EAI Endorsed Transactions
on Scalable Information Systems

Mixed Bayesian Stackelberg Strategies for Robust Adversarial Classifiers

Hakeem Quadri[1,*]

[1]Victoria University Melbourne, Australia.

Abstract

Deep neural networks (DNNs) have achieved state-of-the-art performance in classification tasks; however, they are susceptible to small perturbations that are seemingly imperceptible to the human eye but are enough to fool the network into misclassifying images. To develop more robust DNNs against adversarial attacks, research methods have focused on exploring the interaction between a machine learning classifier and a single adversary. However, these methods do not adequately model the real-world scenarios in which these classifiers are deployed. In this research paper, we address this gap and propose an adversarial learning algorithm with multiple adversaries using Bayesian Stackelberg games to model the interaction between the learner and multiple adversaries. We conclude that the nested Bayesian Stackelberg method is a useful strategy for developing adversarial learning algorithms to improve the robustness of DNNs. This strategy can serve as a benchmark in future defense attempts to create DNNs that resist adversarial attacks.

Keywords: Convolution neural networks (CNN), Game theory, Stackelberg games, Mixed strategies, Adversarial Training

Received on 23 10 2024; accepted on 30 11 2024; published on 4 12 2024

doi:10.4108/eetsis.7635

1. Introduction

With the rapid development of deep learning and artificial intelligence, ensuring the robustness of machine learning classifiers and algorithms against adversarial attacks has become increasingly important [1–4]. The vulnerability of machine learning classifiers to adversarial samples has attracted significant interest from both the machine learning and security communities, raising growing security concerns about the use of machine learning in everyday life [5? –8]. Adversarial attacks can easily mislead a classification model into misclassifying an image with high confidence by introducing small perturbations [9–12]. These perturbations are imperceptible to humans but are sufficient for the model to classify the images incorrectly [13–17].

Common adversarial defense methods include adversarial training, where neural networks are trained with perturbed adversarial samples to improve the robustness of the network [18–21]. This defense enhances robustness against adversarial samples but at the cost of lower accuracy for the network. For instance, the advanced adversarial training algorithm of [22] yielded a CNN with 89% accuracy on the CIFAR-10 dataset; however, standard training easily yields a non-robust network with 96% accuracy on the same dataset [13]. Since attacks do not always occur, it is intuitive for a defender with prior information about the adversaries to mix strategies over a distribution of neural networks rather than select a pure strategy that is only beneficial during an attack. In such domains as object classification, a defender typically has a set of standard and pre-trained CNNs that minimize the classification loss on the input datasets, while the adversary has a set of strategies for optimizing the perturbation vector used to transform the dataset during an attack [23–26]. Thus, a defender who can infer prior knowledge about the adversaries can incorporate this knowledge to derive high-rewarding optimal mixed strategies without compromising robustness [27–32].

Despite the demonstrated efficacy of the Bayesian Stackelberg game for deriving optimal mixed strategies against sophisticated adversaries in adversarial settings[33], significant challenges remain in adapting these strategies to dynamically changing adversary behaviors and evolving attack methodologies. Current adversarial training models, predominantly assume a static nature of adversary strategies or a limited scope of variability, which may not adequately reflect the complexity of real-world scenarios where adversaries continually adapt and optimize their attack vectors. Therefore, this research involves developing a more

*Corresponding author. Email: hakeem.quadri@live.vu.edu.au

flexible and adaptive adversarial training in a Bayesian Stackelberg game model. Such a model needs to efficiently incorporate prior knowledge about adversary actions and strategies, enabling the defender to dynamically adjust their mixed strategies. This adaptive model should not only maintain robustness against known types of attacks but also recalibrate in response to evolving threats, thereby ensuring sustained effectiveness of the defense mechanisms in more uncertain adversarial environments.

Game theory provides a mathematical framework that guides the search for the optimal strategies players can adopt in a two or more-player game. Many existing game theory approaches to adversarial learning focus on the interaction between the adversary and the machine learning algorithm to improve robustness against attacks [34–37]. However, these methods do not adequately reflect the real-world scenarios in which these algorithms are deployed [38–41]. This paper makes the following contributions:

(i) Introduces a framework where a defender (learner) responds to multiple intelligent adversaries, each equipped with diverse attack strategies. This framework is specifically tailored to address intelligent opponents effectively.

