IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 January 2025, accepted 14 February 2025, date of publication 20 February 2025, date of current version 27 February 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3544107

RESEARCH ARTICLE

Evaluating Pretrained Deep Learning Models for Image Classification Against Individual and Ensemble Adversarial Attacks

MAFIZUR RAHMAN[1,2], PROSENJIT ROY[1], SHERRI S. FRIZELL[1], AND LIJUN QIAN[2,3], (Senior Member, IEEE)

[1]Department of Computer Science, Prairie View A&M University, Prairie View, TX 77446, USA
[2]CREDIT Center, Prairie View A&M University, Prairie View, TX 77446, USA
[3]Department of Electrical and Computer Engineering, Prairie View A&M University, Prairie View, TX 77446, USA

Corresponding author: Mafizur Rahman (mrahman13@pvamu.edu)

This work was supported by the Army Research Office under Grant W911NF-23-1-0214 and Grant W911NF-24-2-0133.

ABSTRACT The robustness of Deep Neural Networks (DNNs) against adversarial attacks is an important topic in the area of deep learning. To fully investigate the robustness of DNNs, this study examines four frequently used white box adversarial attack techniques, namely, the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), Basic Iterative Method (BIM), DeepFool, and their effects on DNN models for the image classification task. The results show that ResNet152 and DenseNet201 are less vulnerable comparing to other DNN models to a variety of individual attacks, highlighting their intrinsic strength even in the lack of specific adversarial training. Further, we propose two ensemble adversarial attacks combining three individual attacks for generating adversarial examples from the tiny ImageNet, CIFAR-10, CIFAR-100, and SVHN datasets for DNN model evaluation. It is observed that the performance of the DNNs deteriorate significantly under the proposed ensemble adversarial attacks even after defensive measures have been applied. For instance, the accuracy of the most robust DNN that we tested, namely the defense distillation enhanced DenseNet201, dropped more than 59% under the proposed ensemble adversarial attacks, comparing to only 34% decrease under the individual attacks.

INDEX TERMS Adversarial attack, adversarial training, basic iterative method, ensemble attack, deep neural networks, fast gradient sign method, projected gradient descent.

I. INTRODUCTION

Deep learning models have become highly effective and versatile tools in the present day, especially for image classification. Deep neural networks (DNNs) can discern complex patterns and extract meaningful features from large datasets has revolutionized multiple domains, including automobile, finance, manufacturing, healthcare, and entertainment [1]. These algorithms have showcased incredible accuracy and effectiveness, consistently outperforming conventional methods and setting new benchmarks in diverse applications. However, the widespread adoption of DNNs has also raised concerns concerning their vulnerability to adversarial attacks.

The associate editor coordinating the review of this manuscript and approving it for publication was Alicia Fornés.

The goal of adversarial attacks is to manipulate input data through intentionally designed perturbations to fool machine learning algorithms into incorrectly classifying the data. Therefore, even while DNNs perform well under normal conditions, powerful perturbation may quickly make them vulnerable(see an example of adversarial classification in Figure 1). Consequently, there are still concerns about the security and robustness of these models in critical applications. For instance, this report [2] claims that an autonomous Uber car injured a bicyclist because its systems failed to respond quickly enough. Also, these vehicles occasionally overlooked dynamic agents, real-time objects, and road layout information, which led to collisions among vehicles [3]. Such incidents occur from the complex and multifaceted situations encountered in real-world settings.

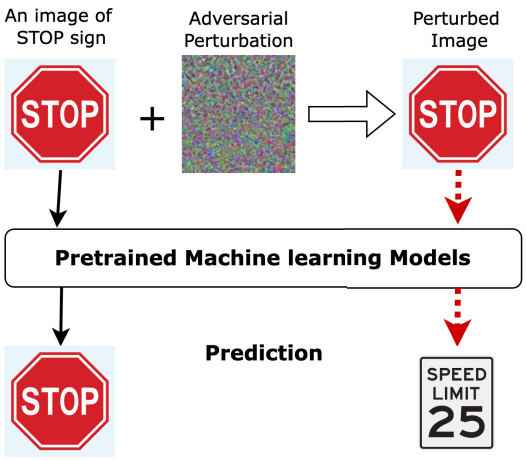


FIGURE 1. An adversarial perturbation example for image classification.

Thus, despite extensive training and evaluation of DNNs on large datasets like ImageNet, CIFAR-100, MS COCO, and Pascal VOC, etc. with the utilization of cutting-edge hardware such as GPUs, TPUs, and NPUs, these models still face challenges in accurately predicting objects in certain real-world scenarios [4]. The reason for focusing on white-box attacks in this analysis lies in their role as the most difficult threat model, where attackers have complete access to model architecture and gradients. This unrestricted access allows for preparing precise, highly targeted adversarial examples that reveal fundamental vulnerabilities in model robustness. Testing against white-box attacks delivers the most comprehensive assessment of model resilience and allows the refinement of defense mechanisms, such as defensive distillation, specially designed to counter such attacks. Although black-box and gray-box attacks are vital in adversarial research, they provide the attacker with limited information, generally resulting in less effective attacks that do not fully demonstrate a model's weaknesses [5], [6]. By focusing on white-box attacks, we set a high standard for robustness, confirming that models can withstand not only extreme adversarial scenarios but also the inherently weaker gray-box and black-box conditions, where the threat level is lower. Thus, we mainly focus on white-box adversarial attacks for image classification in this paper.

Currently, there's been a surge in curiosity about adversarial examples, encouraging researchers to improve the robustness of DNNs through adversarial training, particularly concentrating on image classification tasks. We found key classical algorithms like FGSM [7], PGD [8], DeepFool [9], JSMA [10], and CW [11] are commonly used for creating adversarial examples by introducing noise perturbations into the original images [12]. Additionally, some approaches also utilize patch perturbations for physical real-world attacks. These attack strategies are classified based on the attacker's knowledge about the target model: White-box, Black-box, and Gray-box. While White-box attacks need comprehensive knowledge, Gray-box attacks only require partial information [13]. In contrast, Black-box attacks such as One Pixel Attacks, Boundary Attacks, etc. do not require any prior knowledge about the target model [13]. The majority of earlier research has predominantly focused on employing adversarial attacks in image classification tasks, particularly highlighting techniques such as FGSM, PGD, and One Pixel Attack. Researchers delved into the impact of the ϵ value in the FGSM attack, the effect of single-pixel perturbations on input images, and the ramifications of adapting the step size and number of iterations in the PGD method. Additionally, most studies mainly consider adversarial attacks in regulated laboratory settings utilizing standardized datasets like MNIST or Fashion-MNIST, however, these datasets may not capture real-world complexities.

Our research makes major contributions in the field of adversarial attacks on DNNs for image classification, building upon previous techniques. To test the strength of our proposed ensemble attacks, we examined our method against two popular adversarial attack defense mechanisms, Defensive Distillation and Gradient Masking, on the CIFAR-10 and CIFAR-100 datasets. While defensive distillation performance has already shown to be compelling against an individual attack, our ensemble methods were able to effectively defeat this defensive distillation process and sufficiently damage the model to cause misclassification. In summary, we make the following contributions:

1) **Impact of Attack Method and Noise Evaluation:** We perform a thorough analysis of different attack methods, including FGSM, PGD, Basic Iterative Method (BIM), DeepFool, and two proposed ensemble attacks on several well-known pre-trained deep learning models including Resnet18, Resnet50, Resnet152, Densenet201, Inception_V3, VGG11, VGG16, and MobileNet_V2. We measure the influence of adversarial attacks on model performance by comparing the accuracy of models on tiny ImageNet [14] images before and after these attacks. In addition to traditional gradient-based attacks, we explore the impact of adversarial noise and spatial transformations on model robustness.
2) **Proposed Ensemble Attacks:** We employ two adversarial ensemble attack techniques (Mean and Weighted Ensemble), where we construct an ensemble for generating adversarial examples by utilizing three different individual attacks. Precisely, we present a method for adjusting the weights of different attacks within an ensemble method to examine how these ensemble attacks affect image classification. We receive 50% to 60% accuracy for individual attacks on some DNNs, such as ResNet152 and DenseNet201. However, under the proposed ensemble attacks, all eight DNNs' accuracy decreases to $\leq$27%, resulting in a 23% to 33% decline in performance. Although we concentrate solely on white-box attacks in this work, the proposed

process can be extended to encompass black-box attacks as well.

