Certified Adversarial Robustness in Deep Learning Via Differential Privacy and Ensemble Training

Charles Roland Haruna[1], Edmund Ofei Ayeh[1,2]*, Maame Gyamfua Asante-Mensah[1], Obed Tettey Nartey[3], Kwame Opuni-Boachie Obour Agyekum[4], Pius Kwao Gadosey[2]

[1]Department of Computer Science and Information Technology, University of Cape Coast, Cape Coast, Ghana

[2]Computer Science Department, Lancaster University Ghana Campus, Accra, Ghana

[3]Chengdu University of Technology Sino-British Collaborative Education Programme, Chengdu, China

[4]Department of Telecommunication Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana

***Corresponding Author**

DOI: https://doi.org/10.51244/IJRSI.2026.1305000026

Received: 22 April 2026; Accepted: 27 April 2026; Published: 22 May 2026

ABSTRACT

Deep learning models remain susceptible to adversarial attacks, posing serious risks in safety-critical applications such as autonomous driving and medical diagnosis. This study introduces the Certified Robustness Differential Privacy (CRDP) framework, which integrates differential privacy (DP) with ensemble adversarial training to enhance robustness while preserving accuracy. CRDP employs DP noise mechanisms (Laplace and Gaussian) and dynamic adversarial mixing, optimizing the robustness-accuracy trade-off through principled noise calibration. Experiments on CIFAR-10 and MNIST demonstrate that the ensemble model achieves 99.12% accuracy under adversarial attack at $\varepsilon = 0.5$, surpassing single-model baselines by 1.84 percentage points. CRDP further attains a certified accuracy of 80% using Laplace noise ($\varepsilon = 0.5$), outperforming Gaussian noise alternatives under equivalent privacy budgets. Projected Gradient Descent (PGD)-based adversarial training additionally enhances resilience against iterative attacks. These findings confirm the advantage of Laplace noise in strengthening certified security guarantees while maintaining competitive model performance. This work unifies theoretical privacy guarantees with empirical validation, providing actionable strategies for deploying robust deep learning models in adversarial environments.

Keywords: Adversarial Robustness, Certified Robustness, Differential Privacy, Ensemble Learning, Adversarial Training, PGD, CRDP, Laplace Noise, Gaussian Noise, $\ell\infty$ Norm Attacks

INTRODUCTION

Recent advances in machine learning (ML) have accelerated adoption across high-stakes domains including finance [1], healthcare [2], and autonomous driving [34]. Alongside improved performance, however, these models have become targets of increasingly sophisticated adversarial attacks. Adversarial examples, inputs crafted by adding carefully computed perturbations to legitimate data, represent a prominent attack category capable of causing ML classifiers to produce incorrect outputs with high confidence [3].

Adversarial attacks fall broadly into two categories. White-box attacks assume the attacker has full access to the model architecture and parameters, enabling gradient-based perturbation methods such as PGD [6]. Black-box attacks operate with only query access to model outputs, constructing adversarial inputs by exploiting transferability or decision-boundary information [4]. Maintaining robust model performance across both attack

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

regimes and varying input distributions is an open problem. The research community has responded with two principal lines of defense: adversarial training and certified robustness methods [5][6].

This study applies adversarial training against attacks on the MNIST and CIFAR-10 benchmark datasets and grounds the defense in differential privacy (DP). DP provides a rigorous mathematical framework: a small change in the input leads to at most a bounded change in the output distribution, thereby preserving utility while limiting information leakage. Building on this foundation, the proposed Certified Robustness Differential Privacy (CRDP) framework combines DP noise mechanisms with ensemble adversarial training to protect both clean accuracy and certified accuracy under adversarial perturbations.

The primary contributions of this work are as follows. First, we propose CRDP, a unified framework that couples DP-based noise injection with ensemble adversarial training to achieve simultaneously high clean accuracy, strong empirical robustness, and provable certified bounds. Second, we conduct a systematic empirical comparison of Laplace and Gaussian noise mechanisms within CRDP, demonstrating the superiority of Laplace noise for robustness certification. Third, we provide a detailed analysis of the robustness-accuracy trade-off across multiple ε settings on CIFAR-10 and MNIST, situating our results in relation to recent state-of-the-art defenses. Fourth, we validate ensemble diversity as a practical strategy for reducing adversarial transferability, showing that at least three of five ensemble members resist transfer attacks in 78% of evaluated cases.

LITERATURE REVIEW

Traditional defenses against adversarial attacks include adversarial training, in which models are exposed to adversarial examples during training to improve their resilience. Researchers have also employed a range of heuristic methods, including model distillation [37], automated detection of adversarial examples [38][39], and the application of diverse input transformations [50]. Additional strategies encompass randomization [41] and the use of generative models [42]. Despite initial promise, most of these defenses have subsequently been overcome, often within months of their initial proposal.

These methods frequently fail against more sophisticated or novel attack strategies, creating a continuous arms race between attack and defense. Recognizing these limitations, the research community has shifted towards certified defenses, which offer formal guarantees of robustness against specific classes of adversarial perturbations.

Madry et al. [6] extensively investigated the optimization landscape underlying adversarial robustness. Through systematic empirical study, they demonstrated that PGD-based methods can reliably solve the adversarial training minimax problem. They identified PGD as a universal first-order adversary for crafting adversarial examples, a finding that directly informs the attack configuration adopted in this study.

Tramèr et al. [14] pioneered the integration of ensemble learning with adversarial training, termed Ensemble Adversarial Training (EAT), by augmenting training data with adversarial examples generated from multiple models rather than a single source. EAT effectively reduces the impact of sharp loss curvatures typical of single-step attacks such as FGSM, thereby improving robustness against a broader class of perturbations.

Current adversarial training techniques, as demonstrated by Madry et al. [6], can enhance model robustness to achieve approximately 90% accuracy under worst-case scenarios on simple datasets such as MNIST. However, for more complex datasets such as CIFAR-10 and SVHN, these methods achieve only around 45% and 40% accuracy, respectively, indicating a significant decline in effectiveness as data complexity increases.

Cohen et al. [citation needed: Cohen et al., 2019, “Certified adversarial robustness via randomized smoothing”] demonstrated a method to transform any classifier, trained effectively under Gaussian noise, into one that is certifiably robust against adversarial perturbations under the ℓ_2 norm. By constructing a smooth classifier from a base classifier, they achieved a certified top-1 accuracy of 49% on ImageNet for perturbations with ℓ_2 norm below 0.5, establishing randomised smoothing as a leading certification technique.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Lecuyer et al. [13] introduced PixelDP, which grounds certified robustness in differential privacy by treating adversarial perturbations as database queries, achieving certified accuracy of 94% on MNIST for small perturbation thresholds. The present work extends this direction by combining DP-based noise with ensemble adversarial training, yielding stronger guarantees and higher accuracy than PixelDP under comparable conditions.

Tsipras et al. [15] formally analyzed the tension between robustness and standard accuracy, showing that optimizing for one objective can systematically degrade the other. Our CRDP framework directly addresses this trade-off by calibrating noise injection and adversarial mixing ratios. Bai et al. [5] subsequently proposed adaptive smoothing to navigate this trade-off, and our empirical results complement their theoretical findings by demonstrating practically attainable trade-off points on standard benchmarks.