(ii) Shows that the Bayesian Stackelberg equilibrium model can successfully find an optimal mixed strategy, especially valuable when the defender lacks complete information about the adversaries in real-world scenarios.

(iii) Validates empirically that employing a mixed strategy, which integrates various defensive strategies, provides a significant advantage in dealing with unknown or diverse types of adversaries.

(iv) Focuses on solving payoff matrices for both defender and adversaries, emphasizing accuracy and classification errors, contributing to a refined understanding of strategy optimization within the game framework.

(v) Derives an optimal mixed strategy by modeling the interactions between the defender and adversaries as a Bayesian Stackelberg game, enabling the defender to effectively switch between strategies, such as Convolutional Neural Network (CNN) models.

(vi) The derived optimal mixed strategy enhances the robustness of the defender against both targeted and perturbation attacks, improving the security and reliability of models under adversarial conditions.

Hence, we propose a game theory framework using Bayesian Stackelberg games that models the interaction between a single defender and multiple adversaries. By leveraging prior knowledge, this framework aims to obtain a high-rewarding mixed strategy for a defender uncertain about the type of adversary it may encounter.

2. Related works

Previous works have shown that conventional methods of training may not be sufficient to guarantee the robustness of CNN algorithms [42–45]. Methods such as data augmentation only provide partial solutions to misclassifications [14]. Goodfellow et al. used empirical methods to demonstrate that dimensionality and image complexity impact a classifier's robustness against adversarial attacks in the real world. Hence, adversarial learning is essential for the development of CNN algorithms that are less susceptible to practical attack methods [46–49]. Our study focuses on using adversarial training in a Stackelberg game to find a mixed equilibrium strategy that guarantees optimal accuracy and robustness for a CNN with fixed dimensions [29][28].

Game theory has been used in numerous works to model the interaction between a classifier and adversarial attacks to obtain optimal robust strategies.

To optimize a learner's defense mechanism for resilience towards adversarial attacks, it is important to understand how the attacks are developed [22]. The essence of adversarial data generation is to understand different methods for which adversarial data can be created by a potential adversary [30]. The adversary aims to perturb a valid data sample such that the perturbation is imperceptible to the human eye, but when presented to the machine learner, the data is misclassified to a wrong class [50]. This is achieved by adding just enough perturbation to cross the decision boundary of the learner classifier [51–54]. If the value of the perturbation is too large, the data becomes distorted and nonsensical to the human eye and becomes obviously perturbed. Also, if the perturbation is too small, the data looks normal to the human but is not enough to cross the decision boundary and would not lead to misclassification by the learner [55–58]. Carlini et al. [13] proposed a technique that added a small vector to an input of a model such that the magnitude of the vector is equal to the sign of the gradients of the cost function of the model, which reliably causes a wide variety of classifiers to misclassify their input. The technique showed that by training the model with the worst-case adversarial perturbation rather than itself helps to regularize the model and generally makes it perform better even under adversarial attacks. Goodfellow et al. [29] proposed the fast gradient method (FGSM) to generate perturbations

that are added to examples. The work highlighted the importance of the direction of the gradient of the cost function in deriving appropriate perturbations. Madry et al. [28] investigated the robustness of neural networks through min-max optimization with Projected Gradient Descent (PGD). The min-max formulation reflects adversarial training and attacks against constrained optimization models.

To obtain optimal strategies, attack models need to be defined explicitly. There is no single learning strategy that can be unilaterally implemented for all attack models [30]. Current neural networks and defenses are only effective against a few attacks, keeping the models vulnerable to other types of attacks [59–62]. Indeed, there is a trade-off between accuracy and robustness in the implementation of defense against adversarial samples [28, 63, 64]. The large number of scenarios of attacks and metrics such as L_0, L_1, L_2, and L_∞ makes it difficult to generalize defenses since different levels of perturbations result in varying attack sensitivity and resulting adversarial accuracy. Therefore, there is a need for algorithms that generalize well over multiple attacks without trading off accuracy for the robustness of the network.