3) **Evaluation against Defensive Measures:** Lastly, we assess all attacks against the defensive distillation approach. We use all eight DNNs on the CIFAR-10 dataset and find that again ResNet152 and DenseNet201 perform well against both individual and our proposed ensemble attacks compared to the other models. It is also observed that both ResNet152 and DenseNet201 with the defensive mechanism in place still suffer a significant performance decrease of over 58% under the proposed ensemble adversarial attacks, comparing to their performance drop by just 39% under individual attacks. As both ResNet152 and DenseNet201 demonstrate strong performance, we also test our ensemble models on the CIFAR-100 and SVHN datasets. Similar performance drops are observed, even with the application of gradient masking as a defensive mechanism.

A. RATIONALE FOR PRIORITIZING WHITE-BOX SCENARIOS

We limited our study to white-box attacks for numerous reasons. Black-box or gray-box attacks do not provide a controlled environment [15] like white-box attacks, where all model parameters and gradients are accessible. This confirms that the evaluation of adversarial techniques focuses solely on the effectiveness of the proposed methods without confounding variables like query limits or transferability gaps [16]. In addition, our primary goal is to introduce and validate two novel ensemble attack strategies. Thus, adding black-box or gray-box scenarios would dilute this focus by requiring further adjustments for query budgets [17], surrogate model training, or transferability strategies [18], which are beyond the scope of our intended contribution. Additionally, including black-box or gray-box attacks adds complexity like managing queries or surrogate models [18], which could detract from our focus on fundamental innovations in ensemble attacks and is already well-covered in prior research [19], [20], [21]. Another vital reason, black-box attacks rely on adversarial example transferability, which varies by model and data [22]. Since our study presents new ensemble strategies, relying on transferability assumptions could obscure the evaluation of the presented methods inherent strengths. Also, black-box attacks that rely on query-based methods are computationally intensive and time-consuming [23]. So, incorporating such methods would require significant computational resources, which could detract from the focus on validating ensemble strategies under optimal conditions.

II. RELATED WORK

Different exploration for adversarial attack stood out because of their capability to compromise the robustness and reliability of artificial intelligence frameworks. Doss and Gunasekaran [24] presented the Basic Iterative Method (BIM) for creating intense adversarial examples, especially applied to ResNet50 utilizing the ImageNet Stubs dataset. BIM iteratively noised input images to sidestep defense systems, showing its adequacy in fooling ResNet50's classification. Moreover, they assessed defense systems like adversarial preparation and info pre-processing to alleviate BIM-based attacks. They highlighted the meaning of strong protection systems in shielding deep learning models against adversarial dangers. Beerens and Higham [25] progressed algorithm advancement in adversarial attack writing by refining annoyance estimation, expecting to relieve adversarial influence on deep learning models. Their new calculation, contrasted with DeepFool, limited componentwise relative change, preserving zero-esteemed pixels, subsequently improving human eye subtlety. This experiment showed potential for working on model robustness against adversarial attacks. According to the Jung et al. [26], the effectiveness of adversarial attacks like FGSM, BIM, PGD, Deepfool were shown to be substantially correlated with neural network complexity, with the attacks being less effective on Inception-v3, somewhat effective on VGG16 and ResNet50, and highly effective on MobileNet. It was discovered that the model architecture had a greater impact on adversarial labels than the original image, with the second label of the clean image frequently emerging as the adversarial example's best guess. Attack label comparisons showed greater similarity across attacks using comparable algorithms, and because of its lightweight architecture, MobileNet consistently generated similar hostile labels.

Recently, He and Cui in [27] presented two ensemble adversarial attack methods: magnitude-agnostic bagging ensemble (MABE) and gradient-grouped bagging and stacking ensemble (G^2BASE), to improve attack performance. Each model in MABE contributes to the final gradient direction by normalizing the input gradients based on the gradient amplitude whereas G^2BASE groups models by gradient magnitude, utilizing bagging within groups and stacking between groups to prevent dominance by any single model. Although G^2BASE is time-consuming, it offers more flexibility and effectiveness than MABE, which is efficient. Fu et al. [28] introduced a new black-box adversarial attack model ELAA that utilizes the AutoAttacker framework based on reinforcement learning and ensemble learning for image classification. No internal knowledge about the target network, such as its structure or weights, is required for ELAA to generate adversarial samples. ELAA achieved a success rate of 15% higher than any baseline technique, while attacks using ensemble learning have a success rate of about 35% higher than attacks utilizing a single learning model. Tramèr et al. presented Ensemble Adversarial Training (EAT) [29], a strategy that improves model reliability by using adversarial examples generated from numerous pre-trained models. This method broadens the range of perturbations, thereby strengthening defenses against black-box attacks. In 2022, Serial Minigroup Ensemble Attack (SMGEA) [30], a black-box adversarial attack method

introduced that uses long-term gradient memories to make it easier to move adversarial examples between models. SMGEA performed better than existing methods at tasks like image translation and predicting what parts of an image are most important. Xiong et al. [31] presented a Stochastic variance-reduced ensemble adversarial attack (SVRE). This attack utilizes stochastic variation reduction techniques to make adversarial example generation more viable. The strategy consolidates multiple attack techniques, which diminishes the variety in gradient estimates and raises the pace of assembly. Heterogeneous Prototype Learning (HPL) with its disentangled feature learning [32], can be utilized by isolating identity-specific information to study how adversarial attacks exploit domain variations. This approach enables targeted analysis of cross-domain vulnerabilities and potential defense mechanisms. Adversarial attacks on Explainable GNNs underscore vulnerabilities in interpretability [33], complementing our focus on ensemble attacks by extending insights into robustness challenges in AI systems. Thus, Wang et al. [34] introduce a certified robustness-inspired attack framework for GNNs, concentrating on nodes with smaller certified perturbation sizes to improve the effectiveness of evasion and poisoning attacks. Their framework significantly enhances the performance of existing attack methods by leveraging certified robustness properties.

Above all these analyses evaluated diverse adversarial attacks across a range of models. The majority of research concentrated on how to maintain model robustness when dealing with individual adversarial attacks or multi-layer attacks without iterative stages. Also, the main focus of recent research [27], [28], [29], [30] on ensemble adversarial attacks has been on black box techniques, stochastic variation reduction [31], gradient magnitude conflicts, and mitigating model dominance. However, our study completely differs from these studies since we examined white-box attacks and analyzed the effects of varying the weights of various attacks within an ensemble technique on image classification in addition to single attacks for various DNNs. Furthermore, we tested the robustness of our ensemble methods against the commonly used adversarial attack defense mechanism, defensive distillation. These contributions offer crucial insights into the ways in which adversarial manipulation may impair deep learning models and recommend the optimal model that is more resilient to various forms of adversarial attacks.

III. METHODOLOGY

In this section, we review three individual adversarial attacks as baseline firstly. Then defense distillation is introduced as the defense mechanism to improve the robustness of the machine learning models against adversarial attacks. The details of the proposed ensemble adversarial attacks are given afterwards.

