

Adversarial Robustness Enhancement through Deep Feature Representations-Based CIFAR-10 Image Classification

Malik D. Al-Sharaf

University of Al-Hamdaniya, Nineveh, Iraq.

*Corresponding authors' email: malik.al-sharaf@uohamdaniya.edu.iq

ABSTRACT

Adversarial perturbations pose a significant challenge to the reliability of deep neural networks in image classification. The conventional feature aggregation approaches are still vulnerable to adversarial noise, which inspires the improvement of more powerful feature presentations. In this study, the adversarial robustness of a ResNet34-based architecture is investigated on the CIFAR-10 dataset, where conventional global average pooling (GAP) is compared with a proposed framework combines Global Average Pooling (GAP) and Fractal Average Pooling (FAP) within a ResNet-34. Multiple perturbation budgets were tested on the models via the white-box fast gradient sign method (FGSM) and projected gradient descent (PGD) attacks. The test accuracy of the proposed model was 97.52% on clean data whereas the test accuracy of the GAP baseline was 95.69%, which is a 1.83% improvement. In the FGSM, the adversarial accuracy increased from 59.56% to 81.38% at $\epsilon = 1/255$ and from 15.81% to 70.81% at $\epsilon = 8/255$. With the PGD, the improvement was not significant; the accuracy increased from 0.06% to 14.99% at $\epsilon = 1/255$ but reached only 0.32% versus 0.03% at $\epsilon = 8/255$. The distortion characteristics of any of the above methods are highly similar to the perturbation-quality analysis results of the PSNR, SSIM, SNR, and MSE. Overall, the proposed representation enhances the ability to perform clean classification tasks while also being robust to FGSM attacks, while stronger iterative PGD attacks are still a major challenge.

Received: 21 Sep. 2026

Accepted: 10 Oct. 2026

Published: 12 Oct. 2026

Keywords: CIFAR-10, Resnet34, FGSM, PGD, Fractal Average Pooling

INTRODUCTION

Artificial intelligence (AI)-based classification models are an integral part of modern pattern recognition systems and are used in various applications, such as medical image analysis, biometric identification, autonomous systems, surveillance, and cybersecurity (J. Wang et al., 2026). In particular, deep convolutional neural networks (CNNs) have been proven to be very effective in automatically learning discriminative feature representations from visual data and to be highly accurate in various image recognition tasks (Dargan et al., 2020). However, deep neural networks are vulnerable to adversarial perturbations, which are carefully designed and generally imperceptible alterations to the input that can lead to incorrect predictions by the model and minimal differences in the original image (Chen et al., 2021) (Shehu et al., 2026). In the context of evaluating the security, robustness and reliability of deep learning-based recognition systems, adversarial attacks have thus become a leading concern (Madry et al., 2017). However, if these decision boundaries are learned by a neural network, careful perturbations can cause an input image that is known to be correctly classified to be misclassified (Zhou et al., 2022). The fast gradient sign method (FGSM) is among the most widely researched gradient-based attack techniques that create adversarial examples via an efficient single-step perturbation procedure (Tanay & Griffin, 2016). Projected gradient descent (PGD), on the other hand, uses an iterative optimization procedure, where the perturbation is updated continuously and the new perturbation is kept in a specified constraint region (Goodfellow et al., 2014). Because of its multistep refinement, the PGD is typically considered a more robust and more stringent method for assessing model vulnerability. Thus, the evaluation of deep learning models against FGSM

and PGD attacks can provide complementary information regarding their ability to be manipulated by adversarial inputs and their overall robustness (Villegas-Ch et al., 2024).

In addition to the choice of adversarial attack method, the quality and discriminative power of the learned feature representation can significantly impact the robustness of a classification model (Xu et al., 2021). Global average pooling (GAP) is often used to reduce the size of spatial feature maps to compact feature vectors. While GAP is easy to implement and computationally efficient, it can cause the loss of potentially important information on spatial variability, local structures, and irregular patterns in deep feature maps (Li et al., 2025) (Karjauv, 2026). This constraint has led to research on alternative methods of feature aggregation, such as fractal pooling methods (Florindo, 2024).

Florindo, (2024) added a fractal pooling method. In this method, feature maps output by convolutional architectures are first converted into residual feature maps and then pooled together via differential box-counting estimation of the fractal dimension. Given these as examples, the fractal dimension computed for every feature map is assigned as a weighting component and added to its average activation response to create the last feature representation. The reported results show that the use of fractal information can enhance deep feature representations, particularly for texture-based recognition tasks. The benefits of pairing fractal-based pooling with conventional global average pooling, however, and its potential role in boosting adversarial robustness have not yet been adequately studied.

To fill this research gap, the current study proposes a ResNet34-based classification framework that uses fractal average pooling (FAP) and global average pooling (GAP). This study investigates whether the combination of

complementary features of GAP and FAP can lead to a more informative and discriminative feature representation than the use of GAP alone, instead of proposing FAP as a new pooling technique. The proposed architecture, namely, ResNet34 + GAP + FAP, is evaluated against a baseline ResNet34 + GAP model on the CIFAR-10 dataset in both clean and adversarial settings.

The major findings of this study are summarized below:

- i. A hybrid architecture of feature aggregation is explored by using fractal average pooling (FAP) along with global average pooling (GAP) in a ResNet34 classifier. FAPs are well-known feature aggregations that are based on fractals and are not suggested as a new pooling algorithm.
- ii. The ResNet34 + GAP and ResNet34 + GAP + FAP models are evaluated on the CIFAR-10 dataset for the white-box Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) attacks for various perturbation budgets and then compared for adversarial robustness.
- iii. Clean accuracy, adversarial accuracy, the attack success rate (ASR), and accuracy degradation are used to quantify the correlation between feature aggregation strategies and adversarial classification accuracy.
- iv. Further validation of the perceptual quality and magnitude of the adversarial perturbations generated is provided by the peak signal-to-noise ratio (PSNR), the structural similarity index measure (SSIM), the signal-to-noise ratio (SNR), and the mean squared error (MSE), which offer complementary evidence of the numerical intensity and structural influence.

Overall, the objective of this study is to evaluate whether the classification performance and empirical adversarial robustness can be improved in ResNet34 without adding any explicit adversarial training strategy or dedicated defense mechanism by augmenting the conventional GAP with feature aggregation based on fractal analysis.

Related Work

However, the ability to achieve high accuracy in classification on the original data does not guarantee that it will be robust to deliberate perturbations, making adversarial robustness a key area of recent deep learning research. Adversarial examples take advantage of weaknesses in learned decision boundaries by making small, limited changes to samples of input data that can cause the model to predict incorrectly. The literature classifies such attacks on the basis of the adversary's knowledge of the model and focuses on two threat scenarios, white-box and black-box, with a strong focus on the vulnerability of deep neural networks to gradient attacks. Moreover, white-box attacks are a fundamental testbed for the assessment of robustness; the adversary obtains every piece of information about the target model, such as the architecture, parameters, and gradients of the model, which allows the adversary to directly generate very effective inputs. In this context, two basic gradient-based methods are still prevalent for assessing the robustness of visual recognition models: the fast gradient sign method (FGSM) and the projected gradient descent (PGD) (J. Liu et al., 2024).

