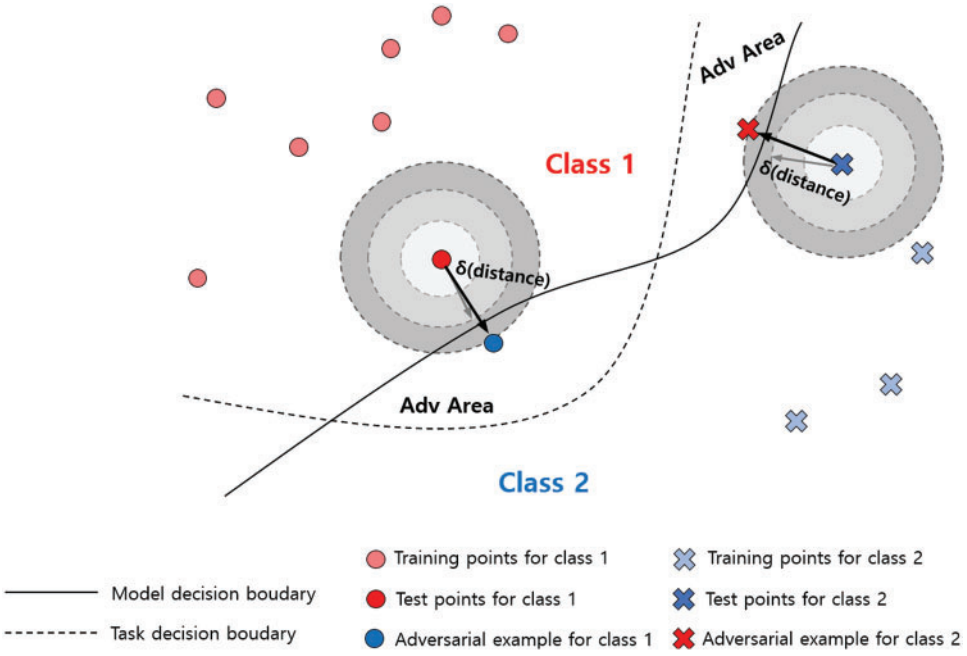

Figure 2: Example of generating adversarial samples

The attack methods can be categorized into white-box and black-box attacks depending on whether they utilize the internal information of the AI model or not. Fast Gradient Sign Method (FGSM) [7] is an adversarial attack technique that computes gradient using weight values within the model and manipulates input values to generate adversarial samples. Projected Gradient Descent (PGD) [20], an attack technique developed from FGSM, repeats as many steps as specified and finds an adversarial sample via the L-infinity norm, namely the largest value of a set vector element, which regulates the changed data so that it does not exceed the maximum permissible value. In image classification, most images contain more than one object, so the image needs to be divided into regions and each region needs to be classified individually. To perform effective and efficient adversarial attacks on such segmentation models, an attack technique called SegPGD [21], which is an improvement of PGD. DeepFool [22] is a technique that attacks nonlinear neural network structures through repetitive queries, projecting perpendicular to the decision boundary at several points and adding an appropriate amount of noise rather than using a gradient calculation like FGSM. Jacobian-based Salience Map Attack (JSMA) [23] is an attack technique that creates adversarial samples by changing inputs so that the model misclassifies by mapping the changes of inputs to outputs into a matrix. ZOO attack (Zeroth Order Optimization-based black-box attack) [24] approaches the optimization problem like the gradient-based attack technique. However, it is an optimization method in which the gradient is estimated and used by assuming a black-box situation in which the information within the model is unknown. Carlini & Wagner's attack (C&W) [25] does not directly solve the problem of finding adversarial samples but carries out optimization by approximation with another objective function. In adversarial attack research, many experiments have been conducted using PGD and C&W attacks.

Methods to defend against adversarial attacks are also being studied. To defend against white-box-based attacks that use information internal to the model, such as FGSM and PGD attacks, a gradient

masking method can be used that hides the internal gradient. In addition, input pre-processing is a method that reduces the effect of noise through noise filtering and pre-processing of the input. Also, as shown in Fig. 3, the adversarial training approach reduces the perturbation effect by including adversarial samples in the training dataset. In Fig. 3a, the adversarial sample succeeded in attacking the traditionally trained model. However, as shown in Fig. 3b, it can be confirmed that the attack failed in the model that conducted adversarial training [26]. This involves training with adversarial samples in advance to solidify the decision boundary and make it difficult for adversarial attacks to succeed. In the case of gradient masking, if an attack technique does not use a gradient, it is impossible to respond to the attack. Also, in the case of input pre-processing, there may be a problem with the performance of the AI model due to special processing. Adversarial training can also improve robustness by making it challenging to generate adversarial samples.