A. WHITE BOX ADVERSARIAL ATTACKS

The **Fast Gradient Sign Method (FGSM)** is being utilized to produce adversarial samples. To maximize the loss, FGSM first calculates the gradients of the loss function with respect to the input data and then modifies the input in that direction. FGSM specifically determines the sign of the gradient of the loss function with respect to the input pixels for a given input image. FGSM creates adversarial samples, which are frequently misclassified by the classifier, by adding a little disturbance to the input picture in the direction of this sign [35]. The following formula describes this process:

$$\text{FGSM} = \text{Input Image} + \epsilon \cdot \text{sign}(\nabla_{\text{Input Image}}\text{Loss})$$

where the perturbation's magnitude is controlled by ϵ. The generated adversarial examples show weaknesses in the classifier's decision boundaries by exhibiting subtle changes that are invisible to human observers but can have a substantial influence on its predictions. Through the use of FGSM in our technique, we calculated accuracy on adversarial attack and the amount of noise introduced to original image. An adversarial example is shown in Figure 2 with a noise amount of 0.85.

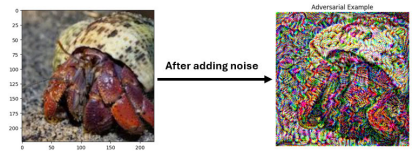


FIGURE 2. An adversarial example from ImageNet after injecting noise.

Using the **Projected Gradient Descent (PGD)** attack, we adhere to the formula's [36] description of an iterative optimization process:

$$x_{\text{adv}}^{t+1} = \text{Image}_{x,\epsilon}\left(x_{\text{adv}}^{t} + \alpha \cdot \text{sign}(\nabla_x J(f(x_{\text{adv}}^{t}), y))\right)$$

The PGD attack is instantiated in the code where x is the input image, with parameters like maximum iterations (t), the step size (α), and maximum perturbation (ϵ), which reflect the optimization goal. In an effort to cause misclassification, the PGD attack is then executed repeatedly in the code, perturbing the input data in order to maximize the loss function J. We learn more about the classifier's vulnerability to adversarial perturbations by assessing the classifier's accuracy on the produced adversarial instances and measuring the amount of noise injected, as implemented in the code.

We also use the **Basic Iterative Method (BIM)** attack, which is an iterative version of the Fast Gradient Sign Method (FGSM). Using a bounded -norm ball, BIM repeatedly modifies the input data in an effort to optimize the loss function and cause misclassification.

The **DeepFool** attack seeks to cause misclassification by iteratively perturbing input data along the direction of minimal change. The primary equation [9] that guides DeepFool's iterative procedure is:

$$x_{\text{adv}}^{t+1} = x + \frac{\left|\nabla_x J(f(x^t), y)\right|}{\|\nabla_x J(f(x^t), y)\|_2^2} \cdot \nabla_x J(f(x^t), y)$$

The original input image is represented by x_{adv}, the adversarial attack, the iteration step by t, the loss function by J, the prediction function by the model by f, and the gradient of the loss function with respect to the input image x^t is $\nabla_x J(f(x^t), y)$. The adversarial examples that are generated are then used to display adversarial images, measure the amount of noise that DeepFool adds to the original image, calculate accuracy on adversarial instances, and assess classifier predictions. This methodology sheds light on the classifier's susceptibility to DeepFool-based adversarial attacks.

B. DEFENSIVE DISTILLATION

Defensive distillation is a sophisticated method for making DNNs more resistant to challenges from other adversarial attacks. For this method, a simplified model is trained. This is a smaller and more reliable version of the original model. At the beginning, the original dataset is used to train a teacher model to make ''soft labels,'' or class probabilities, using a modified softmax function. As a result of these softer probabilities, training goals are set for a student model that acts like the teacher model but is also more resistant to changes made by adversaries.The main goal behind defensive distillation is that it can make the neural network's decision limits smoother. When the student model train on soft labels, the network becomes less sensitive to small changes in the input, which makes adversarial attacks less effective. This smoothing impact complicates the risk of making antagonistic models that can fundamentally change the model's prediction. In defensive distillation, the softmax function is modified to include a temperature parameter T, which controls the softness of the output probabilities [37]. The probability distribution q_i for the teacher model is:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

where z_i is the logit for class i and T is the temperature parameter. During distillation, the student model is trained to minimize the cross-entropy loss between its predictions p_i and the softened labels from the teacher model q_i:

$$L = -\sum_i q_i \log(p_i)$$

The essence of defensive distillation lies in this process where higher temperatures in the softmax function lead to softer probability distributions [38]. That is the main reason to make the model's decision boundaries smoother and less sensitive to small perturbations. This reduces the effectiveness of adversarial attacks.

C. PROPOSED ENSEMBLE ADVERSARIAL ATTACK METHOD

Through the **Mean Ensemble Attack(MEA)**, adversarial samples can be generated independently by applying multiple attack techniques, including the Saliency Map Method, PGD, and FGSM. By introducing perturbations to the input photos, these techniques produce adversarial examples that deceive the classification algorithm.

Let $\mathbf{x}$ be the actual input image, and $\mathbf{x}^A_{\text{adv}}$ be the adversarial sample generated by a specific attack A. We represent the perturbation added to $\mathbf{x}$ by δ_A. The adversarial example created by attack method A is shown by:

$$\mathbf{x}^A_{\text{adv}} = \mathbf{x} + \delta_A$$

Similarly, let $\mathbf{x}^B_{\text{adv}}$ and $\mathbf{x}^C_{\text{adv}}$ be the adversarial examples generated by attack methods B and C, respectively. The Mean Ensemble Method includes averaging the perturbations induced by these attack methods to construct a new adversarial instance $\mathbf{x}_{\text{ens}}$:

$$\mathbf{x}_{\text{ens}} = \frac{\mathbf{x}^A_{\text{adv}} + \mathbf{x}^B_{\text{adv}} + \mathbf{x}^C_{\text{adv}}}{3}$$

This ensemble attack adversarial sample $\mathbf{x}_{\text{ens}}$ is then utilized for evaluation purposes.

The **Weighted Ensemble Attack(WEA)** also uses multiple attack methods, such as FGSM, PGD, and BIM to construct adversarial examples. By altering the weights given to each attack technique, the Weighted Ensemble Method allows for more flexibility in combining the adversarial examples. For example, we can assign weights of FGSM (0.4), PGD (0.3), and BIM (0.3) to create a weighted ensemble-generated adversarial scenario.

Let $\mathbf{x}^A_{\text{adv}}$, $\mathbf{x}^B_{\text{adv}}$, and $\mathbf{x}^C_{\text{adv}}$ represent the adversarial samples generated by attack methods A, B, and C. The weighted ensemble adversarial example $\mathbf{x}_{\text{wens}}$ is shown by:

$$\mathbf{x}_{\text{wens}} = w_A \cdot \mathbf{x}^A_{\text{adv}} + w_B \cdot \mathbf{x}^B_{\text{adv}} + w_C \cdot \mathbf{x}^C_{\text{adv}}$$

where w_A, w_B, and w_C are the weights assigned to each attack method, with $\sum_i w_i = 1$.

Therefore, using the example weights of FGSM (0.4), PGD (0.3), and BIM (0.3), the weighted ensemble adversarial example $\mathbf{x}_{\text{wens}}$ becomes:

$$\mathbf{x}_{\text{wens}} = 0.4 \cdot \mathbf{x}^A_{\text{adv}} + 0.3 \cdot \mathbf{x}^B_{\text{adv}} + 0.3 \cdot \mathbf{x}^C_{\text{adv}}$$

This method allows for fine-tuning the combination of adversarial examples based on the desired impact of each attack method.

In this work, we presented two types of ensemble methods and evaluated their effect on different DNNs for image classification utilizing the Tiny ImageNet dataset. Later, to assess the robustness of our ensemble attack techniques, we tested our both ensemble methods against defensive distillation using previous eight DNNs on the CIFAR-10 dataset. This experiment allowed us to assess the performance of our ensemble model when DNNs undergo adversarial training and distillation processes. Above, we introduce our proposed weighted ensemble attack Algorithm 1 along with defensive distillation.

