RDI: An adversarial robustness evaluation metric for deep neural networks based on model statistical features

Jialei Song[1,3] **Xingquan Zuo**[*1,3] **Feiyang Wang**[1,3] **Hai Huang**[1,3] **Tianle Zhang**[2,3]

[1]Shool of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[2]Shool of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
[3]Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing, China

Abstract

Deep neural networks (DNNs) are highly susceptible to adversarial samples, raising concerns about their reliability in safety-critical tasks. Currently, methods of evaluating adversarial robustness are primarily categorized into attack-based and certified robustness evaluation approaches. The former not only relies on specific attack algorithms but also is highly time-consuming, while the latter due to its analytical nature, is typically difficult to implement for large and complex models. A few studies evaluate model robustness based on the model's decision boundary, but they suffer from low evaluation accuracy. To address the aforementioned issues, we propose a novel adversarial robustness evaluation metric, Robustness Difference Index (RDI), which is based on model statistical features. RDI draws inspiration from clustering evaluation by analyzing the intra-class and inter-class distances of feature vectors separated by the decision boundary to quantify model robustness. It is attack-independent and has high computational efficiency. Experiments show that, RDI demonstrates a stronger correlation with the gold-standard adversarial robustness metric of attack success rate (ASR). The average computation time of RDI is only 1/30 of the evaluation method based on the PGD attack. Our open-source code is available at: `https://github.com/BUPTAIOC/RDI`.

1 INTRODUCTION

With the rapid advancement of deep learning, deep neural networks (DNNs) have achieved unprecedented success in fields such as computer vision Ruiz et al. [2023], natural language processing Alberts et al. [2023], and multimodal tasks Liu et al. [2024]. However, recent studies have shown that even DNNs that perform excellently on standard benchmark datasets may experience significant performance degradation when facing with small perturbations Goodfellow et al. [2015]. These minor disturbances, especially adversarial examples generated by advanced attack algorithms, reveal the vulnerabilities of DNNs in terms of adversarial robustness. Consequently, effectively assessing the adversarial robustness of DNNs has become one of the current research hotspots.

Research on adversarial robustness assessment has primarily focused on two directions: attack-based Moosavi-Dezfooli et al. [2016], Mądry et al. [2018], Bai et al. [2023], Andriushchenko et al. [2020], Williams and Li [2023] and certified robustness evaluation methods Hein and Andriushchenko [2017], Weng et al. [2018], Cohen et al. [2019], Carlini et al. [2023]. The former generates adversarial examples through attack algorithms and measures adversarial robustness by calculating the attack success rate (ASR). However, these methods have notable limitations: on one hand, the process of generating adversarial examples is complex and resource-intensive; on the other hand, such robustness evaluation heavily depend on the selected attack methods, which may lead to biased assessments. Certified robustness employs mathematical methods to analyze the model's structure and activation functions, aiming to determine the lower bound of the perturbation that causes misclassification. However, due to its analytical nature, these methods are often difficult to implement in large and complex models, and the resulting robustness lower bounds are often overly conservative, limiting their practical applicability.

To address the above issues, [Jin et al., 2022] proposed a robustness evaluation metric, ROBY, based on the decision boundary of the model. This metric quantifies robustness by measuring the intra-class and inter-class distances of vectors separated by the model's decision boundary. The evaluation method calculates intra-class and inter-class distances of vectors, and then combines them to produce the final ad-

[*]Corresponding author

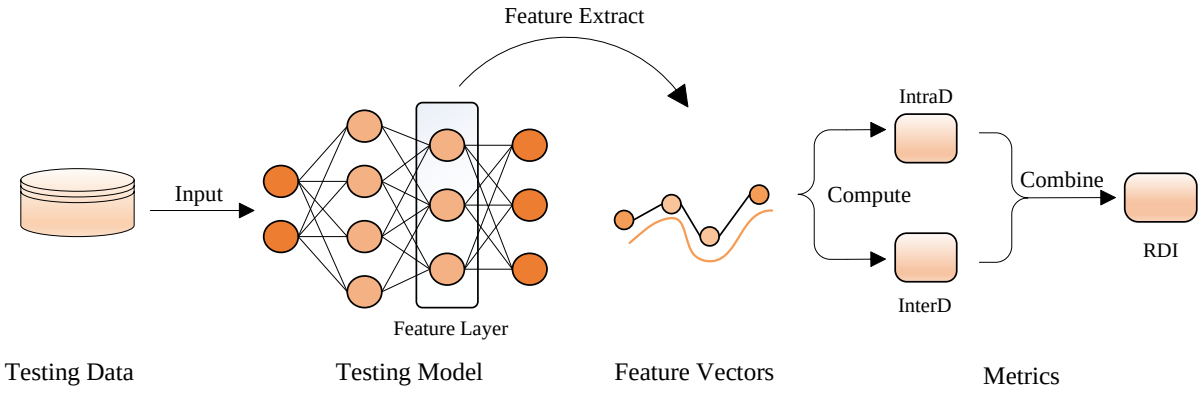

Figure 1: Framework of RDI adversarial robustness evaluation.

versarial robustness metric. Therefore, the accuracy of the robustness assessment depends critically on the distance calculation and the reasonableness of the weighting ratio used to combine them into the final metric. Such evaluation method brings the following issues: First, it overlooks the fact that the average magnitude of inter-class distance is larger than that of intra-class distance, resulting in difficulty to combine them to calculate the final ROBY score. Second, it is difficult to find an appropriate normalization strategy, making it challenging to compare ROBY scores across different models, which leads to a poor correlation with ASR. Finally, the metric calculation is time consuming and has a low efficiency. The reasons for the above issues are explained in Section A of the supplementary material.

To address the limitations of existing evaluation methods, we propose a new adversarial robustness evaluation metric for DNNs, Robustness Difference Index (RDI). RDI extracts feature vectors from the model's feature layers and analyzes the intra-class and inter-class distances of feature vectors separated by the model's decision boundary to quantify the model's robustness. Different from ROBY, RDI adopts a novel approach to calculate intra-class and inter-class distances, as well as the final robustness evaluation metric, ensuring a more reasonable combination of these two distance indicators in the final RDI. RDI does not need a normalization strategy, allowing for effective comparison of RDI values across different models and providing an accurate assessment of adversarial robustness among different models. Moreover, RDI has high computational efficiency.