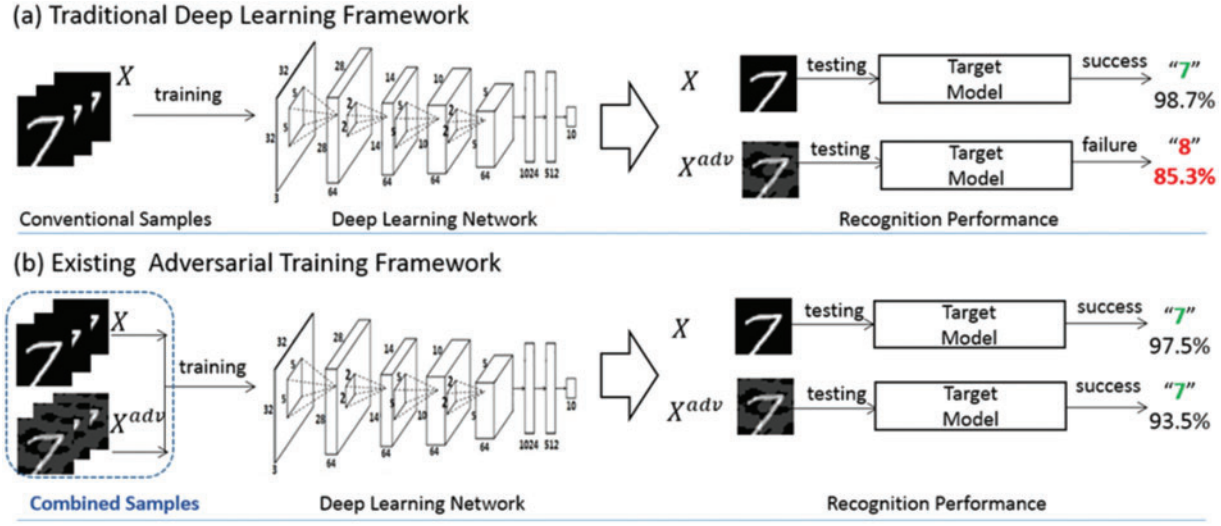

Figure 3: Example of adversarial training [26]

2.2 Robustness Evaluation of AI Models

2.2.1 Traditional Evaluation Indicators of the AI Model

Accuracy, precision, recall, and F1 score are widely used as traditional evaluation indicators of AI models. AI models are evaluated based on how accurate they are in performing their intended function. To operate a malware detection system using AI, malware must be detected in a situation where most of the files are normal. If the AI model predicts that all files are normal, the accuracy will be high because most are normal files. However, not all malware files have been detected so the recall will be calculated as very low. In this environment, the recall will be used as a more important evaluation indicator of AI model performance than accuracy. So, it is necessary to select and evaluate performance indicators suitable for the goals and purposes of the AI model. As such, in addition to the indicators that implement the target performance, indicators are needed to evaluate whether they are safe from adversarial attacks [9].

2.2.2 Robustness Evaluation Indicators of AI Model

Some indicators can evaluate the performance of AI models, such as accuracy and recall, but the robustness of a model cannot be evaluated using only traditional evaluation methods. Therefore,

a new evaluation indicator that can evaluate the robustness of AI models is needed [27]. In Fig. 4, "unmodified" denotes the original data, while C&W, PGD, and FGSM, which are types of adversarial attack techniques, denote samples created through each attack. As seen in Fig. 4, even if the recall is high for general data, it is significantly reduced by adversarial attacks. The high recall does not mean that successful adversarial attacks are challenging for normal data. Therefore, there is a need for an indicator that can specifically evaluate robustness.

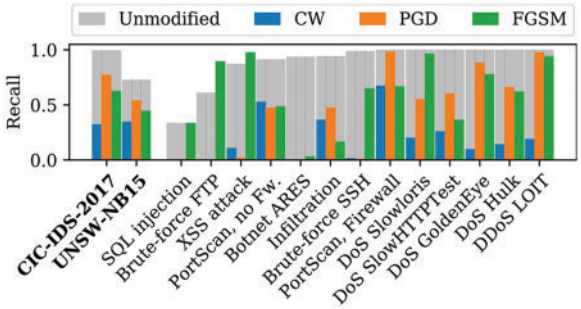

Figure 4: Comparison of recall for original data and adversarial samples by attack type [9]

Several studies have been conducted to evaluate the level of robustness of AI models. Chang et al. propose a scoring method to evaluate the robustness of AI models [28]. Using APIs provided by Adversarial Robustness Toolbox [29], Foolbox [30], CleverHans [31], and so on, 13 attack techniques have been applied to attack AI models, with the variance calculated based on each attack's success rate. The calculated variance is used as an indicator of robustness.

Berghoff et al. propose a verification scheme to evaluate robustness. The evaluation was conducted by subdividing the robustness into several factors [32]. Robustness is evaluated in four categories: robustness to image noise, such as Gaussian noise, robustness to pixel modulation, robustness to geometrical transformations, such as rotation and scaling, and robustness to hue and color transformations. Experiments are conducted while increasing the modulation magnitude for each category, and the degree of accuracy deterioration is evaluated.

Hartl et al. proposed an Adversarial Robustness Score (ARS) as a robustness evaluation indicator for AI models [9]. ARS is a number that indicates how resistant a model is to adversarial attacks. Eq. (1) represents the ARS calculation formula:

$$ARS = \frac{1}{\lceil N/2 \rceil} \sum_{s \in S} d_s \tag{1}$$

where S denotes the set of all samples, s denotes a specific sample, and N is the total number of samples. Moreover, d means the distance between the adversarial sample and the original data, and the distance for unsuccessful samples is replaced by ∞. Thus, ARS is approximately the average distance no greater than the median. If an attack is less than 50% successful, the AI model is judged to be robust.

Robustness against adversarial attacks can be seen as more robust as the difficulty of generating adversarial samples through adversarial attacks increases. The more difficult it is to generate adversarial samples, the higher the degree of modulation becomes. As a result, ARS, which measures the distance from the original, also increases so that the level of robustness can be evaluated based on the difficulty of sample creation.