The fast gradient sign method (FGSM) is a gradient-based attack strategy that offers a simple single-step algorithm to create adversarial examples by making perturbations along the direction of the sign of the loss gradient, which is also low in computation. Projected gradient descent (PGD), on the other hand, learns the adversarial samples by iteratively updating the perturbation, projecting the updated input onto a given feasible perturbation set. PGD is typically a more

rigorous first-order method than single-step methods for assessing adversarial robustness, as it can be iterated to gradually improve perturbations over time. The FGSM and PGD are thus widely used for evaluating the vulnerability of CNN-based classifiers to adversarial manipulations in the "white box" setting (Wong et al., 2020). In recent empirical studies, deep learning models have been compared with adversarial attack models such as the FGSM and PGD and benchmark datasets such as CIFAR-10, and all have shown significant loss of classification accuracy with adversarial perturbation (Riski et al., 2025).

CIFAR-10 has been extensively employed to evaluate the effectiveness of convolutional neural network architectures and feature-representation approaches. (Shin et al., 2025) implemented and compared AlexNet-based and ResNet-based models on the CIFAR-10 dataset, reporting superior classification performance for the ResNet-based architecture, which achieved a final test accuracy of 90.69%. Their findings further supported the capacity of residual networks to learn discriminative visual representations from CIFAR-10 images. However, the study was limited to conventional classification under clean-data conditions and architectural comparison; it did not consider adversarial attacks or examine alternative feature-aggregation mechanisms as a means of improving adversarial robustness.

Mehra et al. (2026) investigated the performance of FGSM adversarial training on CNNs with the CIFAR-10 and MNIST datasets. In their experimental assessment, they tested several magnitudes of perturbations and different ratios of adversarial and clean samples for training. The baseline models were found to be quite vulnerable to FGSM attacks on the CIFAR-10 dataset, whereas adversarial training achieved some level of robustness, which varied with the perturbation budget and the ratio of clean to adversarial samples. Remarkably, a balanced 50:50 split between clean and adversarial samples led to relatively consistent results, with an average combined accuracy of 71%. A ratio of $\epsilon=0.5$ yielded the best results, whereas other ratios led to less uniform results. The authors found that adversarial training based on the FGSM can improve robustness against FGSM attacks but is not very effective against more powerful multistep, iterative adversarial attacks.

Swain et al. (2026) studied the vulnerability of a ResNet-34 model trained with the CIFAR-10 dataset to adversarial perturbations created via a projected gradient descent (PGD) attack. This study determined the effects of increasing levels of perturbation on classification accuracy and image quality in terms of accuracy, the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM), the signal-to-noise ratio (SNR), and the mean squared error (MSE). The results demonstrated that the PGD could significantly impact the accuracy of classification but keep the perceptual similarity between an original image and an adversarial image relatively high. Furthermore, the authors compared the FGSM, PGD, mean ensemble attack (MEA), and weighted ensemble attack (WEA), concluding that the PGD is highly effective in attacking the model. The investigation, however, was focused on the effectiveness of adversarial perturbations and the degradation in image quality that comes with them, without studying the role that feature aggregation mechanisms can play in enhancing adversarial robustness.

Riski et al. (2025). Studied the ability of a ResNet-18 model trained with CIFAR-10 data to resist adversarial examples created with the fast gradient sign method (FGSM). They reported that the more severe the perturbation was, the poorer the classification performance was 43.39 on clean samples at $\epsilon=0.05$. The study also explored adversarial training and

distillation for defense, which is a general training-based defense method, and discussed the difficulty in achieving high classification accuracy on clean data while also ensuring high resistance to adversarial attacks.

Eleftheriadis et al. (2024) investigated adversarial robustness on the CIFAR-10 dataset under several attack and defense configurations. Their experiments involving a ResNet model demonstrated pronounced susceptibility to white-box adversarial attacks, with reported attack success rates of 96.14% for the fast gradient sign method (FGSM) and 100% for the projected gradient descent (PGD). These findings underscore the effectiveness of gradient-based, particularly iterative, attacks in substantially compromising the performance of conventional deep neural classifiers. Nevertheless, the study concentrated primarily on adversarial attack and defense strategies and did not investigate the potential contribution of feature aggregation mechanisms to adversarial robustness.

Hu et al. (2024) introduced a new approach called feature pattern consistency constraint (FPCC) to increase adversarial robustness by ensuring that the latent feature patterns remain consistent under adversarial noise. The approach was tested on both the CIFAR-10 and the CIFAR-100 datasets via several adversarial attacks, such as the fast gradient sign method (FGSM) and projected gradient descent (PGD). The results are very impressive on CIFAR-10, where the FPCC outperforms the baseline by 12.50%. PGD The baseline system that we used on the WRN-28--10 architecture reached only 0.03% under the same evaluation environment. The results suggest that by leveraging improved deep feature representations in terms of both quality and stability, substantial improvements in adversarial robustness can be achieved. FPCC does not introduce an alternative feature-aggregation mechanism to achieve this improvement; rather, it does so explicitly by way of feature-consistency constraints and seeks additional optimization objectives.

Pang et al. (2025) studied adverse robustness on the CIFAR-10 dataset via the ResNet-18 model and compared the model's vulnerability to FGSM and PGD attacks. The standard model results were approximately 95% correct on clean samples, but it decreased to 72% when tested with the FGSM and to 0.7%

when tested with the PGD, showing that the iterative PGD perturbations were much more effective. After adversarial training, the accuracy of the model increased significantly: approximately 89% with respect to the FGSM and 57% with respect to the PGD. The study primarily targeted the adversarial training process instead of exploring how feature aggregation strategies could contribute to adversarial robustness, even though the results demonstrate the effectiveness of adversarial training in making a model more resistant to adversarial manipulation.

Villegas-Ch et al. (2024) assessed the robustness of deep learning models trained on the CIFAR-10 dataset against multiple adversarial attack methods, including the fast gradient sign method (FGSM), projected gradient descent (PGD), and Carlini-Wagner (CW) attack. Their findings revealed substantial reductions in classification performance under adversarial perturbations, thereby highlighting the susceptibility of conventional image-classification models to white-box attacks. Nevertheless, the study focused primarily on the effectiveness of adversarial attacks and defense strategies without directly examining whether alternative feature aggregation mechanisms could enhance adversarial robustness.

MATERIALS AND METHODS

Overall Proposed Framework

In this paper, a fractal-aware feature aggregation framework is introduced to assess adversarial robustness on CIFAR-10 via ResNet-34. The framework is shown in Fig. 1, which consists of preprocessing, feature extraction, feature aggregation, classification, and adversarial evaluation. The baseline ResNet-34 + GAP is a ResNet-34 network with GAP, whereas the proposed network ResNet-34 + GAP + FAP consists of a ResNet-34 network + GAP + FAP, which concatenates features. The same experimental setup is employed for both models for a controlled evaluation of the FAP. The robustness is measured in the context of white-box FGSM and PGD attacks with adversarial accuracy, accuracy degradation, and ASR. The PSNR, SSIM, SNR and MSE are also used for characterizing adversarial perturbations.