Algorithm 1 Weighted Ensemble Attack Pseudocode

Data: CIFAR-10 dataset, pretrained DNN model T, N_{CLASSES}
Result: Ensemble attack accuracy

Loading and Initialization
Load CIFAR-10 dataset & pretrained DNN model T;
Modify final layer for N_{CLASSES};
Copy T to student model S;
Train Teacher Model T
for *each epoch* **do**
  for *each batch b in training data* **do**
    Forward pass through T;
    Compute loss $\mathcal{L}_T$ and backpropagate;
Save trained teacher model T;
Distillation Process
Define distillation loss $\mathcal{L}_D$;
Initialize DistillModel with T and S;
Train Student Model S **with Distillation**
for *each epoch* **do**
  for *each batch b in training data* **do**
    Forward pass through T and S;
    Compute distillation loss $\mathcal{L}_D$ and backpropagate;
Evaluate Student Model S
Set S to evaluation mode;
for *each batch b in test data* **do**
  Forward pass through S and compute accuracy;
Adversarial Attacks
Define FGSM, PGD, and BIM attacks;
Function `Create_adv_examples`(*model*, *img*, *labels*, *device*) **:**
  return FGSM, PGD, BIM adversarial examples;
Function `Ensemble_attack`(*attacks*, *input_batch*, *weights*) **:**
  Initialize *ensemble_attack* $\leftarrow 0$;
  for *each attack a* **do**
    Add weighted adversarial examples to *ensemble_attack*;
  return *ensemble_attack*;
Evaluate Against Ensemble Attack
Set model to evaluation mode;
for *each batch b in test data* **do**
  Generate ensemble adversarial examples;
  Forward pass through model and compute accuracy;
Test original and distilled model against ensemble;

IV. RESULTS AND ANALYSIS

In this section, we report a complete analysis of various attack methods and evaluate the performance of SOTA image classification models against these adversarial attacks. Initially, we delve into a comparative evaluation of different attack strategies, shining a light on the vulnerabilities and strengths of popular models when confronted with these tactics. Following this, we explore how our proposed ensemble adversarial attack mechanisms fool DNN models and compare our approach with individual attack performance with the adversarial defense method. The accuracy values shown in each results table are the averages computed over multiple experimental runs, ensuring the reliability and consistency of the results.

A. INDIVIDUAL ATTACK METHOD EVALUATION

We perform experiments utilizing four different adversarial attack methods, namely FGSM, PGD, BIM, and DeepFool, to assess their influence on different DNN models. In our study with FGSM, we uniformly use a perturbation magnitude (ϵ) of 0.5 across all DNN models to observe their reaction to this level of disturbance. Also, for the BIM attack, we fix ϵ to 0.5, set a step size of 1, and limit the method to a maximum of 10 iterations. For PGD attacks, we set the parameters to an ϵ of 0.5, a step size of 1, and maximum iterations of 20. We explore different values of the perturbation magnitude (ϵ) and find that $\epsilon = 0.5$ delivers a balanced and effective benchmark for assessing model robustness across diverse attack types. Lower values of ϵ result in subtle perturbations that do not significantly challenge the models, while higher values often lead to excessive distortions, which are less representative of natural adversarial scenarios. Thus, we select $\epsilon = 0.5$ as an optimal threshold, providing a fine level of disturbance to test model resilience effectively without overly compromising the natural structure of the input. We evaluate the performance of the classification models on adversarial examples generated from the ImageNet dataset, proposing insights into the models' strength against intentionally crafted inputs aiming to exploit their vulnerabilities. The table below presents a summarized overview of the outcomes of these methods.

TABLE 1. Accuracy of models before and after various attacks.

Model	Baseline Acc(%)	Under Attack Acc.(%)			
		FGSM	PGD	BIM	DeepFool
ResNet18	69	21.33	19.21	47.10	42.29
ResNet50	73	27.12	25.21	54.40	55.29
ResNet152	**82**	**33.05**	31.72	55.90	51.32
VGG11	73	22.24	23.75	43.44	52.92
VGG16	78	25.95	29.67	43.56	49.32
DenseNet201	80	32.82	**36.54**	**60.12**	**61.22**
Inception_V3	60	27.80	26.33	38.19	41.22
MobileNet_V2	73	25.45	30.45	49.91	51.17

According to Table 1, before experiencing adversarial attacks, ResNet152 offers the highest accuracy at 82%, while Inception_V3 has the least accuracy at 60%. All models show a considerable fall in accuracy after the adversarial attacks. When compared to other techniques, the FGSM attack typically results in a smaller drop in accuracy. However, PGD, BIM, and DeepFool attacks prove to be more influential, with DeepFool forcing the most substantial decrease in accuracy for most models. After the attacks,

DenseNet201 displays the maximum accuracy; specifically, after the PGD, BIM, and DeepFool attacks, it records the highest post-attack accuracies at 36.54%, 60.12%, and 61.22%, respectively. This suggests that DenseNet201 is the most robust model among those tested against the adversarial attacks applied. DenseNet201's architecture has direct links from any layer to all subsequent layers, improving gradient flow and leading to better-learned, more robust features. Also, its depth enables the capture of complex patterns and nuances, making it tougher for adversarial attacks to deceive the model.

TABLE 2. Time taken for adversarial attack generation for each model in seconds.

Model	FGSM	PGD	BIM	DeepFool
ResNet18	0.03	0.54	0.17	0.49
ResNet50	0.06	0.56	0.26	1.12
ResNet152	0.45	**1.31**	0.98	2.22
VGG11	0.06	0.66	0.68	1.22
VGG16	0.10	0.69	0.76	1.30
DenseNet201	**0.61**	1.21	**1.26**	**2.43**
Inception_V3	0.09	0.96	0.66	0.92
MobileNet_V2	0.06	0.33	0.34	0.65

Table 2 implies that among all four attacks, DeepFool is the attack that takes the longest to execute for generating adversarial example, taking 2.43 seconds to finish in DenseNet201. Besides this, ResNet152 also exhibits longer generation times, particularly for PGD and DeepFool attacks, taking 1.31 and 2.22 seconds, respectively. Conversely, attacks against ResNet18 and MobileNet_V2 are generated rather quickly; for ResNet18, FGSM takes as little as 0.03 seconds, while for MobileNet_V2, PGD takes only 0.33 seconds. For all models, FGSM generates attacks the fastest overall, while DeepFool often takes the longest.

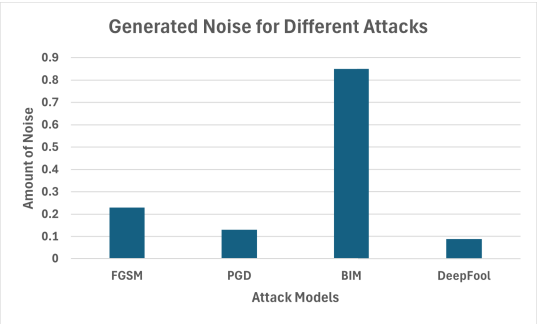


FIGURE 3. Noise level for generating adversarial examples.

Figure 3 shows the relative amount of noise induced by four adversarial attacks FGSM, PGD, BIM, and DeepFool when used with fixed parameter values across different DNN models. Compared to the other methods, the BIM attack produces the most noise, with PGD and DeepFool generating the least. Compared to PGD and DeepFool, FGSM generates a substantial amount of noise, yet much less than BIM. Although the attack model parameters value(such as epsilon, step size, etc.) can be set higher to generate more noise, we have standardized the noise ratio for all DNN models since our goal is to assess how well each model performs when classifying adversarial samples.

