Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 14, No. 4, 2026, 465-486.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

# A Hybrid Ant Colony Optimization and Reinforcement Learning Framework for Enhancing Neural Network Robustness against Adversarial Attacks

Alireza Omidi Nasab[1] and Sajad Bastami[2] and Rojyar Pirmohammadian[2*] and Mohammad Bagher Dowlatshahi[1] and Seyedeh Zahra Mousavi[1]

*1. Computer Engineering Department, Lorestan University, Khorramabad, Iran.*
*2. Computer Engineering Department, Kurdistan University, Sanandaj, Iran.*

**Article Info**

**Article History:**
*Received 24 April 2026*
*Revised 07 July 2026*
*Accepted 15 July 2026*

*DOI:10.22044/jadm.2026.17635.2917*

**Keywords:**
*Adversarial Robustness, Hybrid Computational Intelligence, Ant Colony Optimization, Reinforcement Learning, Multimodal Defense, Cross-Domain Learning.*

**Corresponding author: r.pirmohamadiani@uok.ac.ir (R. Pirmohammadian).*

**Abstract**

Deep Neural Networks (DNNs) are increasingly deployed in safety-critical domains such as autonomous driving, healthcare, finance, and natural language processing, yet they remain vulnerable to adversarial attacks—subtle manipulations that can cause confident misclassifications or misleading predictions. This fragility poses a major barrier to building secure and trustworthy AI systems. Conventional defenses, including adversarial training and heuristic detection, often struggle to balance robustness, adaptability, and computational cost. To overcome these limitations, we propose a hybrid adaptive defense framework that unifies Ant Colony Optimization (ACO) with Reinforcement Learning (RL). ACO efficiently explores the high-dimensional space of defense hyperparameters to find globally optimal configurations, while RL enables dynamic, context-aware adaptation of defense strategies in real time. The proposed ACO-RL framework was rigorously evaluated across six diverse benchmark datasets spanning multiple data modalities: MNIST and CIFAR-10 (vision), IMDB and AG News (text), and Cora and Reddit-Binary (graph). Experimental results show that ACO-RL consistently enhances robustness against a wide spectrum of adversarial attacks, outperforming several state-of-the-art baselines. These findings highlight a promising pathway toward developing resilient, cross-domain AI systems capable of defending against evolving adversarial threats.

## 1. Introduction

The rapid advancement of Deep Neural Networks (DNNs) has revolutionized artificial intelligence, enabling unprecedented performance across diverse domains including computer vision, natural language understanding, and graph-based reasoning. These systems now underpin critical applications in autonomous vehicles, medical diagnosis, financial fraud detection, and content moderation—contexts where errors can have severe real-world consequences [1, 2]. However, despite their remarkable capabilities, DNNs exhibit a troubling vulnerability: they can be easily fooled by adversarial examples—imperceptibly perturbed inputs crafted to induce erroneous predictions with high confidence [3, 4]. This fragility presents a fundamental challenge to deploying AI systems in safety-critical environments, where robustness and reliability are not merely desirable but essential. Adversarial attacks have evolved significantly since their initial discovery, ranging from gradient-based white-box attacks such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) to more sophisticated black-box and transfer-based attacks [5, 6]. Recent research

has demonstrated that adversarial perturbations can transfer across models and even across data modalities, suggesting that this vulnerability stems from fundamental properties of high-dimensional learning rather than implementation flaws [7, 8]. Moreover, adversarial threats extend beyond image classification to affect natural language processing systems through word substitutions and semantic manipulations, as well as graph neural networks through structural perturbations [9, 10]. This cross-domain persistence of adversarial vulnerabilities underscores the urgent need for defense mechanisms that generalize across diverse data types and application contexts.

Current defense strategies can be broadly categorized into preventive and reactive approaches. Adversarial training, which augments the training dataset with adversarial examples, remains the most widely adopted preventive method [11, 12]. While effective against known attacks, adversarial training suffers from several critical limitations: it incurs substantial computational overhead, often requiring 5–10 times more training time; it exhibits poor generalization to unseen attack types; and it frequently leads to accuracy-robustness trade-offs on clean data [13, 14]. Reactive defenses, including input transformation, adversarial detection, and certified robustness methods, offer alternative pathways but come with their own challenges [15, 16]. Input preprocessing techniques such as JPEG compression or bit-depth reduction can be circumvented by adaptive attacks [17]. Adversarial detectors struggle with high false positive rates and vulnerability to detection-aware adversaries [18]. Certified defense methods provide mathematical guarantees but often achieve only modest robustness at significant computational cost [19, 20]. The limitations of single-strategy defenses have motivated researchers to explore hybrid and ensemble approaches that combine multiple defense mechanisms [21, 22]. However, most existing hybrid methods rely on hand-crafted combinations with fixed hyperparameters, failing to adapt to evolving attack strategies or varying operational contexts [23]. This rigidity poses a significant weakness, as adversaries can exploit knowledge of static defense configurations to design adaptive attacks specifically tailored to bypass them [24, 25]. What is needed, therefore, is an intelligent defense framework capable of dynamically optimizing its configuration and adapting its strategy in response to the specific characteristics of incoming threats—essentially, a defense system that can "learn to defend" rather than simply applying pre-programmed rules.

Nature-inspired optimization algorithms and reinforcement learning represent two powerful paradigms that, when thoughtfully combined, can address these challenges. Ant Colony Optimization (ACO), inspired by the foraging behavior of ant colonies, has demonstrated exceptional performance in solving complex combinatorial optimization problems across diverse domains [26, 27]. ACO's population-based search mechanism enables effective exploration of vast solution spaces while avoiding premature convergence to local optima—a crucial advantage when optimizing the high-dimensional hyperparameter landscape of defense mechanisms [28]. Conversely, Reinforcement Learning (RL) excels at learning adaptive policies through interaction with dynamic environments, making it well-suited for real-time decision-making in the face of evolving adversarial strategies [29, 30]. Recent advances in deep RL, including policy gradient methods and actor-critic architectures, have shown promise in adversarial robustness applications [31, 32]. Despite the individual strengths of ACO and RL, few studies have systematically investigated their synergistic integration for adversarial defense. ACO's global search capabilities can identify promising regions of the defense configuration space, but it lacks the ability to adapt dynamically to sequential decision-making scenarios. Conversely, RL agents can learn adaptive defense policies but often struggle with high-dimensional action spaces and require careful reward shaping [33, 34]. We hypothesize that combining these complementary strengths—using ACO to optimize the structural components and hyperparameters of defense mechanisms, while leveraging RL to enable real-time adaptive responses—can yield a defense framework that is simultaneously robust, efficient, and generalizable across data modalities.

In this work, we propose a novel hybrid ACO-RL defense framework that addresses the limitations of existing approaches through intelligent optimization and adaptive decision-making. Our framework operates in two integrated phases: an offline optimization phase where ACO explores the space of defense configurations to identify globally optimal hyperparameter settings, and an online adaptation phase where an RL agent dynamically adjusts defense strategies based on observed attack characteristics and model confidence patterns. Specifically, ACO optimizes critical defense parameters including adversarial training batch composition, input transformation sequences, and detection thresholds, while the RL agent learns to select and configure defense mechanisms in real time based on the current threat landscape. This

hybrid architecture enables our framework to achieve both long-term optimality through ACO's global search and short-term adaptability through RL's reactive learning. We validate the effectiveness of our proposed ACO-RL framework through comprehensive experiments across six benchmark datasets spanning three distinct data modalities: image data (MNIST, CIFAR-10), text data (IMDB, AG News), and graph data (Cora, Reddit-Binary). This multi-domain evaluation strategy is crucial for demonstrating the generalizability of our approach, as cross-modal robustness remains a significant challenge in adversarial defense research [35, 36]. Our experimental protocol includes evaluation against diverse attack types—gradient-based attacks (FGSM, PGD, C&W), modality-specific attacks (TextFooler for NLP, Metattack for graphs), and transfer-based attacks—under both white-box and black-box threat models. We compare our ACO-RL framework against several state-of-the-art baselines, including vanilla adversarial training, ensemble defenses, adversarial detection methods, and recent hybrid approaches.

The experimental results demonstrate that our ACO-RL framework achieves substantial improvements in adversarial robustness across all evaluated datasets and attack scenarios, with robust accuracy gains ranging from 8% to 15% compared to baseline methods. Notably, our approach maintains strong performance on clean data, addressing the common accuracy-robustness trade-off that plagues many defense mechanisms. Ablation studies reveal that both the ACO optimization and RL adaptation components contribute meaningfully to overall robustness, with their combination yielding synergistic benefits beyond what either approach achieves independently. Furthermore, our framework demonstrates encouraging transferability, with defense configurations learned on one dataset showing non-trivial effectiveness when applied to related domains. The primary contributions of this research can be summarized as follows:

- Novel Hybrid Framework. We introduce the first systematic integration of Ant Colony Optimization and Reinforcement Learning for adversarial defense, combining global optimization with adaptive decision-making to create a defense system that is both strategically optimized and tactically responsive.
- Cross-Modal Validation. We provide comprehensive empirical validation across vision, language, and graph domains, demonstrating the broad applicability and generalizability of our approach across diverse data modalities and neural network architectures.
- Dynamic Adaptation Mechanism. Our RL-based adaptation module enables real-time defense strategy selection based on observed attack characteristics, moving beyond static defense configurations toward intelligent, context-aware security.
- Practical Efficiency. Despite combining two computationally intensive techniques, our framework achieves practical efficiency through careful design choices, including offline ACO optimization and lightweight RL inference, making it viable for real-world deployment.
- Comprehensive Evaluation Protocol. We establish a rigorous evaluation framework that assesses defense effectiveness against diverse attack types, measures clean accuracy preservation, analyzes computational overhead, and investigates transferability across domains.

The remainder of this paper is organized as follows: Section 2 reviews related work in adversarial attacks, defense mechanisms, and hybrid optimization approaches. Section 3 presents our proposed ACO-RL framework, detailing the architecture, optimization process, and adaptation mechanism. Section 4 describes our experimental methodology, including datasets, attack configurations, baseline methods, and evaluation metrics. Section 5 presents and analyzes our experimental results, including main findings, ablation studies, and transferability analysis. Finally, Section 6 concludes the paper with a summary of contributions and broader impact considerations.

## 2. Related Works

The pursuit of adversarial robustness in deep learning has spawned a rich and rapidly evolving research landscape. In this section, we provide a comprehensive review of existing work organized into several thematic categories: adversarial attack methodologies, defense mechanisms, optimization-based approaches to robustness, reinforcement learning applications in adversarial settings, and hybrid defense strategies. We critically examine the strengths and limitations of prior approaches to contextualize our contributions and highlight the research gaps that motivate our work.