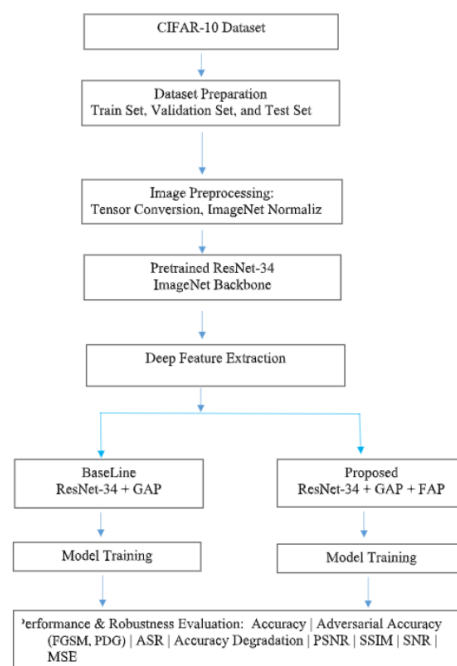


Figure 1: Overall proposed framework

Dataset Preparation and Image Preprocessing

The proposed framework is assessed via the CIFAR-10 dataset, a widely used image-classification benchmark containing RGB images across ten object categories at an original spatial resolution of 32×32 pixels. The dataset is stored separately in files containing the corresponding images and class labels, comprising 60,000 RGB images from 10 classes. was divided into 45,000 training images, 5,000 validation images, and 10,000 test images, with the latter retained solely for final model performance assessment (Krizhevsky & Hinton, 2009).

The preprocessing pipeline includes the original 32×32 RGB images were converted to tensors and resized to 224×224 pixels, followed by ImageNet normalization to match the input requirements of the ImageNet-pretrained backbone the adopted channel wise mean and standard (He et al., 2016).

ResNet-34 Deep Feature Extraction

To obtain deep features, ResNet-34 is used as the backbone of convolutional networks. The network is based on the residual learning architecture from which the network learns hierarchical feature representations effectively through residual connections (Liao et al., 2025). With an input image provided by x , the ResNet-34 backbone generates a feature tensor that is deep convolutional as follows:

$$F = R_{34}(x) \quad (1)$$

Where F is the resulting deep feature tensor and x is the intermediate feature tensor. The extracted tensor is the high-level spatial activation map produced from the final convolutional layer of the network, thus maintaining valuable spatial properties in the learned feature representation. The extracted feature tensor is then passed to the feature aggregation layer of the network and finally to the last classification layer

Fractal Average Pooling (FAP)

The new fractal average pooling (FAP) approach is based on the fractal pooling strategy proposed by (Florindo, 2024). This original framework consists of three main phases: deep feature extraction, residual feature encoding, and fractal pooling. In this work, the underlying fractal pooling principle is used to obtain a deep feature representation from ResNet-34, which is then utilized to create a complexity-aware feature descriptor. Let $F = R_{34}(x)$ in Eq. (1) be the deep feature tensor extracted from an input image x . Next, a residual feature representation is calculated by performing a transfer operation on the extracted feature maps and subtracting the resultant feature representation from the original feature tensor, as defined in Eqs. (2) and (3):

$$F' = \text{ReLU}(\text{BN}(\text{Conv}_{(1 \times 1)}(F))) \quad (2)$$

$$O = F' - F \quad (3)$$

Where F' denotes the transferred feature representation and O represents the resulting residual feature tensor. The new fractal average pooling (FAP) approach is based on the fractal pooling strategy proposed by Florindo. This original framework consists of three main phases: deep feature extraction, residual feature encoding, and fractal pooling. In this work, the underlying fractal pooling principle is used to obtain a deep feature representation from ResNet-34, which is then utilized to create a complexity-aware feature descriptor. Let $F = R_{34}(x)$ be the deep feature tensor extracted from an input image x . Next, a residual feature representation is calculated by performing a transfer operation on the extracted feature maps and subtracting the resultant feature representation from the original feature tensor.

For each of the residual feature maps O_k , the fractal dimension is calculated via the differential box-counting method. In particular, max pooling is performed over several spatial scales, $\delta = 1, 2, \delta_{\max}$, and the quantities in each of the feature maps are summed to obtain the quantity at the scale. $N\delta$ as defined in Eq. (4)

$$N_{\delta} = \sum_{i,j} (\text{MaxPool}_{(\delta \times \delta)}(O_k))_{i,j} \quad (4)$$

Fractal dimension D_k of k . The bilogarithmic relationship between the k -th feature map and Y is then least squares regressed to estimate the k -th feature map. δ and $N\delta$. This allows for the capture of the multiscale complexity of the residual activation distribution and generates a compact fractal descriptor for each feature map, as defined in Eq. (5):

$$D_k = \left(\frac{\delta_{\max} \sum_{\delta=1}^{\delta_{\max}} (\log(\delta) \log(N_{\delta})) - (\sum_{\delta=1}^{\delta_{\max}} \log(\delta)) (\sum_{\delta=1}^{\delta_{\max}} \log(N_{\delta}))}{\delta_{\max} \sum_{\delta=1}^{\delta_{\max}} \log(\delta)^2 - (\sum_{\delta=1}^{\delta_{\max}} \log(\delta))^2} \right) \quad (5)$$

The obtained fractal dimension D_k is subsequently used as a feature weight. The average activation A_k of the corresponding residual feature map is calculated as:

$$A_k = \left(\frac{1}{HW} \right) \sum_{i,j=1}^{H,W} O_k(i,j) \quad (6)$$

The resulting fractal-aware feature is obtained by weighting the average activation with its corresponding fractal dimension:

$$fk = D_k A_k \quad (7)$$

Accordingly, the complete FAP representation for C feature maps is expressed as:

$$f_{\text{FAP}} = [D_{1A_1}, D_{2A_2}, \dots, D_{CA_C}]^T \quad (8)$$

Thus, the original spatial feature tensor is transformed into a compact C -dimensional fractal-aware representation. This representation is subsequently used in the proposed GAP-FAP feature aggregation framework.

GAP-FAP Feature Aggregation

The idea of aggregation combines the traditional global average pooling (GAP) representation and the fractal-aware feature representation obtained via fractal average pooling (FAP). The two representations share the same deep feature extraction backbone and are combined to create a single feature vector for classification. Let $g \times$ denote the GAP representation. f : Representation by FAP. The two feature vectors are joined together along the feature dimension in the following way:

$$z = [g; f] \quad (9)$$

Where z is the fused feature vector. This aggregation approach preserves the global semantic information collected by GAP but also adds additional complexity-aware information collected by FAP. The fused representation is then fed into a fully connected classification layer to generate class logits.