RDI doesn't rely on adversarial examples or attack algorithms, making it an metric independent of attacks. Experiments demonstrate that, across various advanced DNNs (AlexNet Krizhevsky et al. [2012], ResNet50 He et al. [2016], ResNet101 He et al. [2016], DenseNet121 Huang et al. [2017], DenseNet161 Huang et al. [2017], MobileNetV2 Howard et al. [2018], ViT Dosovitskiy [2020]), RDI shows a significant correlation with the robustness gold standard metric ASR in adversarial robustness evaluation. In addition, in terms of computational efficiency, RDI outperforms both attack-based methods and ROBY. The efficiency of attack-based methods is influenced by the model architecture and attack parameter settings, while ROBY's efficiency depends on the number of classes in the dataset. However, RDI is almost unaffected by these factors. For example, the average computation time of RDI across all models on five image classification datasets is only 1/30 of the evaluation method based on PGD attack Mądry et al. [2018], and for the Tiny-ImageNet dataset Deng et al. [2009] with 200 classes, RDI's computation time is only 1/25 of ROBY's. To validate the generalizability of RDI, we conducted experiments on the speech recognition dataset SPEECHCOMMANDS Warden [2018]. The results demonstrate that this metric also exhibits excellent applicability across classification datasets of different modalities.

In summary, we aim to make the following contributions:

- We propose a model-statistical-features-based adversarial robustness evaluation metric, RDI. It is efficient, lightweight, attack-independent, and applicable to nearly all classification models.
- We use a novel approach to calculate intra-class and inter-class distances of feature vectors, and design the final robustness evaluation metric based on this. This approach ensures a reasonable combination of the two distance metrics to calculate RDI.
- We compare the RDI values across different models, as well as between natural models and adversarially trained models, and verify their strong correlation with ASR. Moreover, compared to attack-based evaluation methods and ROBY, RDI significantly improves computational efficiency.

2 RELATED WORKS

Currently, adversarial robustness evaluation methods can be broadly categorized into two types: attack-based evaluation methods and certified robustness evaluation methods. Additionally, decision-boundary-based evaluation methods are also worth attention.

2.1 ADVERSARIAL ATTACK

Adversarial attacks are the primary method for assessing the robustness of DNNs. Many works proposed attack algorithms aiming at generating l_2 or l_∞ norm constraining adversarial perturbations. In the white-box scenario, attackers have full access to the target model's internal information, including the network architecture and gradient. As a result, many works Moosavi-Dezfooli et al. [2016], Carlini and Wagner [2017], Mądry et al. [2018], Tramèr et al. [2018], Wang et al. [2022], Liu et al. [2023], Gong et al. [2024] leveraged gradient information from the model's loss function to generate adversarial examples. In contrast, in black-box scenario, attackers have limited access to the target model, only able to obtain the output probability distribution. In this case, although direct access to gradient information is unavailable, some studies Chen et al. [2017], Ilyas et al. [2018], Su et al. [2019], Zhou et al. [2022], Bai et al. [2023], Williams and Li [2023], Wang et al. [2025] estimated the gradient of the loss function indirectly by querying the model's output and then used gradient-based optimization methods to create adversarial examples. Additionally, some black-box attack methods Andriushchenko et al. [2020], Guo et al. [2019], Hong et al. [2024] do not require gradient estimation but instead optimize through specific random strategies to attack the model.

2.2 CERTIFIED ROBUSTNESS EVALUATION

Certified robustness evaluation methods primarily rely on the structure of DNNs. Through carefully designed metrics or model certifiers, these methods provide provable lower bounds on the adversarial distance required to cause misclassification.

[Szegedy, 2013] showed that DNNs' input-output mapping often exhibits discontinuity by calculating the global Lipschitz constant for each layer, offering an initial explanation of DNN robustness. Building on this, [Hein and Andriushchenko, 2017] provided a lower bound for model robustness using local Lipschitz continuity conditions. Furthermore, [Weng et al., 2018] introduced CLEVER, a metric to evaluate the minimum perturbation needed for effective adversarial examples. [Cohen et al., 2019] provided verifiable adversarial robustness guarantees for classifiers through the method of random smoothing. [Carlini et al., 2023] proposed a more efficient boundary algorithm by integrating pre-trained denoising diffusion models with high-precision classifiers. [Hammoudeh and Lowd, 2024] applied certified robustness to adversarial defense.

2.3 EVALUATION METHODS BASED ON MODEL DECISION BOUNDARY

Besides above two types of robustness evaluation methods, a few studies have shown that the decision boundary is closely related to a model's classification performance on adversarial examples and is a key factor in assessing model robustness Ustun et al. [2019]. Based on this finding, nearly all decision-boundary-based methods focus on exploring the relationship between the model's decision boundary and adversarial robustness.

[Yousefzadeh and O'Leary, 2019] calculated precise points on the model's decision boundary and provided tools to study the surface defining the boundary. [He et al., 2018] further investigated the decision boundary characteristics of model for both adversarial and natural inputs. [Yang et al., 2020] introduced the concept of boundary thickness for evaluating model robustness. However, these decision-boundary-based methods still rely on the generation of adversarial examples. To address this issue, [Jin et al., 2022] proposed a robustness evaluation metric, ROBY, which does not depend on adversarial examples or attack algorithms.

3 PROPOSED EVALUATION METHOD FOR ADVERSARIAL ROBUSTNESS

3.1 OVERVIEW

The framework of the proposed metric RDI is shown in Figure 1. First, the testing data is input into the model to be evaluated, and the features of the model's embedding space are extracted to form feature vectors. Then, for each class label of the data, compute the intra-class distance (IntraD) and inter-class distance (InterD) of the feature vectors. Finally, based on the IntraD and InterD, the adversarial robustness evaluation metric RDI is obtained.

Our method transforms the problem of calculating the distance from test data to the decision boundary in the model's embedding space into the computation of intra-class and inter-class distances of the feature vectors. Specifically, the smaller intra-class distance and the larger inter-class distance, the stronger the model's resistance to adversarial attacks. Figure 2 illustrates the principles behind the RDI calculation. Figure 2(a) presents a two-dimensional visualization of the model's decision space, where solid circles in different colors represent samples from various classes, and the boundary regions indicate the model's decision boundaries. $\bar{E}_k(k \in [1,4])$ represents the center of the feature vectors for each class, while $\bar{E}_0$ denotes the center of all