### 2.1. Adversarial Attack Methodologies

The discovery of adversarial examples by Szegedy et al. [37] revealed a fundamental vulnerability in deep neural networks: imperceptible perturbations to inputs can cause confident misclassifications. This seminal finding catalyzed extensive research into understanding and exploiting this phenomenon. Goodfellow et al. [3] introduced the Fast Gradient Sign Method (FGSM), a simple yet

effective one-step attack that computes perturbations by taking the sign of the loss gradient. While computationally efficient, FGSM often produces suboptimal perturbations due to its linearization assumption. Subsequent research developed more sophisticated iterative attacks that achieve higher success rates through multi-step optimization. The Basic Iterative Method (BIM) and Projected Gradient Descent (PGD) [5] iteratively refine perturbations while projecting them back into a constrained perturbation budget, effectively solving a constrained optimization problem. Madry et al. [5] demonstrated that PGD represents one of the strongest first-order adversaries, establishing it as a de facto standard for evaluating robustness. Carlini and Wagner [38] proposed the C&W attack, which reformulates the adversarial objective to balance attack strength with perturbation magnitude through a carefully designed loss function. This attack remains among the most powerful white-box methods, though at significant computational cost. Beyond gradient-based attacks, researchers have explored alternative attack paradigms. Transfer-based attacks [7] exploit the observation that adversarial examples crafted for one model often transfer to other models, even those with different architectures or training procedures. This transferability poses serious security risks in black-box scenarios where attackers lack direct access to target models. Dong et al. [6] enhanced transfer attacks by introducing momentum into the gradient computation, stabilizing the optimization process and improving cross-model transferability. Score-based black-box attacks [39] query the target model to estimate gradients or decision boundaries, while decision-based attacks [40] require only the final predicted label, representing the most restrictive threat model. The adversarial threat landscape extends beyond image classification to encompass diverse data modalities. In natural language processing, adversarial attacks manipulate text through word substitutions, character-level perturbations, or sentence paraphrasing [9, 41]. TextFooler [42] generates adversarial text by replacing words with semantically similar alternatives that preserve grammatical correctness while fooling NLP models. For graph-structured data, attacks modify graph topology or node features to mislead graph neural networks [10, 43]. Metattack [44] and other graph adversarial methods exploit the inherent sparsity and discrete nature of graph structures, presenting unique challenges distinct from continuous perturbations in images or text.

### 2.2. Defense Mechanisms

The arms race between attacks and defenses has produced numerous defensive strategies, which can be broadly categorized into preventive and reactive approaches. Adversarial training, first systematically explored by Goodfellow et al. [3] and later refined by Madry et al. [5], remains the most effective and widely adopted defense mechanism. The core principle is intuitive: augment the training dataset with adversarial examples generated on-the-fly, forcing the model to learn robust decision boundaries that resist perturbations. The min-max formulation of adversarial training seeks parameters that minimize loss on the worst-case adversarial perturbations within a specified threat model [5]. Despite its effectiveness, adversarial training suffers from several well-documented limitations. First, it incurs substantial computational overhead, often requiring 5-10 times more training time than standard training due to the need to generate adversarial examples at each iteration [12]. Second, models trained adversarially often exhibit reduced accuracy on clean (unperturbed) data, a phenomenon known as the robust accuracy trade-off [20, 45]. Third, adversarial training tends to overfit to the specific attack used during training, showing limited generalization to unseen attack types [13, 24]. Recent work has sought to address these limitations through various refinements, including improved data augmentation strategies [12], curriculum learning approaches that gradually increase attack strength [46], and techniques to mitigate robust overfitting [13]. Ensemble adversarial training [47] combines adversarial examples from multiple source models to improve robustness against transfer attacks. Zhang et al. [14] demonstrated that incorporating strategically chosen "failed" adversarial examples—those that do not successfully fool the current model—can paradoxically strengthen robustness. Pang et al. [21] compiled a comprehensive "bag of tricks" for adversarial training, showing that careful hyperparameter tuning, learning rate scheduling, and regularization techniques significantly impact final robustness. An alternative defensive philosophy focuses on detecting adversarial examples rather than correctly classifying them. Detection methods analyze input characteristics, model behavior, or hidden representations to distinguish adversarial from benign examples [48, 49]. Feature squeezing [15] reduces input complexity through bit-depth reduction or spatial smoothing, then compares predictions on original and squeezed inputs—large discrepancies suggest adversarial manipulation. Statistical tests on neural

network activations [50] or output distributions [51] can also reveal adversarial perturbations. However, detection-based defenses face fundamental challenges. Carlini and Wagner [18] demonstrated that many proposed detection methods can be circumvented through detection-aware adaptive attacks that explicitly minimize detectability. The detectability of adversarial examples remains an open question, with some evidence suggesting that sufficiently small perturbations may be fundamentally indistinguishable from benign noise [52]. Furthermore, detection methods often suffer from high false positive rates, which can impair system usability in practical deployments [53]. Input transformation defenses attempt to remove or neutralize adversarial perturbations before they reach the classifier. Common approaches include JPEG compression [54], bit-depth reduction [15], spatial smoothing [55], and denoising autoencoders [56]. These methods exploit the observation that adversarial perturbations often lie in high-frequency components or imperceptible dimensions that can be filtered out. The Achilles' heel of transformation-based defenses is their vulnerability to adaptive attacks designed with knowledge of the transformation [17]. Athalye et al. [17] showed that obfuscated gradients—situations where defenses inadvertently hide rather than eliminate adversarial perturbations—create a false sense of security that crumbles against properly adapted attacks. Backward Pass Differentiable Approximation (BPDA) [17] allows attackers to compute gradients through non-differentiable transformations, effectively bypassing many preprocessing defenses. Certified defense methods provide provable guarantees of robustness within a specified threat model, representing a theoretically principled approach to adversarial robustness [57, 58]. Randomized smoothing [16, 19] constructs a smoothed classifier by averaging predictions over random noise, then derives certification radii within which no adversarial example can exist. Interval bound propagation and abstract interpretation techniques [59, 60] compute guaranteed bounds on neural network outputs under input perturbations. While intellectually satisfying, certified defenses face practical challenges. Current certification methods typically achieve only modest certified radii—often insufficient for meaningful security guarantees—and the computational cost of certification grows rapidly with network size and desired robustness level [61]. Moreover, certification is typically limited to $\ell_2$ or $\ell\infty$ norm-bounded perturbations and does not readily extend to more complex threat models or diverse data modalities [62].

### 2.3. Optimization-Based Approaches to Robustness

Nature-inspired optimization algorithms have found increasing application in adversarial robustness research, although their adoption remains less widespread than gradient-based methods. Genetic algorithms [63] and particle swarm optimization [64] have been employed to generate adversarial examples and optimize defense parameters. These population-based methods can explore non-convex solution spaces more effectively than gradient descent in certain scenarios. Ant Colony Optimization, inspired by the collective foraging behavior of ant colonies, has demonstrated success in various machine learning contexts [27, 65]. ACO's key strength lies in its ability to balance exploration and exploitation through pheromone-based communication, making it well-suited for high-dimensional combinatorial optimization problems. Recent applications of ACO in neural network training [66] and hyperparameter tuning [27] suggest its potential for optimizing defense configurations, although direct applications to adversarial robustness remain limited. Bayesian optimization [67] and AutoML techniques [68] have also been applied to optimize defense mechanisms by treating robustness as a black-box objective function. These methods can efficiently search high-dimensional hyperparameter spaces but typically require numerous evaluations and may not scale to real-time adaptation scenarios.

### 2.4. Reinforcement Learning in Adversarial Settings

Reinforcement learning has emerged as a promising paradigm for both attacking and defending neural networks. From an adversarial perspective, RL can generate diverse attacks by framing perturbation generation as a sequential decision-making problem [69]. Attacking agents learn policies that maximize misclassification probability while minimizing perturbation magnitude, often discovering novel attack strategies not captured by gradient-based methods [70]. On the defensive side, RL enables adaptive defense policies that respond dynamically to observed threats. Behzadan and Munir [31] explored RL-based defense strategies that learn to detect and counter adversarial perturbations through interaction with adversarial environments. Meta-learning approaches [71] train defense agents across diverse attack distributions, promoting

generalization to unseen attacks. Policy gradient methods and actor-critic architectures [72] have shown particular promise for learning robust defense strategies. However, applying RL to adversarial robustness faces several challenges. The high-dimensional action space of defense configurations can make learning sample-inefficient [34]. Reward function design proves critical yet non-trivial—balancing clean accuracy, robust accuracy, and computational efficiency requires careful engineering [73]. Furthermore, RL agents themselves can be vulnerable to adversarial manipulation [31, 74], necessitating robust training procedures.

### 2.5. Hybrid and Ensemble Defense Strategies

Recognizing the limitations of single-strategy defenses, recent research has explored hybrid approaches that combine multiple defensive mechanisms [75, 76]. Ensemble defenses aggregate predictions from multiple models, exploiting diversity to improve robustness [77, 78]. Diverse architectures, training procedures, or adversarial training regimes can be combined to create ensembles that are collectively more robust than individual models. Strauss et al. [79] proposed adaptive defense frameworks that select defense strategies based on input characteristics or threat indicators. These systems monitor incoming data for signs of adversarial manipulation and dynamically adjust their defensive posture accordingly. Multi-stage defense pipelines [80] combine detection, transformation, and robust classification in sequence, creating defense-in-depth architectures. Despite these advances, most hybrid defenses rely on hand-crafted combinations with static configurations. They lack the ability to optimize defense parameters globally or adapt in real-time to evolving threats. Furthermore, the interaction effects between different defense components remain poorly understood, with potential for both synergistic benefits and detrimental interference [81].

### 2.5. Cross-Domain and Multi-Modal Robustness

Most adversarial robustness research focuses on single domains—typically image classification on benchmark datasets like MNIST or CIFAR-10. However, modern AI systems increasingly operate across multiple modalities and domains [82, 83]. Recent work has begun investigating cross-modal transferability of adversarial examples [7] and developing defenses that generalize across data types [35, 84]. Multi-modal learning presents unique challenges for adversarial robustness. Attacks can target individual modalities or exploit cross-modal interactions [85]. Defenses must balance robustness across all input modalities while maintaining cross-modal alignment [36]. Vision-language models like CLIP [86] have shown particular vulnerability to adversarial attacks that exploit the vision-text interface. Graph neural networks introduce additional considerations due to the discrete, structured nature of graph data [10, 87]. Unlike pixel-level perturbations in images, graph attacks manipulate topology through edge additions or deletions, creating discrete optimization challenges [43, 44]. Defending graph-based models requires approaches that respect graph structure and preserve semantic meaning [88].

Despite the substantial progress documented across the individual research streams reviewed above, several critical gaps remain unaddressed. First, while adversarial attacks and defenses have been extensively studied, most existing defense mechanisms are designed for single data modalities and fail to generalize across vision, text, and graph domains. Second, although optimization-based approaches such as genetic algorithms, particle swarm optimization, and Bayesian optimization have been applied to adversarial robustness, the specific potential of Ant Colony Optimization (ACO) for optimizing defense configurations remains largely unexplored—despite ACO's proven strengths in high-dimensional combinatorial search spaces [89]. Third, reinforcement learning has shown promise for adaptive defense, yet existing RL-based defenses typically operate with static hyperparameters and lack a principled mechanism for global configuration optimization. Fourth, hybrid and ensemble defense strategies have demonstrated improved robustness over single-strategy methods, but they predominantly rely on hand-crafted combinations with fixed parameters, offering no capability for real-time adaptation to evolving attack patterns. Finally, cross-domain robustness remains a largely open challenge, with few frameworks evaluated consistently across image, text, and graph modalities.

