Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Novel Two-Stage Training Detection Mechanism Against Adversarial Attacks on Images

Suman Das

phcs220020@nitsikkim.ac.in

National Institute of Technology Sikkim

Dr. Pankaj Kumar Keserwani

National Institute of Technology Sikkim

Prof. Mahesh Chandra Govil

National Institute of Technology Sikkim

Research Article

Keywords: Adversarial machine learning, Adversarial attacks, Adversarial training, Deep learning, Image classification, and Attack detection.

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8338233/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Novel Two-Stage Training Detection Mechanism Against Adversarial Attacks on Images

Suman Das[a], Dr. Pankaj Kumar Keserwani[a], Prof. Mahesh Chandra Govil[a]

[a]*Computer Science and Engineering National Institute of Technology Sikkim, Ravangla, 737139, Sikkim, India*

Abstract

Deep learning (DL) models are prone to a variety of adversarial attacks, which results in monetary losses and reputational damage for their applications. Detecting adversarial attacks is crucial to take corrective actions for reducing their impact on applications that are based on the DL approach. Many researchers have developed different detection methods that rely on analytically determined parameters, often resulting in inadequate outcomes. The present detection algorithm incurs considerable computational overhead due to complex feature extraction and high-dimensional data processing, which consequently affects the classification performance. To upgrade the performance, a more effective detection mechanism is needed. In this work, a fast and efficient detection method is approached. Since CNN-based models are widely used for classification of images, the proposed detection technique is tested on common CNN architectures, such as ResNet50, VGG16, InceptionV4, and MobileNetV2. The publicly available CIFAR-100 dataset is subjected to various adversarial attacks, such as Basic Iterative Method (BIM), Projected Gradient Descent (PGD), Carlini and Wagner (C&W), SparseFool, and AutoAttack. The performance analysis of different models shows how good the detection method is as the adversarial training-based detection achieves accuracies of 94.16% for ResNet50, 93.81% for VGG16, 91.26% for InceptionV4, and 94.06% for MobileNetV2. The findings show that, in comparison to current techniques, the suggested detection method significantly improves adversarial detection performance.

Keywords: Adversarial machine learning, Adversarial attacks, Adversarial training, Deep learning, Image classification, and Attack detection.

1. Introduction

With the exponential expansion of data, commonly known as big data, DL has emerged as a powerful tool for extracting meaningful insights from massive and diverse datasets. Deep learning (DL) models consist of multiple layers of perceptrons to form the neural networks He et al. (2016), Karras et al. (2019). These models can learn complex patterns from the data and perform well with unstructured data collected from sensors, webcams, video, and document data LeCun et al. (2015). They can also operate well in parallel computing environments. These properties of the DL methods are very effective for cutting-edge applications such as image recognition, predictive models, object detection, speech recognition, and natural language processing (NLP). Nowadays, companies are using big data to improve system operations and customer service to increase profit and revenue. It will provide valuable insights that can be used to refine the marketing strategies Wha. Some widely used DL applications where safety is crucial are listed below:

- **Autonomous Cars:** DL models are essential to safety of autonomous cars Gao et al. (2018). It assists in object detection, traffic sign recognition, pedestrian detection and lane detection, making cars smarter and able to drive safe on the roads.
- **Medical Imaging:** In medical imaging applications for disease detection, abnormalities identification, and analysis of radiological images, DL models are widely used Shin et al. (2016). DL models are capable of being an early diagnostic of diseases for the medical industry, helping doctors make the correct diagnoses and treatment plan.
- **Surveillance and Security Systems:** Surveillance and security systems, for example, in face recognition, person detection, and object detection, also rely on the performance of DL models Muhammad et al. (2019). It helps reduce dangers and looks through public areas for improvements in safety.
- **Content Moderation:** In content moderation systems, DL models are used to detect and filter certain types of uncensored or offensive content Liu et al. (2016). It creates a helpful environment for online social platforms, websites and communities by automatically detecting and flagging inappropriate content.
- **Environmental Monitoring:** DL models are deployed Wang et al. (2019) in many environmental applications including land use, vegetation condition, deforestation and natural disasters. It is helping in the conservation of environment and to take decisions for sustainable exploitation of resources.
- **Retail and E-commerce:** Recently DL models are also adapted to the retail and e-commerce domain for product

recognition, visual search, and recommendation systems by incorporating automatic product tagging, personalized recommendations on visual preference, and improved user experience in online shopping sites Niwa et al. (2016).

As the use of DL applications grows, it is crucial to improve the resilience of DL models against various types of attacks. The adversarial attack is one of them. Adversarial attacks involve making slight modifications of input data, leading DL models to make incorrect predictions, as demonstrated in Szegedy et al. (2014), Goodfellow et al. (2015). For example, in autonomous driving, attackers can manipulate traffic signs to cause a self-driving car to misinterpret a 'stop' sign as a 'left turn,' potentially leading to hazardous traffic incidents Eykholt et al. (2018). Each adversarial attack method uses different strategies to take advantage of weaknesses in deep learning models, leading to compromises in model performance and reliability. There are various perspectives on how to classify adversarial attacks. Here, the main focus is to explore the implementation perspective, which categorizes adversarial attacks into three types Zhang et al. (2022):

1. **Gradient modification-based attack:** In this method, the attackers iteratively perturb the image by making minor changes to the gradients of the loss function to the input image in order to maximize the prediction error. Examples of gradient-based adversarial attacks include the Fast Gradient Sign Method (FGSM) Goodfellow et al. (2015), Basic Iterative Method (BIM) Kurakin et al. (2017), and Projected Gradient Descent (PGD) Madry et al. (2019).

2. **Optimization strategy-based attack:** In these attacks, the central idea is to formulate an optimization problem that seeks to minimize the dissimilarity between the predicted class probability distribution of the original image $p(y|x)$, and the adversarial image $p(y|x')$, while ensuring that the perturbation δ remains bounded by a predefined norm constraint. This can be mathematically expressed as follows:

$$\text{minimize:} \quad J(x', y) \quad \text{subject to} \quad \|\delta\|_p \leq \epsilon \tag{1}$$

   Examples of optimization-based adversarial attacks are the Carlini-Wagner (CW) attack Carlini and Wagner (2017), DeepFool attack Moosavi-Dezfooli et al. (2016), etc.

3. **Decision-based attack:** In a decision-based attack the attackers mainly emphasize crafting perturbations that manipulate the model's decision boundary, without relying on gradients or optimization techniques. These attacks aim to generate perturbed images x' that lead to a different model prediction compared to the original image x, without explicitly computing gradients. The primary goal is to find perturbation δ such that

$$f(x + \delta) \neq f(x) \tag{2}$$

   Here, $f(.)$ represents the model's decision function, and x, x' are the original and perturbed images respectively. Examples of decision-based adversarial attacks are Boundary attack Brendel et al. (2017), Decision-based black-box attack Ilyas et al. (2018), and Decision-based gradient-free attack Weng et al. (2018).

Some commonly used attack examples are the FGSM, CW, BIM, PGD, etc. These attacks pose a significant threat to existing deep-learning models. The challenges associated with these attacks lie in detecting and mitigating them due to their complex and evolving nature, high computational cost, and limited availability of data.