3 Proposed Method

3.1 Overview

AI technology is being actively studied in various fields due to its excellent usability. However, for introduction into real environments, the development of trustworthy AI models is required. It is necessary to verify robustness against adversarial attacks, which is the cause of the unreliability of AI models but is challenging to evaluate with traditional evaluation indicators such as accuracy and recall [9]. Therefore, a separate evaluation indicator to evaluate the level of robustness of the AI model and a method to improve insufficient robustness are needed.

This paper proposes an indicator to evaluate robustness based on adversarial attacks for developing trustworthy AI models. The robustness level is analyzed through the proposed evaluation indicator, and a method to improve the robustness of the data group that lacks robustness is proposed.

Fig. 5 shows the overall structure of our approach to evaluating and improving the robustness of AI models. We measure robustness based on the difficulty of generating adversarial samples through adversarial attacks to evaluate robustness. Then, robustness is improved through adversarial training after identifying vulnerable data groups through robustness evaluation for each data group. Finally, based on the proposed evaluation indicator, we confirm that robustness is improved by comparing robustness before and after adversarial training.

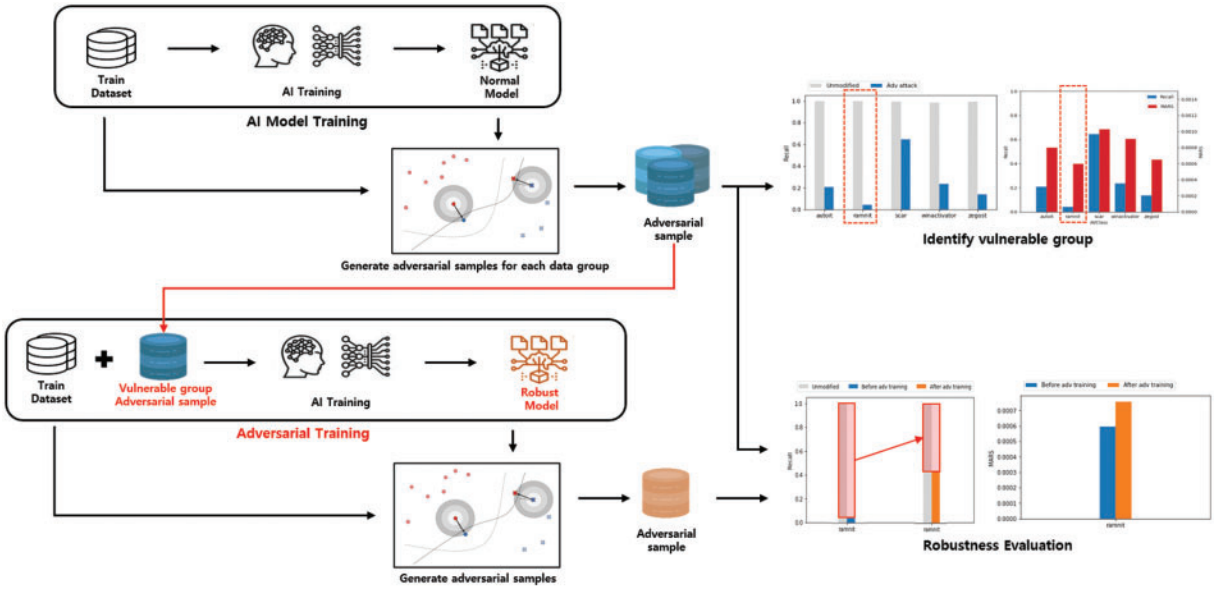

Figure 5: Adversarial attack-based robustness evaluation and improvement method

3.2 Robustness Evaluation Method

The rationale for evaluation robustness based on the difficulty of generating adversarial samples is that accurate evaluation is impossible with traditional indicators such as precision and recall. An adversarial attack aims to deceive an AI model with a slight perturbation that is difficult to perceive. As such, the robustness of AI models against adversarial attacks can be evaluated through the difficulty of generating adversarial samples.

ARS [9] uses the Euclidean distance between the sample and the original data to determine the difficulty of generating adversarial samples. However, this does not consider the number of features used in the input value of the AI model and the scale for each feature, and it is difficult to evaluate with the same criteria in different environments. Moreover, to evaluate the difficulty of generating adversarial samples, a detailed evaluation is required that considers the ratio of perturbated features, the amount of probability change, and so on.

In this paper, we consider different AI model environments for general use and propose SARS indicators for precise measurement. The method proposed in this paper can be computed using only the targeted AI model, original data, and adversarial samples. The robustness of an AI model against adversarial attacks can be evaluated based on the difficulty of the adversarial attack.

Evaluating the difficulty of adversarial attacks with a single criterion such as the distance between the original and adversarial samples, or the success rate of attacks may be inappropriate. The results can vary greatly depending on the parameters set when attempting the attack. if the allowable range of perturbation is greatly increased, the attack success rate may increase, but the distance between the original and adversarial samples will also increase. In addition, differences may occur depending on the type of the data used by the AI. For example, in the case of images, it is difficult to detect small adjustments made to all pixels, but if some pixels are changed by a large amount, it can be easily identified visually. Injecting a perturbation of 10 to all pixels in an image with 1,024 pixels that have values between 0 and 255, and inverting 128 of 80 pixels, will be evaluated equally due to the same amount of perturbation. Therefore, to establish a standardized evaluation indicator, it is necessary to consider various factors in the evaluation criteria.