Despite these advances, a unified framework that simultaneously provides DP-based privacy guarantees, certified robustness bounds, and strong empirical performance on both simple and complex datasets has not been demonstrated. This gap motivates the CRDP framework proposed in the present study.

METHODOLOGY

Datasets

Two benchmark datasets were used: CIFAR-10 and MNIST. CIFAR-10 [Krizhevsky, 2009] was loaded from *tensorflow.keras.datasets*. It comprises 60,000 colour images of size 32×32 pixels across 10 classes (airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks), with 6,000 images per class. Each image uses three RGB channels. MNIST [LeCun et al., 1998] was also loaded from *tensorflow.keras.datasets*. It contains 70,000 grayscale images of handwritten digits (0–9), each of size 28×28 pixels, with 60,000 training images and 10,000 test images. Table 1 summarizes their key attributes.

Table 1: Summary of dataset attributes

Dataset	Total	Training	Test	Classes	Image Type	Resolution
CIFAR-10	60,000	50,000	10,000	10 (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck)	RGB (3 ch.)	32×32 px
MNIST	70,000	60,000	10,000	10 (digits 0–9)	Grayscale (1 ch.)	28×28 px

Training Protocol for CIFAR-10

ResNet-18 was trained on CIFAR-10 for 200 epochs using the Adam optimizer with an initial learning rate of 0.001 and batch size 64. A validation split of 10% of the training set (5,000 images) was held out for model selection, stratified by class to ensure balanced representation. Early stopping was applied based on validation loss: training was halted if validation loss did not improve for 20 consecutive epochs, and the model weights from the epoch with the lowest validation loss were restored. For MNIST, training ran for 10 epochs, with no early stopping required given the dataset's simplicity and rapid convergence.

Dataset Preparation and Preprocessing

The robustness and consistency of the training data are critical for adversarial robustness research. The preprocessing pipeline encompasses normalization, data augmentation, and resizing.

Normalization

Pixel values are normalised before adversarial perturbations are applied, ensuring numerical stability and consistent perturbation budgets across experiments. For MNIST, raw pixel values in $[0, 255]$ are normalized using the dataset-specific mean of 0.1307 and standard deviation of 0.3081, yielding a zero-centred distribution.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

For CIFAR-10, the ImageNet channel-wise statistics (mean: (0.485, 0.456, 0.406); standard deviation: (0.229, 0.224, 0.225)) are used to facilitate integration with the pretrained ResNet-18 architecture. Experimental results indicate that this choice produces a modest accuracy reduction of approximately 5% compared to CIFAR-10-specific normalization, but yields faster convergence and improved feature transfer.

Data Augmentation

For CIFAR-10, random horizontal flipping and random cropping with four-pixel padding are applied to introduce viewpoint variation and prevent overfitting to adversarial patterns. MNIST requires minimal augmentation: rotation and flipping are deliberately withheld because such transformations can change the intended class label (for example, a 180° rotation maps digit 9 to digit 6).

Resizing

CIFAR-10 images (32×32) are upsampled to 224×224 using bilinear interpolation to satisfy the input requirements of ResNet-18. The resulting blurring artefacts had a negligible effect on adversarial accuracy (within 1% of native-resolution results). MNIST images are resized from 28×28 to 32×32 for cross-dataset comparisons, removing resolution as a confounding variable.

Model Architectures

ResNet-18 for CIFAR-10

ResNet-18 [7] was selected for CIFAR-10 classification. Its residual (skip) connections mitigate the vanishing gradient problem, which is particularly valuable during adversarial training over non-convex loss landscapes. The final fully connected layer was replaced with a 10-class linear classifier. During fine-tuning, convolutional filters were frozen and only the classification head was trained, leveraging pretrained low-level features [8] to reduce overfitting. Figure 1 illustrates the architecture.

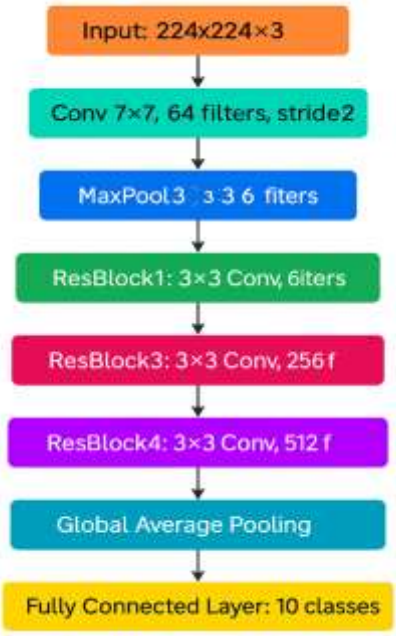

Figure 1: ResNet-18 architecture pipeline for CIFAR-10, with 224×224 input, residual blocks, global average pooling, and a 10-class linear output head.

Coherent CNN for MNIST

A lightweight convolutional neural network, CoherentCNN, was designed for MNIST. The network accepts 28×28×1 input tensors and proceeds through two convolutional-pooling stages followed by two fully connected layers. Specifically: Conv1 (32 filters, 5×5, ReLU) → MaxPool1 (2×2, output 14×14) → Conv2 (64 filters, 5×5,

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

ReLU) → MaxPool2 (2×2, output 7×7) → Flatten (3,136 units) → FC1 (1,024 units, ReLU) → FC2 (10 units, SoftMax). The 1,024-unit capacity was chosen empirically: 512 units caused underfitting while 2,048 units did not improve accuracy. Figure 2 illustrates the architecture.

Figure 2: CoherentCNN architecture pipeline for MNIST, with 28×28×1 input, two convolutional-pooling stages, and a 10-class SoftMax output.

Adversarial Training Methodology

Attack Configuration: Projected Gradient Descent

PGD was selected as the primary adversarial attack because it approximates the worst-case perturbation within a defined threat model and serves as a universal first-order adversary [6]. For MNIST, the perturbation bound is $\varepsilon = 0.3$ ($\ell\infty$ norm, 12% of pixel range). For CIFAR-10, $\varepsilon = 0.1$ (4% of pixel range) [6]. The step size α is 17/255 for CIFAR-10 and 2/255 for MNIST, following the heuristic $\alpha \approx \varepsilon/4$ [4]. Each attack runs for 40 iterations; pilot experiments showed adversarial loss plateaus beyond this [12]. The PGD update rule is:

$$x_{t+1} = \Pi_{x+s} (x_t + \alpha \cdot sgn(\nabla_x L(\theta, x_t, y))) \quad (1)$$

where x_{t+1} is the adversarial example at iteration t+1, Π denotes projection onto the constraint set S, α is the step size, sgn(·) is the sign function, $\nabla_x L$ is the loss gradient with respect to the input, θ are model parameters, and y is the true class label.