These logits are then passed through the Softmax function to obtain the posterior class probabilities:

$$P(y = c|z) = \frac{\exp(o_c)}{\sum_{j=1}^C \exp(o_j)} \quad (10)$$

Where o_c the following shows the logit corresponding to the class. Here the logit corresponding to the class is c , and C is the number of all classes. The class with the highest posterior probability is the one that is selected to obtain the predicted label:

$$y^* = \arg\max_c P(y = c|z) \quad (11)$$

Hence, the architecture of ResNet-34 + GAP + FAP consists of traditional global feature aggregation in conjunction with a complexity-aware fractal representation to further improve the discriminative ability of the final feature descriptor.

Adversarial Attack Generation

To assess the adversarial robustness of the baseline and proposed models, two popular white-box, gradient-based attack methods are considered: the fast gradient sign method (FGSM) and the projected gradient descent (PGD). The attacks offer complementary robustness assessments from a computationally efficient single-step perturbation attack to a more powerful iterative optimization-based attack. The FGSM was designed by Goodfellow et al. (Goodfellow et al., 2014). It operates by basing the adversarial example on the gradient of the classification loss function with respect to the original input image. More specifically, the method uses only one perturbation in the direction that causes the greatest loss in the model. Owing to its "one step" formulation, the FGSM is an efficient method for evaluating the susceptibility of a trained classifier to adversarial perturbations. They introduced an iterative optimization procedure for gradient-based adversarial generation, called the PGD, by (Madry et al., 2017). At each iteration, the adversarial input is modified on the basis of the loss gradient and then mapped onto a specified perturbation set to keep the perturbation within bounds. Therefore, PGD is a significantly more powerful first-order adversarial attack and has been widely adopted to assess the security of deep neural networks.

Experimental Configuration

The main parameters of the experiments used for the training and adversarial attack of the ResNet-34+GAP and the ResNet-34+GAP+FAP models were evaluated with the same attack budgets and evaluation process. Both models were trained with the AdamW optimizer, a learning rate of 0.0001, a batch size of 32, and 25 epochs, using a fixed random seed of 42. In the adversarial evaluation, FGSM was used with perturbation budget values of $\epsilon = 1/255$, $2/255$, $4/255$, and $8/255$. The algorithm (PGD) was applied for 10 iterations, with the step size α equal to $\epsilon/4$ at each iteration. The same attack budgets were used for both models in the evaluation, as well as the same attack configuration.

Evaluation Metrics

We tested the baseline and proposed models with three metrics: classification performance, adversarial robustness and perturbation quality. Together, these measures assess predictive performance, the impact of adversarial attacks on model decisions, and the size and perceptual similarity of the generated adversarial attacks.

Classification accuracy: Classification accuracy is the ratio of the number of samples classified correctly to the number of samples that are evaluated and is expressed as a percentage (Tharwat, 2021).

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \times 100 \quad (12)$$

Attack success rate (ASR): The attack success rate is the ratio between the number of correctly classified samples and the number of misclassified samples after applying an adversarial perturbation (Gupta et al., 2024). When the untargeted attacks studied here are considered, the definition of ASR is as follows:

$$\text{ASR} = \frac{N_{\text{successful attacks}}}{N_{\text{correct}}} \times 100 \quad (13)$$

Where $N_{\text{successful attacks}}$ the number of is correctly classified samples initially but misclassified after perturbation and where N_{correct} is the number of samples that are correctly classified before the attack. The smaller the ASR is, the greater the degree of adversarial robustness. This is a typical success rate measure used to measure the effectiveness of adversarial attacks.

Accuracy Degradation. The reduction in classification accuracy due to one's adversarial attack is quantified and measured as accuracy degradation is calculated as:

$$AD = \text{Accuracy}_{\text{clean}} - \text{Accuracy}_{\text{adv}} \quad (14)$$

Where $\text{Accuracy}_{\text{clean}} - \text{Accuracy}_{\text{adv}}$ refer to the classification accuracy before and after adversarial perturbation, respectively. The lower the accuracy degradation is, the less the model will suffer from attack.

The mean squared error (MSE) is a measure of the average squared error per pixel between a set of input and output images (Aarthy & Keerthi, 2023). I_i and its opposite image, I_i^{adv} , I advise that it is widely adopted to evaluate the severity of visual perturbations in the generation of adversarial images:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (I_i - I_i^{\text{adv}})^2 \quad (15)$$

Where N denotes the number of evaluated pixel values. A lower MSE indicates a smaller numerical difference between the original and adversarial images.

The peak signal-to-noise ratio (PSNR) is used to measure the visual fidelity of an adversarial image; with respect to its original counterpart, the PSNR is a commonly used measure to assess the degree of distortion caused by adversarial perturbations and is defined as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{(\text{MAX}_I)^2}{\text{MSE}} \right) \quad (16)$$

Where MAX_I represents the maximum possible pixel intensity. A higher PSNR indicates lower pixel-level distortion. The PSNR is closely related to the MSE and is widely used for full-reference image-quality assessment (Keleş et al., 2021).

The structural similarity index measure (SSIM) is used to measure the perceptual and structural similarity between an original image and its adversarial counterpart (Z. Wang et al., 2004). This is not only a pixel wise metric but also considers the differences in luminance, contrast, and structure. On the basis of the work the SSIM is given by:

$$\text{SSIM}(xy) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (17)$$

Here, μ_x and μ_y denote the mean intensities at the local points of the two images, and σ_x^2 and σ_y^2 represent the local variations of each. The term σ_{xy} denotes the global covariance value denotes the covariance value of the image at the current point, denoted as the local covariance; C_1 and C_2 are small positive constants, which are added for numerical stability. Higher SSIM values indicate that structural information has been preserved and that there are more perceptual similarities between the original and adversarial images.

Signal-to noise ratio (SNR). The SNR quantifies the strength of the original image signal relative to the introduced perturbation (C. Liu et al., 2023) and is calculated as:

$$\text{SNR} = 10 \log_{10} \left(\frac{\sum_{i=1}^N (I_i^2)}{\sum_{i=1}^N ((I_i - I_i^{\text{adv}})^2)} \right) \quad (18)$$

A higher SNR indicates that the perturbation has a smaller magnitude relative to the original image signal

Algorithm 1. Proposed GAP-FAP Adversarial Robustness Framework

Input: Input: Image dataset Output:

Out Put: Clean and adversarial performance of baseline and proposed models

1. Prepare training, validation, and test sets
2. Apply pre-processing and ImageNet normalization
3. Extract deep features _using ResNet-34: $F \leftarrow R34(x)$
4. Baseline: $zB \leftarrow \text{GAP}(F)$
5. Proposed: $f \leftarrow \text{FAP}(F)$ $zP \leftarrow [zB; f]$