To address adversarial attacks on deep learning models, researchers have proposed many defense-based and detection-based tactics. The defense-oriented solutions focus on training strong models that can correctly classify adversarial images. This method is executed through three main phases. The first phase applies augmentations on input data (both training and testing samples) of the model. Some of the examples that fall into this category include adversarial training Madry et al. (2019), data compression Das et al. (2017), Dziugaite et al. (2016). The second phase consists of approaches that modifies the network architecture and uses optimization based techniques to make the model more robust which includes adding layers in network, modifying loss functions like,['defensive distillation' [8], 'gradient obfuscation' [3]) Papernot et al. (2017a),Papernot et al. (2016) and Athalye et al. (2018a). The third phase uses external modules as supplementary to the original model, e.g., feature squeezing Xu et al. (2018). In contrast, detection-based approaches concentrate on identifying the problems of relying on model augmentation to achieve robustness. The utility of such methods as a detection against adversarial attacks is gaining traction. Several detection-based methods have been proposed to defend adversarial examples. Katzir et al. Katzir and Elovici (2019) observed that adversarial examples and normal samples followed differ-ent spatial distributions in dissimilar activation spaces. Yang et al. Yang et al. (2020) also noted the characteristic attribution graphs of adversarial examples and their originals are different. Feinman et al. Feinman et al. (2017a) detect adversarial examples in the latent space used kernel density estimation and Bayesian neural network uncertainty.

The following challenges we experienced in dealing with the negative aspect of adversarial attacks:

- **Adversarial attacks are difficult to detect:** It is also hard for traditional detection methods to discriminate adversarial samples from original ones, especially around decision boundaries Machado et al. (2021).
- **Adversaries and defenses are moving targets:** There have been new forms of attacks that keep emerging, which overrun current defenses Apruzzese et al. (2019).
- **Limited availability of data:** Obtaining sufficient data for training and testing robust defense mechanisms remains a challenge.
- **Difficulty in generalizing defenses:** A defense effective against one attack may fail against another, limiting generalization.

Various researchers have suggested different approaches to explore and detect adversarial attacks. Some of these approaches focus on creating new attack strategies, while others concentrate on systems to detect adversarial attacks. In this work, some of the challenges, such as the difficulty of adversarial attack detection, the trade-off between accuracy and robustness of the detection method, and data unavailability, are targeted to be addressed. To accomplish these, a comprehensive study is being performed on the vulnerability of DL models: ResNet-50 He et al. (2016), VGG-16 Simonyan and Zisserman (2015), Inception V4 Szegedy et al. (2016), and MobileNet V2Howard et al. (2017) to various adversarial attacks using the image dataset CIFAR-100. The extent of the models' vulnerability is quantified by examining the original image accuracy and the accuracy on generated adversarial attack images. In this work, a novel detection mechanism is proposed that can identify adversarial samples effectively in a fast manner. The detection method first trains the models with clean datasets and then tests them against adversarial samples to find the vulnerable models and identify the most severe attacks. Unlike conventional adversarial training approaches, which are limited to detecting only the specific attacks they were trained on, the proposed approach generalizes across multiple attacks, allowing the models to distinguish between genuine and adversarial samples with high accuracy.

The claimed contributions are:

C1: **Development of a novel adversarial attack detection method:** A novel adversarial detection mechanism is designed and developed to identify perturbed data effectively. It combines feature extraction, deep learning classifiers, adversarial training, and logistic regression for improved accuracy.

C2: **Generation of novel attack datasets:** Novel attack datasets are generated by labeling the non attack images based on seven adversarial attack methods, namely FGSM, BIM, PGD, One Pixel, Sparsefool, AutoAttack, and C&W.

C3: **Analysis of four DL model performances on generated attacked datasets:** Four DL models: ResNet50, VGG16, MobileNetV2, and InceptionV4, are evaluated on the generated attacked datasets. The results indicate that the adversarial attacks are effective in significantly degrading the performance of these models.

C4: **Comparative analysis of the detection mechanism with recent works:** The efficacy of the proposed detection mechanism is illustrated through a comparative analysis with recent works in the field. The results demonstrate that the proposed method is superior.

The remaining sections of this paper are structured as follows: In Section 2, an extensive examination of prior research in the domain of adversarial attacks and defense mechanisms is presented. Section 3 outlines the suggested approach for detecting and defending against adversarial attacks. The outcomes and discussion are covered in Section 4, focusing on the assessment of vulnerability and the effectiveness of the proposed detection and defense methods. Lastly, Section 5 concludes the paper by discussing the findings and their implications.

2. Related Work

This "Related Work" section is structured into two subsections: Summarization of Adversarial Attack based papers, and Summarization of Adversarial Attack Detection based papers.

2.1. Adversarial Attack:

Adversarial attacks were explored by Shamir et al. (2019), Lowd and Meek (2005), Goodfellow et al. (2015). They showed that specific inputs can be carefully designed to fool deep learning models into making wrong predictions. The authors demonstrated the susceptibility of different deep learning models to adversarial examples and investigated the underlying cause of this phenomenon. In addition, Goodfellow et al. (2015) further demonstrated that advanced deep learning models LSTM, RELU or maxout networks, are vulnerable from various adversarial attacks. They showed that carefully crafted inputs with imperceptible perturbations could easily deceive these models. The paper discussed different methods for generating adversarial examples and investigated the transferability of these examples across various models. On the other hand, a methodology for practical black-box attacks on machine-learning models was presented by Papernot et al. (2017b). In this study, the effectiveness of substitute models in generating adversarial examples to deceive target models without detailed knowledge of their internal parameters was demonstrated, as well as the vulnerability of machine learning models to black-box attacks, which was highlighted by emphasizing the importance of robust defenses. Moosavi-Dezfooli et al. (2016) developed a novel Deepfool algorithm to generate minimal perturbations that could deceive deep neural networks and alter their predicted class labels. Carlini and Wagner (2017) focused on evaluating neural network robustness against adversarial attacks. They proposed the Carlini-Wagner attack, which aimed to find adversarial examples with minimal perturbations while evading detection. Again, Kurakin et al. (2017) examined the robustness of deep neural networks on real-world scenarios involving physical objects to investigate the effect of adversarial examples. They showed that adversarial perturbations applied to printed images could fool object identification systems, even when captured with a real camera. In the same year, Universal Adversarial Perturbations were introduced by Fawzi et al. (2017), showing their transferability across datasets and models and also highlighting the existence of common vulnerabilities in deep learning models. In 2018, Athalye et al. (2018b) proposed a novel method for synthesizing robust adversarial examples using different perturbation magnitudes. The perturbation distribution-based framework was introduced to generate adversarial examples. Adversarial attacks targeting time series data were researched by Karim et al. (2021). The study focused on manipulating predictions of time series models, leading to misclassification or evading anomaly detection systems. In 2021, Shehu et al. (2020) proposed adversarial patch attacks that demonstrate the ability

to deceive object recognition models by placing small patches on images. In 2022, Chen et al. (2022) discussed adversarial attacks targeting recommender systems, highlighting the vulnerability of these systems by manipulating user preferences or item features. In 2023, Ling et al. (2023) explored state-of-the-art adversarial attacks to fool the Windows PE malware detection system. In 2024, Cheng et al. (2024) proposed a feature-aware transferable adversarial attack framework to enhance attack capability while maintaining transferability. In 2025, Ran et al. (2025) investigated the vulnerability of No-Reference Image Quality Assessment (NR-IQA) models to black-box adversarial attacks, which uses A novel bi-directional loss function to maximize quality score deviations.