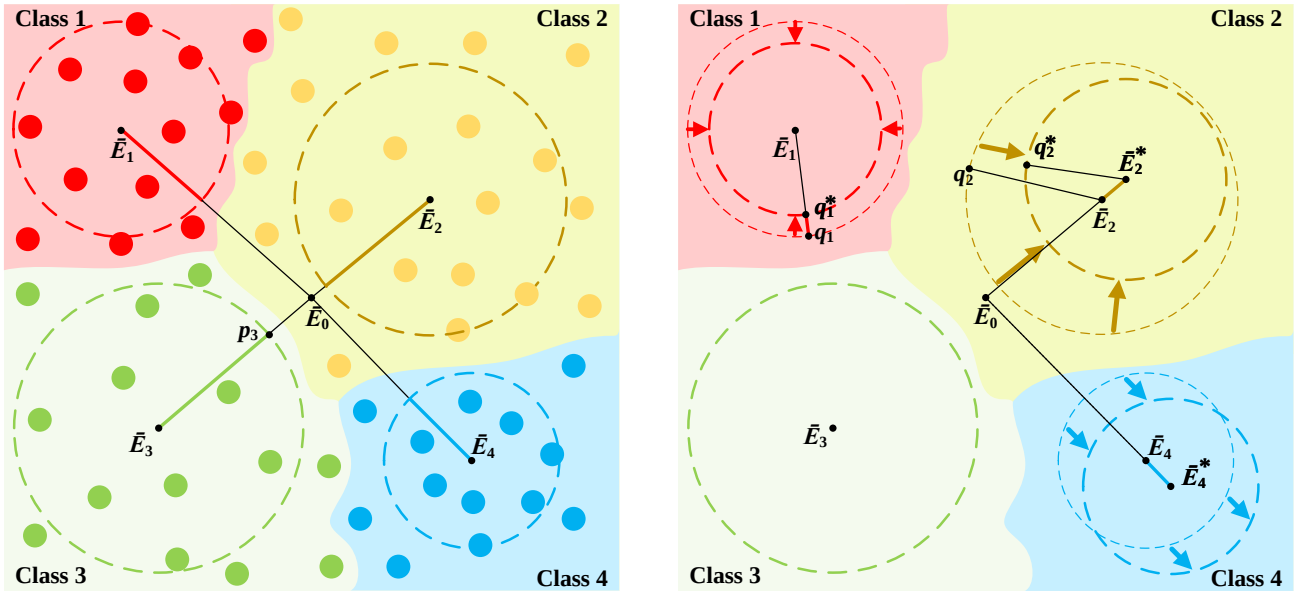

Figure 2: RDI principle schematic.

sample feature vectors. $\bar{E}_k$ and $\bar{E}_0$ are computed in subsections3.4 and 3.5, respectively. Figure 2(b) demonstrates how the model's robustness affects the RDI value. More details will be provided in the subsequent sections.

3.2 FEATURE VECTOR EXTRACTION

Since the feature (penultimate) layer of the model can quickly capture and contain sufficient feature information, we use its output as the feature vectors representation of the testing data in the model's embedding space. Therefore, the dimensionality of each feature vector equals the number of model classes. For example, for an input sample x, we use a K-dimensional embedding $E(x) : (e_1, ..., e_K)$ as its vector representation in the model's embedding space. Suppose the model classifies sample x as $k \in R^K$, we denote the feature vector of this sample as $E_k(x_i)$.

3.3 DISTANCE FUNCTION

We use the Euclidean distance (l_2 norm distance) to represent the distance between two K-dimensional vectors $E(x_i) : (e_{i1}, ..., e_{iK})$ and $E(x_j) : (e_{j1}, ..., e_{jK})$ in the model's embedding space, denoted as $||E(x_i), E(x_j)||_2$:

$$||E(x_i), E(x_j)||_2 = (\sum_{k=1}^{K}(e_{ik} - e_{jk})^2)^{1/2}. \quad (1)$$

3.4 INTRAD METRIC

The intra-class distance (IntraD) of the feature vectors denotes the compactness of samples classified by the model as the same class.

Since the feature vector center of each class reflects the overall characteristics of that class, we first define the center to facilitate the subsequent calculation of IntraD. For all samples classified by the model as class k (denoted as set N_k), the mean feature vectors of this class in the model's embedding space is defined as the center of the class's feature vectors, denoted as vector $\bar{E}_k$, e.g., $\bar{E}_k$ ($k \in [1, 4]$) in Figure 2(a). The radius of each dashed circle in the figure represents the average intra-class distance of all sample feature vectors within that class.

$$\bar{E}_k = \frac{1}{|N_k|} \sum_{x_i \in N_k} E_k(x_i). \quad (2)$$

where $|N_k|$ is the total number of samples classified as class k by the model, x_i is the i-th sample in N_k, and $E_k(x_i)$ denotes the feature vector of sample x_i classified as class k by the model. The IntraD metric is the average distance of the feature vectors of samples within the same class to the center of the class's feature vectors:

$$IntraD_k = \frac{1}{|N_k|} \sum_{i=1}^{|N_k|} ||E_k(x_i), \bar{E}_k||_2. \quad (3)$$

$$IntraD = \frac{1}{K} \sum_{i=1}^{K} IntraD_k. \quad (4)$$

where $IntraD_k$ represents the intra-class distance of the feature vectors for class k, e.g., the distance between $\bar{E}_3$ and p_3 in Figure 2(a), and K is the total number of classes. Equations (3) and (4) ensure that IntraD is always positive.

3.5 INTERD METRIC

The inter-class distance (InterD) of the feature vectors denotes the separation among samples classified by the model as different classes in the model's embedding space.

Since the feature vector centers of all classes reflect the overall distribution features of the samples, to calculate InterD, we first compute the mean of the feature vectors centers for each class as the center of all class feature vectors, denoted as vector $\bar{E}_0$, e.g., $\bar{E}_0$ in Figure 2(a).

$$\bar{E}_0 = \frac{1}{K}\sum_{k=1}^{K}\bar{E}_k. \tag{5}$$

The InterD metric is the average distance from each class's feature vectors center to the center of all class feature vectors, e.g., the average distance between $\bar{E}_0$ and $\bar{E}_k$ ($k \in [1, 4]$) in Figure 2(a).

$$InterD = \frac{1}{K}\sum_{i=1}^{K}||\bar{E}_k, \bar{E}_0||_2. \tag{6}$$

Equation (6) ensures that InterD is always positive.

3.6 RDI METRIC

Generally, models with higher robustness exhibit smaller intra-class distances and larger inter-class distances. Therefore, we define RDI to ensure a positive correlation between RDI and model robustness:

$$RDI = \frac{InterD - IntraD}{\max(InterD, IntraD)}. \tag{7}$$

Equation (7) ensures that RDI is in the range $[-1, 1]$.