The difficulty of adversarial attacks can be determined by considering the attack's success rate, the size and rate of application of perturbations required to change the label, and the probability of changes caused by the perturbations. As the attack's success rate decreases and the size of the perturbations required to change the label increases, the difficulty of the adversarial attack increases. Furthermore, if even small perturbations cause significant changes in the AI's probability for adversarial samples, the difficulty of the adversarial attack decreases. Finally, if changing only a few features is sufficient to change the label instead of changing all features, the difficulty of the adversarial attack decreases. To evaluate robustness, adversarial attacks on the training dataset must be performed. However, the resources and time required for this vary depending on the experimental environment, so these factors cannot be included in robustness evaluations for general use. Based on these factors, the robustness can be evaluated sophisticatedly. Eq. (2) represents the SARS calculation formula:

$$SARS = AVG\left(\sum d(X, X')\right)\frac{1}{S}, d\left(X, X'\right) = \frac{L_0 rate\sqrt{\sum_{i=1}^{n}\frac{(x_i, x_i')^2}{r_i}}}{\triangle P(X, X')} \tag{2}$$

where X represents the original data, X' represents an adversarial sample of X, S represents the success rate of an adversarial attack, i represents the feature index, x_i represents the value of the i-th feature of X, and r represents the size of each feature scale. For example, if the value of feature A can range between -5 and 5, r becomes 10. $L_0 rate$ represents the percentage of features perturbated by adversarial attacks. $\triangle P(\cdot)$ is the change in probability predicted by the AI model from the original data to the adversarial sample.

Eq. (2) was derived by referencing Eq. (1). SARS is configured to be calculated as a higher value as the difficulty of generating an adversarial sample increase. The attack's success rate and the change in probability have an inversely proportional relationship with SARS. In contrast, the ratio of perturbated

features and the perturbation size are directly proportional. The lower the attack success rate, the more difficult it is to create adversarial samples. A small change in probability denotes that the AI model's judgment change is small due to an adversarial attack, indicating that it is challenging to create an adversarial sample. From the perspective of the number of perturbated features, comparing an adversarial sample whose label changes only when 90% of the features are perturbed and an adversarial sample whose label changes even if only one feature is perturbed, it can be seen that the difficulty of generating the adversarial sample is naturally higher in the former. The larger the perturbation size of the adversarial sample, the greater the difficulty of generating the adversarial sample. It is possible to measure robustness through SARS by considering these indicators for sophisticated robustness level evaluation. And, by using r to generalize the diversity of the feature range and $L_0 rate$ to reduce the influence of the number of used features, it makes it possible to compare different AI models with different inputs using a single evaluation indicator.

3.3 Robustness Improvement Method

In this paper, we evaluate and improve the robustness of AI models based on adversarial attacks. The process of improving the robustness of the AI model is shown in the adversarial training segment of Fig. 5. After training the AI model with the training dataset, the training dataset is divided into groups, and an adversarial attack is performed against each group. Based on the generated adversarial samples, SARS is used to evaluate the robustness of each data group. Afterward, the robustness is improved by identifying a data group with vulnerable robustness and performing adversarial training for that data group. This is based on the assumption that adversarial training can improve robustness. We aim to indirectly compare the robustness of adversarial attacks by comparing the recall change and ARS and SARS separately, as a direct comparison between ARS and SARS is not possible.

The adversarial attack technique to be used in the proposed model uses a PGD attack from the attack techniques introduced above. In line with the position of an expert developing a trustworthy AI model, the attack is performed through PGD, a white-box-based attack that can query the AI model internally. the robustness level is then evaluated based on the general adversarial samples to identify data groups vulnerable to attacks. To improve the robustness of the vulnerable data group, adversarial training is performed by adding adversarial samples for the identified data group to the existing training dataset. The adversarial attack is performed again on the model after the completion of adversarial training. Robustness improvement is confirmed through robustness evaluation and comparison based on the generated samples before and after adversarial training.

4 Experimental Results

4.1 Experimental Dataset

In this paper, an experiment was conducted using the 2019 KISA Data Challenge [33] malware dataset. The configuration of the dataset is shown in Table 1. The dataset was classified based on the AVClass of the malware. AVClass is a tool used for malware classification and clustering, which is used to group and label malicious code samples based on their analysis. There are about 800 types of AVClass of malware in the dataset, of which the top five detected were set as experimental subjects: "autoit", "ramnit", "scar", "winactivator", and "zegost." The AVClass configuration of the dataset is shown in Table 2. After learning the AI model, we proceeded with adversarial attacks for each data group, identifying vulnerable groups and conducting adversarial training to confirm that robustness was improved.

Table 1: 2019 KISA data challenge malware dataset

Dataset	Malware	Normal	Total
Train	17,562	11,568	29,130
Test	4,518	4,513	9,301

Table 2: Top five AVClass configuration

Dataset	autoit	ramnit	scar	winactivator	zegost	Total
Train	517	369	281	288	198	1,653
Test	171	57	17	56	37	338

4.2 Experimental AI Model

Information for executing an executable file is recorded in the portable executable structure. Feature extraction proceeds through static analysis of the portable executable file. Information on the header, DLL, API, section entry, string, and entry point of the portable executable file is analyzed and used as features.