B. PROPOSED ENSEMBLE ATTACK PERFORMANCE

1) MEAN ENSEMBLE ATTACK METHOD

We construct three adversarial samples using three different attack methods: FGSM, PGD, and Saliency Map Method. These methods independently generate adversarial examples by applying perturbations to input images, aiming to trick the classification model. We average the perturbations induced by these methods to get a new adversarial example, which is an ensemble attack adversarial example. Later, we assess the effectiveness of the targeted classification model by using this ensemble example as input for classification.

2) WEIGHTED ENSEMBLE ATTACK METHOD

We use FGSM, PGD, and BIM to create adversarial examples. In contrast to the first method, which gives equal weights to each attack method, we combine the adversarial instances in this way using weighted averaging. In this case, the attack techniques can be given weights to determine how much of an impact they have on the overall adversarial situation. By default, equal weights are assigned in the absence of any weights. A weighted ensemble-generated adversarial scenario is shown in Figure 4, illustrating the combined effect of various attack techniques with different weights of FGSM(0.4), PGD(0.3), and BIM(0.3). We choose the weights (0.4, 0.3, 0.3) for FGSM, PGD, and BIM via systematic exploration within the range [0, 1] to attain a balanced evaluation of the model's robustness across various attack strengths. Also, we theoretically justified this weight setup utilizing game theory. Assigning 0.4 to FGSM guarantees that simpler, single-step attacks are well-represented, while the balanced weights of 0.3 each for PGD and BIM capture the model's resilience to stronger, iterative attacks without overemphasizing high-intensity threats. This combination avoids robustness overfitting, where higher weights on PGD or BIM could skew the evaluation toward complex attacks at the expense of generalizability. The resulting weighted robustness score,

$$R = 0.4 \cdot r_{\text{FGSM}} + 0.3 \cdot r_{\text{PGD}} + 0.3 \cdot r_{\text{BIM}}$$

effectively reflects the model's performance across various adversarial conditions, providing a comprehensive and realistic estimation.

FIGURE 4. Combined attack adversarial examples.

3) JUSTIFICATION FOR WEA WEIGHT SELECTION USING GAME THEORY

The preference of weights (0.4, 0.3, 0.3) for FGSM, PGD, and BIM can be theoretically justified using game theory, where our ensemble attack approach aims to achieve a Nash equilibrium [39]. At this equilibrium, the weights convey an optimal trade-off among competing attack techniques to maximize the adversarial effectiveness on the target model while keeping a balance that avoids dominance by any single attack [40], [41]. FGSM is given a higher weight (0.4) to ensure simpler, single-step attacks are adequately defined, as these attacks can often reveal fundamental vulnerabilities in the model that may not be exploited by more computationally intensive, iterative methods. This also enhances the transferability of adversarial examples in real-world systems, where simpler perturbations are more practical. Meanwhile, the balanced weights of 0.3 for PGD and BIM highlight the iterative nature of these attacks, capturing their ability to reveal deeper vulnerabilities in the model's decision boundaries without overemphasizing their impact. We find that overweighting PGD or BIM skew the robustness evaluation toward complex attacks, potentially reducing the generalizability of the ensemble's assessment.

We tested with several alternative weight configurations (e.g., 0.5-0.2-0.3, 0.3-0.4-0.3) to investigate the impact of different trade-offs between attack methods. However, the (0.4, 0.3, 0.3) configuration provided the most consistent and comprehensive results across our benchmarks. This configuration avoided robustness overfitting, balanced attack intensity, and performed a well-rounded adversarial evaluation. The Nash equilibrium perspective validates this choice, as the selected weights ensure that no single attack can unilaterally adjust its weight to enhance the overall adversarial impact significantly. This optimal balance is theoretically sound and empirically supported, indicating that the weighted ensemble provides a robust and fair evaluation of the model's performance under various adversarial situations.

TABLE 3. Comparison of model accuracies for ensemble attacks.

Model	Accuracy(%)	
	Mean Ensemble	Weighted Ensemble
ResNet18	9.50	14.8
ResNet50	13.98	14.54
ResNet152	21.88	**27.08**
VGG11	11.98	12.65
VGG16	13.95	12.40
DenseNet201	**25.43**	26.76
Inception_V3	13.34	17.04
MobileNet_V2	11.23	9.89

From Table 3 results, ResNet152 and DenseNet201 show the highest accuracies in both mean and weighted ensemble methods, displaying their superior stability to ensemble attacks. For other DNNs, accuracies under the mean ensemble method are below 15%. In contrast, the weighted ensemble method exhibits higher precisions for most models except for VGG11 and Inception_V3, compared to their mean ensemble accuracies. This might result from giving FGSM a greater weight of 0.4 and PGD and BIM lower weights of 0.3 in the ensemble. Since FGSM is a less complex attack compared to PGD and BIM, its higher weight contributes less to the overall result. Changing the weights to favor PGD or BIM might lead to a reversal of these outcomes.

C. PROPOSED ENSEMBLE ATTACK WITH DEFENSIVE DISTILLATION

So far, we have analyzed how individual attacks and our proposed ensemble attacks affect the performance of DNNs. In this section, we will discuss the significance and robustness of our proposed ensemble method. We tested our mean and weighted ensemble attack algorithms against the defensive distillation method. Our goal was to evaluate the performance reduction in DNN performance even after adversarial training and distillation processes. We selected prior eight models as our DNN, and each DNN model acts as the teacher model. The table below presents the parameter values used in our testing experiment. To evaluate our proposed ensemble method, we first compute the model performance for individual attacks on CIFAR-10, similar to Table 1, but now including an adversarial defense mechanism (i.e., defensive distillation)(See Table 5). Based on prior studies [37], we set the distillation parameter alpha to 0.7 and the temperature to 3.0, as these values have been found optimal for the DNN model in a similar context. However, these parameters can be adjusted according to the specific model and task requirements.

TABLE 4. Parameter values for testing experiment.

Parameter	Value
Batch Size	64
Learning Rate	0.001
Perturbation Magnitude (ϵ)	0.5
Alpha (for attacks)	0.01
Temperature (for distillation)	3.0
Alpha (for distillation)	0.7
Attack Methods	FGSM, PGD, BIM
Weights for Attacks	[0.4, 0.3, 0.3]

Table 5 shows the performance of different DNN models on the CIFAR-10 dataset, emphasizing their accuracy before and after adversarial attacks, both with and without adversarial defense (A.D.). Notably, the implementation of A.D. led to substantial progress in baseline accuracy across all models, for instance, DenseNet201's accuracy improved from 90.32% to 95.11%, marking a 4.79% progress. Under adversarial attacks, A.D. greatly mitigated accuracy drops. For example, ResNet18's performance against the PGD attack improved from 23.2% to 25.31% with A.D., indicating a 2.11% increase. Similarly, DenseNet201's accuracy against the BIM attack improved from 56.7% to 59.1%, an improvement of 2.4%. Once again, we notice that even after introducing adversarial defense, ResNet152 and DenseNet201 perform the best among all models, with DenseNet201 reaching the

TABLE 5. Model performance with and without adversarial defense (A.D.) on CIFAR-10(Lower acc. % indicates stronger attack).