2.2. Adversarial Attack Detection:

In the realm of adversarial attack detection, numerous research endeavors have introduced methodologies for identifying adversarial perturbations. Papernot et al. (2016) introduced an adaptive noise reduction to identify adversarial examples. A visual foresight based detection, proposed by Metzen et al. (2017a) for robot control task. The other one is Feinman et al. (2017b), who address the detection task by knowing perturbation distortions of adversarial sample generation and increases scrutiny when verifying adversarial samples. An alternative direction is to design detectors that are able to capture any form of new adversarial examples as well with appropriately schedulable thresholds or control limits: Grosse et al. (2017b) introduces a statistical test for detection of adversaries by focusing on the discrepancies of feature distributions which serve as indicative markers. Lu et al. (2017) proposed a radial basis function kernel Support Vector Machine (SVM) classifier which took the difference on the pattern of ReLU activation's output as a feature to discover adversarial examples. One year later, Lee et al. (2018b) presented Gaussian Discriminant Analysis (GDA) which computes the Mahalanobis distance to the advertisement samples in order to anticipate the distribution of features. In the same year, Zheng and Hong (2018b) have made use of the similar idea to model the output distributions of hidden neurons in a DNN classifier, to effectively identify adversarial samples. Agarwal et al. (2018) implemented an SVM classifier with pixel values and principal component decomposition features for adversarial perturbations. Ma et al. (2018) proposed a model of detecting adversarial example by the technique based on Local Intrinsic Dimensionality (LID). The approach captures locally the intrinsic dimensionality of a reference sample, and,thus, provides also means for detection of possible adversarial samples. Further Goswami et al. (2019) introduced a technique based on the feedback light in the hidden layers of DNN models for adversarial attack detection. In the following year, Madry et al. (2019) introduced techniques to train deep learning models that are resistant to adversarial examples. Adversarial training via projected gradient descent (PGD) was adopted to improve robustness against adversarially noisy inputs. Madry et al. (2019) the training procedure also pursued an iterative approach of producing adversarial samples and updating model parameters to promote representation robustness. Cohen et al. (2020) introduced a k-nearest Neighbour (k-NN) model that scores the utility of training samples based on how they affect the validation set and incorporates them into meta-training. After that, a logistic regression model was trained to recognize adversarial instances. In the year of 2022, Yang et al. (2022) developed the ContraNet adversarial detection framework to gauge the discrepancy between retrieved discriminative features and semantic information extracted by the deep neural network (DNN) model while dealing with adversarial examples. One year later, Gong and Wang (2023) contributed a detection mechanism grounded in density ratio estimation by emphasizing the divergence between adversarial and clean data, proposing. In 2024, Qing et al. (2024) proposed ADDNP method to disentangle natural images from adversarial perturbations. The method leverages reconstruction discrepancies in latent space to detect adversarial attacks. In the same year, Li et al. (2024) introduced a fuzzy rule-based method for detecting adversarial attacks, leveraging fuzzification to improve detection performance in neural networks.

An overview of recent detection techniques is summarized in the following TABLE 1

Table 1: An overview of recent detection techniques and their detection accuracies from the authors' original work

Ref. No.	Adversarial Attack	Datasets	Detection Method	Acc. Range (%)
Pertigkiozoglou and Maragos (2018)	BIM, DF	MNIST	Softmax	90-100
Kherchouche et al. (2020)	FGSM, CW, BIM, DF	MNIST, CIFAR10, IMAGENET	NSS	90-100
Lust and Condurache (2020)	FGSM, CW, BIM, JSMA	MNIST, CIFAR10, SVHN	Gradient	90-100
Feinman et al. (2017a)	FGSM, CW, BIM, JSMA	MNIST	Un - certainty	80-89
Zuo and Zeng (2021)	CW, DF	CIFAR10, IMAGENET	E&R	90-100
Grosse et al. (2017a)	FGSM, JSMA	MNIST	MMD	90-100
Li and Li (2017)	L-BFGS	IMAGENET	PCA	80-89
Feinman et al. (2017a)	FGSM, CW, BIM, JSMA	MNIST	KD	80-89
Ma et al. (2018)	FGSM, BIM, JSMA	MNIST, CIFAR10, SVHN	LID	90-100
Lee et al. (2018a)	FGSM	CIFAR10, SVHN	Maha -lanobis	90-100
Metzen et al. (2017b)	FGSM, BIM, DF	CIFAR10	Training	80-89
Eniser et al. (2020)	FGSM, DF, BIM, PGD, CW, JSMA	CIFAR10, MNIST	RAID	90-100
Pertigkiozoglou and Maragos (2018)	BIM, DF	MNIST	Histogram	80-89
Pang et al. (2018)	FGSM, CW, BIM, JSMA	MNIST, CIFAR10	Reverse Cross-Entropy	90-100
Aldahdooh et al. (2023)	FGSM, PGD, CW, DF	MNIST, CIFAR10	SFAD	90-100
Zheng and Hong (2018a)	FGSM	MNIST	GMM	90-100
Liang et al. (2021)	FGSM, CW, DF	MNIST, IMAGENET	Adaptive Noise Reduction	90-100

As depicted from above table the last column represents accuracy in range, it is due to concentration of perturbation in variable ratios for generation of adversarial attacks. Some identified the research gaps such as unavailability of direct research data, model complexity, etc., motivated us to design and develop a new detection mechanism to overcome the mentioned research gaps. The next section discusses the proposed approach based on two-stage training for detecting adversarial samples.

3. Proposed Approach

A novel technique using two-stage adversarial training to detect adversarial attacks is graphically presented below by Fig. 1.

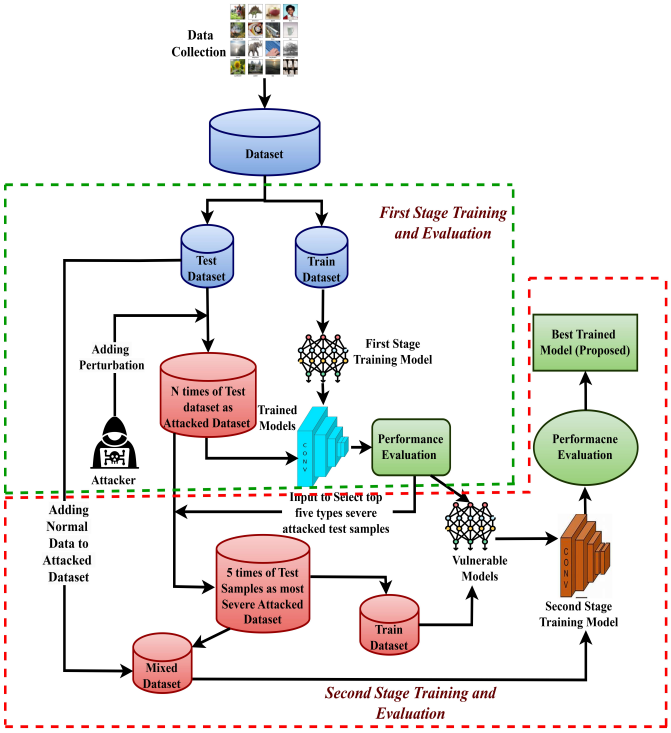


Figure 1: Proposed System Model for Adversarial Attack Detection

For the first stage training, the collected dataset is divided into a training dataset and a testing dataset. Four DL models, namely: VGG16, ResNet50, MobileNet V2, and Inception V4, are trained using the non-attacked training dataset and tested on adversarial samples. The testing dataset is generated using seven adversarial attacks, namely: FGSM, BIM, PGD, OnePixel, SparseFool, CW, and Autoattack. Based on the first stage evaluation results, the five "most severe attacks" are identified and used to create a "most severe attack" dataset. This created attack dataset is again split into a training dataset and a testing dataset. This severe attack training dataset is used in the second stage training of the same DL models to evaluate the performance of the second stage training. This helps in determining the best model to detect adversarial attacks. So the overall detection method is divided into two crucial stages first stage is for finding the vulnerable model, and the second stage presents the detection accuracy of all models. All modules of the proposed approach are discussed in further subsections:

3.1. Data Collection and Preprocessing

The practice of obtaining unprocessed data from multiple sources is known as data collection. In order to effectively train and assess machine learning models, preprocessing of gathered data guarantees the creation of high-quality, diversified, and representative data. The freely accessible CIFAR-100 tor dataset is taken into consideration in this work. An expansion of CIFAR-10, the CIFAR-100 dataset is intended for more detailed categorization tasks. It offers 100 different classes and 60,000 color 32x32 images. Accordingly, there are 600 photos in each class. Following methods are utilized during data preprocessing:

- **Image Resizing:** The input images are usually resized to specific dimensions to ensure consistency across the dataset and match the expected input size of the DL model. In this work, two different resizing dimensions are used, like `transforms.Resize((229,229))` and `transforms.Resize((224,224))` for Inception v4 and ViT model respectively. The image sizes are taken to concatenate with the models input dimension.

- **Image Normalization:** Normalization is a preprocessing technique that ensures standardization of pixels to reduce the influence of varying pixel intensities across different images. The normalization technique helps the models to focus on the relative patterns and features within the images. In this work, the normalization operation transforms each pixel value x by subtracting the mean (0.5) and dividing by the standard deviation (0.5).

$$\text{Normalize}(x) = \frac{x - 0.5}{0.5} \tag{3}$$

  This scales the input data from the range $[0, 1]$ to $[-1, 1]$, which helps stabilize and speed up training in deep learning models.

3.2. First stage training and performance evaluation

In this stage, the collected dataset is divided into a training set and a testing set in a ratio of 80% and 20%. The training dataset is used to train four DL models: VGG16, ResNet50, MobileNetV2, and InceptionV4 on non-attacked images. The testing dataset is utilized to generate an adversarial dataset using seven popular attack methods, namely FGSM, BIM, PGD, OnePixel, SparseFool, CW, and AutoAttack. In this way, the test dataset contains seven types of adversarial attack image datasets and one non-attack image dataset. As a result, the size of the testing dataset is twice that of the training dataset.

In the creation of an adversarial attacked image, small perturbations are added to the normal image. The adversarial image is used to make fool to the trained model so that it predicts the incorrect class. Figure 2 illustrates the process of generating adversarial samples. Each adversarial attack method applies

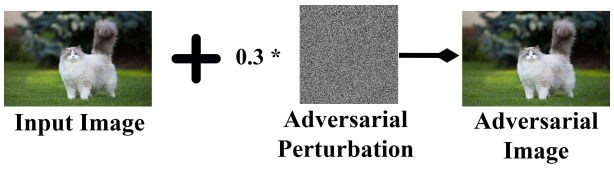


Figure 2: Generation of Adversarial Image

different perturbation strategies to provide a diverse and comprehensive assessment of the vulnerability of the trained model. The attack methods considered are as follows:

- **Fast Gradient Sign Method (FGSM) Goodfellow et al. (2015):** This method computes the gradient of the loss function based on the input data and perturbs the input in the direction that maximizes the loss.
- **Basic Iterative Method (BIM) Kurakin et al. (2017):** The Fast Gradient Sign Method (FGSM) is used iteratively with minor perturbations to produce adversarial examples is called BIM. Given a neural network with parameters θ, an input sample $\mathbf{x}$, and the corresponding label y, the BIM attack aims to find an adversarial example $\mathbf{x}^*$ that maximizes the model's prediction error. The iterative update rule for BIM is given by:

$$\mathbf{x}^{(i+1)} = \mathbf{x}^{(i)} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}^{(i)}, y)) \tag{4}$$

- **PGD Attack Madry et al. (2019):** The Projected Gradient Descent (PGD) attack is an iterative adversarial attack that applies multiple small perturbations to maximize the model's loss while keeping the perturbation within a specified bound.

$$x'_{t+1} = \text{Clip}_{x,\epsilon}(x'_t + \alpha \cdot \text{sign}(\nabla_x \mathcal{L}(f(x'_t), y))) \tag{5}$$

  In Equation 5, x'_t represents the perturbed image at iteration t, α is the step size, $\nabla_x \mathcal{L}$ is the gradient of the loss function $\mathcal{L}$, and ϵ controls the perturbation limit.

- **Carlini-Wagner (CW) attack Carlini and Wagner (2017):** The CW attack formulates an optimization problem to find the minimum perturbation that causes misclassification while also satisfying certain constraints.

$$\begin{aligned} &\text{minimize} \quad \|\delta\|_p + c \cdot f(x+\delta) \\ &\text{subject to} \quad x + \delta \in [0,1]^n, \end{aligned} \tag{6}$$

- **One-Pixel Attack Su et al. (2019):** The one-pixel attack perturbs a single pixel in an image to generate adversarial examples while keeping all other pixels unchanged. This minimal modification can significantly mislead deep learning models.

$$x'_{i,j} = \text{Clip}(x_{i,j} + \delta, 0, 255) \tag{7}$$

  In Equation 7, $x'_{i,j}$ represents the modified pixel value at position (i, j), $x_{i,j}$ is the original pixel value, and δ is the perturbation applied.

- **SparseFool Attack Modas et al. (2019):** The SparseFool attack generates adversarial examples by perturbing only a few pixels in the image while maintaining imperceptibility.

$$x' = x + \eta \cdot v \tag{8}$$

  In Equation 8, x' is the perturbed image, x is the original image, η is the perturbation magnitude, and v is a sparse vector indicating the modified pixels.

- **AutoAttack Liu et al. (2022):** AutoAttack is an ensemble of multiple strong adversarial attacks, including PGD, FAB, Square Attack, and DeepFool, ensuring a reliable evaluation of model robustness.

$$x' = \mathcal{A}(x, f, \epsilon) \tag{9}$$

  In Equation 9, x' is the adversarial image, x is the original input, f is the target model, and $\mathcal{A}$ represents the AutoAttack framework applying multiple attack methods within the perturbation bound ϵ.