A higher RDI value typically indicates better separation among samples from different classes and stronger grouping of samples within the same class in the model's embedding space. This results in each sample being farther from the decision boundary, meaning that larger adversarial perturbations are needed to shift the samples into adversarial regions. As a result, the model exhibits stronger robustness. Figure 2(b) visually illustrates the relationship between the RDI value and model robustness, revealing 3 cases leading to an increase in the RDI value. For class 1 in Figure 2(b), the inter-class distance remains unchanged while the intra-class distance decreases (from $[\bar{E}_1, q_1]$ to $[\bar{E}_1, q_1^*]$). For class 4, the intra-class distance remains unchanged while the inter-class distance increases (from $[\bar{E}_0, \bar{E}_4]$ to $[\bar{E}_0, \bar{E}_4^*]$). For class 2, both the intra-class distance decreases and the inter-class distance increases (the intra-class distance decreases from $[\bar{E}_2, q_2]$ to $[\bar{E}_2^*, q_2^*]$, and the inter-class distance increases from $[\bar{E}_0, \bar{E}_2]$ to $[\bar{E}_0, \bar{E}_2^*]$). Apparently, all the 3 cases increase the average distance from samples to the decision boundary, resulting in a higher RDI value. In other words, an increase in RDI indicates a larger average distance from the samples to the decision boundary, demonstrating greater robustness. Models with higher RDI values require larger perturbation to generate adversarial samples, demonstrating greater robustness.

The calculation of RDI only requires natural samples, the property makes it an attack-independent metric. Note that the model evaluation uses samples with predicted label, not the true label. The reason is that the predicted labels reflect the model's partitioning of the data, representing the decision boundary it has truly learned, while the true labels do not capture the model's decision-making behavior. Robustness evaluation is meaningful only for models with a certain level of classification accuracy. In other words, it makes no sense to evaluate the robustness of a model with a very low classification accuracy.

The pseudocode of RDI calculation is presented in Algorithm 1 of the supplementary material. First, feature vectors of the testing samples are extracted. Then, the feature centers for each class and for all samples are calculated, followed by the computation of intra-class and inter-class distances. Finally, the RDI metric is computed from these distances.

4 EXPERIMENTS

We conducted four experiments to verify the effectiveness and efficiency of the RDI, focusing on the following research questions: (1) Can RDI effectively address the issues of existing evaluation methods and accurately assess adversarial robustness across different models? (2) Can RDI evaluate the robustness of adversarial trained models? (3) Does RDI demonstrate high computational efficiency? Our experiments are conducted on a server with an Intel Xeon Gold 6330 CPU, NVIDIA 4090 GPUs using PyTorch 1.13.0 on the Python 3.9.0 platform.

4.1 DATASETS AND MODELS

For image classification tasks, we conducted experiments on MNIST LeCun et al. [1998], Fashion-MNIST Xiao et al. [2017], CIFAR10 Krizhevsky et al. [2009], CIFAR100 Krizhevsky et al. [2009], and Tiny-ImageNet Deng et al. [2009]. For the first four datasets, we select six DNNs: AlexNet Krizhevsky et al. [2012], ResNet50 He et al. [2016], ResNet101 He et al. [2016], DenseNet121 Huang et al. [2017], DenseNet161 Huang et al. [2017], and MobileNetV2 Howard et al. [2018]. For the Tiny-ImageNet dataset, to further validate the effectiveness of RDI in assessing the robustness of state-of-the-art models, we additionally included the ViT Dosovitskiy [2020] model, as ViT is not well-suited for smaller datasets.

To show the broad applicability of RDI, we applied it to a speech recognition task. We conducted exper-

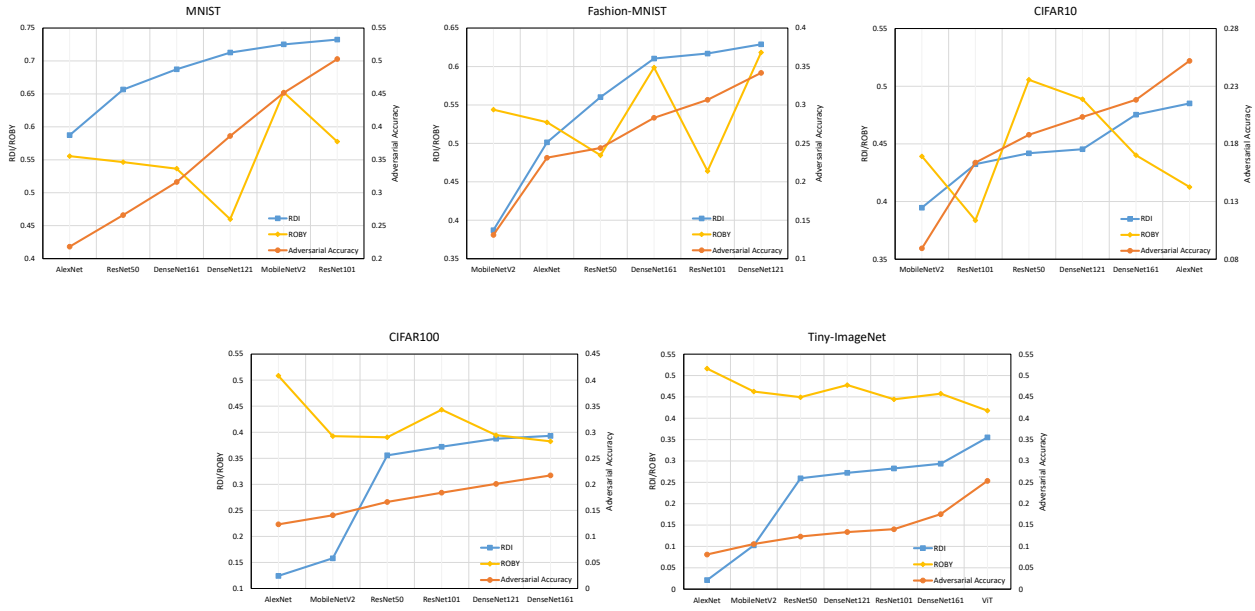

Figure 3: The relationship among RDI, ROBY, and Adversarial Accuracy across different image classification models.

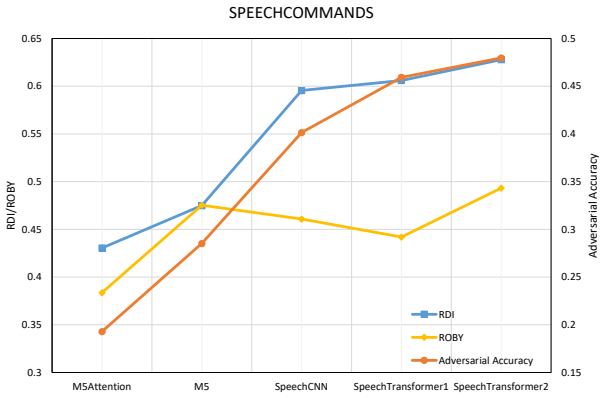

Figure 4: The relationship among RDI, ROBY, and Adversarial Accuracy across different speech recognition models.

iments on SPEECHCOMMANDS Warden [2018] and evaluated the robustness metrics on M5 model Arik et al. [2017], M5 model with an attention mechanism (M5Attention) Arik et al. [2017], CNN model based on DeepSpeech (SpeechCNN) Abdel-Hamid et al. [2014], and two transformer-based speech recognition models (SpeechTransformer) Dong et al. [2018].