Grosse et al. [30] model a machine learning scenario as an interaction between a learner and an adversary. The learner's objective is to correctly predict the input data, while the adversary transforms the data to make the learner misclassify them to a wrong label or output. Adversarial learning presents a considerable level of cybersecurity threat in the domains of machine learning classifiers, including automated email spam filters, image classification algorithms for self-driving cars, medical imaging applications, etc [65, 66]. Kantarcioglu et al. solved a classification problem using Stackelberg equilibrium with a simulated annealing algorithm to obtain an optimal set of attributes. Fiez et al. [32] also conducted similar work, but rather than assuming both players knew one another's payoff function, they showed that it's enough to know only the adversary's payoff function. Both works modeled the adversary as the leader who stochastically chooses his strategy, while the classifier is the follower and searches for an equilibrium after observing the adversary's choice. Madry et al. [28] investigated the robustness of machine learning classifiers through robust optimization of mini-max theoretical frameworks. The optimization method reflected the essence of adversarial training and attack methods against constrained optimization [67, 68].

2.1. Preliminaries

Given a classifier $f_\theta : \mathcal{X} \longrightarrow \mathcal{Y}$ and a dataset $(x_i, y_i)_{i=1}^N \in \mathcal{X} \times \mathcal{Y}$, the adversary finds a perturbation d that changes x from its original class to adversarial data, yet the changes on the adversarial data $x^{'}$ are imperceptible to the human eye. This action is called an adversarial attack. To ensure the attack is undetectable, the adversary constrains the perturbation within a defined budget $\epsilon > 0$ in a boundary ball around x such that $B_\epsilon(x) = \{x^{'} : d(x, x_i) \leq \epsilon\}$. While the classifier is pre-trained on x by reducing the empirical loss function $\ell(x, y; \theta)$, the adversary aims to increase the classifier's loss on the adversarial data $x^{'}$.

2.2. Game theory perspective

In this game, the defender is the row player and the adversary is the column player. q denotes the defender's strategies consisting of a vector of pure strategies, in this case, a pre-trained model and an adversarially trained model. The value of q_i is the proportion of time the defender uses the strategy i in their set q. Similarly, p denotes the vector of possible strategies deployed by the adversary. Q and P represent the sets of both the adversary's and defender's pure strategies. The payoff matrices D and R are defined such that D_{ij} represents the accuracy of the classifier and R_{ij} is the misclassification rate of the classifier when the defender chooses a classifier q_i and the adversary deploys an attack j. Given an adversary, the defender maximizes their payoff by selecting the optimal classifier to attack p_j as follows:

$$\max \sum_{q \in Q} \sum_{p \in P} D_{ij} p_i q_j \tag{1}$$

$$s.t \sum_{q \in Q} q_i = 1$$

The objective function maximizes the defender's expected payoff given q, while the constraints ensure a mixed strategy j for the defender. The adversary maximizes their payoff function given the policy q of the defender by selecting a pure strategy p_j in response. The adversary solves the following objective function.

$$\max \sum_{p \in P} \sum_{q \in Q} R_{ij} q_i p_j \tag{2}$$

$$s.t \sum_{p \in P} p_j = 1$$

2.3. Stackelberg game

Similar to adversarial training , the defender solves its objective function to minimise the empirical loss for a classifier $q \in Q$ which is either pre-trained on natural data x or retrained on adversarial data $x^{'}$ depending on the strategy $p \in P$ deployed by the adversary. The solution for the set of strategies $q \in Q$ converge to an equilibrium that minimizes the expectation of

adversarial loss on the dataset. Q denotes the set of possible strategies by the defender as shown

$$Q = \begin{cases} \min_\theta \frac{1}{n} \sum_{i=1}^{n} (l(f_\theta(x_i), y_i)) \\ \min_\theta \frac{1}{n} \sum_{i=1}^{n} \left\{ \max_{x_i^{'} \in B_\epsilon[x_i]} l(f_\theta(x_i^{'}), y_i) \right\} \end{cases} \tag{3}$$

The classifier $q \in Q$ selected by the defender updates its learning parameters θ to the minimising the adversarial loss across all data points to improve accuracy. The adversary aiming to increase the loss or mis-classification rate of the selected classifier, perturbs the natural data $(x_i, y_i)_{i=1}^{N}$. To achieve the attack, the adversary finds an optimal pure strategy $p \in P, p_j = \{x^{'} : x + \delta\}$ which is the best response to θ that maximizes the loss. The maximum perturbation δ is derived using projected gradient descent (PGD) algorithm [28][33].

$$P := \sum_{i=1}^{n} \max_{x_i^{'} \in B_\epsilon(x_i, \delta)} \left(l(f_\theta(x_i^{'}), y_i) \right) \tag{4}$$

where $B_\epsilon(x_i, \delta) := \{x_i^{'} : d(x_i, x_i^{'}) \leq \epsilon\}$ denotes the ϵ-ball around x_i. The adversary selects a best response pure strategy q_j that guarantees a high payoff after observing the defenders selection .