In the portable executable header, 37 features considered meaningful are extracted and used. The 37 features used are shown in Table 3. This feature analyzes the data distribution characteristics for each header value of malware and normal files and allocates a different feature value according to the distribution. DLL and API each proceed with hashing after extraction. After hashing, each number is counted by mapping to 512 values through a modulo operation. DLL and API are each mapped to 512 hash maps and use a total of 1,024 features. The string is used by extracting 525 features after mapping to a hash map by performing hash and modulo operations in units of string length selected through feature analysis by string length. Fifty bytes are extracted from the entry point, setting 1 byte as one feature for a total of 50 features. A total of 1,764 features were extracted and used through the static analysis process of these portable executable files.

Table 3: Feature extraction target in the header information

No.	Feature	No.	Feature
1	SizeOfInitializedData	20	IMAGE_FILE_BYTES_REVERSED_HI
2	DllCharacteristics	21	IMAGE_FILE_BYTES_REVERSED_LO
3	MajorImageVersion	22	IMAGE_FILE_RELOCS_STRIPPED
4	CheckSum	23	MajorLinkerVersion[H1]
5	NumberOfSections	24	MinorLinkerVersion[H1+L1]
6	Known_Sections_por	25	MinorOperatingSystemVersion[H2]
7	Unknown_Sections_por	26	SizeOfUninitializedData
8	Rdata_VirtualSize_por	27	AddressOfEntryPoint_size
9	Text_VirtualSize_por	28	RawSize
10	Reloc_VirtualSize_por	29	VirtualSize
11	Subsystem	30	PointerToLinenumbers

(Continued)

Table 3 (continued)

No.	Feature	No.	Feature
12	AddressOfEntryPoint	31	IMAGE_SCN_CNT_UNINITIALIZED_DATA
13	ImageBase	32	SectionEntropy
14	e_lfanew	33	IMAGE_SCN_CNT_CODE
15	Characteristics	34	IMAGE_SCN_CNT_UNINITIALIZED_DATA
16	MajorLinkerVersion	35	IMAGE_SCN_LNK_NRELOC_OVFL
17	MajorOperatingSystemVersion	36	IMAGE_SCN_MEM_DISCARDABLE
18	MajorSubsystemVersion	37	IMAGE_SCN_MEM_WRITE
19	SizeOfStackReserve		

Next, scaling is performed for each feature group, and 1,764 pieces of 1-dimensional data are converted into $42 * 42$ two-dimensional image data blocks for use. Each block has a value between 0 and 1. The AI model was trained using Convolutional Neural Networks (CNN), which are widely used in processing image data. The training results for the AI model are shown in Table 4.

Table 4: 2019 KISA dataset training results with CNN model

Accuracy	Precision	Recall	F1 score
0.9767	0.9770	0.9767	0.9767

4.3 Adversarial Attack and Robustness Evaluation of AI Model

To identify data groups with vulnerable robustness, an adversarial attack is performed for each data group in the training dataset. Based on the generated adversarial sample, the difficulty of generating the adversarial sample is then evaluated by SARS. The data group with the lowest SARS is identified, and robustness improvement is performed.

In this experiment, a white-box-based PGD attack is applied by assuming an environment where the model can be easily accessed from the perspective of developing a robust model. The parameters of the PGD attack used to attack the CNN model are shown in Table 5. The PGD attack for each data group in the training dataset was repeated three times to generate 4,959 adversarial samples for 1,653 original data. Fig. 6 compares the recall of the AI model predicting the original data and the adversarial samples. Although the AI model shows an almost perfect detection rate for the original data, it can be seen that recall has significantly decreased due to adversarial attacks. This is not problematic at the moment but indicates that the AI model is very vulnerable to possible attacks in the future, demonstrating that the robustness of AI models cannot be adequately assessed with typical evaluation indicators such as recall.

Table 5: Parameters used for PGD attack

Parameter	Value
Norm	2
Eps	0.01

(Continued)

Table 5 (continued)

Parameter	Value
Epx_step	0.00001
Max_iter	10,000
Num_random_init	5

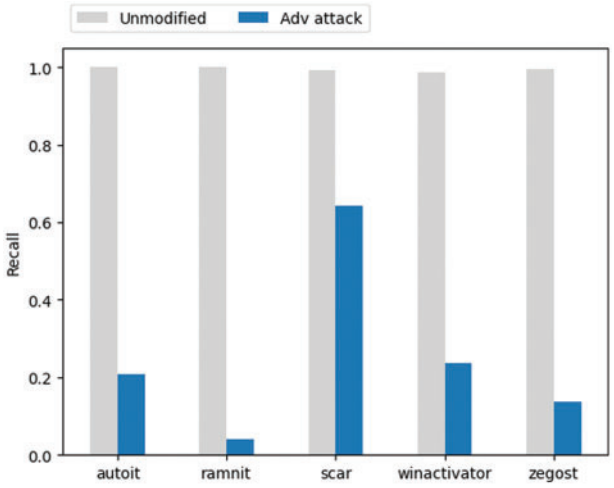

Figure 6: Comparison of recall as reduced by adversarial attacks

Using SARS, we try to identify the data groups that are most vulnerable in terms of robustness. Table 6 shows the robustness evaluation indicators for each data group. When comparing the changes in the recall, it can be seen that ramnit was the most vulnerable to attack, while scar showed the highest resistance. The change in the recall can indirectly show the success rate of adversarial attacks and, thus, the difficulty of generating adversarial samples. It is impossible to directly compare ARS and SARS since there is no standardized evaluation indicator. We would like to indirectly confirm the appropriateness of the robustness evaluation of ARS and SARS by comparing the trend of the change in the recall by each data group.