Model	Baseline(Acc. %)		Under Attack without A.D.(Acc. %)				Under Attack with A.D.(Acc. %)			
	Without A.D.	With A.D.	PGD	FGSM	BIM	DeepFool	PGD	FGSM	BIM	DeepFool
ResNet18	80.74	84.75	25.02	25.31	44.12	43.14	25.81	29.5	48.6	47.01
ResNet50	84.4	87.90	29.06	29.19	45.76	44.45	29.50	31.05	50.09	52.58
ResNet152	87.67	90.95	27.2	**35.81**	47.89	45.4	28.11	**36.15**	51.86	49.91
VGG11	82.65	83.76	24.24	25.43	39.45	39.94	26.43	25.45	44.6	42.01
VGG16	84.66	88.51	25.26	30.3	49.22	50.41	**38.14**	32.12	50.63	53.71
DenseNet201	**90.32**	**95.11**	**39.21**	29.12	**56.76**	**54.23**	29.3	30.12	**59.1**	**60.45**
InceptionV3	83.18	85.23	30.76	32.23	40.06	43.2	30.91	35.41	43.66	43.34
MobileNetV2	81.29	85.01	23.23	19.74	33.05	40.24	21.34	22.83	35.84	43.2
Avg. Acc.	**84.36**	**87.65**	**28.00**	**28.39**	**44.54**	**45.13**	**28.70**	**30.33**	**48.05**	**49.03**

TABLE 6. Model performance for our proposed ensemble attacks with and without adversarial defense (A.D.) on cifar-10 (Lower acc. % indicates stronger attack).

Model	Accuracy (%)			
	MEA without A.D.	MEA with A.D.	WEA without A.D.	WEA with A.D.
ResNet18	10.33	11.23	10.03	13.92
ResNet50	16.66	17.12	19.45	19.4
ResNet152	21.13	**32.55**	16.33	29.34
VGG11	10.53	10.08	9.99	11.27
VGG16	12.67	16.92	12.12	17.14
DenseNet201	**27.24**	27.45	**27.89**	**35.23**
Inception_V3	15.90	16.35	16.15	21.76
MobileNet_V2	11.89	15.93	9.93	14.01
Avg. Acc.	**15.79**	**18.45**	**15.24**	**20.26**

TABLE 7. Performance Drop (%) under different adversarial conditions (Higher drop rate % indicates stronger attack).

Condition	Proposed Ensemble Attack		Individual Attack			
	MEA	WEA	PGD	FGSM	BIM	DeepFool
Without A.D.	81.29%	81.91%	66.81%	66.34%	47.19%	46.50%
With A.D.	78.95%	76.91%	67.26%	65.41%	45.18%	44.07%

highest individual accuracy of 60.45% on DeepFool with the defensive mechanism. As DenseNet201 achieved a baseline accuracy of 95.11% without an attack (with the defensive mechanism), the accuracy dropped by 34.66% after the attack (with the defensive mechanism). The average accuracy across all eight DNNs is summarized in the last row of the table. This metric provides expected performance when the exact DNN model is not known to the attacker.

Now, we evaluate how each DNN performs after executing our two proposed ensemble attacks with and without defensive mechanisms. Table 6 indicates that, along with the defensive mechanism, the mean ensemble attack (MEA) with ResNet152 reached the highest accuracy of 32.55%, while the weighted ensemble attack (WEA) with DenseNet201 achieved the highest accuracy of 35.23%. Considering their baseline model accuracy with the defensive mechanism, which were 90.95% for ResNet152 and 95.11% for DenseNet201, we see that their accuracy declined by around 58% for ResNet152 and 59.88% for DenseNet201 after our ensemble adversarial attacks. This result indicates that even after adversarial training and distillation, DNNs still misclassify significantly under adversarial conditions, highlighting the strength of our proposed ensemble adversarial attacks.

Average Attack Effectiveness: The performance drop is computed using the following formula:

$$\text{Dropped (\%)} = \left(\frac{\text{Baseline Accuracy (\%)} - \text{Attack Accuracy (\%)}}{\text{Baseline Accuracy (\%)}}\right) \times 100$$

Our MEA and WEA result in higher performance drops both with and without adversarial defenses (A.D.)(see Table 7). For example, MEA/WEA with A.D. causes a performance reduction of 78.95% and 76.91%, respectively. Both are much higher than the drops yielded by individual attacks with A.D. This indicates that the proposed MEA and WEA attacks can exploit vulnerabilities in DNN models more effectively, thereby they are less affected by adversarial defenses.

1) EFFECTIVENESS OF WEA AND MEA ON CIFAR-100 AND SVHN

Based on the above analysis, we find both ResNet-152 and DenseNet-201 most robust against our ensemble attack methods, WEA and MEA, even without adversarial defenses. Therefore, we extend our evaluation utilizing two additional datasets, CIFAR-100 and SVHN, using ResNet-152 and DenseNet-201 models in order to further validate our

TABLE 8. Model performance for our proposed ensemble attacks with and without adversarial defense (A.D.) on CIFAR-100 (Lower acc. % indicates stronger attack).

Model	Baseline Acc(%)	Accuracy (%)			
		MEA without A.D.	MEA with A.D.	WEA without A.D.	WEA with A.D.
ResNet152	**78.21**	**25.67**	**33.35**	**24.92**	28.12
DenseNet201	75.92	23.45	28.23	23.89	**31.20**

TABLE 9. Model performance for our proposed ensemble attacks with and without adversarial defense (A.D.) on SVHN (Lower acc. % indicates stronger attack).

Model	Baseline(Acc %)	Accuracy (%)			
		MEA without A.D.	MEA with A.D.	WEA without A.D.	WEA with A.D.
ResNet152	**97.02**	33.12	**35.10**	**31.32**	**35.81**
DenseNet201	96.90	**33.23**	35.05	29.89	31.90

attack strategies. These two datasets capture real-world complexities and are widely used for testing adversarial attacks, providing a comprehensive assessment of our attack methods.

The extended analysis in Tables 8 and 9 highlights the impact of MEA and WEA attacks on DNN accuracy for CIFAR-100 and SVHN datasets. Without adversarial defense (A.D.), these attacks significantly degrade accuracy (e.g., 25.67% and 24.92% for ResNet152 on CIFAR-100). Although A.D. slightly improves accuracy (e.g., MEA to 33. 35% and WEA to 28.12%), it struggles to effectively mitigate attack impacts. Similar patterns are observed on SVHN, where A.D. provides minimal progress in DNN accuracy under both attack types, emphasizing the challenges in defending against such strong adversarial attacks.

D. RESILIENCE OF WEA AND MEA AGAINST GRADIENT OBFUSCATION

To assess our ensemble models, we use another popular adversarial defense mechanism, namely Gradient Masking(G.M.). This method obfuscates gradient information, which is vital for adversarial attack generation, effectively limiting the attacker's ability to craft optimized perturbations [42]. By disrupting the direct exploitation of gradients, it works as a strong barrier against gradient-based attacks, improving the resilience of deep learning models [42]. However, the results in Table 10 indicate that, even with the application of gradient masking, our ensemble methods maintain their effectiveness in significantly reducing model accuracy. For both ResNet152 and DenseNet201, MEA and WEA without G.M. lead to considerable accuracy drops, showcasing their strength as adversarial attacks. Even after using gradient masking, both ensemble attacks continue to achieve notable reductions in accuracy. For instance, WEA with G.M. achieves an accuracy of 29.32% on ResNet152, compared to its baseline of 78.21%. Similarly, MEA with G.M. maintains its adversarial impact, reducing DenseNet201's accuracy to 29.82%, compared to its baseline of 75.92%. These results highlight the robustness and effectiveness of our proposed methods, as they remain impactful even in the presence of strong defense mechanisms like gradient masking.

From the experiment, we find that the discrepancy in effectiveness between WEA and MEA arises from multiple factors beyond the greater weight assigned to FGSM in WEA. While the 0.4 weight for FGSM provides the representation of single-step attacks to reveal fundamental vulnerabilities, its weaker perturbations reduce the overall adversarial intensity compared to iterative attacks like PGD and BIM. Also, iterative attacks often generate aligned gradients, reinforcing their perturbations, whereas FGSM's single-step gradient can diverge, introducing conflicts that reduce the combined adversarial strength. In MEA, equal weighting allows iterative attacks to dominate, leading to greater adversarial effectiveness. We find another factor is the sensitivity of the target model to different types of perturbations. Models trained with adversarial defenses often exhibit gradient masking, making them less vulnerable to FGSM-like attacks [43]. The increased reliance on FGSM in WEA might thus contribute less to the ensemble's overall effectiveness against such models. In contrast, MEA benefits from a higher combined contribution of stronger, iterative attacks that can bypass gradient masking [44]. This highlights the trade-off in WEA's design, balancing attack contributions but potentially limiting overall impact.