4.2 ATTACK METHODS

For image classification models, we used three white-box attack methods (PGD Mądry et al. [2018], RFGSM Tramèr et al. [2018] and C&W Carlini and Wagner [2017]) and one black-box attack method (Square Attack Andriushchenko et al. [2020]). For the MNIST and Fashion-MNIST, PGD and RFGSM ran for 40 iterations with a step size of 0.01 and perturbation size $\epsilon = 0.3$, C&W ran for 50 iterations with a step size of 0.01 and the confidence parameter $c = 1$, while Square Attack ran for 5000 iterations with $\epsilon = 0.3$. For the CIFAR10, PGD and RFGSM ran for 10 iterations with a step size of 0.0025 and perturbation size $\epsilon = 0.1$, C&W ran for 10 iterations with a step size of 0.01 and the confidence parameter $c = 0.5$, and Square Attack ran for 200 iterations with $\epsilon = 0.1$. For the CIFAR100 and Tiny-ImageNet, PGD and RFGSM ran for 10 iterations with a step size of 0.001 and perturbation size $\epsilon = 0.01$, C&W ran for 10 iterations with a step size of 0.01 and the confidence parameter $c = 0.1$, while Square Attack ran for 200 iterations with $\epsilon = 0.01$. For speech recognition models, we employed the adversarial audio attack method SirenAttack Du et al. [2020], which ran for 50 iterations with a step size of 0.05 and perturbation size $\epsilon = 0.5$.

4.3 BASELINE METRIC

We used ASR, Adversarial Accuracy, and ROBY Jin et al. [2022] as baseline comparisons with RDI.

ASR and Adversarial Accuracy:

$$ASR = N_{err}/N. \tag{8}$$

$$Adversarial\ Accuracy = 1 - ASR. \tag{9}$$

where N_{err} represents the number of classification errors on adversarial examples, and N denotes the total number of adversarial samples. Since a larger RDI value indicates a

more robust model, it is inversely proportional to ASR. Thus we introduced Adversarial Accuracy as one of the baseline metrics for a more intuitive comparison.

ROBY: ROBY is another metric which quantifies the model's robustness via the intra-class and inter-class distances of vectors, the detailed calculation method is provided in Section A of the supplementary material. A smaller ROBY value indicates higher model robustness.

4.4 RESULTS OF RDI AND BASELINE METRICS FOR EVALUATING MODEL ROBUSTNESS

To validate the precision of RDI in evaluating model adversarial robustness and highlight the issues of ROBY, we conducted experiments across multiple datasets. Specifically, we reported the average Adversarial Accuracy, ROBY, and RDI metrics for five image classification datasets under four different attack methods as well as a speech recognition dataset under SirenAttack. The relationships among these metrics are illustrated in Figure 3 and 4.

For each dataset in the figures, the models are arranged from left to right in increasing order of Adversarial Accuracy, resulting in an increasing curve for Adversarial Accuracy. Notably, RDI also forms an increasing curve. The similar trend of both metrics indicates a significant positive correlation between RDI and Adversarial Accuracy. Additionally, for more challenging datasets (such as CIFAR100 and Tiny-ImageNet), the RDI values of all models are generally lower. This phenomenon aligns with our expectations for the RDI metric, as models tend to learn ambiguous decision boundaries on challenging datasets, leading to degraded classification performance. Moreover, such ambiguous boundary result in increased intra-class distances and reduced inter-class distances in the model's embedding space, making the model more susceptible to attack. However, for models with reasonable classification capability, the inter-class distances is typically greater than the intra-class distances. Consequently, we did not observe any instances where the RDI values fell below zero in our experiments.

ROBY should be negatively correlated with Adversarial Accuracy. However, as shown in Figure 3 and 4, ROBY exhibits clear outliers in each dataset, and its trend does not strictly follow an inverse relationship with Adversarial Accuracy. This result indicates that ROBY exhibits limited accuracy in evaluting model robustness.

Finally, we presented a more detailed evaluation results of model robustness under each dataset in Table 3 and Table 4 in the supplementary material. For image classification datasets, we reported the accuracy, ASR under different attacks, average ASR, average Adversarial Accuracy, ROBY, and RDI values for each model, arranged in descending order of average ASR for each dataset. For the speech recognition dataset, models are ranked by descending ASR under SirenAttack, with reported metrics including accuracy, ROBY, RDI, ASR under SirenAttack, and corresponding Adversarial Accuracy.

Table 1: The computation time of RDI and ROBY.

Dataset	Category	Time (s)	
		RDI	ROBY
MNIST	10	3.42	3.95
Fashion-MNIST	10	3.38	4.02
CIFAR10	10	3.43	3.96
CIFAR100	100	4.12	19.84
Tiny-ImageNet	200	8.82	220.22

Table 2: The computation time for evaluating model robustness using RDI metric versus PGD attack method.

Model	Time (s)	
	RDI	PGD Attack
AlexNet	3.38	11.50
MobileNetV2	3.05	65.77
ResNet50	3.61	108.52
DenseNet121	4.63	138.66
ResNet101	5.36	162.91
DenseNet161	6.57	217.96
Avg	4.43	133.39

4.5 RESULTS OF RDI METRIC FOR NATURAL AND ADVERSARIAL TRAINING MODELS

Previous studies Guo et al. [2019], Szegedy [2013] have shown that adversarial training increases the distance between the training (testing) data and the decision boundary. In our experiments, we adopted the widely adopted PGD adversarial training method Mądry et al. [2018] for models adversarial training. We evaluated the effectiveness of the RDI in assessing the adversarial robustness of both natural models and adversarially trained models on five image classification datasets. The experimental results are shown in Figure 5.

As observed, in all five datasets, almost all adversarially trained models have higher RDI values than their natural models. This result further validates the effectiveness of the robustness evaluation metric RDI. We believe that the higher adversarial robustness of adversarially trained models is due to the increased distance between samples of different classes and the decision boundary, while the distribution of samples within the same class becomes more concentrated. As a result, the model becomes less susceptible to attack. This provides a reasonable explanation for the increase in the values of RDI after adversarial training.