Table 6: Comparison of robustness evaluation indicators by data group

	autoit	ramnit	scar	winactivator	zegost
Recall change	0.7911	0.9593	0.3488	0.7488	0.8578
ARS	0.223638	0.217001	0.229269	0.260872	0.218290
SARS	0.000199	0.000102	0.000219	0.000165	0.000134

Since there are substantial differences in the values calculated for each evaluation indicator, we convert and compare the figures through min-max scaling for each evaluation indicator. Fig. 7 shows the result of converting each evaluation index through min-max scaling and comparing them. The ramnit, in which recall decreased the most, yielded the lowest robustness scores for both ARS and

SARS. However, in the case of the scar with the minor decrease in the recall, SARS evaluated the robustness as the highest, but ARS evaluated the winactivator as the most robust, while the scar was evaluated as very low. The overall change in recall was consistent with SARS rather than ARS. This means that SARS, calculated through several variables in addition to perturbation size, evaluates the level of robustness more accurately. Through SARS, it was confirmed that it is necessary to improve the robustness of ramnit, which was the most vulnerable.

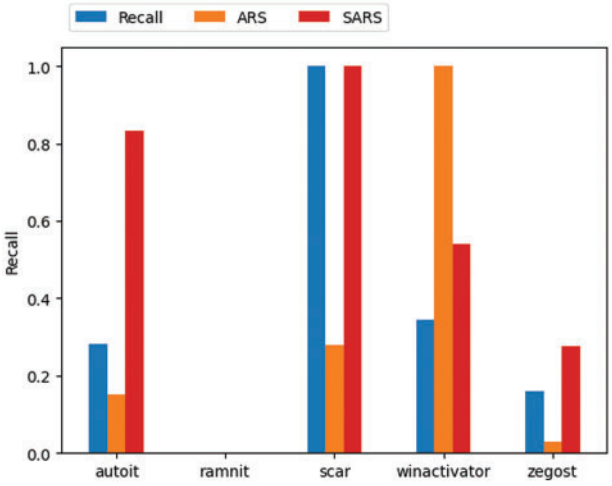

Figure 7: Comparison of robustness evaluation indicators and recall results

4.4 Improving Robustness Through Adversarial Training

Having confirmed through previous experiments that ramnit had the lowest SARS value, adversarial training is conducted to improve the vulnerable robustness of ramnit. The data added to the training dataset for adversarial training predicted malware as normal after a successful adversarial attack targeting the original data using the ramnit malware AVClass. A total of 1,107 samples were generated by attacking three times with 369 original data. Among these, the labels of 1,062 samples whose labels were changed after successful attacks are changed back to the malware versions and added to the training dataset before proceeding with training. Training the existing CNN model is performed in the same way to create a new, improved model. Learning results for the improved AI model are shown in Table 7.

Table 7: Training results for the improved AI model

Accuracy	Precision	Recall	F1 score
0.9767	0.9770	0.9767	0.9767

Next, the PDG attack was performed three times using the same parameters as before, again targeting the ramnit of the improved AI model. Comparing the change in the recall due to adversarial attacks of the AI model before and after improvement, it can be seen that resistance to adversarial attacks has increased, as shown in Fig. 8. Of the 1,107 attacks, 1,062 attacks succeeded in the AI model before improvement. However, only 664 attacks succeeded in the improved AI model, reducing the attack success rate from 95.9% to 62.3%. SARS also improved from 0.000102 to 0.000174.

After adversarial training, SARS increased by 70.59%, and the recall reduction rate improved by 64.96%. On the other hand, as shown in Fig. 9, the improvement rate of ARS is relatively very low compared to recall and SARS. This is because SARS evaluated the difficulty of adversarial attacks by comprehensively evaluating various factors. Through this experiment, it is possible to evaluate whether an AI model is vulnerable to adversarial attacks and to identify vulnerable data types based on SARS. After adversarial training, it was confirmed that SARS also increased. It is expected to contribute to establishing more objective and reliable standards for robustness evaluation based on various factors that can evaluate the difficulty of adversarial attacks.