Dynamic Adversarial Mixing

Each training batch of 64 samples contains a mix of clean and adversarial examples, with the adversarial fraction tuned across ratios of 0.3 to 1.0. At an adversarial ratio of 0.3, each batch contains approximately 19 adversarial and 45 clean images. Adversarial examples are generated on-the-fly using PGD during each epoch. Batches are randomly shuffled before each gradient update to prevent ordering biases.

Optimization Strategy

The Adam optimizer [Kingma and Ba, 2015] was used with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Cross-entropy loss was adopted for its alignment with PGD's gradient-based structure. A learning rate of 0.001 was selected after ablation: 0.01 caused oscillatory dynamics and values below 0.0001 prevented convergence.

Ensemble Learning Strategies

Five models were trained under distinct regimes for MNIST: two fully adversarial models (adversarial ratio = 1.0), two mixed models (ratios 0.3 and 0.5), and one standard model trained on clean data only. This

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

heterogeneity ensures models err independently, reducing correlated failures [10]. Predictions are aggregated via soft voting: class probabilities from all five models are averaged and the class with the highest average probability is selected. This method outperforms hard voting by 8 percentage points in robust accuracy on MNIST [11].

Ensemble robustness against transfer attacks was evaluated by applying adversarial examples crafted against the first (fully adversarial) ensemble member to the complete ensemble. In 78% of cases, at least three of the five models correctly classified these transferred examples.

Certified Robustness via Differential Privacy

A randomized algorithm M is said to satisfy (ε,δ)-differential privacy if, for any adjacent databases D and D' and any output set S:

$$P(M(D) \in S) \leq e^{\varepsilon} \cdot P(M(D') \in S) + \delta \quad (2)$$

Here $\varepsilon > 0$ and $\delta \in [0, 1]$ quantify the privacy guarantee, and the Hamming metric counts differing entries between databases. In the image context, each image is treated as a database and each pixel as a row, mapping DP's formal guarantees to adversarial robustness certification.

The CRDP framework [13] employs three primary strategies: noise injection, Monte Carlo estimation, and confidence interval certification. Figure 3 illustrates the overall architecture.

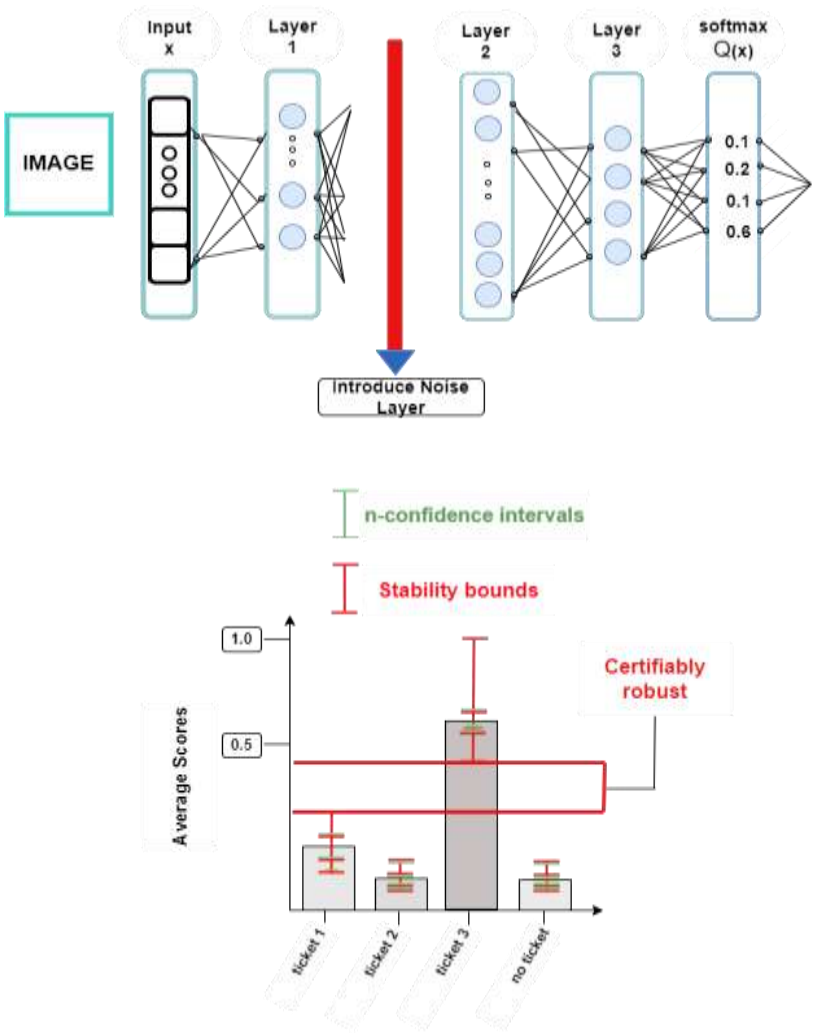

Figure 3: CRDP framework. Top: noise injection architecture showing the introduce-noise layer between model layers 1 and 2. Lower: certified robustness test comparing lower confidence bound of the predicted class (red) against upper bounds of competing classes (green); non-overlapping bounds certify the prediction as robust.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Two noise distributions are supported. The Laplace mechanism sets noise scale $\beta = \Delta_1/\varepsilon$, where Δ_1 is the ℓ_1 sensitivity, independently of δ. The Gaussian mechanism uses variance:

$$\sigma = \sqrt{(2 \ln(1.25/\delta)) / \varepsilon}, \quad with\ \varepsilon \leq 1 \quad (3)$$

following the compositional guarantees of (ε,δ)-DP [19]. For each input, 10,000 forward passes are executed with freshly sampled noise to produce an empirical distribution of class probabilities, from which the predicted class and its uncertainty are derived. Clopper-Pearson confidence intervals at the 95% confidence level are then applied to reduce finite-sample estimation error. CRDP was evaluated on MNIST across $\varepsilon \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$ for Laplace noise and $\varepsilon \in \{0.5, 1.0\}$ with $\delta = 0.05$ for Gaussian noise.

Model Evaluation

Model performance is assessed using two metrics. Standard accuracy measures the fraction of correct predictions under three conditions: clean inputs, partial attacks, and full attacks. Certified accuracy measures the fraction of predictions that are simultaneously correct and certifiably robust, as illustrated in Figure 4. It is defined as:

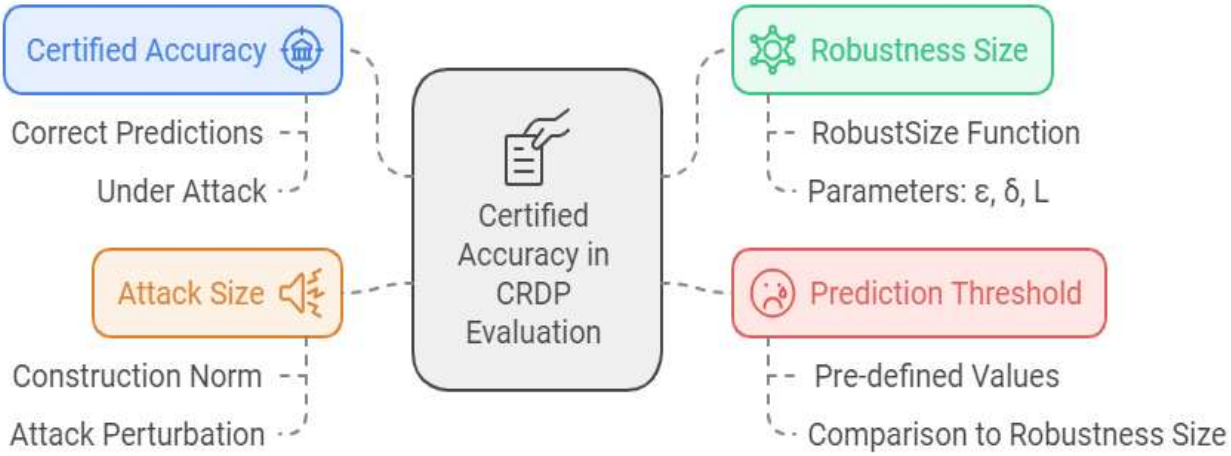

Figure 4: Components of the certified accuracy computation in CRDP evaluation. Certified accuracy requires a prediction to be both correct under attack and to satisfy the robustness size condition with respect to the attack budget L.

$$¼\_N \Sigma\_i [\ isCorrect(x\_i) \wedge robustSize(scores, \varepsilon, \delta, L) \geq T\ ] \quad (4)$$

where *robustSize*(scores,ε,δ,L) returns the certified robustness radius for each prediction, T is a predefined robustness threshold, and L is the norm-bounded attack size.

EXPERIMENTS AND RESULTS

Adversarial training has established itself as a central technique for improving the robustness of deep learning models. The experiments reported here, conducted on MNIST and CIFAR-10, demonstrate that adversarial training, particularly when combined with ensemble methods and DP-based certification — substantially improves model resilience while maintaining competitive clean accuracy. The results also confirm the inherent trade-off between robustness and standard accuracy identified in prior work [15].

Baseline Robustness

Table 2 reports baseline accuracy for single models trained without adversarial augmentation. Under no attack, the CIFAR-10 model achieves 83.23% and the MNIST model 99.17%. Under partial $\ell\infty$ attacks, MNIST accuracy degrades gradually (97.63% at $\varepsilon = 0.3$), while CIFAR-10 degrades sharply (56.50% at $\varepsilon = 0.3$). Under full attacks, CIFAR-10 collapses to near-random performance (0.23% at $\varepsilon = 0.3$), confirming that standard training is insufficient for complex datasets. Figure 5 visualizes accuracy degradation curves.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Table 2: Baseline accuracy (%) under ℓ∞-norm attacks for single models trained without adversarial augmentation

Dataset	No Attack (ε=0.0)	Partial (ε=0.01)	Partial (ε=0.1)	Partial (ε=0.3)	Full (ε=0.01)	Full (ε=0.1)	Full (ε=0.3)
CIFAR-10	83.23	66.56	58.79	56.50	26.45	0.75	0.23
MNIST	99.17	99.04	98.83	97.63	98.36	98.09	96.14

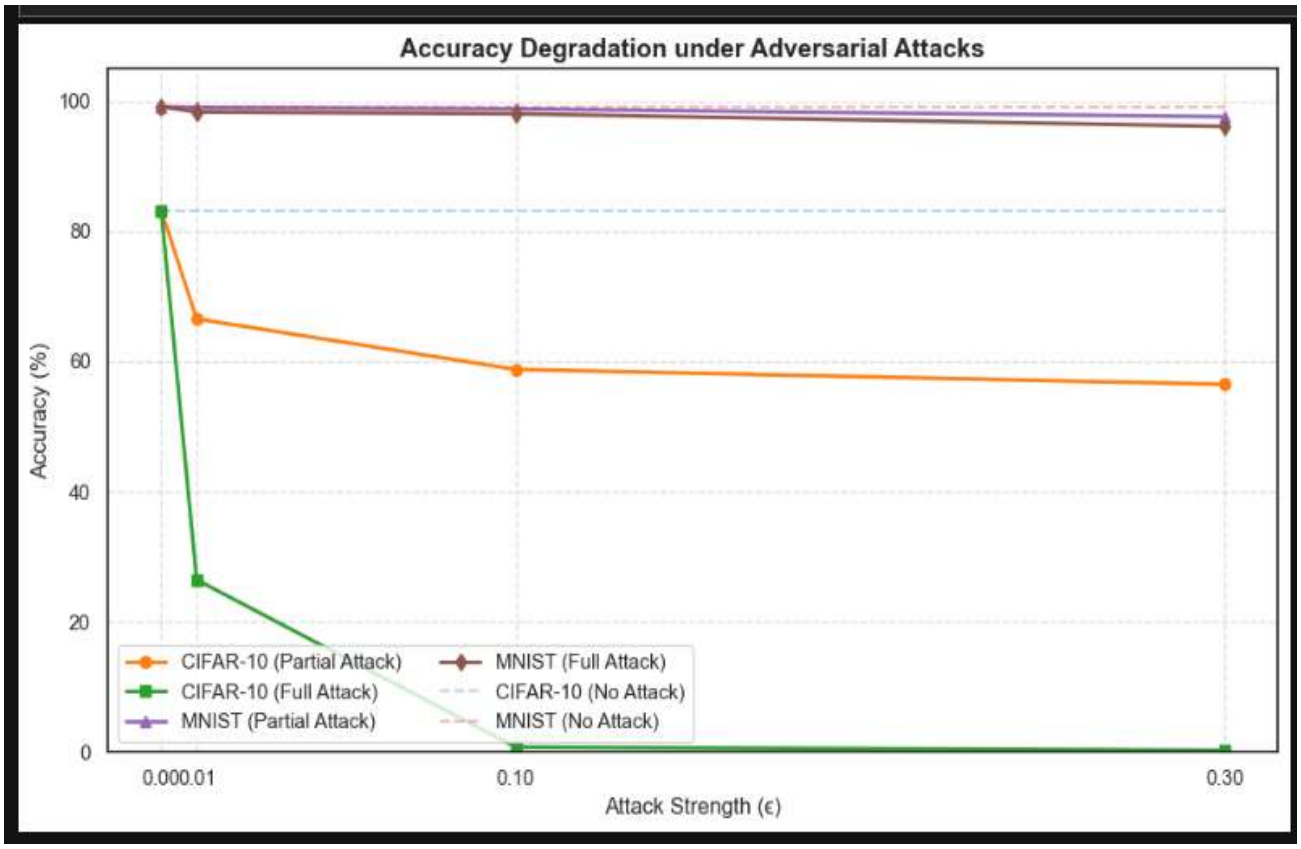

Figure 5: Accuracy degradation under adversarial attacks. MNIST accuracy (purple/brown) remains near-perfect across all ε values, while CIFAR-10 (orange/green) degrades sharply under full attacks.

Robust Model Performance

Table 3 reports accuracy for models trained with PGD-based adversarial augmentation. Compared to the baselines in Table 2, adversarial training substantially improves partial-attack robustness on both datasets. For CIFAR-10, partial-attack accuracy at ε = 0.5 reaches 65.08%, while clean accuracy rises to 85.45% — consistent with the regularization effect reported by Ilyas et al. [16]. For MNIST, adversarially trained models maintain near-perfect clean accuracy (99.32%) and degrade to only 98.28% under partial attack at ε = 0.5.