Algorithm 1 First Stage Training and Attacked Dataset Generation Algorithm

Input: CIFAR-100 dataset ($D_{\text{Orgtrain}}, D_{\text{Orgtest}}$);
DL Models $\mathcal{M}$;
Attack methods $\mathcal{A}$ = {BIM, FGSM, PGD, SparseFool, CW, OnePixel, AutoAttack}
Output: Set of severe attacks $\mathcal{A}_{\text{severe}}$

Initialize: attack frequency counter $F[a] \leftarrow 0$ for all $a \in \mathcal{A}$
for each model $m \in \mathcal{M}$ **do**
  Train m on D_{train}
  Initialize: accuracy list $\mathcal{L}_m \leftarrow []$
  for each attack $a \in \mathcal{A}$ **do**
    Generate adversarial samples D_a from D_{test} using a
    Predict $\hat{y}$ for D_a using m and compute TP, TN, FP, FN
    Compute Accuracy$_{m,a} \leftarrow \frac{TP+TN}{TP+TN+FP+FN}$
    Append $(a, \text{Accuracy}_{m,a})$ to $\mathcal{L}_m$
  end for
  Sort $\mathcal{L}_m$ by accuracy in ascending order
  Select lowest-5 accuracy attacks $\mathcal{A}_m^{\text{low}}$
  for each $a \in \mathcal{A}_m^{\text{low}}$ **do**
    $F[a] \leftarrow F[a] + 1$
  end for
end for
return Top-5 attacks from $F[a]$ with highest frequency as $\mathcal{A}_{\text{severe}}$

The training phase begins with model initialization, where each DL model is initialized without pre-trained weights to ensure unbiased training. A structured training loop is followed, in which the data is processed in mini-batches. During each iteration, forward propagation is carried out to compute predicted outputs and the respective loss. The loss is used in the backpropagation process to compute gradients for updating model weights. The optimization algorithm Adam was used to get optimized weights. Training progress is continuously monitored by tracking loss and accuracy to prevent overfitting or underfitting. After the training phase, the performance of each trained deep learning (DL) model: VGG16, ResNet50, MobileNetV2, and InceptionV4 is rigorously evaluated on an updated test dataset, also termed as adversarially attacked dataset,

using accuracy as the primary metric to determine the vulnerability of each model against various adversarial attacks. For each class (one non attacked and seven adversarial attacks) accuracy of each model is computed, and the five most severe attacks (BIM, C&W, SparseFool, AutoAttack, and PGD) are identified based on the descending drop in accuracy. A new dataset, termed as the most severe adversarial attack dataset, is generated that contains only these five severe adversarial attack images. This most severe adversarial attacked dataset acts as the input for the second stage of training of the proposed approach.

3.3. Second Stage Training and Performance Evaluation

In this stage, the process begins by splitting the most severe adversarial attack dataset into training and testing subsets in 80% and 20% ratio. The training subset is fed to DL models so that critical patterns and features of adversarial perturbations can be learned during the second stage training phase.

Algorithm 2 Second Stage Training and Detection Algorithm

Input: Adversarial dataset $\mathcal{A}_{\text{severe}}$ (split into $\mathcal{A}_{\text{severe_train}}, \mathcal{A}_{\text{severe_test}}$), original test dataset $\mathcal{D}_{\text{OrgTest}}$, pretrained models $\mathcal{M}$, number of epochs E, batch size B, learning rate η, optimizer.
Output: Detection Accuracy (DA) and Percent of Attacked Images (PAI) for each model $m \in \mathcal{M}$

Initialize: Set random seed, batch size B, learning rate η, optimizer = Adam.
Combine Test Sets: $\mathcal{D}_{\text{test_combined}} \leftarrow \mathcal{A}_{\text{severe_test}} \cup \mathcal{D}_{\text{OrgTest}}$
for each model m in $\mathcal{M}$ **do**
    Initialize m with pretrained weights
    Define loss function $\mathcal{L}$
    for epoch := 1 to E **do**
        for each batch (x_b, y_b) in $\mathcal{A}_{\text{severe_train}}$ **do**
            $\hat{y}_b \leftarrow m(x_b)$
            $\mathcal{L} \leftarrow \text{CrossEntropy}(\hat{y}_b, y_b)$
            Backpropagate and update model weights
        end for
        Print average training loss for current epoch
    end for
    Initialize: AttackedCount $\leftarrow 0$, AttackedTotal $\leftarrow 0$, TotalTestSamples $\leftarrow 0$
    for each (x_t, y_t, a_t) in $\mathcal{D}_{\text{test_combined}}$ **do**
        TotalTestSamples $\leftarrow$ TotalTestSamples + 1
        $\hat{y}_t \leftarrow m(x_t)$
        if $a_t = 1$ **then** ▷ Sample is adversarial
            AttackedTotal $\leftarrow$ AttackedTotal + 1
            if $\hat{y}_t \neq y_t$ **then**
                AttackedCount $\leftarrow$ AttackedCount + 1
            end if
        end if
    end for
    $DA \leftarrow \left(\frac{\text{AttackedCount}}{\text{AttackedTotal}}\right) \times 100$
    Store DA for model m
end for

This algorithm describes the training of four deep learning models—VGG16, ResNet50, InceptionV4, and MobileNetV2—on a combined severe adversarial training dataset generated from five distinct attacks.

After the adversarial training process, the trained DL models are evaluated on the mixed dataset. The mixed dataset is created by adding non-attacked images to the testing images most severe attack dataset. In this way, the mixed dataset contains only two classes: non-attacked and attacked, leading us to a binary classification using the parameter Percentage of Attacked Images (PAI) and Detection Accuracy (DA). This mixed dataset is crucial for evaluating the ability of the detection model to distinguish between clean and adversarial inputs. The model achieving the highest PAI and DA is selected as the proposed model. Interestingly, it is determined that the model that was initially most vulnerable to adversarial attacks has achieved the highest PAI and DA on testing phase.

4. Results Overview and Analysis

This section delves into the experimental setup for our proposed model and examines the detection results. Following that, we perform a comparative analysis, contrasting the outcomes of our detection model with those of four other state-of-the-art detection models. Notably, our proposed model demonstrates commendable performance in identifying adversarial examples across three real-world datasets.

4.1. Configuration for Results Generation

The proposed model is implemented on Jupyter notebook platform using the Python programming language. The GPU machine with various features GPU support with CUDA, offering substantial hardware resources, including 32GB of RAM, a 2TB HDD, and a robust 16GB NVIDIA GeForce graphics card are used for implementation. This powerful setup ensures efficient execution of the proposed model on image dataset, providing the necessary computational capabilities for the experimental phase.

4.2. Performance Metrics

This work employs the following evaluation metrics:

- Classification Accuracy: The fraction of correctly identified samples, including true positives (TP) and true negatives (TN), over the total number of samples determined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- Percentage of Attacked Images (PAI): It is calculated as the ratio of misclassified images to the total number of images in the test set, multiplied by 100. The percentage of attacked images is calculated as

$$PAI = \left(\frac{\text{Total Attacked Count}}{\text{Total images}}\right) \times 100$$

- Detection Accuracy (DA): It is calculated as the ratio of the total number of correctly identified adversarial examples to the total number of actual adversarial examples, multiplied by 100. The detection accuracy is calculated as

$$DA = \left(\frac{\text{Attacked Count}}{\text{Attacked Total}}\right) \times 100$$

4.3. Performance Outcomes from First Stage Training

This subsection presents a comprehensive analysis of the performance outcomes from the initial training phase, where various deep learning (DL) models were trained on clean CIFAR-100 data and subsequently evaluated on both clean and adversarially perturbed test samples. Table 2 reports the accuracy of each model on unmodified test data, while Table 3 summarizes their performance under seven different adversarial attacks.

Performance on Clean Data.. As shown in Table 2, all models perform reasonably well on clean CIFAR-100 test samples. Among them, **Inception V4** achieves the highest accuracy of **75.59%**, followed by **ResNet50** at **61.99%**, **MobileNet V2** at **60.54%**, and **VGG16** at **59.79%**. Figure 3 shows how well the models perform on clean data before any attacks are applied.