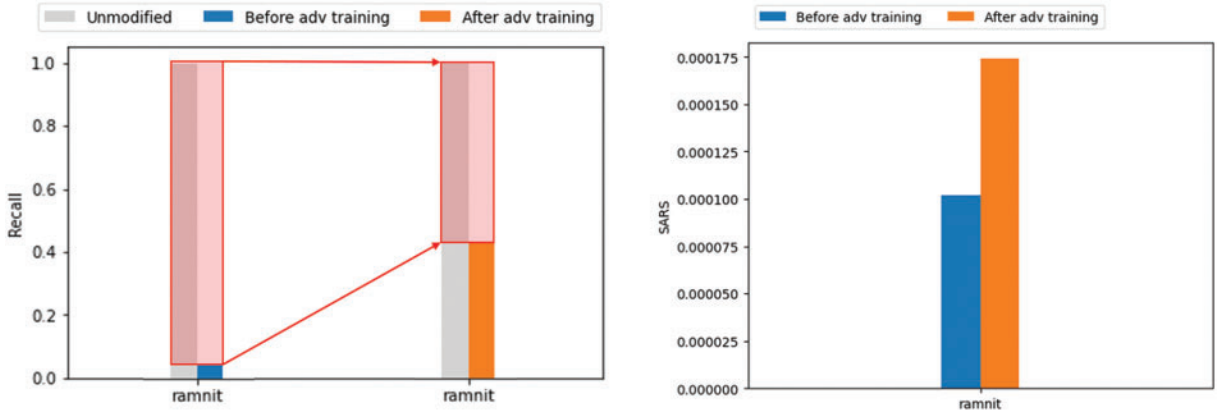

Figure 8: Robustness improvement before and after adversarial training

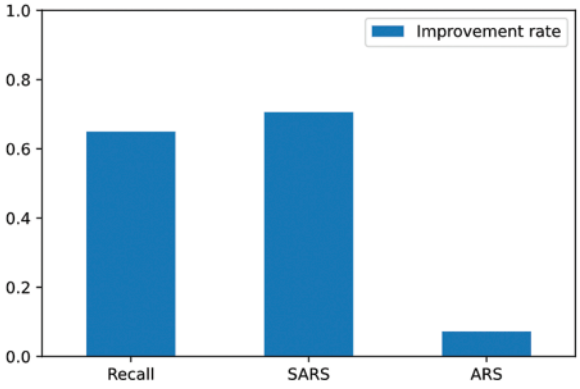

Figure 9: Improvement rate by robustness evaluation indicator after adversarial training

5 Conclusions and Future Work

As AI technology advances and its role expands, the need for trustworthy AI is underscored. Robustness issues may arise due to the nature of AI models, which are dependent on the training dataset and can be biased by its quality and representativeness; it is impossible to include in the dataset all kinds of inputs that may occur in the real environment or to be sure that the AI model can respond to adversarial attacks without errors. In the case of malware, most attacks are caused by attackers who mass-produce variants of existing malware to evade detection systems. Adversarial attacks also generate adversarial samples in the form of slight perturbations to the original data to induce AI

misclassification. Although various adversarial attack and defense methods are actively being studied, there is a lack of research on robustness evaluation indicators that serve as standards for determining whether AI models are safe and trustworthy against adversarial attacks. It is impossible to evaluate robustness against these attacks accurately with traditional AI evaluation indicators such as accuracy and recall, so robustness evaluation and improvement methods are needed. These problems manifest as limitations in trusting and using AI.

This paper proposes an adversarial attack-based robustness evaluation and improvement method for trustworthy AI. via elaborate evaluation through SARS, proposed as a measure of the robustness of AI models that cannot be evaluated with traditional AI evaluation indicators, the robustness level of each data group can be evaluated and improved through adversarial training for data groups with vulnerable robustness.

To verify the effectiveness of the method proposed for trustworthy AI development in this paper, an experiment was conducted using the 2019 KISA malware dataset. As a result of training the AI model for malware detection, the recall was calculated as high as 0.9767, but the recall was easily decreased due to adversarial attacks. This means that traditional evaluation indicators cannot be used to evaluate the robustness of AI models.

The robustness of an AI model can be evaluated through the difficulty of adversarial attacks. This paper proposes SARS to evaluate the robustness by comprehensively considering various factors such as attack success rate, perturbation size and ratio, and probability variation. Through SARS, it is possible to identify whether an AI model is vulnerable to adversarial attacks and the vulnerable data type. It is expected that improved AI models can be operated through robustness enhancement methods such as adversarial training.

Through eXplainable Artificial Intelligence (XAI), it is possible to identify features that have contributed significantly to the AI model predicting the data. The attacker could utilize this information to focus on high-contributing features and perform effective adversarial attacks by perturbating them. In future research, we will study adversarial attack methods using XAI and methods to defend against them.

Funding Statement: This work was supported by an Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. 2022-0-00089, Development of Clustering and Analysis Technology to Identify Cyber-Attack Groups Based on Life-Cycle) and MISP (Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW (2019-0-01834) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) (2019-0-01834).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] *Ethics guidelines for trustworthy AI*, EU: European Commission, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[2] A. A. Diro and N. Chilamkurti, "Distributed attack detection scheme using deep learning approach for Internet of Things," *Future Generation Computer Systems*, vol. 82, pp. 761–768, 2018.

[3] J. Singh and J. Singh, "A survey on machine learning-based malware detection in executable files," *Journal of Systems Architecture*, vol. 112, pp. 101861, 2021.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan *et al.,* "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.

[5] K. Chen, H. Zhu, L. Yan and J. Wang, "A survey on adversarial examples in deep learning," *Journal of Big Data*, vol. 2, no. 2, pp. 71–84, 2020.

[6] J. Gu, "Explainability and robustness of deep visual classification models," arXiv preprint arXiv:2301.01343, 2023.

[7] I. J. Goodfellow, J. Shlens and C. Szegedy, "Explaining and harnessing adversarial examples," in *The Int. Conf. on Learning Representations*, San Diego, United States, 2015.

[8] S. Qiu, Q. Liu, S. Zhou and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," *Applied Sciences*, vol. 9, no. 5, pp. 909, 2019.