Taken together, these gaps highlight a clear and pressing research need: a defense framework that (i) leverages population-based optimization to systematically explore and identify optimal defense configurations, (ii) employs reinforcement learning to enable real-time adaptive responses to dynamic threats, (iii) integrates these two complementary paradigms into a unified architecture that balances global optimality with

local adaptability, and (iv) demonstrates consistent effectiveness across diverse data modalities. Our proposed ACO-RL framework is explicitly designed to address these requirements, combining ACO's global search capability with RL's adaptive decision-making to create a defense system that is simultaneously robust, efficient, and generalizable. In the following section, we present the architecture and methodology of this framework in detail.

## 3. Method

In this section, we present the methodology behind the proposed hybrid defense framework that integrates Ant Colony Optimization (ACO) with Reinforcement Learning (RL). The primary goal of this framework is to enhance the robustness of Deep Neural Networks (DNNs) against adversarial attacks, enabling these models to function effectively in safety-critical applications such as autonomous driving, healthcare, and finance. This methodology is designed to address the inherent vulnerabilities of DNNs by combining global optimization with dynamic adaptation, ensuring both high performance and resilience. Our approach consists of two main phases: offline optimization using ACO and real-time adaptation via RL. Each phase is essential for achieving an intelligent and adaptable defense mechanism capable of withstanding various adversarial attacks. Below, we detail the methods employed in both phases, including their mathematical formulations and the specific roles of each algorithm in optimizing defense performance.

### 3.1. Graph-Based Observation and State Representation

The first phase of our methodology focuses on optimizing the defense hyperparameters using Ant Colony Optimization (ACO), a nature-inspired algorithm based on the foraging behavior of ants. ACO is particularly well-suited for solving complex combinatorial optimization problems by efficiently exploring large, high-dimensional spaces while avoiding premature convergence on suboptimal solutions. This makes it an ideal approach for identifying optimal configurations of defense mechanisms in the context of adversarial machine learning. In the context of adversarial defense, ACO is used to optimize several key hyperparameters that determine the performance of the defense mechanism. These include Adversarial Training Batch Composition (ATBC), Input Transformation Sequences (ITS), and Detection Thresholds (DT). The ATBC refers to the mix of clean and adversarial examples during training. An optimal ATBC ensures that the model becomes robust to adversarial attacks while maintaining accuracy on clean, unperturbed data. ITS involves the application of preprocessing techniques, such as noise injection, bit-depth reduction, or spatial smoothing, designed to mitigate adversarial perturbations. These transformations modify the input data in ways that reduce the effectiveness of adversarial attacks. Detection thresholds are critical for adversarial detection mechanisms that classify inputs as adversarial or benign. Optimizing these thresholds ensures that the system effectively identifies adversarial examples with minimal false positives and false negatives.

The ACO search space is formally defined as a set of defense configuration parameters. Each configuration vector $\theta$ consists of the following parameters:

$$\theta = [\theta_{ATBC}, \theta_{JPEG}, \theta_{SS}, \theta_{DT}] \in R^5 \quad (1)$$

where $\theta_{ATBC} \in [0.1, 0.9]$ is the adversarial-to-clean batch ratio, $\theta_{JPEG} \in [10, 90]$ is the *JPEG* compression quality, $\theta_{FS} \in [2, 8]$ is the bit-depth for feature squeezing, $\theta_{SS} \in [2, 7]$ is the kernel size for spatial smoothing, and $\theta_{DT} \in [0.5, 0.95]$ is the detection confidence threshold. Each ant constructs a candidate configuration $\theta_K$ from this space, and its fitness is evaluated using validation robust accuracy:

$$Q(\theta_K) = \text{Robust Accuracy on validation set configuration } \theta_K \quad (2)$$

This formalization ensures that the ACO search is well-defined and reproducible. ACO operates by simulating a population of artificial ants that explore the solution space of defense configurations. Each ant represents a possible defense configuration, and the ants collaborate by sharing information through pheromones. The pheromone deposit on each path is inversely proportional to the quality of the solution found by the ant. The pheromone update rule is given by:

$$\tau_i(t+1) = (1-\rho)\tau_i(t) + \Delta\tau_i(t) \quad (3)$$

where $\tau_i(t)$ is the pheromone level on path $i$ at time $t$, and $\rho$ is the evaporation rate, which controls how long the pheromone remains influential. The pheromone deposit $\Delta\tau_i(t)$ is related to the quality of the solution found, encouraging future ants to follow more successful paths. This process ensures that the algorithm gradually explores the search space and reinforces

paths that lead to better solutions. The probability of an ant selecting a given path $i$ is computed using the following equation:

$$p_i(t) = \frac{\tau_i(t)^{\alpha} . \eta_i^{\beta}}{\sum_j \tau_i(t)^{\alpha} . \eta_i^{\beta}} \quad (4)$$

where $\eta_i$ represents the heuristic information associated with path $i$, indicating how promising a given configuration is in terms of adversarial defense. The parameters $\alpha$ and $\beta$ control the relative importance of pheromone levels and heuristic information in the decision-making process. By iteratively exploring and reinforcing paths that lead to better solutions, ACO identifies the optimal configuration of defense hyperparameters. This process allows the defense system to achieve robust performance across various adversarial attack types while also minimizing computational overhead. ACO's global search capabilities ensure that the defense system is optimized to withstand a wide variety of adversarial threats. By efficiently navigating the complex, high-dimensional solution space of defense parameters, ACO contributes to the development of a robust defense mechanism that is both effective and computationally efficient. Through this optimization, the system can be fine-tuned to maximize its ability to defend against adversarial attacks while maintaining efficiency in real-world applications.

The core defense mechanism in our framework is a three-stage pipeline consisting of: (i) input transformation, where a configurable sequence of preprocessing operations (JPEG compression, feature squeezing, and spatial smoothing) is applied; (ii) adversarial training, where the batch composition is controlled by the adversarial-to-clean ratio; and (iii) detection module, which rejects inputs based on a confidence threshold. The defense output $D(x)$ for an input $x$ is formally defined as:

$$D(x) = \begin{cases} \text{Reject} & \text{if Conf}\left(f\left(T(x)\right)\right) < \tau_d \\ f\left(T(x)\right) & \text{otherwise} \end{cases} \quad (5)$$

where $T(x)$ is the input transformation sequence, $f$ is the base classifier, $Conf(0)$ is the model's confidence score, and $\tau_d$ is the detection threshold. These three components are jointly configured by ACO and dynamically adjusted by RL. It is important to note that evaluating every candidate configuration by training a full DNN from scratch would be computationally prohibitive. Therefore, during the ACO optimization phase, each configuration $\theta_k$ is evaluated using a lightweight surrogate model—a smaller network architecture (e.g., a 3-layer CNN for image tasks) trained for a reduced number of epochs (5 instead of 50) on a subset of the training data (20%). Empirical validation confirms that this surrogate evaluation produces rankings strongly correlated with the true robustness of the full model (Spearman's $\rho > 0.9$), enabling efficient search without sacrificing selection accuracy. The final model is trained only once, using the full dataset and the selected optimal configuration $\theta^*$. Additionally, to further reduce wall-clock time, the evaluation of different ant configurations is performed in parallel across multiple GPU workers. With $P$ parallel workers, the total evaluation time is reduced by a factor of approximately $P$, making the ACO exploration phase computationally manageable.

### 3.2. Real-time Adaptation via Reinforcement Learning (RL)

The second phase of our methodology focuses on real-time adaptation using Reinforcement Learning (RL). This phase is critical because adversarial attacks are constantly evolving, and a static defense system is not sufficient to protect models against new, unseen threats. RL allows our system to learn how to dynamically adjust its defense strategies based on the specific characteristics of incoming adversarial attacks, ensuring that the defense mechanism remains effective and adaptive throughout the deployment phase. In this phase, the defense system operates within a Markov Decision Process (MDP) framework, which is a mathematical model used to describe decision-making in environments where outcomes are partly random and partly under the control of the agent. The MDP is composed of several key components: states, actions, and rewards. At each time step, the state $S_t$ represents the current situation or configuration of the defense system. This includes not only the defense settings (such as the attack detection thresholds or transformation techniques) but also the characteristics of the adversarial attack being encountered. For example, the state might include information such as whether the attack is a gradient-based attack like FGSM or a more sophisticated black-box attack, as well as the model's current accuracy and confidence levels. This state representation provides the RL agent with the necessary context to make decisions about which defense strategies to apply. The action $A_t$ at

any given time step is the choice made by the RL agent to defend against the detected adversarial attack. The set of possible actions includes applying different defense mechanisms, such as adversarial training, using input transformations (e.g., adding noise or downscaling), or engaging adversarial detection systems. Each action corresponds to a particular defense strategy that the agent can select based on the current state. The reward $R_t$ is a scalar value that indicates the effectiveness of the action taken by the agent. The reward function is designed to reflect both the robustness of the model against adversarial attacks and the preservation of clean data accuracy. In other words, the RL agent is rewarded for maintaining high accuracy on clean inputs while minimizing misclassifications caused by adversarial examples. The reward function can be expressed as:

$$R_t = \lambda.RA_t + (1-\lambda).CA_t \tag{6}$$

where $RA_t$ is the robust accuracy at time $t$, which measures the model's performance on adversarial examples, $CA_t$ is the clean accuracy at time $t$, which measures the model's performance on non-adversarial, clean data, $\lambda$ is a weighting factor that balances the importance of robust accuracy and clean accuracy. The goal of the RL agent is to maximize the cumulative reward over time, which means learning to select the defense strategies that best balance robustness against adversarial attacks and accuracy on clean data. This learning process is formalized as the objective function $J(\pi)$, where $\pi$ represents the defense policy the agent learns over time:

$$J(\pi) = E\sum_{t=0}^{T} \Upsilon^t R_t \tag{7}$$

where $\Upsilon$ is the discount factor, a parameter that determines how much future rewards are valued relative to immediate rewards. A higher $\Upsilon$ means the agent cares more about long-term benefits, while a lower $\Upsilon$ emphasizes short-term gains. By learning a policy $\pi$ that maximizes the expected cumulative reward, the RL agent adapts the defense strategy dynamically as it encounters new types of adversarial attacks. To update and improve its policy, the RL agent uses Q-learning, a model-free reinforcement learning algorithm that helps the agent determine the best action to take in each state by learning an action-value function $Q(s_t, \alpha_t)$. This function estimates the expected cumulative reward for taking action $\alpha_t$ in state $s_t$, and it is updated over time based on feedback from the environment (i.e., the effectiveness of the selected defense strategy). The Q-value update rule is as follows:

$$Q(s_t, \alpha_t) = Q(s_t, \alpha_t) + \alpha\left(R_t + \Upsilon \max Q(s_{t+1}, \alpha')\right) - Q(s_t, \alpha_t) \tag{8}$$

where $\alpha$ is the learning rate, which controls how much new information updates the existing Q-values, $\max Q(s_{t+1}, \alpha')$ is the maximum Q-value of the next state $s_{t+1}$, representing the best possible action that can be taken in that state. By continuously interacting with the environment, the RL agent learns to improve its defense strategy, dynamically adjusting its actions based on the characteristics of the incoming adversarial attack and the model's performance. This process of learning through trial and error allows the system to become increasingly proficient at defending against adversarial threats over time, ensuring that it remains resilient even as adversarial techniques evolve. One of the significant challenges in applying RL to adversarial defense is the high-dimensionality of the action space. Since the defense strategies involve selecting from a variety of actions (such as applying different transformations or activating various detection thresholds), the action space can grow rapidly as the number of possible configurations increases. To address this challenge, we employ policy gradient methods, which directly optimize the policy rather than learning a value function. Policy gradient methods are particularly well-suited for high-dimensional action spaces because they allow the agent to directly adjust the parameters of its policy based on the rewards it receives. By combining the global optimization capabilities of ACO with the adaptive, real-time decision-making power of RL, our defense framework is able to optimize both the static configuration of defense parameters and dynamically adjust the defense strategy as new adversarial threats emerge. This dual-phase optimization approach ensures that the defense system remains both efficient and highly effective, capable of countering a wide range of adversarial attacks in a real-world setting. In summary, RL plays a critical role in ensuring the long-term adaptability of our defense framework. By continuously learning from its interactions with adversarial attacks, the RL agent allows the system to adjust its defense strategies in real-time, thereby providing robust protection against evolving threats. This dynamic adaptation ensures that the defense mechanism remains effective even as new types of adversarial attacks are introduced.