In a Stackelberg game the defender seeks a mixed strategy of q that maximizes his payoff, given that the adversary selects an optimal response $p\,(q)$, hence the defender solves the following optimization

$$\max_q \sum_{q \in Q} \sum_{p \in P} D_{ij} p\,(q)\, q_i \tag{5}$$

$$s.t \sum_{q \in Q} q_i = 1$$

$$q_i \in [0...1]$$

$$p_j \in \{0, 1\}$$

2.4. Payoff for the Defender and adversary

A Bayesian Stackelberg game models the interaction between a defender and multiple adversaries, where the defender only knows the prior probabilities p of the different types of adversaries $t \in T$. The prior probability that an adversary of type t will appear is p^t, while the probability of encountering another type of adversary is $1 - p^t$. We assume that each adversary t has two attack strategies: a selective strategy p_1 that focuses solely on the impact of adversarial data x' on the classifier selected by the defender, and a universal strategy p_2 that targets the overall accuracy of the attack on both natural x and adversarial data x'.

With the PGD attack, we can model a range of attack types by varying k to adjust the strength of the attack. A small k value results in a small perturbation corresponding to a weak attack, while a large k value leads to a larger perturbation, indicating a stronger attack. The payoff of strategy p_1 is the classification error caused by the perturbed data D' on the classifier $q \in Q$. Thus, for a classifier q_i with accuracy A on dataset D', the payoff R of adversary t_n using the selective strategy p_1 is given as

$$R_1 = 1 - A \tag{6}$$

The universal adversary strategy p_2 also attacks a classifier using varying values of k in the PGD attack. However, p_2 considers both the classification error of the selected classifier $q \in Q$ by the defender on adversarial data D' and natural data D. The intuition behind this is that the more a classifier is adapted to the adversarial data D', the less accurately it predicts on the natural data D. For instance, a classifier retrained on D' will be less accurate on D because the distribution of the datasets varies due to the perturbations added to D'. Hence, along with the classification error on D', strategy p_2 also accounts for the classification error of the pre-trained classifier on D. The payoff R of strategy p_2 for a classifier q_i selected by the defender, given that the accuracy of q_i on dataset D is $A_{q_i}(D)$, is given by

$$R_2 = A_{q_i}(D, D') = 2 - (A_{q_i}(D) + A_{q_i}(D')) \tag{7}$$

An adversary $t \in T$ changes the value of k in the projected gradient descent (PGD) attack to vary the intensity of the attack. A small value of k yields a small perturbation δ, and vice versa. Therefore, a spectrum of adversary types can be specified, ranging from least aggressive to most aggressive. Using the payoff matrices of the classifier and the adversaries, a single defender with T possible types of adversaries can be modeled using decomposed multiple integral linear programming to obtain an optimal strategy for the leader as follows:

2.5. Mixed Bayesian Strategy for Multiple Adversaries

When multiple types of adversaries are considered in adversarial training, the adversary chooses an optimal pure strategy after observing the defender's strategy. This formulation can be solved using Bayesian Stackelberg Equilibrium. The defender's strategy Q is a vector probability distribution of the defender's pure strategies q, where q_i represents the proportion of times strategy i is used. Q^t denotes the vector of strategies for adversary type $t \in T$, and the corresponding payoffs for the adversary and defender are given as D_{ij}^t and R_{ij}^t, respectively. M is a large constant, and r^t is the upper bound corresponding to the highest payoff obtainable by the adversary.

$$\max_{q,p,r} \sum_{q\in Q}\sum_{t\in T}\sum_{j\in J} p^t D_{ij} p_i j_j^t \tag{8}$$

$$s.t \sum_{q\in Q} p_i = 1$$

$$\sum_{q\in Q} p_j^t = 1$$

$$0 \le \left(r^t - \sum_{p\in P} A_{ij}^t q_1 \right) \le \left(1 - p_i^t\right) M$$

$$q_i \in [0...1]$$

$$p_j \in \{0,1\}$$

$$r^t \in \mathbb{R}$$

The prior probability of the occurrence of an adversary type t is denoted by p^t. p_i denotes the probability that the defender selects a mixed strategy i. p_j^t represents the probability that the adversary of type t adopts a pure strategy. Constraints 1 and 4 enforce a feasible mixed strategy for the defender, while constraints 2 and 5 enforce a feasible pure strategy for the adversary. Constraint 3 ensures the feasibility of the adversary's problem by guaranteeing an optimal pure strategy with a maximum payoff of $a = \sum_{q\in Q} R_{ij} p_i$ when $p^t = 1$. The quadratic programming problem in (8) can be linearized by combining the terms $p_i q_j^t$ such that $z_{ij}^t = p_i q_j^t$, leading to the following equations [33].