6. Train ResNet34 + GAP and ResNet34 + GAP + FAP
7. Select the best models using validation performance
8. Evaluate both models on:
 - i. Clean test images
 - ii. FGSM adversarial images
 - iii. PGD adversarial images
 - iv. Compute accuracy and robustness measures

RESULTS AND DISCUSSION

Clean Classification Performance

The performance of the models evaluated in terms of classification accuracy on clean images is shown in Table 1. The ResNet34+ global average pooling (GAP) model was used as the baseline model with training, validation, and test

accuracies of 97.69%, 95.76%, and 95.69%, respectively. The proposed ResNet34 + GAP + fractal average pooling (FAP) model achieved a higher training accuracy of 99.68%, a validation accuracy of 97.73% and a test accuracy of 97.52%. In particular, the proposed model achieves an increase in the clean-test accuracy of 1.83% compared with the ResNet34 + GAP baseline. Additionally, the gap between the validation and test performance of the proposed GAP+FAP architecture was 0.21 percentage points, which demonstrates that the network was stable and learned well under the test conditions of new samples. The training-to-validation accuracy differences were also very similar between the two models (1.95 percentage points for the GAP+FAP model and 1.93 percentage points for the GAP model). These findings suggest that there was no notable increase in the generalization gap with improved classification performance when the FAP was used.

Table 1: Clean Classification Performance

Model	Train Acc	Validation Acc	Clean Test Acc
ResNet34 + GAP	97.47%	95.67%	95.69%
ResNet34 + GAP + FAP	99.68%	97.73%	97.52%

Robustness against the FGSM Attack

A summary of the FGSM evaluation results is presented in Table 2, and the trends in adversarial accuracy are shown in Figure 1. For all the considered perturbation sizes, the proposed ResNet34 + GAP + FAP model proved to be more resistant to FGSM-generated perturbations than the GAP model, which served as the baseline model.

At $\epsilon = 1/255$, the adversarial accuracy increased from 59.56% to 81.38% for the proposed GAP+FAP configuration, with an increase of 21.82% over the GAP baseline. Similarly, gains of 14.01 and 24.70 percentage points were obtained for improvements at $\epsilon = 2/255$ and $\epsilon = 4/255$, respectively. The greatest disparity was observed at $\epsilon = 8/255$. In the case of the GAP+FAP model, the accuracy was 70.81% in the adversarial setting, whereas the GAP baseline was 15.81%. This represents a considerable improvement of 55.00 percentage points.

In addition, the advantages of robustness offered by the proposed feature representation are further confirmed by the attack success rate (ASR) and the accuracy-degradation

results presented in Table 2. At $\epsilon = 8/255$, the ASR decreased from 84.32% to 27.54% with the incorporation of FAP in the GAP model ($\epsilon=8/255$). Similarly, the accuracy degradation decreased markedly from 79.88 to 26.71 percentage points. Overall, the results indicate that adding FAP significantly reduces the robustness of the classifier to FGSM perturbations, especially when the perturbation budget is large.

As illustrated in Figure 2, the adversarial accuracy of the GAP baseline is shown to be no monotonic with empirical behavior: its adversarial accuracy is 59.56% at $\epsilon = 1/255$ to 66.43% at $\epsilon = 2/255$, which then decreases precipitously at higher values of perturbation. The reverse, however, is true for the GAP+FAP model, which shows a more systematic and stable decline in the adversarial accuracy from 81.38% to 70.81% as ϵ increases. This more stable tendency suggests that the representation is more robust, as it is attacked by increasingly stronger FGSM attacks.

Table 2: Adversarial Robustness of ResNet34 with GAP and GAP+FAP under FGMS

Epsilon	GAP Adv. Accuracy (%)	GAP+FAP Adv. Accuracy (%)	GAP ASR (%)	GAP+FAP ASR (%)	GAP Accuracy Degradation (pp)	GAP+FAP Accuracy Degradation (pp)
1/255	59.56	81.38	37.77	16.55	36.13	16.14
2/255	66.43	80.44	30.93	17.51	29.26	17.08
4/255	55.31	80.01	43.21	17.96	40.38	17.51
8/255	15.81	70.81	84.32	27.54	79.88	26.71

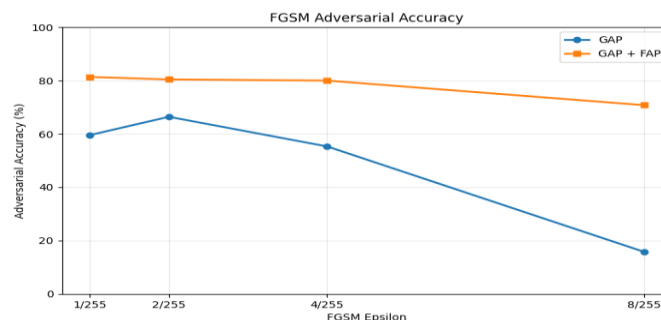


Figure 2: Adversarial accuracy of ResNet34 + GAP and ResNet34 + GAP + FAP under the FGSM

Robustness against PGD

The PGD evaluation results are reported in Table 3, and compared with the FGSM, the iterative PGD attack resulted in a greater decrease in the performance of both models, suggesting that the PGD is more effective at exploiting weaknesses in the learned decision boundaries.

At the baseline, the adversarial accuracy was reduced to 0.06% at $\epsilon = 1/255$, with an ASR of 99.94%. Increasing the perturbation budget resulted in only small additional changes: the adversarial accuracy reached 0.03% at both $\epsilon = 4/255$ and $\epsilon = 8/255$, whereas the ASR increased to 99.97%. This result shows the high sensitivity of the GAP model to the PGD at the lowest level of perturbation.

The robustness of the proposed GAP+FAP model was better than that of the other two models under lower perturbations. When $\epsilon = 1/255$, an adversarial accuracy of 14.99% was reached, whereas the gap baseline reached only 0.06%, and the ASR was reduced from 99.94% to 84.63%. However, these benefits diminished with increasing perturbation strength. The adversarial accuracy decreased to 3.92%,

1.13%, and 0.30% at $\epsilon = 2/255$, $\epsilon = 4/255$, and $\epsilon = 8/255$, respectively, whereas the ASR increased to 95.98%, 98.84%, and 99.69%, respectively.

The accuracy-degradation values given in Table 3 follow a similar pattern. At $\epsilon = 1/255$, the performance gap decreased from 95.63 percentage points to 82.53 by GAP+FAP. At $\epsilon = 2/255$, the degradation was also lower for GAP+FAP (93.60 percentage points) than for GAP (95.63 percentage points). At higher perturbation levels, however, there was some slight degradation in the proposed model, to 96.39 percentage points at $\epsilon = 4/255$ and 97.22 percentage points at $\epsilon = 8/255$. As shown in Figure 3, both models had a monotonic trend of decreasing adversarial accuracy with increasing PGD perturbation budgets. On the whole, GAP+FAP had greater robustness when the PGD perturbation was low; however, this robustness was reduced when the PGD perturbation was stronger during the iterative attacks. Compared with the FGSM evaluation, the FAP evaluation also achieved a greater and more stable level of robustness improvement.