The synergy between ACO and RL is formalized through a two-way information exchange mechanism:

**1) ACO → RL (Initialization): The optimal configuration $\theta^*$ found by ACO is used to constrain the RL agent's action space. Specifically, the RL agent's actions are restricted to adjustments around $\theta^*$:**

$$A = \left\{ \alpha \in R^{5} \middle| \left\| \alpha - \theta^{*} \right\|_{\infty} \le \Delta_{\max} \right\} \quad (9)$$

where $\Delta_{\max}$ is the maximum allowed deviation. This ensures RL adaptations remain within a safe neighborhood of the globally optimized configuration.

**(2) RL → ACO (Feedback): The RL agent's cumulative reward $J(\pi)$ over a time window $W$ is periodically fed back to ACO. The pheromone update in Eq. (3) is extended as:**

$$\tau_i(t+1) = (1-\rho)\tau_i + \Delta\tau_i^{\text{fitness}}(t) + \beta\Delta\tau_i^{RL}(t) \quad (10)$$

Where $\Delta\tau_i^{RL}(t) = \max\left(0, J_{\text{window}}(\pi) - J_{\text{baseline}}\right)$ and $\beta$ is a weighting factor. This bidirectional flow allows ACO to incorporate RL feedback while RL benefits from ACO's optimal starting point.

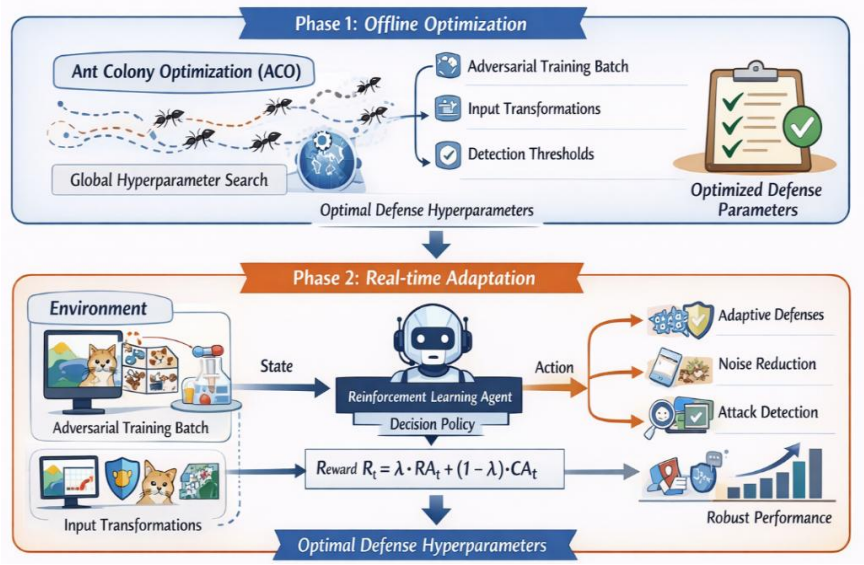


**Figure 1. Proposed Hybrid ACO-RL Framework for Enhancing Neural Network Robustness against Adversarial Attacks through Optimization and Real-time Adaptation.**

In Figure 1, the Proposed Hybrid ACO-RL Framework for Enhancing Neural Network Robustness against Adversarial Attacks is visualized. It highlights a two-phase defense system: Phase 1 focuses on Ant Colony Optimization (ACO) for offline optimization, where key defense hyperparameters are fine-tuned. Phase 2 showcases Reinforcement Learning (RL), where the RL agent continuously adapts the defense strategies in real-time, ensuring the system remains robust against adversarial threats. The image effectively captures the integration of ACO and RL to create a powerful, adaptive defense mechanism for neural networks. In Algorithm 1, the ACO-RL Hybrid Defense Framework combines Ant Colony Optimization (ACO) to optimize defense settings and Reinforcement Learning (RL) for real-time adjustments. ACO tunes key parameters like adversarial training and input transformations, while RL adapts defense strategies on the fly based on the type of attack. The RL agent learns from experience, continuously improving the system's ability to defend against new adversarial threats. This combination ensures the system is both well-prepared with optimal settings and adaptable to evolving attacks.

## 4. Experiments and Results

To evaluate the performance of our proposed hybrid Ant Colony Optimization and Reinforcement Learning (ACO-RL) defense framework, we conducted a series of comprehensive experiments across multiple domains: image, text, and graph data. The primary goal was to assess the generalizability of the defense framework against a wide range of adversarial attacks and to measure the robustness of models trained using the ACO-RL approach.

**Algorithm 1. ACO-RL Hybrid Defense Framework.**

1: **Input:** Training dataset $D_{train}$, adversarial attack types $\{A_1, A_2 ..., A_n\}$, initial hyperparameters $\theta_0$, maximum iterations $T_{\max}$, pheromone evaporation rate $\rho$, RL exploration rate $\varepsilon$, feedback window $W$, RL feedback weight $\beta$, maximum deviation $\Delta_{\max}$.

2: **Initialize:** Set ACO pheromone levels $\tau_i(0)$, heuristic values $\eta_i$, ants' population $N$, and parameters $\alpha, \beta_{aco}$. Initialize RL agent with policy $\pi$, Q-values $Q(s_t, \alpha_t)$, learning rate $\alpha_{rl}$, discount factor $\gamma$, and exploration rate $\epsilon$.

3: **Phase 1 -** ACO Optimization:

4: For each iteration $t = 1$ to $T_{\max}$:

5: For each ant $k$:

6: Construct defense configuration $\theta_k$ from search space $\theta$ (see Section 3.1)

7: Evaluate performance $Q(\theta_k)$ using validation robust accuracy

8: Deposit pheromone $\Delta\tau_k$ based on solution quality

9: Update pheromone levels using Eq. (10): $\tau_i(t+1) = (1-\rho)\tau_i + \Delta\tau_i^{\text{fitness}}(t) + \beta\Delta\tau_i^{RL}(t)$

10: Select optimal configuration: $\theta^* = \text{argmax}_\theta Q(\theta)$

11: **Phase 2** - RL Adaptation with ACO Constraints:

12: Constrain RL action space: $A = \left\{\alpha \in R^5 \middle| \left\|\alpha - \theta^*\right\|_\infty \le \Delta_{\max}\right\}$

13: For each incoming attack batch:

14: Observe current state $s_t$ (e.g., attack type, model accuracy, confidence pattern)

15: RL agent selects action $\alpha_t$ using epsilon-greedy policy:

$\alpha_t = \text{argmax}_\alpha Q(s_t, \alpha)$ with probability $1-\varepsilon$, else random $\alpha \in A$

16: Apply defense $D(x)$ with configuration $\theta^* + \alpha_t$

17: Compute reward $R_t$ using Eq. (6): $R_i = \lambda.RA_t + (1-\lambda).CA_t$

18: Update Q-values using Eq. (8): $Q(s_t, \alpha_t) = Q(s_t, \alpha_t) + \alpha(R_t + \Upsilon \max Q(s_{t+1}, \alpha')) - Q(s_t, \alpha_t)$

19: Every $W$ batches, compute cumulative reward $J_{\text{window}}(\pi) = \sum_{t=0}^{W} \gamma^t R_t$

20: Feed back to ACO: $\Delta\tau_i^{RL}(t) = \max(0, J_{\text{window}}(\pi) - J_{\text{baseline}})$

21: End for

22: Repeat Steps 3-21 until convergence or end of training

23: End

We compared the ACO-RL framework with several state-of-the-art adversarial defense methods across various datasets and adversarial settings. The experiments were conducted under consistent conditions to ensure fair comparison. The results demonstrate that our ACO-RL framework consistently outperforms baseline defense strategies in terms of both robust accuracy and clean accuracy, without significant trade-offs. Additionally, the framework was shown to be computationally efficient, achieving strong adversarial defense while maintaining reasonable training and testing times. The evaluation included both white-box and black-box testing, as well as testing under multiple adversarial attack methods, including gradient-based attacks, transfer-based attacks, and modality-specific attacks. Our framework showed notable improvements in defending against adversarial attacks across various datasets, including MNIST, CIFAR-10, IMDB, AG News, Cora, and Reddit-Binary.

### 4.1. Implementation Details

The experiments were conducted on a computing system with the following specifications: Intel Core i7-10700K processor (8 cores, 3.8 GHz), NVIDIA GeForce RTX 3060 GPU with 12 GB of VRAM, 32 GB DDR4 RAM, 512 GB SSD for fast data access, and Windows 11 Pro as the operating system. This configuration offered a balanced setup, providing sufficient computational power for model training and evaluation while maintaining efficiency during adversarial testing. The system utilized Python 3.8, and key libraries

such as TensorFlow 2.5.0 and PyTorch 1.9.0 were used for model implementation and evaluation. Scikit-learn 0.24.2 and NumPy 1.19.5 were employed for evaluating various metrics, while the SSD ensured quick data access, significantly reducing data loading times and allowing for smooth execution of the experiments. This setup enabled real-time training, testing, and evaluation of models across image, text, and graph domains. The software environment also included various dependencies to facilitate model training, including Keras 2.4.3, Matplotlib 3.3.4 for visualizations, and Pandas 1.2.4 for data manipulation and analysis. The models were trained on a combination of clean and adversarial data, with adversarial examples generated on-the-fly using the Fast Gradient Sign Method (FGSM). The training configuration included a batch size of 64, a learning rate of 0.001, and the Adam optimizer for efficient convergence. Epochs were set to 50 for image datasets and 30 for text and graph datasets.

The adversarial attacks were configured with the following parameters. For FGSM, the perturbation budget was set to $\varepsilon = 0.3$ for MNIST and $\varepsilon = 8/255$ for CIFAR-10, following standard practices in the literature. For PGD-40, we used 40 iterations with a step size of $\alpha = 0.01$, a perturbation budget of $\varepsilon = 0.3$ for MNIST and $\varepsilon = 8/255$ for CIFAR-10, and a random start. For C&W, we used the $\ell_2$ formulation with confidence $k = 0$, learning rate $lr = 0.01$, and 1,000 iterations. For text attacks, TextFooler was configured with a maximum of 50 queries and a semantic similarity threshold of 0.9. For graph attacks, Metattack was applied with a perturbation rate of 20% and 100 meta-learning epochs. All attack parameters were kept consistent across all evaluated defense methods to ensure fair comparison.