$$\max_{q,p,r} \sum_{q\in Q}\sum_{t\in T}\sum_{j\in J} p^t D_{ij} z_{ij}^t \tag{9}$$

$$s.t \sum_{q\in Q}\sum_{p\in P} z_{ij}^t = 1$$

$$\sum_{p\in P} z_{ij}^t \le 1$$

$$\sum_{q\in Q} p_i = 1$$

$$0 \le \left(r^t - \sum_{p\in P} A_{ij}^t \left(\sum_{p\in P} z_{ij}^t \right) \right) \le \left(1 - p_i^t\right) M$$

$$\sum_{p\in P} z_{ij}^t = \sum_{p\in P} z_{ij}^1$$

$$z_{ij}^t \in [0...1]$$

$$p_j \in \{0,1\}$$

$$r^t \in \mathbb{R}$$

3. Experiment

3.1. Discussion

In this experiment, we use the CIFAR-10 dataset as the test data to be perturbed by the adversary and evaluate the impact of adversarial attacks on four different CNN classifiers: MobileNet, ResNet, VGG13BN, and ShuffleNet. The original CIFAR-10 dataset is evaluated on each of the pre-trained models to obtain the initial accuracy A of the models. The perturbations added to the natural dataset are derived using the Projected Gradient Descent (PGD) algorithm, with varying k values to adjust the strength of the attack. A higher value of k corresponds to a higher attack strength, and vice versa. The attack algorithm takes in the natural dataset and returns adversarial datasets generated with respect to the corresponding pre-trained model and bounded by epsilon ϵ. The pre-trained models are then evaluated with the generated adversarial dataset to observe the accuracy A_k of the models after the PGD attack, which is lower than the initial accuracy A, as shown in Table 1. Using adversarial training, the pre-trained models are retrained to obtain models robust to perturbed adversarial data. The accuracy results show a significant improvement from the pre-trained models. The accuracy $A_k^{'}$ of the retrained models is also shown in Table 1. The accuracy of the model decreases with the strength of the PGD attack, which can be varied by changing the value of k. Increasing the value of k in the PGD algorithm produces more perturbed CIFAR-10 datasets, leading to more misclassifications of the pre-trained models. For the pre-trained ResNet-53 model, the accuracy reduced from 94.24% to 10.24% with a PGD k value ranging from 1 to 7 ($k = \{1, 3, 5, 7\}$). Similarly, ShuffleNetv2, MobileNetv2, and VGG13BN also show reduced accuracy as k increases, as depicted in Figure 2.

To observe the impact of adversarial data on the robust retrained model, the retrained model is evaluated on the natural dataset. We find that the accuracy $A_k^{'}$ of the retrained model on the natural dataset is significantly lower than the accuracy of the pre-trained model on the natural dataset.

We performed experiments on four pre-trained classifiers: MobileNet, ResNet56, VGG13BN, and ShuffleNet. Using the PGD attack, we modeled two pairs of attacks: a mild adversarial perturbation and a strong perturbation attack, corresponding to a weak attacker g and a stronger attacker G by varying the k value in the PGD algorithm. The pairs of attacks represent the adversary type; a lower value of k denotes a weak adversary g, while a higher value of k denotes a stronger adversary G. In an attack scenario, adversary type t_1, which is the weak adversary g, will have a lower k value compared to attack t_2, which is the stronger adversary G. In addition to these, each adversary has

Table 1. Mixed Bayesian Stackelberg Accuracy A^* for Multiple Adversary Types $k = (1, 3)$

Models	A	k	$A_k\%$	$A_k^{'}\%$	$A^*_{min}\%$	$A^*_{max}\%$
vgg13_bn	94.24	1	53.41	17.93	38.05	43.81
		3	14.38	16.97	–	
mobilenetv2_x1_4	93.88	1	52.31	10.54	43.29	46.39
		3	21.72	10.78	–	
shufflenetv2_x2_0	93.63	1	53.76	20.80	39.20	43.08
		3	15.14	20.09	–	
ResNet-56	94.46	1	54.06	32.00	45.84	48.69
		3	22.93	32.32	–	