Table 2: Different Models' accuracies on CIFAR100 Image dataset

Model Name	VGG16	ResNet50	Inception V4	MobileNet V2
Accuracy	59.79%	61.99%	75.59%	60.54%

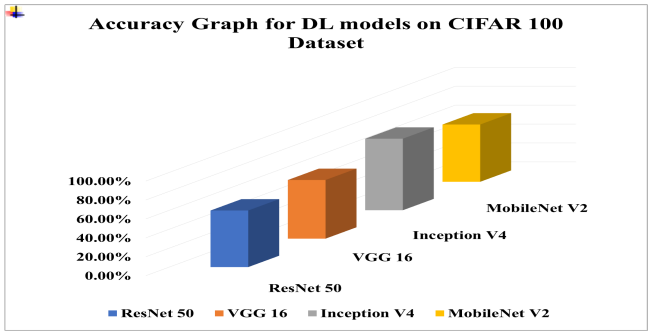


Figure 3: Accuracy on non-attacked Image for CIFAR100-dataset

Performance under Adversarial Attacks. Upon exposure to adversarial samples, the performance of all models degrades drastically, as shown in Table 3 as well as in Fig 4. Attacks such as **SparseFool**, **AutoAttack**, **C&W**, **PGD**, and **BIM** significantly impair model accuracy, often reducing it to near-zero values. For example, SparseFool reduces the accuracy of ResNet50 and MobileNet V2 to just **0.0003%** and **0.0004%**, respectively. AutoAttack produces similar devastation, with Inception V4 accuracy dropping to merely **0.0052%**. Even relatively simple attacks, such as FGSM and OnePixel, cause significant performance degradation. Notably, the **OnePixel** attack yields model-specific effects — Inception V4, while severely impacted, still maintains the highest accuracy at **0.1560%**, whereas VGG16 drops to just **0.09%** and MobileNet V2 to **0.0403%**. ResNet50 achieves an extremely low **0.0349%** accuracy, indicating that this attack, although subtle, is highly effective.

Table 3: Classification Accuracies of Different DL Models on Attacked Data

Attack \ Model	Classification Accuracy			
	ResNet50	VGG16	Inception V4	MobileNet V2
BIM	0.66%	7.50%	10.11%	6.50%
FGSM	5.54%	10.28%	20.56%	9.89%
OnePixel	0.0349%	0.09%	0.1560%	0.0403%
C&W	0.10%	0.90%	0.50%	0.50%
PGD	1.20%	1.80%	2.02%	1.50%
SparseFool	0.0003%	0.00032%	0.00035%	0.0004%
AutoAttack	0.00046%	0.001%	0.0052%	0.00059%

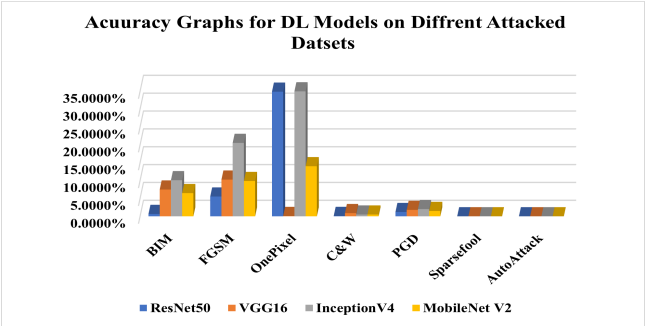


Figure 4: Graphical representation of Classification Accuracies of Different DL models

Most Severe Attacks Identified.. From the observed results, the five most impactful attacks across all DL models are:

- **SparseFool**
- **AutoAttack**
- **Carlini & Wagner (C&W)**
- **Projected Gradient Descent (PGD)**
- **One Pixel**

These attacks consistently reduce classification performance to critical levels, thereby validating their strength and adversarial effectiveness.

Model-wise Vulnerability. A cross-model comparison shows that **ResNet50** is the most vulnerable to adversarial examples. It achieves the lowest classification accuracy in almost all attack cases e.g., only**0.0349%** under OnePixel, **0.0003%** under SparseFool, **0.00046%** under AutoAttack and just **0.10% for C&W attack**. Even for FGSM and BIM (deemed less complex), ResNet50 is only able to achieve **5.54%** and **0.66%** accuracies respectively. On the other hand, **Inception V4** always presents superior robustness and also requires the state-of-the-art accuracy against most adversarial attacks like FGSM (**20.56%**) and OnePixel (**0.1560%**). This indicates that adversarial robustness of a model may be significantly influenced by the architectural structure.

Insights and Implications.. Several important observations can be made based on the first stage-trained results:

- DL models that are trained on only clean data are extremely susceptible to adversarial perturbations, even by ostensibly weak attacks.
- Adversarial attacks such as SparseFool and AutoAttack show a remarkable prowess in taking the model accuracy to nearly zero.

By setting a solid performance benchmark, this analysis highly encourages further development and testing of the detection approach.

4.4. Performance Results from Second Stage Training

In the second layer of our pipeline, each trained deep learning model is used for training a logistic regression classifier to test their own ability in detecting adversarial samples. The goal here wasn't classification of objects but rather binary detection – is the input sample normal or adversarial

Each DL model (ResNet50, VGG16, Inception V4, MobileNet V2) was previously trained on the combined adversarial dataset. From each trained model, we extracted feature representations from the penultimate layer, which were then used to train a logistic regression classifier for adversarial detection. This classifier was evaluated on the hybrid test dataset described above. Table 4 presents the detection accuracies for each model, while Figure 5 provides a visual comparison.

Table 4: Detection Accuracy of Different DL models on Hybrid Dataset

Adversarial Attacks	DL Models	Detection Method	Accuracy%
OnePixel, PGD, C&W, SparseFool, AutoAttack	ResNet50 (Proposed)	Adversarial Training + LR	94.16
OnePixel, PGD, C&W, SparseFool, AutoAttack	VGG16	Adversarial Training + LR	93.81
onePixel, PGD, C&W, SparseFool, AutoAttack	MobileNet V2	Adversarial Training + LR	91.26
OnePixel, PGD, C&W, SparseFool, AutoAttack	Inception V4	Adversarial Training + LR	94.06

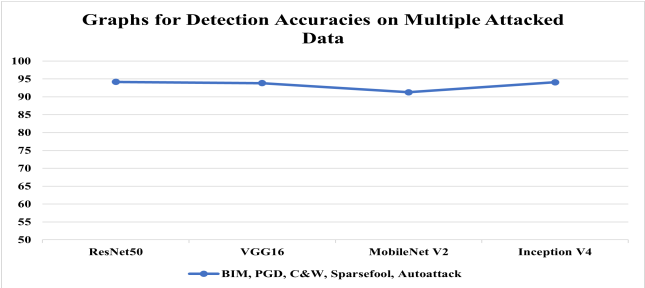


Figure 5: Graphical Representation of Detection Accuracies of Different DL Models on the Hybrid Dataset

Key Findings:

- All deep learning models, when combined with logistic regression on top of adversarially trained features, showed high detection accuracy on the hybrid dataset.
- **ResNet50**, which was previously identified as the most vulnerable model to adversarial attacks in the first stage, surprisingly delivered the best detection performance with an accuracy of **94.16%**.
- ResNet50 is therefore adopted as the **proposed model** for adversarial detection due to its superior performance in this stage.
- The results reinforce that combining adversarial training with lightweight classifiers such as logistic regression produces strong and scalable detection systems.

The second stage showcases a robust detection framework in which even the most vulnerable model from stage one can serve as a powerful adversarial detector, underscoring the importance of re-purposing feature extractors for security tasks beyond classification.