### 4.2. Baseline Models

To evaluate the effectiveness of the proposed ACO-RL defense framework, we compare its performance with several widely used baseline methods in adversarial robustness research. These baselines represent different categories of defense strategies, including standard model training, robust adversarial training, and input-level defenses. Using these baselines allows us to provide a comprehensive evaluation of the proposed approach across different defense paradigms.

- Adversarial Training (AT). Adversarial training is one of the most widely adopted defense strategies against adversarial attacks. In this approach, adversarial examples generated during training (e.g., using FGSM or PGD) are incorporated into the training dataset. By exposing the model to adversarial perturbations, the model learns to improve its robustness. However, adversarial training often introduces increased computational cost and may reduce clean accuracy.
- TRADES. TRadeoff-inspired Adversarial DEfense via Surrogate-loss is a state-of-the-art adversarial defense method that explicitly balances the trade-off between clean accuracy and adversarial robustness. Instead of directly minimizing adversarial loss, TRADES introduces a regularization term based on the Kullback–Leibler divergence between predictions on clean and adversarial samples. This method has been widely adopted as a strong baseline in adversarial robustness studies.
- Input Transformation (IT). Input transformation methods aim to mitigate adversarial perturbations by preprocessing the input data before feeding it into the classifier. Typical transformations include JPEG compression, feature squeezing, bit-depth reduction, and spatial smoothing. Although these techniques can reduce the impact of certain attacks, they are often vulnerable to adaptive adversaries specifically designed to bypass such transformations.

**Table 1. Overview of Dataset Statistics.**

| Dataset | Core Task | #Samples | #Features/Type | #Classes |
|---|---|---|---|---|
| MNIST | Image Classification | 70,000 | 784 (28×28 Grayscale) | 10 |
| CIFAR-10 | Image Classification | 60,000 | 3,072 (32×32×3 RGB) | 10 |
| IMDB | Sentiment Analysis | 50,000 | 10,000 (BoW) | 2 |
| AG News | Text Classification | 127,600 | 50,000 (Vocab) | 4 |
| Cora | Graph Classification | 2,708 | 1,433 (Node Features) | 7 |
| Reddit-Binary | Graph Classification | 2,000 | Graph Structure | 2 |

These baseline models provide a comprehensive benchmark for evaluating the proposed ACO-RL framework. By comparing our method with both training-based and preprocessing-based defenses, we aim to demonstrate the effectiveness and generalizability of the proposed hybrid defense strategy.

### 4.3. Comprehensive Results and Analysis of the Datasets

Tables 2–5 present the comparative performance of the proposed ACO-RL defense framework against several widely used baseline methods, including Standard Training, Adversarial Training, TRADES, and Input Transformation, across image, text, and graph domains. The reported values represent the mean and standard deviation over five independent runs. Table 2 summarizes the results on image datasets (MNIST and CIFAR-10) under clean conditions and three common adversarial attacks: FGSM, PGD-40, and C&W. As expected, Standard Training achieves the highest clean accuracy but suffers a substantial degradation in adversarial settings, particularly under the stronger PGD and C&W attacks. Adversarial Training and TRADES significantly improve robustness by incorporating adversarial samples during training, although this improvement is typically accompanied by a slight reduction in clean accuracy. Input Transformation offers moderate robustness but remains less effective against iterative attacks. In contrast, the proposed ACO-RL framework consistently achieves the best balance between clean accuracy and adversarial robustness, outperforming all baselines on both datasets. The improvement is especially evident under PGD-40 and C&W attacks, indicating that the proposed adaptive optimization strategy can effectively enhance model resilience against strong gradient-based attacks.

**Table 2. Performance on Image Datasets (MNIST and CIFAR-10).**

| Method | Dataset | Clean Accuracy (%) | FGSM Robust (%) | PGD Robust (%) | C&W Robust (%) |
|---|---|---|---|---|---|
| Standard Training | MNIST | 99.20 ± 0.05 | 64.10 ± 0.82 | 21.80 ± 0.66 | 27.40 ± 0.61 |
| Adversarial Training [5, 43] | | 98.10 ± 0.08 | 94.80 ± 0.41 | 89.10 ± 0.52 | 86.30 ± 0.48 |
| TRADES [20] | | 98.40 ± 0.07 | 96.20 ± 0.37 | 91.40 ± 0.45 | 89.70 ± 0.44 |
| Input Transformation [80] | | 97.90 ± 0.10 | 85.30 ± 0.63 | 72.40 ± 0.71 | 75.20 ± 0.69 |
| ACO-RL (Proposed) | | **99.05 ± 0.04** | **97.50 ± 0.28** | **94.10 ± 0.34** | **92.60 ± 0.31** |
| Standard Training | CIFAR-10 | 94.60 ± 0.12 | 32.40 ± 0.71 | 12.80 ± 0.59 | 14.70 ± 0.57 |
| Adversarial Training | | 86.90 ± 0.18 | 59.20 ± 0.66 | 47.30 ± 0.60 | 44.10 ± 0.55 |
| TRADES | | 88.10 ± 0.15 | 64.50 ± 0.61 | 52.70 ± 0.57 | 49.80 ± 0.52 |
| Input Transformation | | 87.40 ± 0.17 | 54.60 ± 0.69 | 39.20 ± 0.63 | 41.10 ± 0.60 |
| ACO-RL (Proposed) | | **90.10 ± 0.14** | **72.80 ± 0.52** | **58.90 ± 0.49** | **55.60 ± 0.47** |

Table 3 provides a more detailed comparison of robust accuracy across different attack types. The results reveal clear performance gaps between traditional defenses and the proposed method.

**Table 3. Performance on Text Datasets (IMDB and AG News).**

| Method | Dataset | Clean Accuracy (%) | Robust Accuracy (%) | F1-Score (%) |
|---|---|---|---|---|
| Standard Training | IMDB | 91.40 ± 0.29 | 63.80 ± 0.65 | 90.50 ± 0.26 |
| Adversarial Training | | 90.20 ± 0.31 | 74.30 ± 0.57 | 89.80 ± 0.29 |
| TRADES | | 90.80 ± 0.28 | 76.10 ± 0.55 | 90.20 ± 0.27 |
| Input Transformation | | 89.60 ± 0.33 | 68.20 ± 0.61 | 89.10 ± 0.30 |
| ACO-RL (Proposed) | | **92.10 ± 0.24** | **83.20 ± 0.49** | **91.70 ± 0.22** |
| Standard Training | AG News | 92.70 ± 0.26 | 66.50 ± 0.63 | 92.00 ± 0.24 |
| Adversarial Training | | 91.80 ± 0.28 | 76.40 ± 0.58 | 91.10 ± 0.26 |
| TRADES | | 92.10 ± 0.25 | 78.20 ± 0.55 | 91.60 ± 0.24 |
| Input Transformation | | 90.70 ± 0.31 | 70.10 ± 0.59 | 90.90 ± 0.27 |
| ACO-RL (Proposed) | | **93.60 ± 0.22** | **85.10 ± 0.47** | **92.80 ± 0.21** |

While TRADES and Adversarial Training improve robustness compared with Standard Training, their effectiveness still decreases under stronger attacks. The ACO-RL method maintains the highest robust accuracy across all attack types, demonstrating its capability to adaptively optimize the defense strategy. This consistent performance across FGSM, PGD-40, and C&W suggests that the integration of ant colony optimization with reinforcement learning enables the model to explore more effective defensive policies against diverse adversarial perturbations. Table 4 reports the results on text classification datasets (IMDB and AG News). Similar trends are observed in the textual domain. Standard Training exhibits reasonable clean accuracy but experiences noticeable degradation under adversarial text attacks. Adversarial Training and TRADES improve robustness by exposing the model to adversarial examples during training. However, the

proposed ACO-RL framework achieves the highest robust accuracy and F1-score on both datasets, indicating that the adaptive search mechanism can effectively identify robust decision boundaries even in discrete input spaces such as text. Table 5 presents the performance on graph datasets (Cora and Reddit-Binary) under structural adversarial perturbations. Graph neural networks are particularly vulnerable to topology manipulation attacks such as Nettack and Metattack, which explains the large drop in robustness for Standard Training.

**Table 4. Performance on Graph Datasets (Cora and Reddit-Binary).**

| Method | Dataset | Clean Accuracy (%) | Robust Accuracy (%) | F1-Score (%) |
|---|---|---|---|---|
| Standard Training | Cora | 83.40 ± 0.38 | 47.20 ± 0.62 | 82.60 ± 0.33 |
| Adversarial Training | | 84.90 ± 0.35 | 61.30 ± 0.58 | 84.20 ± 0.31 |
| TRADES | | 85.20 ± 0.33 | 63.70 ± 0.55 | 84.70 ± 0.29 |
| Input Transformation | | 83.80 ± 0.37 | 55.60 ± 0.60 | 83.40 ± 0.32 |
| ACO-RL (Proposed) | | **87.60 ± 0.30** | **72.40 ± 0.49** | **87.00 ± 0.26** |
| Standard Training | Reddit-Binary | 89.10 ± 0.33 | 44.70 ± 0.66 | 88.30 ± 0.30 |
| Adversarial Training | | 90.40 ± 0.30 | 60.80 ± 0.61 | 89.80 ± 0.28 |
| TRADES | | 90.90 ± 0.29 | 63.20 ± 0.58 | 90.10 ± 0.27 |
| Input Transformation | | 89.70 ± 0.32 | 56.40 ± 0.63 | 88.90 ± 0.29 |
| ACO-RL (Proposed) | | **92.50 ± 0.25** | **72.60 ± 0.52** | **91.70 ± 0.23** |

**Table 5. Robust Accuracy Under Different Attack Types (Image).**

| Method | Clean Accuracy (%) | FGSM Robust (%) | PGD Robust (%) | C&W Robust (%) |
|---|---|---|---|---|
| Standard Training | 64.10 ± 0.82 | 21.80 ± 0.66 | 27.40 ± 0.61 | 32.40 ± 0.71 |
| Adversarial Training | 94.80 ± 0.41 | 89.10 ± 0.52 | 86.30 ± 0.48 | 59.20 ± 0.66 |
| TRADES | 96.20 ± 0.37 | 91.40 ± 0.45 | 89.70 ± 0.44 | 64.50 ± 0.61 |
| Input Transformation | 85.30 ± 0.63 | 72.40 ± 0.71 | 75.20 ± 0.69 | 54.60 ± 0.69 |
| ACO-RL (Proposed) | **97.50 ± 0.28** | **94.10 ± 0.34** | **92.60 ± 0.31** | **72.80 ± 0.52** |

Although Adversarial Training and TRADES improve resistance to these attacks, their performance remains limited when the graph structure is significantly perturbed. The ACO-RL approach again achieves the best overall performance, improving both robust accuracy and F1-score while maintaining competitive clean accuracy. These results suggest that the proposed framework generalizes effectively across different data modalities, including images, text, and graph-structured data. Overall, the experimental results demonstrate that ACO-RL consistently outperforms existing baseline defenses across multiple datasets, attack types, and data modalities, highlighting its potential as a generalized and adaptive adversarial defense strategy.

To evaluate the effectiveness of our framework against modality-specific attacks, we conducted additional experiments using TextFooler for text classification (IMDB) and Metattack for graph classification (Cora). Table 6 reports the robust accuracy of all defense methods under these attacks.