Table 2. Mixed Bayesian Stackelberg Accuracy A^* for Multiple Adversary Types $k = (5, 7)$

Models	A	k	$A_k\%$	$A_k^{'}\%$	$A^*_{min}\%$	$A^*_{max}\%$
vgg13_bn	94.24	5	25.17	15.66	35.96	37.59
		7	10.24	17.58	–	
mobilenetv2_x1_4	93.88	5	32.37	10.67	40.99	42.23
		7	16.94	10.72	–	
shufflenetv2_x2_0	93.63	5	25.96	20.77	37.06	38.56
		7	10.69	20.25	–	
ResNet-56	94.46	5	33.50	34.53	44.66	46.14
		7	17.85	33.65	–	

two strategies to choose from to maximize their payoff. The payoff for each strategy is derived from Equations (6) and (7) to confront a defender that chooses between deploying a pre-trained or retrained model.

As an illustration, a defender deploys a pre-trained model with an accuracy of 94.24% on the CIFAR-10 dataset. After an adversary uses PGD with $k = 1$ to perturb the dataset, the pre-trained model's accuracy drops to 53.41%. However, by using adversarial training to retrain the pre-trained model on the perturbed dataset, the accuracy improves from the previous 53% to 63%. On evaluating the retrained model on the original CIFAR-10 dataset, we observe that even though the retrained model has improved accuracy on the adversarial data, its accuracy on the original data dropped to 17.93%. The accuracy of the retrained model facing an adversary t_2 with $k = 5$ is even lower. The adversarial training accuracy is 46.92%, while the retrained accuracy on CIFAR-10 is 16.97%.

To obtain a model that performs well on both natural and adversarial datasets, a mixed Bayesian Stackelberg algorithm is employed. The problem is modeled with two types of adversaries using two different strategies: a global strategy and a direct strategy. The payoffs for both adversary strategies are given by Equations (6) and (7). The optimal mixed strategy of the defender is obtained by solving the mixed integer quadratic equation (9) and the corresponding accuracy payoff. The goal is to develop a randomized classifier selection strategy such that the adversary cannot deploy a perturbed dataset to undermine the accuracy of the selected classifier. The relationship between the

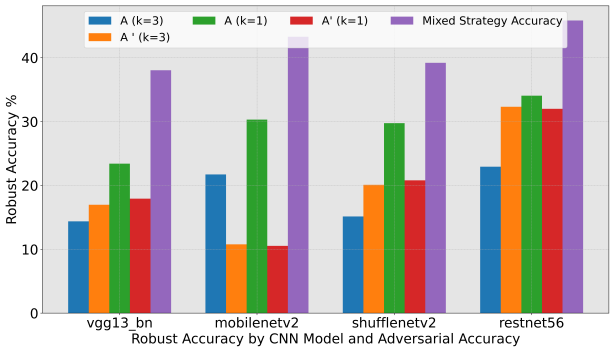

Figure 1. Robust Accuracy for CNN Models Considering Adversary Types $k = (1, 3)$

defender and the adversary is framed as a Bayesian Stackelberg game consisting of t adversary types, $1, \ldots, t$. The defender's set of pure strategies includes two CNN models: a pre-trained model and a retrained model. The defender can choose a mixed strategy such that the adversary is uncertain about which CNN model is being deployed, although the adversary may be aware of the mixed strategy the defender is implementing. For instance, the adversary can observe how often each CNN model is deployed over time and then select an attack strategy that guarantees maximum impact. The adversary will receive a lower payoff if it uses a direct attack targeted at a pre-trained model while the defender deploys a retrained model. Conversely, the

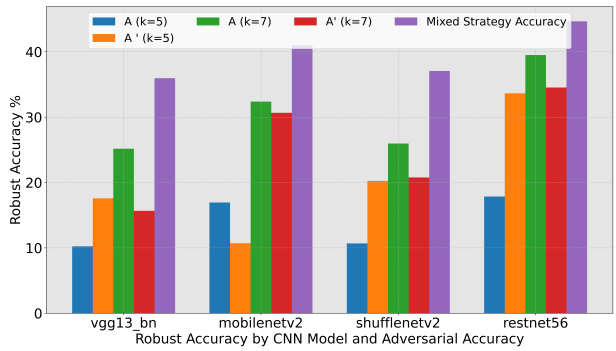

Figure 2. Robust Accuracy for CNN Models Considering Adversary Types $k = (5, 7)$