4.5. Comparative Analysis with State-of-the-Work

The comparison results of various adversarial detection approaches on CIFAR-100 are shown in Table 5. It is observed that the traditional methods, including DSVDD and LID+KDE have detection accuracies of 61.2% and 64.3%, suggesting their restriction in the presence of adversarial distortions. Mahalanobis Distance with KDE further improves the performance at 72.3% through considering statistical feature distributions, and HEAD+KDE achieves a remarkable edge over that of 89.7%, indicating the effectiveness of hierarchical deep feature representations for discriminating adversarial examples.

Table 5: Comparison Table of the Detection Accuracies with Other State-of-the-Art Methods

Attacks	Detection Method	Accuracy %
PGD, CW, SparseFool, One Pixel, AutoAttack	DSVDD Feinman et al. (2017a)	61.2
	LID+KDE Feinman et al. (2017a)	64.3
	Mah.+KDE Li et al. (2024)	72.3
	HEAD+KDE Liang et al. (2021)	89.7
	Proposed	**94.16**

The proposed method outperforms all baselines with a detection accuracy of 94.16%, showing its robustness against various attack strategies, including PGD, FGSM, BIM, CW, Sparse Fool, One Pixel, and Auto Attack. The superior performance suggests that our approach effectively captures adversarial perturbations, achieving better generalization across multiple attack types. These results highlight the limitations of existing methods and the importance of advanced detection strategies for improved adversarial robustness.

5. Discussion

One noteworthy finding from our experimental pipeline is that ResNet-50 emerged as the most effective adversarial attack detector in the final stage, even though it was the most susceptible model during initial training. Three complementary factors are responsible for this behaviour. First, small adversarial perturbations cause disproportionately large changes in the internal representations of vulnerable models, which are naturally more sensitive to gradients and have sharper loss landscapes. This increased reactivity is not a disadvantage; rather, it enhances the "adversarial footprint" in the feature space, making it easier to distinguish between clean and disturbed samples. Second, a vulnerable network retains high-frequency adversarial artifacts across its intermediate layers, in contrast to robust architectures that typically smooth out or suppress perturbations. Consequently, the detector receives more distinct cues for classifying the attack. Third, adversarial techniques such as OnePixel, PGD, SparseFool, AutoAttack, and C&W exploit the classifier's weakest curvature directions. These perturbations

produce stronger, more learnable deviations because ResNet-50 exhibits more pronounced weak directions during early training, enabling the detector to consistently capture attack-induced distortions.
On the other hand, despite being slightly vulnerable, VGG16, MobileNetV2, and Inception-V4 showed relatively smoother feature responses to adversarial perturbations. A portion of the adversarial signal was suppressed by their more stable representations, reducing the difference between clean and attacked embeddings and somewhat limiting the detector's ability to distinguish perturbation patterns across various attack algorithms. As a result, despite their competitive performance, these models were unable to match ResNet-50's detection capability. Overall, our results support the idea that models with higher early-stage vulnerability can, paradoxically, be better feature extractors for adversarial detection downstream because their instability amplifies perturbation effects in ways that detection modules can efficiently exploit.

6. Summary and conclusions

In this study, a novel adversarial attack detection framework is proposed. Using this framework, an improved DL model is developed to detect common adversarial attacks such as FGSM, BIM, OnePixel, PGD, CW, and AutoAttack. A novel attack dataset is created by applying different attack methods to the CIFAR-100 dataset, which was then used to test the performance of the proposed model. The proposed model can be applied in real-world areas like medical imaging, Surveillance and Security, and Autonomous vehicles. However, the limitation of this work is that the detection was done using binary classification, meaning the model could only tell whether an image was attacked or not. This approach was used to keep the system simple and establish a strong baseline. Future work will focus on extending the model to perform multiclass classification so that each type of attack can be recognised individually.

Acknowledgment

- The research project did not get any specific funding for a place.
- The publication of the article has been approved, either expressly or implicitly, by all of the authors and the relevant authorities in the area where the work was done.
- Each author has contributed in three ways: (a) creating the concept or blueprint; (b) writing the work and significantly altering its intellectual content; and (c) approving the polished and updated version.
- Every author claims that there is no perceived conflict of interest.

References

What is big data and why is it important? https://www.techtarget.com/searchdatamanagement/definition/big-data. (Accessed on 09/22/2023).

A. Agarwal, R. Singh, M. Vatsa, and N. Ratha. Are image-agnostic universal adversarial perturbations for face recognition difficult to detect? In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–7. IEEE, 2018.

A. Aldahdooh, W. Hamidouche, and O. Déforges. Revisiting model's uncertainty and confidences for adversarial example detection. *Applied Intelligence*, 53:509–531, 2023. ISSN 1573-7497. doi: 10.1007/s10489-022-03373-y.

G. Apruzzese, M. Colajanni, L. Ferretti, and M. Marchetti. Addressing adversarial attacks against security systems based on machine learning. In *2019 11th International Conference on Cyber Conflict (CyCon)*, volume 900, pages 1–18, 2019. doi: 10.23919/CYCON.2019.8756865.

A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples, 2018a.

A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok. Synthesizing robust adversarial examples, 2018b.

W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *ArXiv*, abs/1712.04248, 2017. URL https://api.semanticscholar.org/CorpusID:2410333.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017. doi: 10.1109/SP.2017.49.

L. Chen, J. Li, J. Peng, T. Xie, Z. Cao, K. Xu, X. He, Z. Zheng, and B. Wu. A survey of adversarial learning on graphs, 2022.

S. Cheng, P. Li, K. Han, and H. Xu. Feature-aware transferable adversarial attacks against image classification. *Applied Soft Computing*, 161:111729, 2024. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2024.111729. URL https://www.sciencedirect.com/science/article/pii/S1568494624005039.

G. Cohen, G. Sapiro, and R. Giryes. Detecting adversarial samples using influence functions and nearest neighbors. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 14441–14450, 2020. ISSN 1063-6919. doi: 10.1109/CVPR42600.2020.01446.

N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *ArXiv*, abs/1705.02900, 2017. URL https://api.semanticscholar.org/CorpusID:9377405.

G. K. Dziugaite, Z. Ghahramani, and D. M. Roy. A study of the effect of jpg compression on adversarial images, 2016.

H. F. Eniser, M. Christakis, and V. Wüstholz. Raid: Randomized adversarial-input detection for neural networks. *ArXiv*, abs/2002.02776, 2020. URL `https://api.semanticscholar.org/CorpusID:211066398`.

K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard. The robustness of deep networks: A geometrical perspective. *IEEE Signal Processing Magazine*, 34(6):50–62, 2017.

R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts. *ArXiv*, abs/1703.00410, 2017a. URL `https://api.semanticscholar.org/CorpusID:5157373`.

R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner. Detecting adversarial samples from artifacts, 2017b.

H. Gao, B. Cheng, J. Wang, K. Li, J. Zhao, and D. Li. Object classification using cnn-based fusion of vision and lidar in autonomous vehicle environment. *IEEE Transactions on Industrial Informatics*, 14(9):4224–4231, 2018. doi: 10.1109/TII.2018.2822828.

Z. Gong and W. Wang. Adversarial and clean data are not twins. In *Proceedings of the Sixth International Workshop on Exploiting Artificial Intelligence Techniques for Data Management*. Association for Computing Machinery, 2023. doi: 10.1145/3593078.3593935.

I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples, 2015.

G. Goswami, A. Agarwal, N. Ratha, R. Singh, and M. Vatsa. Detecting and mitigating adversarial perturbations for robust face recognition. *International Journal of Computer Vision*, 127:719–742, 2019.