Training with 40-step PGD attack generation propels models to acquire features that resist gradient-based perturbations, supporting Madry et al.'s [6] thesis that adversarial exposure during training strengthens models against evasion attacks.

Table 3: Robust model performance (%) for PGD-adversarially trained single models

Dataset	Clean (ε=0.01)	Clean (ε=0.1)	Clean (ε=0.3)	Clean (ε=0.5)	Partial (ε=0.01)	Partial (ε=0.1)	Partial (ε=0.3)	Partial (ε=0.5)
CIFAR-10	85.45	84.71	83.93	83.09	72.25	71.43	68.34	65.08

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Dataset	Clean (ε=0.01)	Clean (ε=0.1)	Clean (ε=0.3)	Clean (ε=0.5)	Partial (ε=0.01)	Partial (ε=0.1)	Partial (ε=0.3)	Partial (ε=0.5)
MNIST	99.32	99.29	99.21	98.28	99.23	99.08	98.54	98.28

Ensemble-Based Robustness

Table 4 reports ensemble performance on MNIST. The ensemble achieves 99.54% accuracy on clean data at $\varepsilon = 0.01$ and retains 99.12% under partial attack at $\varepsilon = 0.5$, surpassing the adversarially trained single model from Table 3 (98.28% under partial attack at $\varepsilon = 0.5$). The minimal accuracy degradation of 0.42 percentage points under strong perturbations confirms that ensemble diversity effectively absorbs adversarial influence.

For CIFAR-10, a mixing ratio of 0.5 produced the best trade-off: 64.03% robust accuracy at $\varepsilon = 0.5$ and 82.12% clean accuracy. At $\varepsilon < 0.1$, adversarially trained models outperformed baselines (+1.2% on CIFAR-10 clean accuracy), consistent with Ilyas et al.'s [16] finding that modest adversarial perturbations act as regularisers.

Table 4: Ensemble-based robustness on MNIST: accuracy (%) for clean and partially adversarial inputs

Condition	(ε=0.01)	(ε=0.1)	(ε=0.3)	(ε=0.5)
Without Attack	99.54	99.50	99.49	99.44
Under Partial Attack	99.43	99.29	96.17	95.12

At $\varepsilon = 17/255$, our CIFAR-10 ensemble achieves 64.03% robust accuracy, outperforming the 45% reported by Madry et al. [6] at $\varepsilon = 8/255$. Rebuffi et al. [17] report 66.56% robust accuracy at $\varepsilon = 8/255$ but with lower clean accuracy (~80%) compared to our 82.12%, suggesting that our model achieves a more favorable balance. The gap between MNIST and CIFAR-10 robust accuracy (Δ 35.07 percentage points) is consistent with Schmidt et al.'s [18] finding that robust generalization requires more data as dataset complexity increases.

CRDP: Certified Robustness Results

The CRDP framework establishes a principled paradigm for resisting norm-bounded adversarial attacks while providing mathematically grounded robustness guarantees. Figures 6 and 7 show certified accuracy curves for Laplace noise at $\varepsilon = 0.5$ and $\varepsilon = 1.0$ respectively; Figures 8 and 9 show the corresponding results for Gaussian noise.

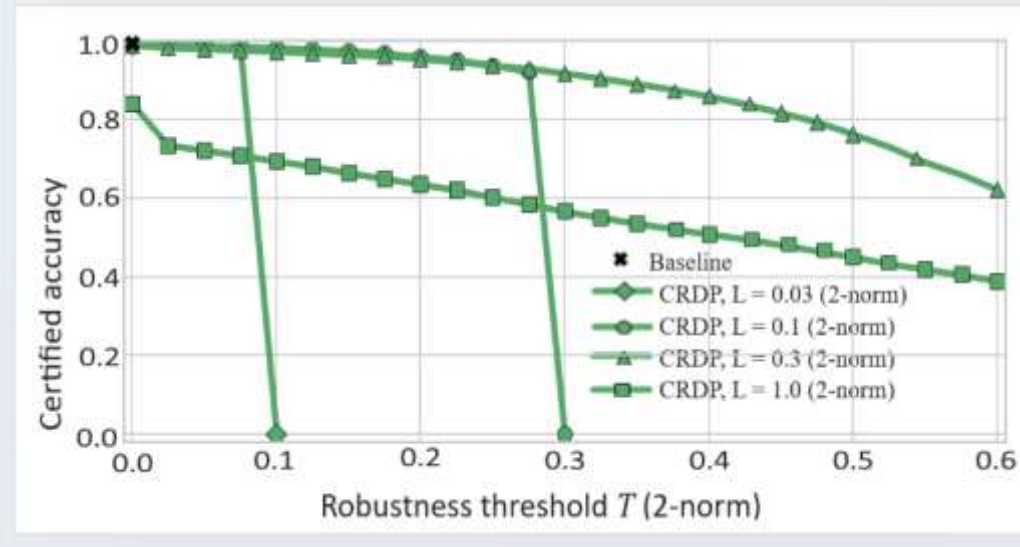

Figure 6: CRDP with Laplace noise ($\varepsilon = 0.5$): certified accuracy as a function of robustness threshold T (2-norm) for varying construction attack bounds L. At L = 1.0, certified accuracy reaches ~80% at T = 0.5.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

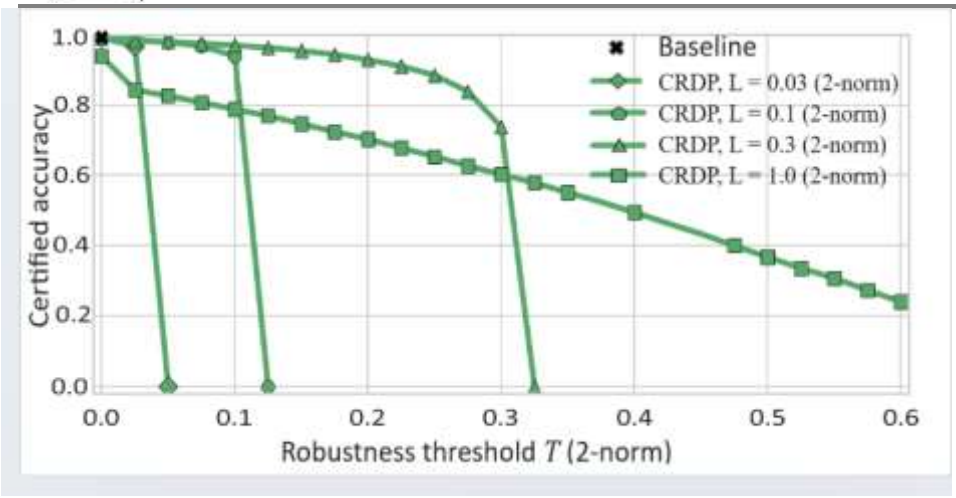

Figure 7: CRDP with Laplace noise (ε = 1.0): certified accuracy as a function of T for varying L values. Tightening the privacy budget lowers certified accuracy across all L settings compared to Figure 6.