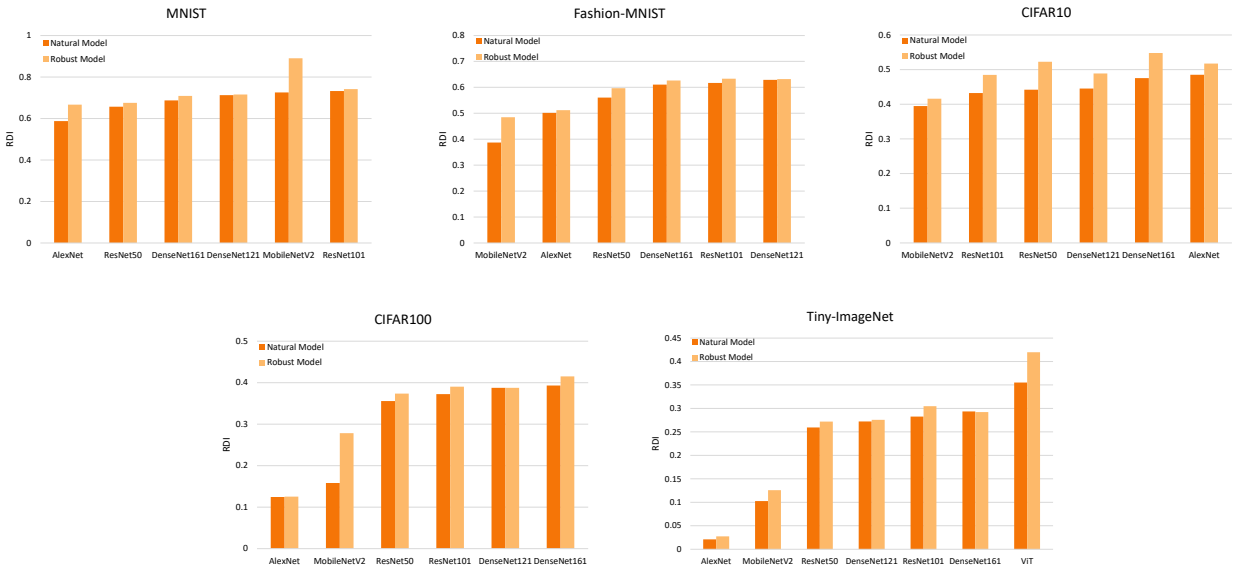

Figure 5: Comparison of RDI between natural and adversarial training models.

4.6 COMPUTATION TIME OF RDI VS. ROBY AND ATTACK-BASED EVALUATION METHOD

This section analyzes the time overhead of RDI, ROBY, and the PGD attack-based method to assess RDI's efficiency. Table 1 shows the computational cost of RDI and ROBY metrics on different image classification datasets. The reported times represent the average computation times for different models on each dataset, covering the entire process from model inference, feature vector extraction, and final metric calculation. Table 2 presents the time overhead of RDI compared to the PGD attack-based evaluation method for different models in the experiment. The reported times are the average computation times of each model across different datasets. Since the ViT model is used only on the Tiny-ImageNet dataset, it is excluded from Table 2 to ensure fairness in the averaging process across datasets.

As shown in Table 1, the computational efficiency of ROBY significantly decreases as the number of classes in the dataset increases, while the efficiency of RDI is nearly unaffected by changes in the number of classes. In datasets with 10 classes, the computational efficiency of ROBY and RDI is comparable; however, as the number of classes increases, the efficiency advantage of RDI becomes more pronounced. For example, for the 100-class dataset, the computational efficiency of RDI is approximately 5 times that of ROBY, and for the 200-class dataset, RDI's efficiency is 25 times that of ROBY. Table 2 shows that the computational efficiency of attack-based methods significantly decreases as the model complexity increases. The average computation time of RDI across all models (except ViT) on five datasets is only 1/30 of that for the PGD attack-based evaluation method. Additionally, the evaluation efficiency of attack-based methods is particularly affected by attack parameters, especially the number of iterations.

The computation of RDI only requires natural samples and does not require generating adversarial examples, significantly reducing the computational time compared to attack-based evaluation methods.

5 CONCLUSION AND FUTURE WORK

This paper proposes a sample-clustering-features-based adversarial robustness evaluation metric, RDI. RDI transforms the evaluation of model robustness into the calculation of intra-class and inter-class distances of sample feature vectors in the model's embedding space, completely eliminating the reliance on adversarial examples and making it an attack-independent metric. Experimental results demonstrate that RDI shows a high correlation with Adversarial Accuracy across different models and accurately reflects the adversarial robustness differences between natural models and adversarially trained models. Additionally, RDI offers significant advantages in computational efficiency.

We believe RDI has broad applicability for various classification tasks. In future research, we will explore the potential applications of RDI in more classification scenarios and investigate its value in other robustness-related problems. For example, RDI's efficiency enables its direct integration into the model training process, allowing for effective enhancement of model robustness without relying on adversarial training methods. This approach not only significantly reduces the overhead of adversarial training but also provides

an efficient and innovative solution for robustness research. Therefore, we believe that the RDI will play a significant role in both model robustness research and applications.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NO.62272117). The authors express their gratitude to the reviewers for their insightful and constructive feedback on the initial version of this paper.

References

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on audio, speech, and language processing*, 22(10): 1533–1545, 2014.

Ian L Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, and Ali Afshar-Oromieh. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? *European journal of Nuclear Medicine and Molecular Imaging*, 50 (6):1549–1552, 2023.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. *Proceedings of the European Conference on Computer Vision*, pages 484–501, 2020.

Sercan O Arik, Markus Kliegl, Rewon Child, Joel Hestness, Andrew Gibiansky, Chris Fougner, Ryan Prenger, and Adam Coates. Convolutional recurrent neural networks for small-footprint keyword spotting. *arXiv preprint arXiv:1703.05390*, 2017.

Yang Bai, Yisen Wang, Yuyuan Zeng, Yong Jiang, and Shu-Tao Xia. Query efficient black-box adversarial attack on deep neural networks. *Pattern Recognition*, 133:109037, 2023.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *Proceedings of the IEEE Symposium on Security and Privacy*, pages 39–57, 2017.

Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (Certified!!) adversarial robustness for free! *Proceedings of the International Conference on Learning Representations*, 2023.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, pages 15–26, 2017.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. *Proceedings of the Conference on International Conference on Machine Learning*, pages 1310–1320, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

Linhao Dong, Shuang Xu, and Bo Xu. Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888, 2018.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Tianyu Du, Shouling Ji, Jinfeng Li, Qinchen Gu, Ting Wang, and Raheem Beyah. SirenAttack: Generating Adversarial Audio for End-to-End Acoustic Systems. *Proceedings of the 15th ACM Asia Conference on Computer and Communications Security*, (13):357–369, 2020.