Table 3: Adversarial Robustness of ResNet34 with GAP and GAP+FAP under the PGD

Epsilon	GAP Adv. Accuracy (%)	GAP+FAP Adv. Accuracy (%)	GAP ASR (%)	GAP+FAP ASR (%)	GAP Accuracy Degradation (pp)	GAP+FAP Accuracy Degradation (pp)
1/255	0.06	14.99	99.94	84.63	36.13	16.14
2/255	0.06	3.92	99.94	95.98	29.26	17.08
4/255	0.03	1.13	99.97	98.84	40.38	17.51
8/255	0.03	0.30	99.97	99.69	79.88	26.71

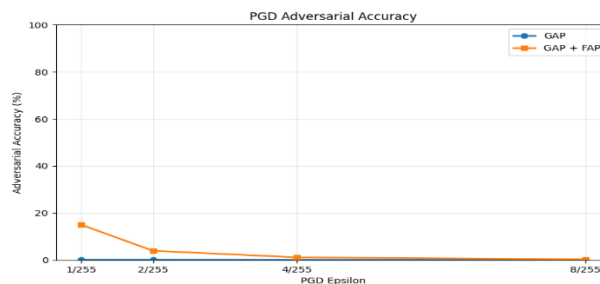


Figure3: Adversarial accuracy of ResNet34 + GAP and ResNet34 + GAP + FAP under PGD

Accuracy Degradation under Adversarial Attacks

To complement the robustness evaluation of the models, the accuracy-degradation metric is provided in the models shown in Figure 4 and is quantitatively reported in Tables 2 and 3. The results clearly show separation between the performance of the GAP+FAP model under FGSM and that under PGD attacks.

The GAP+FAP model consistently performed significantly better in terms of accuracy loss than did the GAP baseline under FGSM perturbations. For example, at $\epsilon = 8/255$, the reduction in accuracy degradation was 53.17 percentage points from GAP (79.88 percentage points) to GAP+FAP

(26.71 percentage points). These results show the large robustness gain of FAP against the single-step FGSM attack. By comparison, the PGD attack yielded significantly greater degradation for both of these models. At $\epsilon = 8/255$, the accuracy loss was 95.66 percentage points for the GAP baseline and 97.22 percentage points for the GAP+FAP model. Thus, the robustness benefit of the FAP at the highest perturbation budgets of the FGSM is not apparent. Rather, the results indicate that the advantage of the FAP holds for FGSM perturbations and is restricted to the lower perturbation levels for PGD attacks.

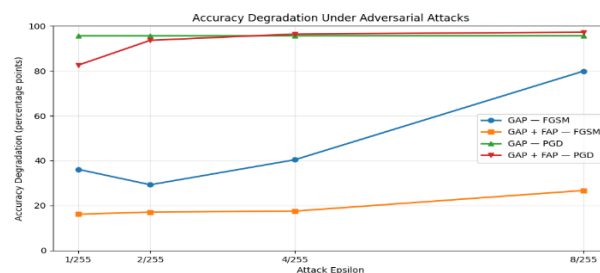


Figure 4: Accuracy degradation of ResNet34 + GAP and ResNet34 + GAP + FAP under FGSM and PGD attacks

Adversarial Perturbation Quality Analysis

The created adversarial examples that were evaluated to check if the observed differences in robustness were caused by differences in the sizes of the perturbations utilized four assessment metrics: the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), signal-to-noise ratio (SNR) and mean squared error (MSE).

FGSM Perturbation Quality

The four perturbation-quality metrics are almost comparable for both the GAP and GAP+FAP models, with all the perturbation budgets evaluated, as reported in Table 4 and shown in Figure 5. For instance, at $\epsilon = 8/255$, the PSNR values are 30.1409 dB and 30.1410 dB for GAP and GAP+FAP, respectively, whereas the corresponding SSIM values are 0.981193 and 0.981194. Likewise, the SNR and MSE values are not significantly different for the two models.

Table 4: FGSM Perturbation Quality

Eps.	GAP PSNR	GAP+FAP PSNR	GAP SSIM	GAP+FAP SSIM	GAP SNR	GAP+FAP SNR	GAP MSE	GAP+FAP MSE
1/255	48.151543	48.151588	0.999692	0.999692	42.191608	42.191658	0.000015	0.000015
2/255	42.143528	42.143562	0.998774	0.998774	36.183598	36.183632	0.000061	0.000061
4/255	36.138981	36.139011	0.995144	0.995144	30.179052	30.179083	0.000243	0.000243
8/255	30.140924	30.140959	0.981193	0.981194	24.180994	24.181028	0.000969	0.000969

As observed in Table 4 and Figure 5, the perturbation-quality measurements for GAP and GAP+FAP using the FGSM are close to each other throughout all the perturbation levels evaluated. At $\epsilon = 8/255$, for example, the PSNR is 30.1409 dB for GAP and 30.1410 dB for GAP+FAP, whereas the SSIM values are 0.981193 and 0.981194, respectively. The same agreement is found for the SNR and MSE. The results

show that both models are subjected to perturbations with almost the same signal distortion and perceptual distortion properties. Therefore, the size difference in robustness to the FGSM is unlikely to be due to a significant difference in the size of the perturbations but rather to the different feature representations created by the two architectures.

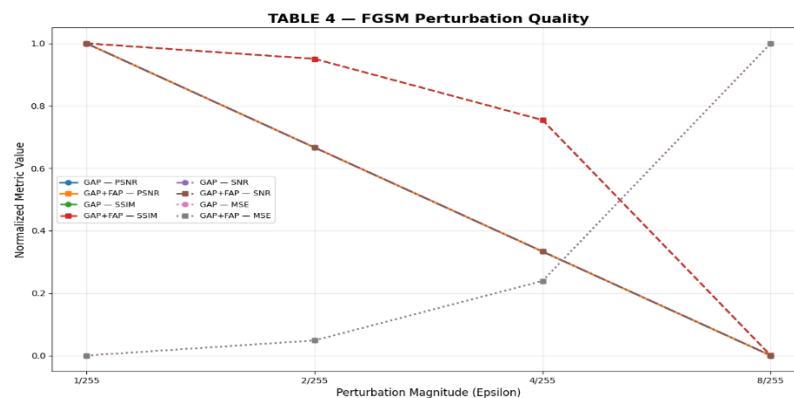


Figure 5: FGSM perturbation quality measured via PSNR, SSIM, SNR, and MSE

PGD Perturbation Quality

The results of the perturbation quality are reported in Table 5 and Figure 6 for the PGD. The models have very similar perturbation characteristics at all budgets considered. The PSNR and SSIM of GAP+FAP are slightly lower, whereas the MSE values are slightly higher than those of GAP; however, the variations are very insignificant. At $\epsilon = 8/255$, the PSNR value of GAP was 33.9975 dB, the PSNR value of

GAP+FAP was 33.9538 dB, and the MSE value of GAP increased slightly from 3.98×10^{-4} to 4.02×10^{-4} . Hence, both architectures have been tested under similar conditions regarding the quality of the perturbations, meaning that the advantage in robustness in both architectures at lower budgets for the PGD cannot be explained by a large difference in the size of the perturbations.