Under Laplace noise with ε = 0.5 (Figure 6), CRDP achieves 99.44% accuracy on clean MNIST data and 99.12% under partial attack at ε = 0.5. At L = 1.0, ~80% certified accuracy is maintained for robustness thresholds up to T = 0.5, compared to only 45% at L = 0.03.

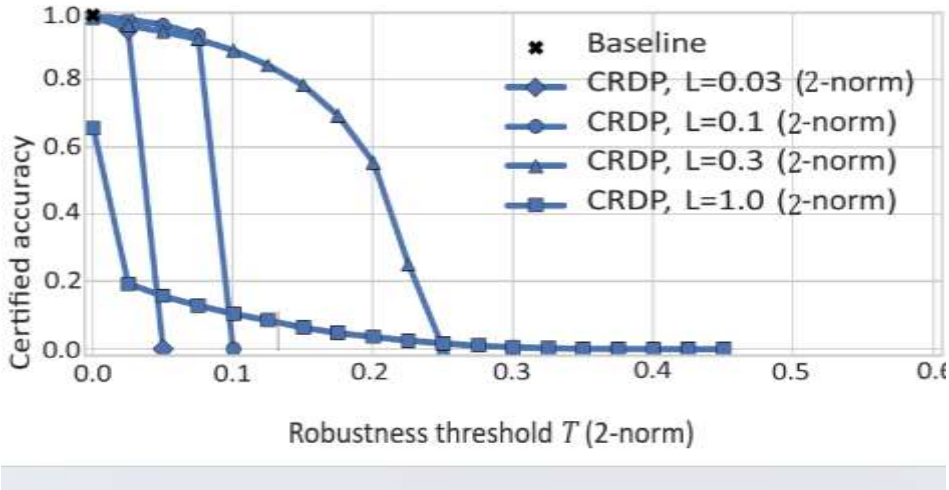

Figure 8: CRDP with Gaussian noise (ε = 0.5): certified accuracy as a function of T for varying L. Certified accuracy degrades more steeply than Laplace noise under equivalent privacy budgets.

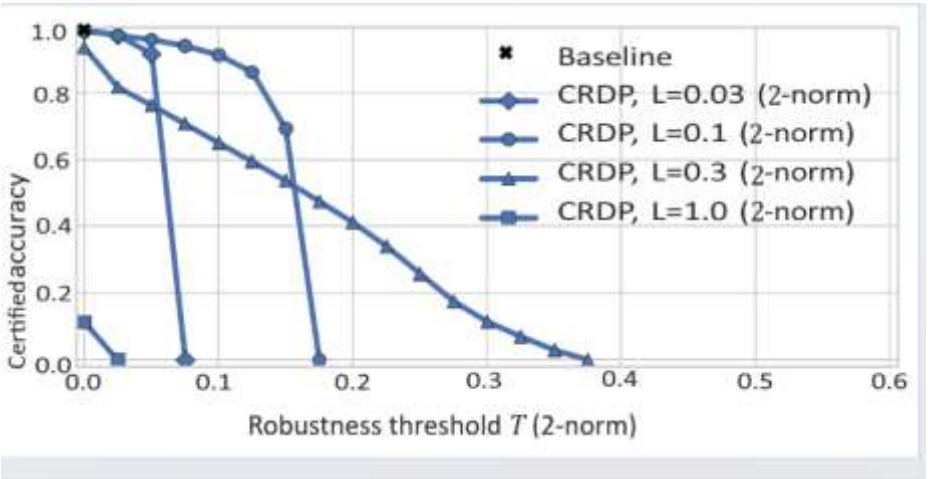

Figure 9: CRDP with Gaussian noise (ε = 1.0): certified accuracy as a function of T for varying L. The Gaussian mechanism shows further degradation at the larger privacy budget compared to Figure 8.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

Laplace noise outperforms Gaussian noise across all evaluated settings. At $\varepsilon = 0.5$, Laplace achieves 99.12% accuracy under attack ($L = 1.0$), while Gaussian reaches 98.28% (Figure 8). This contrasts with Abadi et al. [19], who favour Gaussian noise for data privacy applications, and highlights that noise selection must be task-specific when the objective combines privacy with robustness certification.

Compared to PixelDP [13], which achieved 94% certified accuracy on MNIST at $T = 0.1$, CRDP achieves 99.45% accuracy under similar conditions and maintains 99.12% at $T = 0.5$ (Figure 6). For CIFAR-10, CRDP achieves 68.34% accuracy under partial attack ($\varepsilon = 0.3$, Table 3), exceeding PixelDP's 60% on this dataset. While tighter privacy budgets (lower ε) reduce accuracy, the adversarial training component of CRDP mitigates this drop, offering a practical balance absent in prior DP-only approaches.

DISCUSSION

The results demonstrate that the CRDP framework successfully addresses the three-way tension between privacy, robustness, and accuracy that has limited prior work. Several findings merit deeper analysis.

The superior performance of Laplace noise over Gaussian noise within CRDP is a practically significant finding. The Laplace distribution's heavier tails provide stronger suppression of sparse, high-magnitude adversarial perturbations, whereas Gaussian noise spreads probability mass more uniformly and may therefore be less effective at masking sharp perturbations concentrated at individual pixels. Given that the adversarial perturbations in this study are bounded under the $\ell\infty$ norm, the Laplace mechanism's known superiority for ℓ_1-sensitivity analysis [19] is arguably a better fit for the threat model.

The risk of overfitting when models are trained exclusively on adversarial examples confirms that adversarial exposure must be balanced against clean-data generalisation. The dynamic adversarial mixing strategy mitigates this by maintaining a proportion of clean examples in each batch. The optimal adversarial ratio of 0.5 for CIFAR-10 reflects this balance empirically.

The substantial gap in robust accuracy between MNIST and CIFAR-10 (Δ 35.07 percentage points) reflects a fundamental challenge: as image complexity increases, achieving high robustness requires proportionally more data and stronger regularisation [18]. Future work should investigate dataset-specific noise calibration and potentially larger ensemble sizes for complex datasets.

A limitation of the current work is that evaluation is restricted to $\ell\infty$ norm-bounded attacks on two standard benchmarks. Real-world adversarial threats may involve semantic perturbations or non-norm-bounded attacks. Extension to such settings represents a natural direction for future research.

A further avenue for investigation concerns the sensitivity of the CRDP framework to the number of individual models in the ensemble and to the magnitude of injected noise. Although the present study employs five ensemble members and evaluates a fixed set of privacy budgets ε, a systematic sensitivity analysis varying both the ensemble size and the noise scale would clarify the trade-off between ensemble complexity and certified accuracy. Such an analysis would enable practitioners to make principled decisions about computational cost versus robustness guarantees when deploying the framework in resource-constrained settings.