K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. Mcdaniel. On the (statistical) detection of adversarial examples. *ArXiv*, abs/1702.06280, 2017a. URL `https://api.semanticscholar.org/CorpusID:16863734`.

K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel. On the (statistical) detection of adversarial examples, 2017b.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017.

A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2137–2146. PMLR, 2018.

F. Karim, S. Majumdar, and H. Darabi. Adversarial attacks on time series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3309–3320, 2021.

T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. doi: 10.1109/CVPR.2019.00453.

Z. Katzir and Y. Elovici. Detecting adversarial perturbations through spatial behavior in activation spaces. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9, 2019. doi: 10.1109/IJCNN.2019.8852285.

A. Kherchouche, S. A. Fezza, W. Hamidouche, and O. Déforges. Detection of adversarial examples in deep neural networks with natural scene statistics. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2020. doi: 10.1109/IJCNN48605.2020.9206959.

A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world, 2017.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521, 2015. doi: 10.1038/nature14539.

K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 7167–7177. Curran Associates Inc., 2018a.

K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018b.

X. Li and F. Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5775–5783, 2017. doi: 10.1109/ICCV.2017.615.

Y. Li, P. Angelov, and N. Suri. Adversarial attack detection via fuzzy predictions. *IEEE Transactions on Fuzzy Systems*, 32 (12):7015–7024, 2024. doi: 10.1109/TFUZZ.2024.3473768.

B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Transactions on Dependable and Secure Computing*, 18(1):72–85, 2021. doi: 10.1109/TDSC.2018.2874243.

X. Ling, L. Wu, J. Zhang, Z. Qu, W. Deng, X. Chen, Y. Qian, C. Wu, S. Ji, T. Luo, et al. Adversarial attacks against windows pe malware detection: A survey of the state-of-the-art. *Computers & Security*, page 103134, 2023.

W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. Ssd: Single shot multibox detector. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing, 2016.

Y. Liu, Y. Cheng, L. Gao, X. Liu, Q. Zhang, and J. Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15105–15114, 2022.

D. Lowd and C. Meek. Adversarial learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, page 641–647, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 159593135X. doi: 10.1145/1081870.1081950.

J. Lu, T. Issaranon, and D. Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE international conference on computer vision*, pages 446–454, 2017.

J. Lust and A. P. Condurache. Gran: An efficient gradient-norm based detector for adversarial and misclassified examples. In *28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2020, Bruges, Belgium, October 2-4, 2020*, pages 7–12, 2020. URL `https://www.esann.org/sites/default/files/proceedings/2020/ES2020-159.pdf`.

X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, M. E. Houle, D. Song, and J. Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=B1gJ1L2aW`.

G. R. Machado, E. Silva, and R. R. Goldschmidt. Adversarial machine learning in image classification: A survey toward the defender's perspective. *ACM Comput. Surv.*, 55(1), 2021. ISSN 0360-0300. doi: 10.1145/3485133.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations, 2017a.

J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff. On detecting adversarial perturbations. In *International Conference on Learning Representations*, 2017b. URL `https://openreview.net/forum?id=SJzCSf9xg`.

A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard. Sparsefool: A few pixels make a big difference. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9079–9088, 2019. doi: 10.1109/CVPR.2019.00930.

S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016. doi: 10.1109/CVPR.2016.282.

K. Muhammad, J. Ahmad, Z. Lv, P. Bellavista, P. Yang, and S. W. Baik. Efficient deep cnn-based fire detection and localization in video surveillance applications. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(7):1419–1434, 2019. doi: 10.1109/TSMC.2018.2830099.

H. Niwa, K. Nagata, M. Ohta, and K. Yamashita. A product identification method for a mixed-reality web shopping system. In *2016 IEEE 5th Global Conference on Consumer Electronics*, pages 1–2, 2016. doi: 10.1109/GCCE.2016.7800481.

T. Pang, C. Du, Y. Dong, and J. Zhu. Towards robust detection of adversarial examples. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597, 2016. doi: 10.1109/SP.2016.41.

N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, ASIA CCS '17. Association for Computing Machinery, 2017a. ISBN 9781450349444. doi: 10.1145/3052973.3053009.

N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning, 2017b.

S. Pertigkiozoglou and P. Maragos. Detecting adversarial examples in convolutional neural networks. *ArXiv*, abs/1812.03303, 2018. URL `https://api.semanticscholar.org/CorpusID:54461830`.

Y. Qing, T. Bai, Z. Liu, P. Moulin, and B. Wen. Detection of adversarial attacks via disentangling natural images and perturbations. *IEEE Transactions on Information Forensics*

and Security, 19:2814–2825, 2024. doi: 10.1109/TIFS.2024.3352837.

Y. Ran, A.-X. Zhang, M. Li, W. Tang, and Y.-G. Wang. Black-box adversarial attacks against image quality assessment models. *Expert Systems with Applications*, 260:125415, 2025. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2024.125415. URL `https://www.sciencedirect.com/science/article/pii/S0957417424022826`.

A. Shamir, I. Safran, E. Ronen, and O. Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance, 2019.

H. A. Shehu, W. Browne, and H. Eisenbarth. An adversarial attacks resistance-based approach to emotion recognition from images using facial landmarks. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1307–1314, 2020.

H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016. doi: 10.1109/TMI.2016.2528162.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019. doi: 10.1109/TEVC.2019.2890858.

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

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

S. Wang, L. Liu, L. Qu, C. Yu, Y. Sun, F. Gao, and J. Dong. Accurate ulva prolifera regions extraction of uav images with superpixel and cnns for ocean environment monitoring. *Neurocomputing*, 348:158–168, 2019. doi: https://doi.org/10.1016/j.neucom.2018.06.088. Advances in Data Representation and Learning for Pattern Analysis.

T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. *ArXiv*, abs/1801.10578, 2018.

W. Xu, D. Evans, and Y. Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proceedings 2018 Network and Distributed System Security Symposium*. Internet Society, 2018. doi: 10.14722/ndss.2018.23198.

P. Yang, J. Chen, C.-J. Hsieh, J.-L. Wang, and M. Jordan. Ml-loo: Detecting adversarial examples with feature attribution. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04), 2020. doi: 10.1609/aaai.v34i04.6140.

Y. Yang, R. Gao, Y. Li, Q. Lai, and Q. Xu. What you see is not what the network+infers: Detecting adversarial examples based on semantic contradiction. In *Proceedings 2022 Network and Distributed System Security Symposium*. Internet Society, 2022. doi: 10.14722/ndss.2022.24001. URL `http://dx.doi.org/10.14722/ndss.2022.24001`.

S. Zhang, S. Chen, X. Liu, C. Hua, W. Wang, K. Chen, J. Zhang, and J. Wang. Detecting adversarial samples for deep learning models: A comparative study. *IEEE Transactions on Network Science and Engineering*, 9(1):231–244, 2022. doi: 10.1109/TNSE.2021.3057071.

Z. Zheng and P. Hong. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018a.

Z. Zheng and P. Hong. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. *Advances in Neural Information Processing Systems*, 31, 2018b.

F. Zuo and Q. Zeng. Exploiting the sensitivity of l2 adversarial examples to erase-and-restore. In *Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security*, ASIA CCS '21, page 40–51. Association for Computing Machinery, 2021. ISBN 9781450382878. doi: 10.1145/3433210.3437529.