Table 5: PGD Perturbation Quality

Eps.	GAP PSNR	GAP+FAP PSNR	GAP SSIM	GAP+FAP SSIM	GAP SNR	GAP+FAP SNR	GAP MSE	GAP+FAP MSE
1/255	51.917469	51.368130	0.999871	0.999855	45.957531	45.408199	0.000006	0.000007
2/255	45.966091	45.729767	0.999493	0.999467	40.006165	39.769836	0.000025	0.000027
4/255	39.979996	39.892857	0.997989	0.997954	34.020069	33.932930	0.000100	0.000103
8/255	33.997467	33.953751	0.992115	0.992039	28.037533	27.993818	0.000398	0.000402

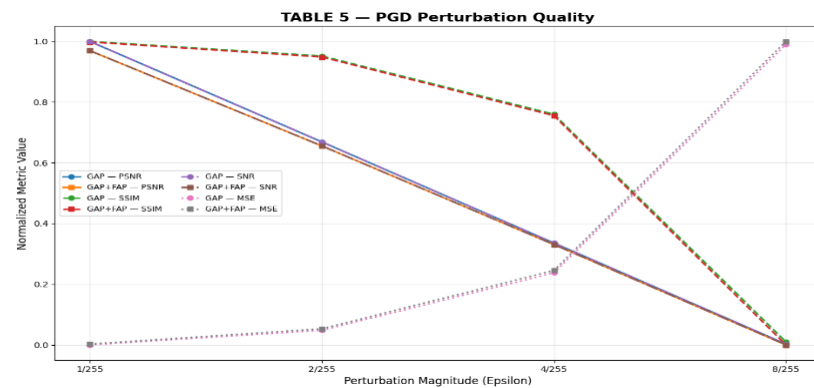


Figure 6: PGD perturbation quality measured via PSNR, SSIM, SNR, and MSE

Overall Robustness Discussion

The overall performance is summarized in Table 6. The proposed ResNet34 + GAP + FAP model performs better than the GAP baseline in terms of clean accuracy (97.52%) and is consistently better than the GAP baseline on FGSM attacks. The highest change is at $\epsilon = 8/255$, with GAP+FAP achieving an adversarial accuracy of 70.81% whereas GAP achieves 15.81% adversarial accuracy.

The robustness gain is more restricted under the PGD. At $\epsilon = 1/255$. As the perturbation strength increases, GAP+FAP increases the adversarial accuracy from 0.06% to 14.99%

with perturbation strength $\epsilon=1/255$ but decreases to 0.32% at $\epsilon=8/255$. Thus, the FAP is robust against the FGSM and low-budget PGD attacks, but it is not robust against strong iterative white-box attacks. Therefore, GAP+FAP provides more empirical robustness against single-step FGSM perturbation while against strong iterative PGD is still limited. Overall, GAP+FAP performs better in terms of clean classification performance and resistance to FGSM perturbations than GAP baseline, but not as well against stronger iterative perturbations.

Table 6: Summary of Clean and Adversarial Performance at Selected Perturbation Budgets

Model	Clean Accuracy	FGM Adv. Accuracy(1/255)	FGM Adv. Accuracy(8/255)	PGD Adv. Accuracy(1/255)	PGD Adv. Accuracy(8/255)
GAP (Baseline)	95.69	59.56	15.81	0.06	0.03
GAP + FAP (Proposed)	97.52	81.38	70.81	14.99	0.32

Comparison with Previous Studies

We compare the proposed ResNet34 + GAP + FAP model with two previous studies with ResNet architectures against the FGSM and PGD attacks published by CIFAR-10, as summarized in Table 7. ResNet34 + GAP is used as a baseline, while FAP is added to ResNet34 with GAP.

Under the FGSM, GAP+FAP achieved 70.81% accuracy at $\epsilon = 8/255$ (0.03137), whereas 43.39% accuracy was reported for the ResNet18 baseline at $\epsilon = 0.05$. Notably, under PGD, a

previous ResNet34 study reported 0% accuracy at $\epsilon = 0.03$, whereas GAP+FAP retained 0.32% accuracy at $\epsilon = 8/255$. The results indicate that the iterative attacks are more effective and that these attacks achieve a significant improvement in robustness with the FGSM. When interpreting cross-study comparisons, however, it should be remembered that there are differences in perturbation budgets and evaluation methods.

Table 7: Comparison with Previous CIFAR-10 Studies under Comparable Adversarial Settings

Study	Backbone	Pooling Configuration	Attack	ϵ	Accuracy
(Riski et al., 2025)	ResNet18	Baseline	FGSM	0.05	43.39
(Swain et al., 2026)	ResNet34	Standard	PGD	0.03	0.00
Present study	ResNet34	GAP (Baseline)	FGSM	8/255(0.03137)	15.81
Present study	ResNet34	GAP + FAP (Proposed)	FGSM	8/255(0.03137)	70.81
Present study	ResNet34	GAP (Baseline)	PGD	8/255(0.03137)	0.03
Present study	ResNet34	GAP + FAP (Proposed)	PGD	8/255(0.03137)	0.32

CONCLUSION

In this paper, the authors explored the fusion of fractal average pooling (FAP) with global average pooling (GAP) in ResNet34 for CIFAR-10 image classification and adversarial robustness. The proposed ResNet34 + GAP + FAP model achieves a clean-test accuracy of 97.52%, which is higher than the GAP baseline accuracy of 95.69%.

In the case of FGSM attacks, adding FAP significantly enhanced the adversarial robustness in all perturbed budgets evaluated. The greatest improvements were made with the children in $\epsilon = 8/255$, with the proposed model having a 55.00 percentage-point improvement over the GAP baseline. Compared with that of the PGD attacks, the performance

degradation was significantly greater in the case of the PGD attacks. While GAP+FAP provided a small robustness benefit at the lowest PGD perturbation level, the benefit became significantly lower at higher levels. ϵ increased, with both models showing almost complete degradation at increased perturbation budgets. To assess the perturbation quality, two metrics were used: the PSNR, SSIM, SNR and MSE, and the perturbation quality analysis revealed that the two architectures GAP and GAP+FAP had very similar perturbation characteristics. In conclusion, the robustness gain that is achieved with the FGSM is mostly due to the proposed FGSM feature representation and not because the

FGSM creates more significant perturbations or perceptually better perturbations.

Overall, the results show that the GAP+FAP approach effective for improving the clean classification performance and empirical robustness, especially when facing FGSM attacks, while it is less effective to resist strong iterative PGD attacks. These findings do not represent a thorough-going adversarial defense. We plan to explore the trade-offs of adversarial training, more powerful adversarial attacks, repeated-seed tests, and testing on other datasets and architectures.