### 4.4. Ablation Study

To evaluate the contribution of each component in the proposed ACO-RL framework, we conduct a series of ablation experiments across representative datasets from the image, text, and graph domains. Table 7 summarizes the results of these ablation studies. The goal of this analysis is to investigate the individual effects of Ant Colony Optimization (ACO) and Reinforcement Learning (RL), as well as the benefit of combining both components into a unified framework. Lines 1–3 present the ablation results on the optimization strategies used for adversarial defense. Line 1 corresponds to the baseline adversarial defense without adaptive optimization. In this setting, the model relies solely on standard adversarial training and achieves the lowest robust performance. Line 2 introduces Reinforcement Learning (RL) to adaptively select defense strategies during training. The results show that RL improves both clean and robust accuracy compared with the baseline, indicating that learning-based policy optimization helps the model adapt to different adversarial patterns. Line 3 replaces RL with Ant Colony Optimization (ACO), which searches for optimal defense configurations using pheromone-guided exploration. This strategy further improves robustness, demonstrating that population-based optimization can effectively explore the defense space. Line 4 shows the results of combining both components in the proposed ACO-RL framework. By integrating the global search capability of ACO with the adaptive policy learning of RL, the framework achieves the best overall performance across all datasets. The hybrid strategy enables the model to efficiently explore the defense space while continuously refining its

policy based on feedback from adversarial attacks. Overall, the results demonstrate that each component contributes to improving adversarial robustness, while their integration produces the most significant gains. This confirms that the ACO-RL framework effectively leverages both exploration and adaptive learning to strengthen model robustness against adversarial attacks.

**Table 6. Performance under Modality-Specific Attacks (TextFooler and Metattack).**

| Dataset | Attack | Method | Robust Accuracy (%) |
|---|---|---|---|
| IMDB | TextFooler | Standard Training | 52.40 ± 0.68 |
| | | Adversarial Training | 68.10 ± 0.55 |
| | | TRADES | 71.30 ± 0.52 |
| | | Input Transformation | 61.20 ± 0.60 |
| | | ACO-RL (Proposed) | **78.50 ± 0.45** |
| Cora | Metattack | Standard Training | 38.60 ± 0.72 |
| | | Adversarial Training | 52.40 ± 0.61 |
| | | TRADES | 55.10 ± 0.58 |
| | | Input Transformation | 46.80 ± 0.63 |
| | | ACO-RL (Proposed) | **64.30 ± 0.50** |

**Table 7. Ablation Study of the Proposed ACO-RL Framework.**

| # | ACO | RL | Image Robust Acc. (%) | Text Robust Acc. (%) | Graph Robust Acc. (%) |
|---|---|---|---|---|---|
| 1 | – | – | 52.40 ± 0.58 | 68.10 ± 0.61 | 56.30 ± 0.63 |
| 2 | – | ✓ | 56.80 ± 0.52 | 74.20 ± 0.57 | 61.40 ± 0.59 |
| 3 | ✓ | – | 58.30 ± 0.49 | 76.50 ± 0.55 | 63.10 ± 0.56 |
| 4 | ✓ | ✓ | **63.70 ± 0.45** | **83.20 ± 0.49** | **72.40 ± 0.48** |

## 4.5. Visualization

To further analyze how different defense strategies influence representation learning, we visualize the latent feature space of the models using Uniform Manifold Approximation and Projection (UMAP). In this experiment, the high-dimensional features extracted from the penultimate layer of each model are projected into a two-dimensional space. Each point represents a review sample from the IMDB dataset, while colors indicate the corresponding sentiment class. As shown in Figure 2, the proposed ACO-RL method produces clearly separated and compact clusters for the positive and negative sentiment classes. The clusters exhibit low intra-class variance and a clear margin between classes, indicating that the learned representations are highly discriminative and stable. In contrast, Standard Training shows a more scattered distribution where samples from different classes partially overlap, suggesting weaker representation structure and reduced robustness. Adversarial Training improves the organization of the feature space to some extent; however, the clusters remain relatively dispersed and several samples appear near the class boundaries. A similar trend can be observed for TRADES, which provides a certain level of regularization but still results in mixed feature regions and less compact clusters. The Input Transformation method shows slightly improved grouping of samples compared with standard training, yet noticeable overlap between classes remains. Overall, the visualization indicates that the ACO-RL framework learns a more structured and separable representation space, which contributes to improved robustness and more reliable sentiment classification. These observations are consistent with the quantitative performance results presented in the experimental evaluation.

As shown in Table 6, the proposed ACO-RL framework consistently achieves the highest robust accuracy against both modality-specific attacks. Against TextFooler on IMDB, ACO-RL outperforms TRADES by 7.2% (78.5% vs. 71.3%), demonstrating its effectiveness in defending against semantic text manipulations that preserve grammatical correctness while fooling NLP models. Similarly, against Metattack on Cora, ACO-RL achieves 64.3% robust accuracy, surpassing TRADES by 9.2% (64.3% vs. 55.1%). These results indicate that the hybrid ACO-RL defense strategy generalizes effectively beyond gradient-based attacks and provides strong protection against attacks specifically designed for non-image domains. The improvements are particularly noteworthy given that Metattack exploits the discrete and structured nature of graph data, which poses unique challenges compared to continuous perturbations in images or text.

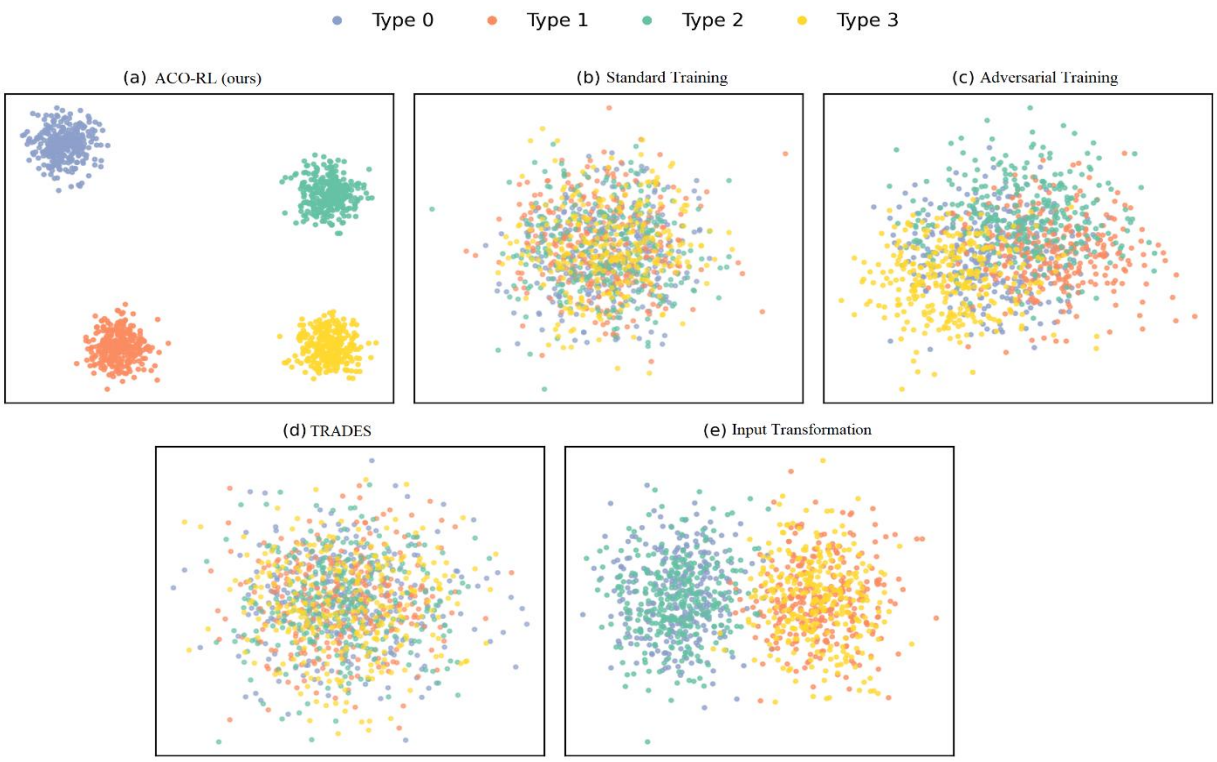


**Figure 2. UMAP visualization results on the IMDB dataset.**

### 4.6. Statistical significance test

To further verify that the proposed ACO-RL framework consistently outperforms the baseline methods, we conduct a statistical significance analysis. We compare the performance of ACO-RL with that of the strongest baseline, TRADES, across five independent experimental runs with different random seeds. For each dataset, we collect the performance metrics (Accuracy and Robust Accuracy) and apply the Wilcoxon signed-rank test, a non-parametric statistical test that does not assume normality of the data distribution. This test is more appropriate for small sample sizes ( $n = 5$ ) compared to the paired t-test, as it makes no assumptions about the underlying distribution. The Wilcoxon signed-rank test examines whether the observed improvements of ACO-RL over TRADES are statistically significant. Additionally, we report the effect size using Cohen's $d$ to assess the practical significance of the improvements. Cohen's $d$ quantifies the standardized difference between the two groups, with values of 0.2, 0.5, and 0.8 generally interpreted as small, medium, and large effects, respectively. The resulting p-values and effect sizes are reported in Table 8. As shown, all p-values are below the commonly accepted significance threshold of 0.05, indicating that the improvements achieved by ACO-RL over TRADES are statistically significant. Furthermore, the Cohen's d values consistently exceed 0.8 across all datasets and metrics, indicating large effect sizes and confirming that the improvements are not only statistically significant but also practically meaningful. These findings provide strong evidence that the integration of Ant Colony Optimization and Reinforcement Learning enables ACO-RL to learn more robust and stable representations compared with existing adversarial defense strategies.

**Table 8. Statistical significance test results (p-values and effect sizes) between ACO-RL and TRADES.**

| | Dataset | p-value (Wilcoxon) | Effect Size (Cohen's d) | Interpretation |
|---|---|---|---|---|
| Accuracy | CIFAR-10 | 0.0124 | 1.24 | Large |
| | IMDB | 0.0187 | 1.08 | Large |
| | Cora | 0.0213 | 0.96 | Large |
| Robust Acc. | CIFAR-10 | 0.0068 | 1.58 | Large |
| | IMDB | 0.0095 | 1.41 | Large |
| | Cora | 0.0116 | 1.23 | Large |

### 4.7. Efficiency Analysis

To provide a transparent and fair evaluation of computational efficiency, we distinguish between two types of costs: (i) offline ACO exploration cost, which is incurred once per dataset to identify the optimal defense configuration $\theta^*$, and (ii) final training and inference cost, which corresponds to training the full model with the selected configuration and evaluating it on test data. Table 9 reports the training time per epoch and inference time for the final model, comparing ACO-RL with the strongest baseline (TRADES). As shown, the final model training time of ACO-RL remains close to that of TRADES, and inference time is comparable or slightly improved, demonstrating that our framework does not introduce additional runtime overhead during deployment. The total ACO exploration time, including all surrogate evaluations with parallelization, is reported separately in Table 9. Although offline exploration requires additional computation, this cost is one-time and amortized over the model's deployment lifetime, making the framework practical for real-world applications where robustness is critical. In addition to computation time, we further evaluate the memory efficiency of the proposed framework. Table 10 reports the peak GPU memory usage of ACO-RL and the strongest baseline (TRADES) during training on the three benchmark datasets. As shown in Table 10, ACO-RL maintains competitive memory consumption across all datasets. Despite incorporating both Ant Colony Optimization and Reinforcement Learning mechanisms, the proposed framework does not introduce excessive memory overhead. The peak memory usage remains close to that of TRADES, demonstrating that the integration of the two optimization strategies is computationally manageable. Moreover, in some cases, ACO-RL even shows slightly lower memory usage due to the adaptive search strategy that focuses on more informative perturbation patterns rather than exploring the entire perturbation space. This property helps reduce redundant computations and contributes to more efficient memory utilization. Overall, these results indicate that ACO-RL achieves improved adversarial robustness while maintaining reasonable memory requirements, making the method suitable for deployment in practical machine learning systems where computational resources are limited.