Yunpeng Gong, Zhun Zhong, Yansong Qu, Zhiming Luo, Rongrong Ji, and Min Jiang. Cross-modality perturbation synergy attack for person re-identification. *arXiv preprint arXiv:2401.10090*, 2024.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *Proceedings of the International Conference on Learning Representations*, 2015.

Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. *Proceedings of the International Conference on Machine Learning*, pages 2484–2493, 2019.

Zayd Hammoudeh and Daniel Lowd. Provable Robustness against a Union of L_0 Adversarial Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19): 21134–21142, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

Warren He, Bo Li, and Dawn Song. Decision boundary analysis of adversarial examples. *Proceedings of the International Conference on Learning Representations*, 2018.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Proceedings of the Conference on Neural Information Processing Systems*, pages 2263–2273, 2017.

Hanbin Hong, Xinyu Zhang, Binghui Wang, Zhongjie Ba, and Yuan Hong. Certifiable Black-Box Attacks with Randomized Adversarial Examples: Breaking Defenses with Provable Confidence. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 600–614, 2024.

Andrew Howard, Andrey Zhmoginov, Liang-Chieh Chen, Mark Sandler, and Menglong Zhu. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. *Proceedings of the International Conference on Machine Learning*, pages 2137–2146, 2018.

Haibo Jin, Jinyin Chen, Haibin Zheng, Zhen Wang, Jun Xiao, Shanqing Yu, and Zhaoyan Ming. ROBY: Evaluating the adversarial robustness of a deep model by its decision boundaries. *Information Sciences*, 587:97–122, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Proceedings of the Conference on Neural Information Processing Systems*, pages 1097–1105, 2012.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Boundary thickness and robustness in learning models. *IEEE*, 86(11):2278–2324, 1998.

Hongying Liu, Zhijin Ge, Zhenyu Zhou, Fanhua Shang, Yuanyuan Liu, and Licheng Jiao. Gradient correction for white-box adversarial attacks. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12):18419–18430, 2023.

Xinyang Liu, Dongsheng Wang, Bowei Fang, Miaoge Li, Yishi Xu, Zhibin Duan, Bo Chen, and Mingyuan Zhou. Patch-Prompt Aligned Bayesian Prompt Tuning for Vision-Language Models. *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2024.

Aleksander Mądry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *Proceedings of the International Conference on Learning Representations*, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *Proceedings of the International Conference on Learning Representations*, 2018.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 33:10–19, 2019.

Feiyang Wang, Xingquan Zuo, Hai Huang, and Gang Chen. ADBA: Approximation Decision Boundary Approach for Black-Box Adversarial Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7628–7636, 2025.

Yixiang Wang, Jiqiang Liu, Xiaolin Chang, Jelena Mišić, and Vojislav B Mišić. IWA: integrated gradient-based white-box attacks for fooling deep neural networks. *International Journal of Intelligent Systems*, 37(7):4253–4276, 2022.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *Proceedings of the International Conference on Learning Representations*, 2018.

Phoenix Neale Williams and Ke Li. Black-box sparse adversarial attack via multi-objective optimisation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12291–12301, 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Yaoqing Yang, Rajiv Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. Boundary thickness and robustness in learning models. *Proceedings of the Conference on Neural Information Processing Systems*, 33: 6223–6234, 2020.

Roozbeh Yousefzadeh and Dianne P O'Leary. Investigating decision boundaries of trained neural networks. *arXiv preprint arXiv:1908.02802*, 2019.

Linjun Zhou, Peng Cui, Xingxuan Zhang, Yinan Jiang, and Shiqiang Yang. Adversarial eigen attack on black-box models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15254–15262, 2022.

RDI: An adversarial robustness evaluation metric for deep neural networks based on model statistical features (Supplementary Material)

Jialei Song[1,3] **Xingquan Zuo**[*1,3] **Feiyang Wang**[1,3] **Hai Huang**[1,3] **Tianle Zhang**[2,3]

[1]Shool of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China
[2]Shool of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, China
[3]Key Laboratory of Trustworthy Distributed Computing and Service, Ministry of Education, Beijing, China

A SPECIFIC CAUSES OF PROBLEMS WITH ROBY

The formulas for ROBY are as follows:

$$FSA_k = \frac{1}{|N_k|}\text{norm}(\sum_{i=1}^{|N_k|} dist(f_{x_i,k}, f_{c_k})).$$

$$FSD_{k,k+1} = \text{norm}(dist(f_{c_k}, f_{c_{k+1}})).$$

$$ROBY_{k,k+1} = FSA_k + FSA_{k+1} - FSD_{k,k+1}.$$

$$ROBY = \frac{\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\text{norm}(ROBY_{ij})}{K(K-1)/2}.$$

where $f_{x_i,k}$ represents the feature vector of a sample classified as class k, f_{c_k} denotes the center of the feature vectors for class k, FSA_k represents the intra-class distance for class k, $FSD_{k,k+1}$ denotes the inter-class distance between classes k and $k+1$, and $ROBY_{k,k+1}$ represents the ROBY metric value between classes k and $k+1$, norm$(\cdot)$ is the min-max normalization function, and dist$(\cdot)$ represents the l_2 norm distance. From the above equation, it can be seen that a smaller ROBY value indicates higher model robustness, and this metric is positively correlated with ASR.

Based on these formulas, we analyze the three main issues with the ROBY metric. First, the intra-class distance (FSA) of feature vectors in ROBY is measured by calculating the distance from the data of the same class to its class center, reflecting the "dispersion of data from the center." In contrast, the inter-class distance (FSD) is directly computed as the distance between the centers of different class data. The inconsistency in the measurement methods of these two distances leads to an unequal weight distribution when combining them into the final ROBY metric, resulting in biased evaluation outcomes. Second, ROBY uses the normalization function norm$(\cdot)$ for calculating FSA, FSD, and itself, employing min-max normalization. As shown in the formulas, this normalization is performed within the independent solution space of each model, making it impossible to directly compare the normalized ROBY values across different models to assess their robustness. Moreover, finding a suitable normalization strategy for this calculation is challenging. Finally, due to the many unnecessary normalization operations, ROBY also suffers from inefficiency in terms of computational cost.