E. LIMITATIONS OF SURROGATE MODEL COMPARISONS IN ENSEMBLE ATTACKS

Our proposed ensemble adversarial attacks, which combine three different attack methods, show a substantial degradation in DNN performance across benchmark datasets such as Tiny ImageNet, CIFAR-10, CIFAR-100, and SVHN. The results highlight that even SOTA defensive mechanisms, like the defense distillation-enhanced DenseNet201, are inadequate to mitigate the compounded adversarial impact of our ensemble methods. However, our approach fundamentally diverges from existing popular ensemble attack methods like Magnitude-Agnostic Bagging Ensemble (MABE) [27], Transferable Adversarial Perturbation (TAP) [45], and Query-Efficient Black-Box Ensemble (QE-BBE) [46]. All these strategies primarily target black-box or gray-box

TABLE 10. Model performance for our proposed ensemble attacks with and without gradient masking (G.M.) on CIFAR-100 (Lower acc. % indicates stronger attack).

Model	Baseline Acc(%)	Accuracy (%)			
		MEA without G.M.	MEA with G.M.	WEA without G.M.	WEA with G.M.
ResNet152	**78.21**	**25.67**	**29.91**	**24.92**	**29.32**
DenseNet201	75.92	23.45	29.82	23.89	28.01

scenarios, concentrating on creating adversarial perturbations that generalize across multiple surrogate models or datasets with limited access to the target model. For instance, MABE emphasizes perturbation diversity without explicit knowledge of model gradients [27], while TAP leverages transferability to craft perturbations that are effective across different architectures [45].

In contrast, our work precisely focuses on white-box scenarios to provide a direct evaluation of ensemble strategies under optimal conditions. Unlike prior methods that rely on approximate model knowledge or indirect methods, WEA and MEA fully exploit gradient-level access to maximize the adversarial impact. This allows for precise benchmarking of DNN vulnerabilities in controlled environments, which is not feasible with black-box or transfer-based approaches. Additionally, our ensemble attack incorporates various individual attacks synergistically, prioritizing maximal degradation of accuracy rather than perturbation generalization. Thus, the reliance of existing ensemble techniques on transferability and query efficiency makes them fundamentally incomparable to our approach, as they operate under entirely different assumptions and objectives.

V. CONCLUSION

This study examines the performance of deep neural network based machine learning models for image classification against adversarial attacks and proposes a new ensemble adversarial attack method. Firstly, we consider models that have not undergone adversarial training, and evaluate their robustness against four different types of adversarial attacks. Then we propose two ensemble methods that integrating three different attacks to form a single, more complex adversarial example, and then evaluate the models' performance under the proposed ensemble adversarial examples. Our results indicate that DenseNet201 and ResNet152 are extremely robust, outperforming other models even when confronted with these sophisticated ensemble attacks. This is due to their complex structures, which can capture patterns with more detail. To further evaluate the proposed ensemble attack, we employed defensive distillation and gradient masking methods to determine if this defensive strategy is effective against individual attacks as well as the proposed ensemble attacks. Our results indicate that the proposed ensemble attacks deeply compromises the DNNs' performance, even after adversarial training and the distillation process, and their impacts are more damaging to DNN models comparing to individual attacks.

For future work, we believe several extensions and advancements to improve the robustness and resilience of DNN models under adversarial conditions. First, we can analyze additional ensemble-based adversarial strategies applying more complex, multi-step attacks to gain deeper insights into model vulnerabilities. We can also explore the impacts of hybrid attacks, combining aspects of both white-box and black-box methods, to broaden our understanding of model constraints. Finally, we can test our approach on real-world, high-dimensional datasets and use our methods in other fields, such as autonomous driving and medical imaging, to validate the generalizability of our findings and strengthen DNN models across various applications.

ACKNOWLEDGMENT

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] M. N. Q. Bhuiyan, S. K. Rahut, R. A. Tanvir, and S. Ripon, "Automatic Acute Lymphoblastic Leukemia detection and comparative analysis from images," in *Proc. 6th Int. Conf. Control, Decis. Inf. Technol. (CoDIT)*, Apr. 2019, pp. 1144–1149.

[2] C. Macrae, "Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk," *Risk Anal.*, vol. 42, no. 9, pp. 1999–2025, Sep. 2022.

[3] E. Candela, Y. Feng, D. Mead, Y. Demiris, and P. Angeloudis, "Fast collision prediction for autonomous vehicles using a stochastic dynamics model," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 211–216.

[4] J. Chen and X. Ran, "Deep learning with edge computing: A review," *Proc. IEEE*, vol. 107, no. 8, pp. 1655–1674, Aug. 2019.

[5] F. Wang, R. Song, Q. Li, and C.-G. Wang, "Efficient black-box adversarial attacks with training surrogate models towards speaker recognition systems," in *Proc. Int. Conf. Algorithms Archit. Parallel Process.*, Jan. 2024, pp. 257–276.

[6] X. Wu, S. Ma, C. Shen, C. Lin, Q. Wang, Q. Li, and Y. Rao, "KENKU: Towards efficient and stealthy black-box adversarial attacks against ASR systems," in *Proc. 32nd USENIX Secur. Symp. (USENIX Secur.)*, 2023, pp. 247–264.

[7] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[8] T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *Proc. AAAI Conf. Artif. Intell.*, Jul. 2019, vol. 33, no. 1, pp. 2253–2260.

[9] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2574–2582.

[10] A. U. H. Qureshi, H. Larijani, M. Yousefi, A. Adeel, and N. Mtetwa, "An adversarial approach for intrusion detection systems using Jacobian saliency map attacks (JSMA) algorithm," *Computers*, vol. 9, no. 3, p. 58, Jul. 2020.

[11] Z. Yao, A. Gholami, P. Xu, K. Keutzer, and M. W. Mahoney, "Trust region based adversarial attack on neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11350–11359.

[12] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting adversarial image examples in deep neural networks with adaptive noise reduction," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 1, pp. 72–85, Jan. 2021.

[13] K. Ren, T. Zheng, Z. Qin, and X. Liu, "Adversarial attacks and defenses in deep learning," *Engineering*, vol. 6, no. 3, pp. 346–360, Mar. 2020.

[14] Hugging Face. (2023). *Zh-plus/tiny-imagenet · Datasets at Hugging Face*. [Online]. Available: https://huggingface.co/datasets/zh-plus/tiny-imagenet

[15] S. Yaghoubi and G. Fainekos, "Gray-box adversarial testing for control systems with machine learning components," in *Proc. 22nd ACM Int. Conf. Hybrid Sys., Comput. Control*, Apr. 2019, pp. 179–184.

[16] E. Shayegani, M. A. A. Mamun, Y. Fu, P. Zaree, Y. Dong, and N. Abu-Ghazaleh, "Survey of vulnerabilities in large language models revealed by adversarial attacks," 2023, *arXiv:2310.10844*.

[17] J. Yang, Y. Jiang, X. Huang, B. Ni, and C. Zhao, "Learning black-box attackers with transferable priors and query feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2020, pp. 12288–12299.

[18] A. T. W. Min, Y.-S. Ong, A. Gupta, and C.-K. Goh, "Multiproblem surrogates: Transfer evolutionary multiobjective optimization of computationally expensive problems," *IEEE Trans. Evol. Comput.*, vol. 23, no. 1, pp. 15–28, Feb. 2019.

[19] Z. Huang and T. Zhang, "Black-box adversarial attack with transferable model-based embedding," 2019, *arXiv:1911.07140*.