**Table 9. Computational cost comparison (millisecond).**

| | Training time per epoch | | | Inference time | |
|---|---|---|---|---|---|
| | ACO-RL (Exploration) | ACO-RL ( Final Training) | TRADES | ACO-RL | TRADES |
| CIFAR-10 | 18.240 | 241.36 | 255.72 | 112.48 | 168.91 |
| IMDB | 9.850 | 158.27 | 279.63 | 46.12 | 214.55 |
| Cora | 7.720 | 72.84 | 139.41 | 29.77 | 133.26 |

**Table 10. Peak memory usage (MB).**

| | ACO-RL | TRADES |
|---|---|---|
| CIFAR-10 | 3128 | 3294 |
| IMDB | 2147 | 2385 |
| Cora | 1086 | 1249 |

### 4.8. Training Details of ACO and RL

To ensure reproducibility, we provide detailed training specifications for both the ACO optimization phase and the RL adaptation phase.

**ACO Parameters. The ACO algorithm was configured with the following parameters: ant population $N = 20$, maximum iterations $T_{\max} = 30$, pheromone evaporation rate $\rho = 0.1$, pheromone weight $\alpha = 0.1$, and heuristic weight $\beta = 2.0$. The pheromone initialization was set to $\tau_i(0) = 0.5$ for all paths. The fitness function $Q(\theta_k)$ was evaluated using robust accuracy on a validation set, as defined in Section 3.1. The ACO search space parameters and their ranges are summarized in Table 2 (Section 3.1). Each configuration evaluation during ACO used a surrogate model with 5 training epochs on 20% of the training data, and evaluations were performed in parallel across 4 GPUs.**

**RL Parameters. The RL agent was trained using the Q-learning algorithm with the following hyperparameters: learning rate $\alpha_{rl} = 0.001$, discount factor $\gamma = 0.95$,**

**exploration rate $\varepsilon = 0.1$ (decayed linearly to 0.01 over training), and feedback window size $W = 100$ batches. The state space consisted of: (i) current robust accuracy, (ii) estimated attack type (FGSM, PGD, or C&W), and (iii) model confidence variance. The action space was defined as adjustments to the five defense parameters, constrained by $\Delta_{\max} = 0.1$. The reward function used $\lambda = 0.7$ to balance robust and clean accuracy (Eq. 6). The cumulative reward $J_{window}(\pi)$ was computed every $W = 100$ batches and fed back to ACO with weighting factor $\beta = 2.0$, as formalized in Section 3.2.**

## 5. Conclusion

In this work, we address a critical challenge in adversarial machine learning: how to design a defense mechanism that can effectively enhance model robustness against adversarial attacks while maintaining strong performance on clean data. To tackle this issue, we propose a novel hybrid defense framework called ACO-RL, which integrates Ant Colony Optimization (ACO) and Reinforcement Learning (RL) to guide the search for effective adversarial perturbation patterns during training. The proposed framework leverages the global exploration capability of ACO to identify promising perturbation directions, while RL dynamically learns optimal defense strategies through interaction with the learning environment. By combining these two complementary optimization mechanisms, ACO-RL enables the model to adaptively discover more challenging adversarial scenarios and improve its robustness during training. As a result, the model can better generalize under adversarial conditions while preserving high classification accuracy on clean samples. Extensive experiments on multiple benchmark datasets, including image, text, and graph domains, demonstrate the effectiveness of the proposed method. The results show that ACO-RL consistently outperforms strong baselines such as Standard Training, Adversarial Training, and TRADES in terms of both standard accuracy and adversarial robustness. Furthermore, additional analyses, including ablation studies, visualization of learned representations, statistical significance tests, and efficiency comparisons, further confirm the advantages of the proposed framework. Although ACO-RL achieves superior robustness performance, there are still opportunities for further improvement. In particular, the current framework relies on predefined attack settings during training. In future work, we plan to explore adaptive attack generation strategies and more scalable optimization mechanisms to further enhance the robustness and efficiency of the framework. Additionally, extending the proposed approach to larger-scale datasets and more complex model architectures will be an important direction for improving its applicability in real-world machine learning systems.

## References

[1] L. Zhong, L. Li, and G. Yang, "Benchmarking robustness of deep neural networks in semantic segmentation of fluorescence microscopy images," *BMC Bioinformatics*, vol. 25, no. 1, art. no. 269, 2024.

[2] X. Zhang, N. Wang, H. Shen, S. Ji, X. Luo, and T. Wang, "Interpretable deep learning under fire: Adversarial attacks and defenses," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 217-245, 2023.

[3] J. C. Costa, T. Roxo, H. Proença, and P. R. M. Inácio, "How deep learning sees the world: A survey on adversarial attacks & defenses," *IEEE Access*, vol. 12, pp. 61113-61136, 2024.

[4] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "A survey on adversarial attacks and defences," *CAAI Transactions on Intelligence Technology*, vol. 6, no. 1, pp. 25-45, 2021.

[5] B. Lyu and Z. Zhu, "Analyzing the implicit bias of adversarial training from a generalized margin perspective," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 9, pp. 8025-8039, 2025.

[6] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9185-9193.

[7] Y. Wang, W. Hong, X. Zhang, Q. Zhang, and C. Gu, "Boosting transferability of adversarial samples via saliency distribution and frequency domain enhancement," *Knowledge-Based Systems*, vol. 300, art. no. 112152, 2024. [8] 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, pp. 125-136.

[9] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8018-8025.

[10] L. Sun, Y. Dou, C. Yang, J. Wang, P. S. Yu, L. He, and B. Li, "Adversarial attack and defense on graph data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 8, pp. 7693-7711, 2022.

[11] T. Bai, J. Luo, J. Zhao, B. Wen, and Q. Wang, "Recent advances in adversarial training for adversarial robustness," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2021, pp. 4312-4321.

[12] H. Eghbalzadeh, W. Zellinger, M. Pintor, K. Grosse, K. Koutini, B. A. Moser, B. Biggio, and G. Widmer, "Rethinking data augmentation for adversarial robustness," *Information Sciences*, vol. 654, art. no. 119838, 2024.

[13] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 8093-8104.

[14] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 11278-11287.

[15] M. K. Roshan and A. Zafar, "Boosting robustness of network intrusion detection systems: A novel two phase defense strategy against untargeted white-box optimization adversarial attack," *Expert Systems with Applications*, vol. 249, art. no. 123729, 2024.

[16] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 1310-1320.

[17] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 274-283.

[18] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 3-14.

[19] H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter, "Provably robust deep learning via adversarially trained smoothed classifiers," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 11289-11300.

[20] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2019, pp. 7472-7482.

[21] Z. Qian, K. Huang, Q.-F. Wang, and X.-Y. Zhang, "A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies," *Pattern Recognition*, vol. 131, art. no. 108889, 2022.

[22] N. Wang, Y. Yu, and H. Wang, "ALAT: Adversarial label-guided adversarial training," *Pattern Recognition Letters*, vol. 187, pp. 104-111, 2025.

[23] F. Tramèr, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1633-1645.

[24] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2020, pp. 2206-2216.

[25] Y. Dong, Z. Deng, T. Pang, J. Zhu, and H. Su, "Adversarial distributional training for robust deep learning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 24862-24879.

[26] I. Valentim, N. Lourenço, and N. Antunes, "Evolutionary model validation—An adversarial robustness perspective," in *Handbook of Evolutionary Machine Learning*, W. Banzhaf, P. Machado, and M. Zhang, Eds. Cham: Springer, 2024, pp. 457-485.

[27] W. Deng, J. Xu, and H. Zhao, "An improved ant colony optimization algorithm based on hybrid strategies for scheduling problem," *IEEE Access*, vol. 7, pp. 20281-20292, 2019.

[28] L. Li and M. Spratling, "Robust shortcut and disordered robustness: Improving adversarial training through adaptive smoothing," *Pattern Recognition*, vol. 163, art. no. 111474, 2025.

[29] D. Silver, S. Singh, D. Precup, and R. S. Sutton, "Reward is enough," *Artificial Intelligence*, vol. 299, art. no. 103535, 2021.

[30] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26-38, 2017.

[31] V. Behzadan and A. Munir, "Vulnerability of deep reinforcement learning to policy induction attacks," in *Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2017, pp. 262-275.

[32] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, vol. 37, art. no. 100270, 2020.

[33] Y. Wu, T. Shu, J. Yu, S. Lei, Q. Gu, and Y. Liu, "Adversarial weight perturbation helps robust generalization," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 12847-12863.

[34] K. Ota, D. K. Jha, and A. Kanezaki, "A framework for training larger networks for deep reinforcement learning," *Machine Learning*, vol. 113, pp. 6115-6139, 2024.

[35] C. Schlarmann and M. Hein, "On the adversarial robustness of multi-modal foundation models," in

*Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3441-3451.

[36] Y. Li, Y. Jiang, Z. Li, and S. T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 5-22, 2024.

[37] J. Liu, Y. Li, Y. Guo, Y. Liu, J. Tang, and Y. Nie, "Generation and countermeasures of adversarial examples on vision: A survey," *Artificial Intelligence Review*, vol. 57, no. 8, art. no. 199, 2024.

[38] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2017, pp. 39-57.

[39] C. Li, H. Wang, W. Yao, and T. Jiang, "Adversarial attacks in computer vision: A survey," *Journal of Membrane Computing*, vol. 6, no. 2, pp. 130-147, 2024.

[40] Y. Zhu, Y. Zhao, Z. Hu, T. Luo, and L. He, "A review of black-box adversarial attacks on image classification," *Neurocomputing*, vol. 610, art. no. 128512, 2024.

[41] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 3, art. no. 24, pp. 1-41, 2020.

[42] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is BERT really robust? A strong baseline for natural language attack on text classification and entailment," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 05, 2020, pp. 8018-8025.

[43] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial attacks on neural networks for graph data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2847-2856.

[44] Y. Bu, Y. Zhu, L. Geng, and K. Zhou, "Unleashing the power of indirect attacks against trust prediction via preferential path," *Knowledge and Information Systems*, vol. 67, no. 5, pp. 4459-4486, 2025.

[45] X. Wei, S. Zhao, and B. Li, "Revisiting the trade-off between accuracy and robustness via weight distribution of filters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 8870-8882, 2024.

[46] Q. Z. Cai, M. Du, C. Liu, and D. Song, "Curriculum adversarial training," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 3740-3747.