*Corresponding author

B THE PSEUDOCODE OF RDI CALCULATION

Algorithm 1 RDI Calculation

Input: The number of classes K, the samples with K classes.
Output: RDI value

Initialization: $center\_list \leftarrow \{\}$, $IntraD\_list \leftarrow \{\}$
// feature vectors extract
// $feature\_vectors[i][j]$ represents the feature vector of sample j classified by the model as category i
$feature\_vectors \leftarrow$ model($samples$)
for $k \leftarrow 1$ **to** K **do**
    // N_k represents the set of feature vectors of all samples classified as k by the model
    $N_k \leftarrow feature\_vectors[k]$
end for
for $k \leftarrow 1$ **to** K **do**
    // calculate the feature centers for each class
    $\bar{E}_k \leftarrow 0$
    for $i \leftarrow 1$ **to** $|N_k|$ **do**
        $\bar{E}_k \leftarrow \bar{E}_k + N_k[i]$
    end for
    $\bar{E}_k \leftarrow \bar{E}_k / |N_k|$
    $center\_list[k] \leftarrow \bar{E}_k$
    // calculate the intra-class distance of class k
    $IntraD_k \leftarrow 0$
    for $i \leftarrow 1$ **to** $|N_k|$ **do**
        $IntraD_k \leftarrow IntraD_k + ||N_k[i], \bar{E}_k||_2$
    end for
    $IntraD\_list[k] \leftarrow IntraD_k / |N_k|$
end for
// calculate the overall inter-class distance
$IntraD \leftarrow mean(IntraD\_list)$
// calculate the overall feature center for all samples
$\bar{E}_0 \leftarrow mean(center\_list)$
// calculate the inter-class distance
$InterD \leftarrow 0$
for $k \leftarrow 1$ **to** K **do**
    $InterD \leftarrow InterD + ||center\_list[k], \bar{E}_0||_2$
end for
$InterD \leftarrow InterD / K$
// calculate RDI
$RDI \leftarrow (InterD - IntraD) / \max(InterD, IntraD)$
return RDI

C EVALUATION RESULTS OF MODEL ROBUSTNESS

Table 3: Robustness evaluation of image classification models based on ASR, Adversarial Accuracy (AA), RDI and ROBY.

Dataset	model	ACC	ROBY	RDI	ASR (PGD)	ASR (RFGSM)	ASR (Square Attack)	ASR (C&W)	ASR (avg)	AA (avg)
MNIST	AlexNet	0.9916	0.5555	0.5874	0.7981	0.8064	0.8121	0.7108	0.7819	0.2181
	ResNet50	0.9895	0.5464	0.6566	0.7408	0.7461	0.7502	0.6988	0.7340	0.2660
	DenseNet161	0.9906	0.5365	0.6873	0.6931	0.6967	0.7071	0.6382	0.6838	0.3162
	DenseNet121	0.9911	0.4598	0.7127	0.6379	0.6447	0.6431	0.5304	0.6140	0.3860
	MobileNetV2	0.9851	0.6523	0.7251	0.5621	0.5738	0.5813	0.4768	0.5485	0.4515
	ResNet101	0.9868	0.5775	0.7324	0.5116	0.5250	0.5247	0.4274	0.4972	0.5028
Fashion-MNIST	MobileNetV2	0.9124	0.5439	0.3872	0.9026	0.9146	0.9098	0.7491	0.8690	0.1310
	AlexNet	0.9125	0.5275	0.5014	0.7966	0.8092	0.8043	0.6642	0.7686	0.2314
	ResNet50	0.9121	0.4846	0.5603	0.7743	0.7841	0.7768	0.6886	0.7560	0.2440
	DenseNet161	0.8877	0.5987	0.6104	0.7362	0.7413	0.7388	0.6503	0.7167	0.2833
	ResNet101	0.9123	0.4640	0.6169	0.7081	0.7167	0.7227	0.6258	0.6933	0.3067
	DenseNet121	0.9040	0.6183	0.6289	0.6713	0.6803	0.6894	0.5923	0.6583	0.3417
CIFAR10	MobileNetV2	0.8134	0.4392	0.3947	0.9064	0.9072	0.9224	0.9063	0.9106	0.0894
	ResNet101	0.8340	0.3837	0.4324	0.8208	0.8237	0.8673	0.8330	0.8362	0.1638
	ResNet50	0.8422	0.5056	0.4419	0.7974	0.7986	0.8394	0.8131	0.8121	0.1879
	DenseNet121	0.8401	0.4888	0.4454	0.7832	0.7856	0.8286	0.7895	0.7967	0.2033
	DenseNet161	0.8214	0.4402	0.4754	0.7622	0.7652	0.8301	0.7695	0.7818	0.2182
	AlexNet	0.8083	0.4126	0.4851	0.7270	0.7299	0.7879	0.7471	0.7480	0.2520
CIFAR100	AlexNet	0.6545	0.5084	0.1242	0.9254	0.9305	0.8487	0.8024	0.8768	0.1232
	MobileNetV2	0.6651	0.3925	0.1580	0.8821	0.8890	0.8547	0.8115	0.8593	0.1407
	ResNet50	0.6928	0.3904	0.3557	0.8607	0.8652	0.8282	0.7812	0.8338	0.1662
	ResNet101	0.7032	0.4433	0.3722	0.8392	0.8462	0.8114	0.7681	0.8162	0.1838
	DenseNet121	0.7096	0.3943	0.3876	0.8206	0.8281	0.7977	0.7498	0.7991	0.2009
	DenseNet161	0.7362	0.3825	0.3931	0.8072	0.8176	0.7773	0.7294	0.7829	0.2171
Tiny-ImageNet	AlexNet	0.6452	0.5163	0.0211	0.9345	0.9477	0.8983	0.8949	0.9189	0.0811
	MobileNetV2	0.6555	0.4627	0.1025	0.9145	0.9299	0.8564	0.8773	0.8945	0.1055
	ResNet50	0.6920	0.4493	0.2595	0.8882	0.9016	0.8562	0.8616	0.8769	0.1231
	DenseNet121	0.6810	0.4778	0.2723	0.8762	0.8879	0.8516	0.8501	0.8665	0.1335
	ResNet101	0.7277	0.4445	0.2826	0.8728	0.8838	0.8420	0.8403	0.8597	0.1403
	DenseNet161	0.7418	0.4576	0.2936	0.8415	0.8571	0.7979	0.8016	0.8245	0.1755
	ViT	0.7988	0.4179	0.3553	0.8042	0.8389	0.6738	0.6689	0.7465	0.2535

Table 4: Robustness evaluation of speech recognition models based on ASR, Adversarial Accuracy (AA), RDI and ROBY.

Dataset	model	ACC	ROBY	RDI	ASR (SirenAttack)	AA (SirenAttack)
SPEECHCOMMANDS	M5Attention	0.9067	0.3837	0.4305	0.8071	0.1929
	M5	0.9575	0.4753	0.4749	0.7148	0.2852
	SpeechCNN	0.9251	0.4608	0.5955	0.5986	0.4014
	SpeechTransformer1	0.9116	0.4419	0.6060	0.5408	0.4592
	SpeechTransformer2	0.9028	0.4932	0.6278	0.5204	0.4796