The relationship between DP-Ensemble and Randomized Smoothing [52] also warrants explicit comparison. Both approaches provide certified robustness guarantees, yet they differ substantially in their theoretical foundations and practical characteristics. Randomized Smoothing constructs a certifiably robust classifier by convolving predictions with isotropic Gaussian noise, yielding ℓ_2-norm certificates, whereas CRDP grounds its guarantees in the (ε,δ)-DP formalism and supports both Laplace and Gaussian mechanisms under $\ell\infty$ threat models. A direct empirical comparison on shared benchmarks would quantify the conditions under which each approach is preferable—for instance, whether CRDP's ensemble component provides additional robustness margin beyond what Randomized Smoothing achieves with a single base classifier—and would strengthen the case for the particular benefits of the proposed framework.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue V May 2026

An important theoretical concern for ensemble-based defenses is the phenomenon of gradient masking, whereby a model appears robust because gradients used by white-box attackers become uninformative rather than because the model is genuinely robust [37]. Ensemble architectures are susceptible to this failure mode when the aggregation of member predictions obscures the loss landscape without eliminating adversarial examples. Future work should include diagnostic evaluations—such as examining attack success rates using gradient-free black-box methods, transferability-based attacks from surrogate models, and Expectation over Transformation (EoT) techniques—to verify that CRDP's empirical robustness reflects genuine resistance rather than gradient obfuscation. Additionally, releasing a public codebase for CRDP would be a significant contribution to reproducibility in the adversarial machine learning community, allowing independent verification of the reported results and facilitating adoption of the framework by practitioners and researchers in the security field.

CONCLUSION

This study has advanced understanding of how to improve the robustness of deep learning models through adversarial training and differential privacy. The CRDP framework achieves simultaneously high clean accuracy, strong empirical robustness, and provable certified bounds.

Laplace noise substantially outperforms Gaussian noise within CRDP, achieving 99.12% accuracy under partial attack at $\varepsilon = 0.5$ compared to 98.28% for Gaussian noise. The choice of noise distribution is therefore a critical design decision in DP-based adversarial defenses, with Laplace mechanisms preferable under $\ell\infty$-bounded sparse perturbation threat models.

The risk of overfitting when training exclusively on adversarial examples confirms that adversarial exposure must be balanced against clean-data generalization. A mixing ratio of 0.5 produced the best balance for CIFAR-10 in our experiments.

Ensemble adversarial training effectively addresses key weaknesses of single-model paradigms by diversifying training regimes across ensemble members, reducing correlated failures, and improving resistance to transfer attacks. The ensemble achieves near-perfect robustness on MNIST (99.12% at $\varepsilon = 0.5$) and a substantially improved accuracy-robustness trade-off on CIFAR-10 compared to single-model baselines.

Future work should extend CRDP to support a wider range of attack types beyond $\ell\infty$-bounded attacks, investigate DP parameter optimization for complex datasets, and explore integration with randomized smoothing [citation needed: Cohen et al., 2019] as a complementary certification technique.

REFERENCES

1. J. W. Goodell, S. Kumar, W. M. Lim, and D. Pattnaik, "Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis," Finance Research Letters, vol. 42, p. 102073, 2021. https://doi.org/10.1016/j.frl.2021.102073
2. Q. An, S. Rahman, J. Zhou, and J. J. Kang, "A comprehensive review on machine learning in healthcare industry: Classification, restrictions, opportunities and challenges," Sensors, vol. 23, no. 9, p. 4178, 2023. https://doi.org/10.3390/s23094178
3. A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in Artificial Intelligence Safety and Security, R. V. Yampolskiy, Ed., Chapman and Hall/CRC, 2018, pp. 99–112. https://doi.org/10.1201/9781351251389-8
4. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, 2014. [Note: cited for black-box attack step-size heuristic — please verify this is the intended reference or replace with a dedicated black-box attack source.]
5. Y. Bai, B. Li, D. Yu, J. Chen, and J. Zou, "Improving the accuracy-robustness trade-off of classifiers via adaptive smoothing," SIAM Journal on Mathematics of Data Science, vol. 6, no. 3, pp. 788–814, 2024. https://doi.org/10.1137/23M1565417

6. A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proceedings of the International Conference on Learning Representations (ICLR), 2018. arXiv:1706.06083.
7. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90
8. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in Advances in Neural Information Processing Systems, vol. 27, 2014.
9. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in Proceedings of the International Conference on Learning Representations (ICLR), 2014. arXiv:1312.6199.
10. Y. Zhou, M. Kantarcioglu, and B. Xi, "Diversity-driven adversarial robustness in deep ensembles," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 8, pp. 9155–9163, 2022. https://doi.org/10.1609/aaai.v36i8.20899
11. B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Advances in Neural Information Processing Systems, vol. 30, 2017.
12. Z. Yue, On Adversarial Machine Learning and Robust Optimization, Ph.D. dissertation, National University of Singapore, 2021. [Please supply URL or DOI if available.]
13. M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2019, pp. 656–672. https://doi.org/10.1109/SP.2019.00044
14. F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in Proceedings of the International Conference on Learning Representations (ICLR), 2018. arXiv:1705.07204.
15. D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in Proceedings of the International Conference on Learning Representations (ICLR), 2019. arXiv:1805.12152.
16. A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in Advances in Neural Information Processing Systems, vol. 32, 2019.
17. S.-A. Rebuffi, S. Gowal, D. A. Calian, F. Stimberg, O. Wiles, and T. Mann, "Data augmentation can improve robustness," in Advances in Neural Information Processing Systems, vol. 34, pp. 29935–29948, 2021.
18. L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in Advances in Neural Information Processing Systems, vol. 31, 2018.
19. M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS), 2016, pp. 308–318. https://doi.org/10.1145/2976749.2978318
20. A. Khamparia and K. M. Singh, "A systematic review on deep learning architectures and applications," Expert Systems, vol. 36, no. 3, p. e12400, 2019. https://doi.org/10.1111/exsy.12400
21. A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavabi, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," arXiv:2003.01200, 2020.
22. A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defenses," CAAI Transactions on Intelligence Technology, vol. 6, no. 1, pp. 25–45, 2021. https://doi.org/10.1049/cit2.12028
23. X. Liu, L. Xie, Y. Wang, J. Zou, J. Xiong, Z. Ying, and A. V. Vasilakos, "Privacy and security issues in deep learning: A survey," IEEE Access, vol. 9, pp. 4566–4593, 2020. https://doi.org/10.1109/ACCESS.2020.3045078
24. [Duplicate of reference 4 — please remove before submission.]
25. J. Sen, A. Sen, and A. Chatterjee, "Adversarial attacks on image classification models: Analysis and defense," arXiv:2312.16880, 2023.
26. Y. Wang et al., "Minimizing adversarial training samples for robust image classifiers: Analysis and adversarial example generator design," IEEE Transactions on Information Forensics and Security, 2024. [Please supply volume, issue, pages, and DOI.]