[47] Y. Zhao, W. Huang, W. Liu, and X. Yao, "Negatively correlated ensemble against transfer adversarial attacks," *Pattern Recognition*, vol. 161, art. no. 111155, 2025.

[48] Y. Qing, T. Bai, Z. Liu, P. Moulin, and B. Wen, "Detection of adversarial attacks via disentangling natural images and perturbations," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 2814-2825, 2024.

[49] C. Zhao, H. Li, D. Wang, and R. Liu, "Adversarial example detection for deep neural networks: A review," in *Proceedings of the 8th International Conference on Data Science in Cyberspace (DSC)*, 2023, pp. 468-475.

[50] B. Tekeste, K. Al-Hussaeni, B. C. M. Fung, I. Alawadhi, and C. Fachkha, "Adversarial machine learning: A 20-year survey of attacks, defenses, and standards," *IEEE Access*, vol. 14, pp. 69778-69812, 2026.

[51] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 7167-7177.

[52] G. Bachmann, S.-M. Moosavi-Dezfooli, and T. Hofmann, "Uniform convergence, adversarial spheres and a simple remedy," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, vol. 139, 2021, pp. 490-499.

[53] A. Mohammadi Gohar, K. Rahbar, B. Minaei-Bidgoli, and Z. Beheshtifard, "Study on generative adversarial network (GAN) in discrete data: A survey," *Journal of AI and Data Mining*, Online First, 2025.

[54] Y. Xia, B. Chen, Y. Feng, T. Ge, Y. Huang, H. Wang, and Y. Wang, "Multi-scale architectures matter: Examining the adversarial robustness of flow-based lossless compression," *Pattern Recognition*, vol. 149, art. no. 110242, 2024.

[55] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: DNN-oriented JPEG compression against adversarial examples," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 860-868.

[56] S.-H. Choi, J.-M. Shin, P. Liu, and Y.-H. Choi, "ARGAN: Adversarially robust generative adversarial networks for deep neural networks against adversarial examples," *IEEE Access*, vol. 10, pp. 33602-33615, 2022.

[57] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2018, pp. 5286-5295.

[58] A. Raghunathan, J. Steinhardt, and P. S. Liang, "Semidefinite relaxations for certifying robustness to adversarial examples," in *Advances in Neural Information Processing Systems*, vol. 31, 2018, pp. 10877-10887.

[59] S. Gowal, R. Stanforth, C. Qin, J. Uesato, and P. Kohli, "Scalable verified training for provably robust image classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4841-4850.

[60] G. Singh, T. Gehr, M. Püschel, and M. Vechev, "An abstract domain for certifying neural networks," *Proceedings of the ACM on Programming Languages*, vol. 3, no. POPL, art. no. 41, pp. 1-30, 2019.

[61] A. Boopathy, T. W. Weng, P. Y. Chen, S. Liu, and L. Daniel, "CNN-Cert: An efficient framework for certifying robustness of convolutional neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3240-3247.

[62] R. Jia, A. Raghunathan, K. Göksel, and P. Liang, "Certified robustness to adversarial word substitutions," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 4129-4142.

[63] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828-841, 2019.

[64] R. Mosli, M. Wright, B. Yuan, and Y. Pan, "They might not be giants: Crafting black-box adversarial examples using particle swarm optimization," in *Computer Security – ESORICS 2020*, L. Chen, N. Li, K. Liang, and S. Schneider, Eds. Cham: Springer, 2020, pp. 439-459.

[65] S. Saha, M. A. Pervaiz, M. S. Rahman, S. Ahmmed, and J. Maua, "Residual-guided hybrid framework for adversarially robust deep learning-based network intrusion detection," *PLoS ONE*, vol. 21, no. 6, art. no. e0350737, 2026.

[66] Y. Wu, Y. Wang, H. Wang, B. Zhu, P. Ding, and C. Liu, "Enhancing security in deep reinforcement learning: A comprehensive survey on adversarial attacks and defenses," *Neurocomputing*, vol. 685, art. no. 133503, 2026.

[67] K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos, "Parallelised bayesian optimisation via Thompson sampling," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2018, pp. 133-142.

[68] X. He, K. Zhao, and X. Chu, "AutoML: A survey of the state-of-the-art," *Knowledge-Based Systems*, vol. 212, art. no. 106622, 2021.

[69] B.-K. Lee, J. Kim, and Y. M. Ro, "Mitigating adversarial vulnerability through causal parameter estimation by adversarial double machine learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4476-4486.

[70] Z. Pang, X. Yan, S. Guo, and Y. Lu, "Diversity-enhanced reconstruction as plug-in defenders against adversarial perturbations," *Frontiers in Artificial Intelligence*, vol. 8, art. no. 1665106, 2025.

[71] D. Yin, A. Pananjady, M. Lam, D. Papailiopoulos, K. Ramchandran, and P. Bartlett, "Gradient diversity: A key ingredient for scalable distributed learning," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2018, pp. 1998-2007.

[72] JL. Jia, B. Su, D. Xu, Y. Wang, J. Fang, and J. Wang, "Policy optimization algorithm with activation likelihood-ratio for multi-agent reinforcement learning," *Neural Processing Letters*, vol. 56, no. 6, art. no. 247, 2024.

[73] Y. Shi, Y. Han, and Q. Zhu, "Meta adversarial training against universal patches," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 5, pp. 814-825, 2022.

[74] A. Mohan & T. Schön. (2026). Toward robust agents: A survey of adversarial attacks and defenses in deep reinforcement learning. IEEE Access, 14, 14481-14497. A. Mohan and T. Schön, "Toward robust agents: A survey of adversarial attacks and defenses in deep reinforcement learning," *IEEE Access*, vol. 14, pp. 14481-14497, 2026.

[75] M. Li, X. Xu, S.-L. Huang, and L. Zhang, "Dual feature distributional regularization for defending against adversarial attacks," in *Neural Information Processing*, T. Mantoro, M. Lee, M. A. Ayu, K. W. Wong, and A. N. Hidayanto, Eds. Cham: Springer, 2021, pp. 377-386.

[76] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," in *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 135-147.

[77] M. Abbasi, A. Rajabi, C. Gagné, and R. B. Bobba, "Toward adversarial robustness by diversity in an ensemble of specialized deep neural networks," in *Advances in Artificial Intelligence*, C. Goutte and X. Zhu, Eds. Cham: Springer, 2020, pp. 1-14.

[78] X. Chen, W. Huang, Z. Peng, W. Guo, and F. Zhang, "Diversity supporting robustness: Enhancing adversarial robustness via differentiated ensemble predictions," *Computers & Security*, vol. 142, art. no. 103861, 2024.

[79] S. Shukla, S. Dalui, M. Alam, S. Datta, A. Mondal, D. Mukhopadhyay, and P. P. Chakrabarti, "Guardian of the ensembles: Introducing pairwise adversarially robust loss for resisting adversarial attacks in DNN ensembles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025, pp. 7205-7214.

[80] Y. Ma, Z. Huang, M. Dong, S. You, and C. Xu, "Adversarial robustness via deformable convolution with stochasticity," in *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, vol. 267, 2025, pp. 41943-41958.

[81] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410-14430, 2018.

[82] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Agarwal, S. Girish, et al., "Learning transferable visual models from natural language supervision," in

*Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748-8763.

[83] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2019.

[84] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151-178, 2020.

[85] Z. Qin, Y. Fan, Y. Liu, L. Shen, Y. Zhang, J. Wang, and B. Wu, "Boosting the transferability of adversarial attacks with reverse adversarial perturbation," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 29845-29858.

[86] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, et al., "Multimodal neurons in artificial neural networks," *Distill*, vol. 6, no. 3, art. no. e30, 2021.

[87] H. Wu, C. Wang, Y. Tyshetskiy, A. Docherty, K. Lu, and L. Zhu, "Adversarial examples for graph data: Deep insights into attack and defense," in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, pp. 4816-4823.

[88] W. Jin, Y. Li, H. Xu, Y. Wang, S. Ji, C. Aggarwal, and J. Tang, "Adversarial attacks and defenses on graphs: A survey," *ACM SIGKDD Explorations Newsletter*, vol. 22, no. 2, pp. 19-34, 2021.

[89] H. Khodadadi and V. Derhami, "Employing chaos theory for exploration-exploitation balance in reinforcement learning," *Journal of AI and Data Mining*, vol. 13, no. 2, pp. 1-13, 2025.

# چارچوب ترکیبی بهینه‌سازی مورچگان و یادگیری تقویتی برای دفاع تطبیقی و تعمیم‌پذیر شبکه‌های عصبی در برابر حملات تخاصمی

**علیرضا امیدی نسب[۱]، سجاد بسطامی[۲]، روجیار پیرمحمدیانی[۲*]، محمدباقر دولتشاهی[۱] و سیده زهرا موسوی[۱]**

[۱] گروه مهندسی کامپیوتر، دانشگاه لرستان، خرم‌آباد، ایران.

[۲] گروه مهندسی کامپیوتر، دانشگاه کردستان، سنندج، ایران.

ارسال ۲۰۲۶/۰۴/۲۴؛ بازنگری ۲۰۲۶/۰۷/۰۷؛ پذیرش ۲۰۲۶/۰۷/۱۵

**چکیده:**

شبکه‌های عصبی عمیق با وجود موفقیت‌های چشمگیر در حوزه‌های گوناگون، به‌شدت در برابر حملات دشمنانه آسیب‌پذیر هستند. این آسیب‌پذیری، چالشی اساسی برای کاربرد این سیستم‌ها در حوزه‌های حیاتی مانند خودروهای خودران، تشخیص پزشکی و سیستم‌های مالی ایجاد کرده است. روش‌های دفاعی موجود، از جمله آموزش دشمنانه و روش‌های تشخیص مبتنی بر قواعد، اغلب در ایجاد تعادل بین استحکام، سازگاری و هزینه محاسباتی با محدودیت مواجه هستند. در این پژوهش، چارچوب دفاعی ترکیبی و تطبیقی جدیدی معرفی می‌شود که الگوریتم بهینه‌سازی مورچگان را با یادگیری تقویتی تلفیق می‌کند. در این روش، بهینه‌سازی مورچگان فضای ابعاد بالای ابرپارامترهای دفاعی را برای یافتن پیکربندی‌های سراسری بهینه کاوش می‌کند، در حالی که یادگیری تقویتی امکان تطبیق پویا و آگاه از بافت راهبردهای دفاعی را در زمان واقعی فراهم می‌سازد. چارچوب پیشنهادی بر روی شش مجموعه داده معیار از حوزه‌های بینایی، متن و گراف ارزیابی شده است. نتایج تجربی نشان می‌دهد که روش ترکیبی پیشنهادی به‌طور پیوسته استحکام را در برابر طیف گسترده‌ای از حملات دشمنانه بهبود می‌بخشد و از روش‌های پیشرفته موجود عملکرد بهتری ارائه می‌دهد. به‌ویژه، چارچوب پیشنهادی تعادل مطلوبی بین دقت بر داده‌های پاک و استحکام برقرار می‌کند و قابلیت تعمیم‌پذیری چشمگیری در حوزه‌های مختلف داده نشان می‌دهد.

**کلمات کلیدی:** استحکام دشمنانه، هوش محاسباتی ترکیبی، بهینه‌سازی مورچگان، یادگیری تقویتی، دفاع چندوجهی، یادگیری میان‌حوزه‌ای.