REFERENCES

- Aarthy, B., & Keerthi, B. S. (2023). Enhancement of various images using coefficients obtained from a class of Sakaguchi type functions. *Scientific Reports*, 13(1), 18722. <https://doi.org/10.1038/s41598-023-45938-y>
- Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, 13(22), 4712. <https://doi.org/10.3390/rs13224712>
- Dargan, S., Kumar, M., Ayyagari, M. R., & Kumar, G. (2020). A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning. *Archives of Computational Methods in Engineering*, 27(4), 1071–1092. <https://doi.org/10.1007/s11831-019-09344-w>
- Eleftheriadis, C., Symeonidis, A., & Katsaros, P. (2024). Adversarial robustness improvement for deep neural networks. *Machine Vision and Applications*, 35(3), 35. <https://doi.org/10.1007/s00138-024-01519-1>
- Florindo, J. B. (2024). Fractal pooling: A new strategy for texture recognition using convolutional neural networks. *Expert Systems with Applications*, 243, 122978. <https://doi.org/10.1016/j.eswa.2023.122978>
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *ArXiv Preprint ArXiv*: 1412.6572. <https://doi.org/10.48550/arXiv.1412.6572>
- Gupta, A., Rajendhran, R., Stringham, N., Srikumar, V., & Marasović, A. (2024). Whispers of doubt amidst echoes of triumph in nlp robustness. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5533–5590. <https://doi.org/10.18653/v1/2024.naacl-long.310>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Hu, J., Ye, J., Feng, Z., Yang, J., Liu, S., Yu, X., Jia, L., & Song, M. (2024). Improving adversarial robustness via feature pattern consistency constraint. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 848–856. <https://doi.org/10.24963/ijcai.2024/94>
- Karjauv, A. (2026). Rethinking Global Average Pooling: Your Classifier Is Secretly a Multi-Instance Learner. *ArXiv Preprint ArXiv*: 2606.14555.
- Keleş, O., Yılmaz, M. A., Tekalp, A. M., Korkmaz, C., & Doğan, Z. (2021). On the computation of PSNR for a set of images or video. *2021 Picture Coding Symposium (PCS)*, 1–5. <https://doi.org/10.1109/PCS50896.2021.9477470>
- Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.
- Li, F., Li, K., Wu, H., Tian, J., & Zhou, J. (2025). Toward robust learning via core feature-aware adversarial training. *IEEE Transactions on Information Forensics and Security*, 20, 6236–6251. <https://doi.org/10.1109/TIFS.2025.3581021>
- Liao, J., Guo, L., Jiang, L., Yu, C., Liang, W., Li, K., & Pop, F. (2025). A machine learning-based feature extraction method for image classification using ResNet architecture. *Digital Signal Processing*, 160, 105036. <https://doi.org/10.1016/j.dsp.2025.105036>
- Liu, C., Zhang, J., Fang, H., Ma, Z., Zhang, W., & Yu, N. (2023). Dear: A deep-learning-based audio re-recording resilient watermarking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11), 13201–13209. <https://doi.org/10.1609/aaai.v37i11.26550>
- Liu, J., Li, Y., Guo, Y., Liu, Y., Tang, J., & Nie, Y. (2024). Generation and Countermeasures of adversarial examples on vision: a survey. *Artificial Intelligence Review*, 57(8), 199. <https://doi.org/10.1007/s10462-024-10841-z>
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *ArXiv Preprint ArXiv*: 1706.06083.
- Mehra, N., Thakar, U., & Tokekar, V. (2026). Adversarial Training Using FGSM Attack for Convolutional Neural Networks. *International Conference on Recent Advancement and Modernization in Sustainable Intelligent Technologies & Applications (RAMSITA-2026)*, 148–161.
- Pang, Z., Yan, X., Guo, S., & Lu, Y. (2025). Diversity-enhanced reconstruction as plug-in defenders against adversarial perturbations. *Frontiers in Artificial Intelligence*, 8, 1665106. <https://doi.org/10.3389/frai.2025.1665106>
- Riski, M. J., Noviandri, K. A., Hanggara, Y., Utama, N. P., & Purwarianti, A. (2025). Analysis of Defense Mechanisms Against FGSM Adversarial Attacks on ResNet Deep Learning Models Using the CIFAR-10 Dataset. *Jurnal Sistem Cerdas*, 8(2), 169–187. <https://doi.org/10.37396/jsc.v8i2.527>
- Shehu, A., Kirubakaran, P., Muhammad-Bello, B. L., & Ibrahim, N. (2026). Deep Learning for Image Steganalysis: A Structured Comparative Review of Dataset, Domain, and Robustness Coverage. *FUDMA Journal of Sciences*, 10(16), 551–556. <https://doi.org/10.33003/fjs-2026-1016-5739>
- Shin, S., Shin, D., & Kim, J. (2025). Implementation and comparison of CNN models on CIFAR-10 dataset. *2025 IEEE 20th Conference on Industrial Electronics and Applications (ICIEA)*, 1–4. <https://doi.org/10.1109/ICIEA65512.2025.11148839>
- Swain, S., Pattnaik, P. K., Jayasingh, S. K., Mohanty, S., Nayak, M., & Dash, B. B. (2026). Adversarial Image Quality Assessment against PGD Attack on ResNet-34. *2026 7th International Conference on Computational Intelligence and*

- Networks (CINE), <https://doi.org/10.1109/CINE68769.2026.11502668>
- Tanay, T., & Griffin, L. (2016). A boundary tilting perspective on the phenomenon of adversarial examples. ArXiv Preprint ArXiv: 1608.07690.
- Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, 17(1), 168–192. <https://doi.org/10.1016/j.aci.2018.08.003>
- Villegas-Ch, W., Jaramillo-Alcázar, A., & Luján-Mora, S. (2024). Evaluating the robustness of deep learning models against adversarial attacks: An analysis with fgsm, pgd and cw. *Big Data and Cognitive Computing*, 8(1), 8. <https://doi.org/10.3390/bdcc8010008>
- Wang, J., Wang, Z., Dong, Q., Zhang, S., Liu, M., & Zhang, J. (2026). A comprehensive survey of pattern recognition classification algorithms. *Third International Conference on Power Electronics and Artificial Intelligence (PEAI 2026)*, 14136, 172–188.
- Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <https://doi.org/10.1109/TIP.2003.819861>
- Wong, E., Rice, L., & Kolter, J. Z. (2020). Fast is better than free: Revisiting adversarial training. ArXiv Preprint ArXiv: 2001.03994. <https://doi.org/10.48550/arXiv.2001.03994>
- Xu, Y., Li, F., Chen, Z., Liang, J., & Quan, Y. (2021). Encoding spatial distribution of convolutional features for texture representation. *Advances in Neural Information Processing Systems*, 34, 22732–22744.
- Zhou, S., Liu, C., Ye, D., Zhu, T., Zhou, W., & Yu, P. S. (2022). Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. *ACM Computing Surveys*, 55(8), 1–39. <https://doi.org/10.1145/3547330>