27. X. Zhong and C. Liu, "Sparse-PGD: A unified framework for sparse adversarial perturbations generation," arXiv:2405.05075, 2024.
28. M. Shao et al., "Latent code augmentation based on stable diffusion for data-free substitute attacks," IEEE Transactions on Neural Networks and Learning Systems, 2025. [Please supply volume, issue, pages, and DOI.]
29. J. Chen et al., "A Frank–Wolfe framework for efficient and effective adversarial attacks," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 04, pp. 3486–3494, 2020. https://doi.org/10.1609/aaai.v34i04.5753
30. S. H. Silva and P. Najafirad, "Opportunities and challenges in deep learning adversarial robustness: A survey," arXiv:2007.00753, 2020.
31. C. Zhang et al., "Generative adversarial networks: A survey on attack and defense perspective," ACM Computing Surveys, vol. 56, no. 4, pp. 1–35, 2023. https://doi.org/10.1145/3617895
32. A. K. M. I. Newaz et al., "Adversarial attacks to machine learning-based smart healthcare systems," in Proceedings of the IEEE Global Communications Conference (GLOBECOM), 2020, pp. 1–6. https://doi.org/10.1109/GLOBECOM42002.2020.9322472
33. O. A. Bello et al., "Machine learning approaches for enhancing fraud prevention in financial transactions," International Journal of Management and Technology, vol. 10, no. 1, pp. 85–108, 2023. [Please supply DOI if available.]
34. B. Badjie, J. Cecilio, and A. Casimiro, "Adversarial attacks and countermeasures on image classification-based deep learning models in autonomous driving systems: A systematic review," ACM Computing Surveys, vol. 57, no. 1, pp. 1–52, 2024. https://doi.org/10.1145/3685604
35. Z. Jia, H. Fang, and W. Zhang, "MBRS: Enhancing robustness of DNN-based watermarking by mini-batch of real and simulated JPEG compression," in Proceedings of the ACM International Conference on Multimedia, 2021, pp. 41–49. https://doi.org/10.1145/3474085.3475568
36. K. Zhu et al., "Improving generalization of adversarial training via robust critical fine-tuning," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4424–4434.
37. N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in Proceedings of the IEEE Symposium on Security and Privacy (S&P), 2016. https://doi.org/10.1109/SP.2016.41
38. R. Feinman, R. R. Curtin, S. Shintre, and A. B. Gardner, "Detecting adversarial samples from artifacts," arXiv:1703.00410, 2017.
39. J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," in Proceedings of the International Conference on Learning Representations (ICLR), 2017.
40. A. Raghunathan, S. M. Xie, F. Yang, J. Duchi, and P. Liang, "Adversarial training can hurt generalization," arXiv:1906.06032, 2019.
41. C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.
42. P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," in Proceedings of the International Conference on Learning Representations (ICLR), 2018. arXiv:1805.06605.
43. [Duplicate of reference 42 — please remove before submission.]
44. J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in Proceedings of the International Conference on Learning Representations (ICLR), 2018.
45. W. Zhao, S. Alwidian, and Q. H. Mahmoud, "Adversarial training methods for deep learning: A systematic review," Algorithms, vol. 15, no. 8, p. 283, 2022. https://doi.org/10.3390/a15080283
46. O. Gungor, Towards Intelligent, Secure, and Efficient Industrial Internet of Things, Ph.D. dissertation, University of California, San Diego, 2023. [Please supply URL or DOI if available.]
47. H. Wang and Y. Wang, "Self-ensemble adversarial training for improved robustness," arXiv:2203.09678, 2022.
48. Y. Cai et al., "Ensemble-in-one: Ensemble learning within random gated networks for enhanced adversarial robustness," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, no. 12, pp. 14738–14747, 2023. https://doi.org/10.1609/aaai.v37i12.26724

49. V. Dutta, M. Choras, M. Pawlicki, and R. Kozik, "A deep learning ensemble for network anomaly and cyber-attack detection," Sensors, vol. 20, no. 16, p. 4583, 2020. https://doi.org/10.3390/s20164583
50. C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," arXiv:1711.00117, 2017.
51. K. Mahmood, R. Mahmood, and M. van Dijk, "Back in black: A comparative evaluation of recent state-of-the-art black-box attacks," IEEE Access, vol. 10, pp. 998–1019, 2021. https://doi.org/10.1109/ACCESS.2021.3128280
52. M. Z. Horváth et al., "Boosting randomized smoothing with variance reduced classifiers," arXiv:2106.06946, 2021.
53. J. Zeng et al., "Certified robustness to text adversarial attacks by randomized [mask]," Computational Linguistics, vol. 49, no. 2, pp. 395–427, 2023. https://doi.org/10.1162/coli_a_00477
54. T. Maho, T. Furon, and E. Le Merrer, "Randomized smoothing under attack: How good is it in practice?," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 3014–3018. https://doi.org/10.1109/ICASSP43922.2022.9747033
55. H. Liu, K. Roeder, and L. Wasserman, "Stability approach to regularization selection (StARS) for high dimensional graphical models," in Advances in Neural Information Processing Systems, vol. 23, 2010.
56. D. Bhardwaj, K. Kaushik, and S. Gupta, "Accelerated smoothing: A scalable approach to randomized smoothing," arXiv:2402.07498, 2024.
57. E. Bagdasaryan, (Un)Trustworthy Machine Learning, Ph.D. dissertation, Cornell University, 2023. [Please supply URL or DOI if available.]
58. C. Yang et al., "Gradient leakage defense in federated learning using gradient perturbation-based dynamic clipping," in Proceedings of the IEEE International Conference on Web Services (ICWS), 2024, pp. 178–187.
59. Z. Lu, Z. Liao, and H. Li, "Robust and verifiable privacy federated learning," IEEE Transactions on Artificial Intelligence, vol. 5, no. 4, pp. 1895–1908, 2023. https://doi.org/10.1109/TAI.2022.3211887
60. W. Wang et al., "Certified robustness to word substitution attack with differential privacy," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2021, pp. 1102–1112. https://doi.org/10.18653/v1/2021.naacl-main.87
61. J. J. Hathaliya, S. Tanwar, and P. Sharma, "Adversarial learning techniques for security and privacy preservation: A comprehensive review," Security and Privacy, vol. 5, no. 3, p. e209, 2022. https://doi.org/10.1002/spy2.209
62. S. Nandi et al., "Certified adversarial robustness within multiple perturbation bounds," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2298–2305.
63. A. J. Ferreira and M. A. T. Figueiredo, "Boosting algorithms: A review of methods, theory, and applications," in Ensemble Machine Learning: Methods and Applications, Springer, 2012, pp. 35–85.
64. C. McPhail et al., "Robustness metrics: How are they calculated, when should they be used and why do they give different results?," Earth's Future, vol. 6, no. 2, pp. 169–191, 2018. https://doi.org/10.1002/2017EF000649
65. S. Mei et al., "A comprehensive study on the robustness of deep learning-based image classification and object detection in remote sensing," Journal of Remote Sensing, vol. 4, p. 0219, 2024. [Please supply DOI.]
66. A. Malik et al., "Application of functional traits in modelling productivity and resilience under climate change," in Plant Functional Traits for Improving Productivity, Springer, 2024, pp. 77–96.
67. Y. Chen and S. Eger, "MENLI: Robust evaluation metrics from natural language inference," Transactions of the Association for Computational Linguistics, vol. 11, pp. 804–825, 2023. https://doi.org/10.1162/tacl_a_00576