[9] A. Hartl, M. Bachl, J. Fabini and T. Zseby, "Explainability and adversarial robustness for RNNs," in *2020 IEEE Sixth Int. Conf. on Big Data Computing Service and Applications (BigDataService)*, Oxford, United Kingdom, pp. 148–156, 2020.

[10] H. Cao, C. Wang, L. Huang, X. Cheng and H. Fu, "Adversarial DGA domain examples generation and detection," in *2020 1st Int. Conf. on Control, Robotics and Intelligent System*, New York, United States, pp. 202–206, 2020.

[11] M. Xu, T. Zhang, Z. Li, M. Liu and D. Zhang, "Towards evaluating the robustness of deep diagnostic models by adversarial attack," *Medical Image Analysis*, vol. 69, pp. 101977, 2021.

[12] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang *et al.,* "LAS-AT: Adversarial training with learnable attack strategy," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, United States, pp. 13398–13408, 2022.

[13] R. Singhal, M. Soni, S. Bhatt, M. Khorasiya and D. C. Jinwala, "Enhancing robustness of malware detection model against white box adversarial attacks," in *Distributed Computing and Intelligent Technology: 19th Int. Conf., ICDCIT 2023*, Bhubaneswar, India, pp. 181–196, 2023.

[14] F. A. Yerlikaya and S. Bahtiyar, "Data poisoning attacks against machine learning algorithms," *Expert Systems with Applications*, vol. 208, pp. 118101, 2022.

[15] T. Bekman, M. Abolfathi, H. Jafarian, A. Biswas, F. Banaei-Kashani *et al.,* "Practical black box model inversion attacks against neural nets," in *Machine Learning and Principles and Practice of Knowledge Discovery in Databases: Int. Workshops of ECML PKDD 2021, Virtual Event, September 13–17, 2021, Proc. Part II*, pp. 39–54, 2022.

[16] J. Wen, S. M. Yiu and L. C. Hui, "Defending against model inversion attack by adversarial examples," in *2021 IEEE Int. Conf. on Cyber Security and Resilience (CSR)*, Rhodes, Greece, pp. 551–556, 2021.

[17] X. Xian, M. Hong and J. Ding, "A framework for understanding model extraction attack and defense," arXiv preprint arXiv:2206.11480, 2022.

[18] B. Wu, X. Yang, S. Pan and X. Yuan, "Model extraction attacks on graph neural networks: Taxonomy and realization," in *2022 ACM on Asia Conf. on Computer and Communications Security*, Nagasaki, Japan, pp. 337–350, 2022.

[19] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, pp. 25–45, 2021.

[20] A. Kurakin, I. Goodfellow and S. Bengio, "Adversarial machine learning at scale," in *The Int. Conf. on Learning Representations*, Toulon, France, 2017.

[21] J. Gu, H. Zhao, V. Tresp and P. H. Torr, "SegPGD: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness," in *European Conf. on Computer Vision*, Tel Aviv, Israel, pp. 308–325, 2022.

[22] S. M. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, United States, pp. 2574–2582, 2016.

[23] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik *et al.,* "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symp. on Security and Privacy (EuroS&P)*, Saarbruecken, Germany, pp. 372–387, 2016.

[24] P. Chen, H. Zhang, Y. Sharma and C. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. of the 10th ACM Workshop on Artificial Intelligence and Security*, Dallas, United States, pp. 15–26, 2017.

[25] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symp. on Security and Privacy (sp)*, San Jose, United States, pp. 39–57, 2017.

[26] G. Jin, S. Shen, D. Zhang, F. Dai and Y. Zhang, "APE-GAN: Adversarial perturbation elimination with GAN," in *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, United Kingdom, pp. 3842–3846, 2019.

[27] Y. Wang, B. Dong, K. Xu, H. Piao, Y. Ding *et al.,* "A geometrical approach to evaluate the adversarial robustness of deep neural networks," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 6, pp. 14410, 2023.

[28] C. Chang, J. Hung, C. Tien, C. Tien and S. Kuo, "Evaluating robustness of AI models against adversarial attacks," in *Proc. of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, New York, United States, pp. 47–54, 2020.

[29] M. Nicolae, M. Sinn, M. N. Tran, B. Buesser, A. Rawat *et al.,* "Adversarial robustness toolbox v1.0.0," arXiv preprint arXiv:1807.01069, 2018.

[30] J. Rauber, W. Brendel and M. Bethge, "Foolbox: A python toolbox to benchmark the robustness of machine learning models," arXiv preprint arXiv:1707.04131, 2017.

[31] N. Papernot, F. Faghri, N. Carlini, I. Goodfellow, R. Feinman *et al.,* "Technical report on the CleverHans v2.1.0 adversarial examples library," arXiv preprint arXiv:1610.00768, 2018.

[32] C. Berghoff, P. Bielik, M. Neu, P. Tsankov and A. V. Twickel, "Robustness testing of ai systems: A case study for traffic sign recognition," in *Artificial Intelligence Applications and Innovations: 17th IFIP WG 12.5 Int. Conf., AIAI 2021*, Hersonissos, Greece, pp. 256–267, 2021.

[33] Korea Internet & Security Agency, KR, 2019. [Online]. Available: http://datachallenge.kr/challenge19/rd-datachallenge/malware/introduction/