[20] J. Hang, K. Han, H. Chen, and Y. Li, "Ensemble adversarial black-box attacks against deep learning systems," *Pattern Recognit.*, vol. 101, May 2020, Art. no. 107184.

[21] A. MaungMaung and H. Kiya, "Ensemble of key-based models: Defense against black-box adversarial attacks," in *Proc. IEEE 10th Global Conf. Consum. Electron. (GCCE)*, Oct. 2021, pp. 95–98.

[22] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.

[23] Y. Bai, Y. Wang, Y. Zeng, Y. Jiang, and S.-T. Xia, "Query efficient black-box adversarial attack on deep neural networks," *Pattern Recognit.*, vol. 133, Jan. 2023, Art. no. 109037.

[24] P. L. M. Doss and M. Gunasekaran, "Securing ResNet50 against adversarial attacks: Evasion and defense using BIM algorithm," in *Proc. 7th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2023, pp. 1381–1386.

[25] L. Beerens and D. J. Higham, "Componentwise adversarial attacks," in *Proc. Int. Conf. Artif. Neural Netw.* Cham, Switzerland: Springer, Jan. 2023, pp. 542–545.

[26] J. Jung, N. Akhtar, and G. Hassan, "Analysing adversarial examples for deep learning," in *Proc. 16th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2021, pp. 585–592.

[27] Z. He, W. Wang, J. Dong, and T. Tan, "Revisiting ensemble adversarial attack," *Signal Process., Image Commun.*, vol. 107, Sep. 2022, Art. no. 116747.

[28] Z. Fu and X. Cui, "ELAA: An ensemble-learning-based adversarial attack targeting image-classification model," *Entropy*, vol. 25, no. 2, p. 215, Jan. 2023.

[29] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–22.

[30] Z. Che, A. Borji, G. Zhai, S. Ling, J. Li, X. Min, G. Guo, and P. Le Callet, "SMGEA: A new ensemble adversarial attack powered by long-term gradient memories," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 1051–1065, Mar. 2022.

[31] Y. Xiong, J. Lin, M. Zhang, J. E. Hopcroft, and K. He, "Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14963–14972.

[32] M. Pang, B. Wang, M. Ye, Y.-M. Cheung, Y. Zhou, W. Huang, and B. Wen, "Heterogeneous prototype learning from contaminated faces across domains via disentangling latent factors," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 1, 2024, doi: 10.1109/TNNLS.2024.3393072.

[33] J. Li, M. Pang, Y. Dong, J. Jia, and B. Wang, "Graph neural network explanations are fragile," 2024, *arXiv:2406.03193*.

[34] B. Wang, M. Pang, and Y. Dong, "Turning strengths into weaknesses: A certified robustness inspired attack framework against graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16394–16403.

[35] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[36] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proc. 4th ACM Workshop Secur. Artif. Intell.*, Oct. 2011, pp. 43–58.

[37] E. Yilmaz and H. Y. Keles, "Adversarial sparse teacher: Defense against distillation-based model stealing attacks using adversarial examples," 2024, *arXiv:2403.05181*.

[38] R. Mehta and K. Jhajharia, "Layered distillation training: A study of adversarial attacks and defenses," in *Proc. 3rd Int. Conf. Innov. Technol. (INOCON)*, Mar. 2024, pp. 1–7.

[39] J. F. Nash Jr., "Equilibrium points in *n*-person games," *Proc. Nat. Acad. Sci. USA*, vol. 36, no. 1, pp. 48–49, 1950.

[40] L. Dritsoula, P. Loiseau, and J. Musacchio, "A game-theoretic analysis of adversarial classification," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 12, pp. 3094–3109, Dec. 2017.

[41] M.-F. Balcan, R. Pukdee, P. Ravikumar, and H. Zhang, "Nash equilibria and pitfalls of adversarial training in adversarial robustness games," in *Proc. Int. Conf. Artif. Intell. Statist.*, Jan. 2022, pp. 9607–9636.

[42] H. Lee, H. Bae, and S. Yoon, "Gradient masking of label smoothing in adversarial robustness," *IEEE Access*, vol. 9, pp. 6453–6464, 2021.

[43] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155161–155196, 2021.

[44] M. Naseer, S. Khan, and F. Porikli, "Local gradients smoothing: Defense against localized adversarial attacks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1300–1307.

[45] W. Zhou, X. Hou, Y. Chen, M. Tang, X. Huang, X. Gan, and Y. Yang, "Transferable adversarial perturbations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2018, pp. 452–467.

[46] Y. Dong, S. Cheng, T. Pang, H. Su, and J. Zhu, "Query-efficient black-box adversarial attacks guided by a transfer-based prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9536–9548, Dec. 2022.

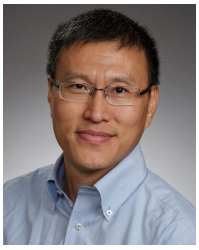
MAFIZUR RAHMAN received the B.Sc. degree in computer science from East West University, Dhaka, Bangladesh. He is currently pursuing the master's degree in computer science with Prairie View A&M University, Prairie View, TX, USA. He is also a Graduate Research Assistant with the CREDIT Center, Prairie View A&M University. To date, he has published over 15 papers in prominent journals and conferences, primarily in the fields of machine learning and deep learning. His current research interests include computer vision, neuromorphic computing, and deep learning.

PROSENJIT ROY received the bachelor's degree in computer science and engineering from Daffodil International University, Bangladesh, in 2019. He is currently pursuing the master's degree in computer science with Prairie View A&M University (PVAMU).

He is interested in computing biology, machine learning, and genomics analyzing. His main job is to look at biology data to find genetic and molecular associations. As a Graduate Research Assistant with PVAMU, he creates algorithms to handle big amounts of biological data, which helps genomics and proteomics make progress. Earlier, he was a Senior Assistant IT Officer with Bangladesh's Aayan Plastic Industry. He was in charge of the IT plan and the network systems. He made machine learning models and did complicated data analysis for Silicon Orchard Software Solution as part of his job. He has involved in several significant projects, including creating a model that combines BERT and GloVe to help find writers, finding fake reviews on Amazon and Facebook, and building advanced search tools for business databases.

SHERRI S. FRIZELL received the B.S. degree from Jackson State University, and the M.C.S.E. and Ph.D. degrees from Auburn University, USA.

She is currently a Professor and the Associate Department Head of the Department of Computer Science, Prairie View A&M University (PVAMU), Prairie View, TX, USA. Her research interests include human–computer interaction, educational technology, social computing, and engineering education, with a focus on health technology applications and multimodal user interfaces. She also has expertise in instructional design and factors influencing the success of African-American students in STEM. She has significant non-academic experience, having worked as a Pre-Professional Engineer with the IBM T.J. Watson Research Center, Yorktown Heights, NY, USA, and a Computer Scientist with the National Security Agency. She is actively involved in promoting the academic and career success of women in STEM, serving as an Advisor to the PVAMU Chapter for the Society of Women Engineers (SWE).



LIJUN QIAN (Senior Member, IEEE) received the B.S. degree from Tsinghua University, Beijing, China, the M.S. degree from the Technion—Israel Institute of Technology, Haifa, Israel, and the Ph.D. degree from Rutgers University, USA.

He was a Member of Technical Staff of Bell-Labs Research, Murray Hill, NJ, USA. He was a Visiting Professor with Aalto University, Finland. He is currently a Regents Professor and holds the AT&T Endowment with the Department of Electrical and Computer Engineering, Prairie View A&M University (PVAMU), Prairie View, TX, USA. He is also the Founder and the Director of the Center of Excellence in Research and Education for Big Military Data Intelligence (CREDIT Center). His research interests include artificial intelligence, machine learning, big data analytics, wireless communications and mobile networks, network security, and computational systems biology.