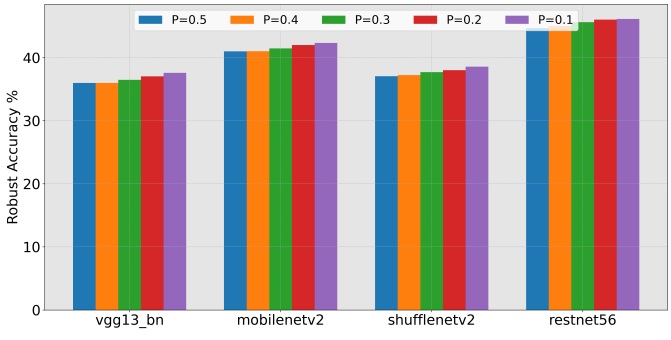

Figure 3. Accuracy of CNN Models based on the Prior Probability of Adversary Type

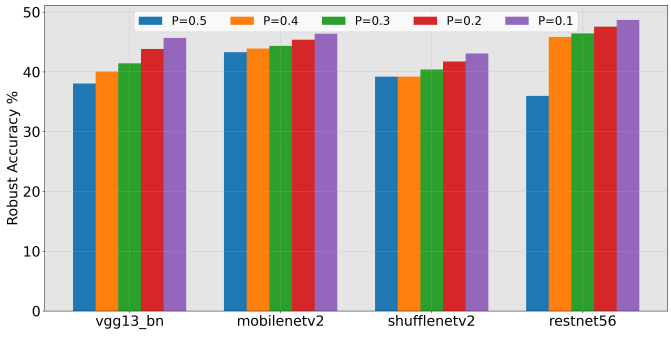

Figure 4. Accuracy of CNN Models based on the Prior Probability of Adversary Type $k = (5, 7)$

adversary will achieve a higher payoff if it uses the global attack while the defender chooses a pre-trained model.

To reconcile the effect of the significant reduction in accuracy, the Bayesian Stackelberg algorithm finds a mixed strategy, as shown in Fig. 1, for the defender. This strategy ensures that the accuracy after retraining

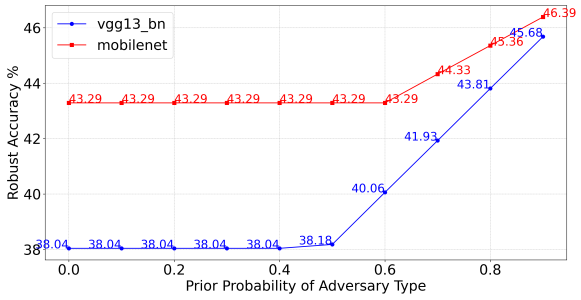

Figure 5. Accuracy of CNN Models based on the Prior Probability of Adversary Type $k = (1, 3)$

the model is consistently better than the accuracy of the pre-trained model when attacked by the strongest adversary, and also better than the accuracy of the retrained model on the original CIFAR-10 dataset. The pre-trained VGG13BN model experienced the highest impact from adversarial attacks, with a notable reduction in accuracy after perturbation for both $k = 3$ and $k = 7$. Figure 1 shows that the pre-trained accuracy A_k and the retrained accuracy $A_k^{'}$ after the attack are 25.17% and 15.66% for $k = 3$, respectively, and even lower, at 10.24% and 17.58% for $k = 7$, as shown in Fig. 2. However, the mixed strategy for the defender, which combines both pre-trained and retrained models, achieves an accuracy of 35.96% as shown in Fig. 1. Similar results are observed for MobileNetV2, ShuffleNetV2, and ResNet-56.

Before committing to a mixed strategy, the defender considers the prior probability P of encountering either type of adversary. With varying probabilities P that a strong adversary G may not appear, the defender only begins to see a notable increase in accuracy when there is at least 60% certainty that they will confront a weaker adversary g, as shown in Fig. 7. This indicates that, with the knowledge that the models are more susceptible to a strong attack, the mixed strategy accuracy is conservative and only improves when there is a higher likelihood that a strong attack will not occur. As shown in Fig. 5 and Fig. 6, the knowledge of the prior probability of an adversary type perturbing the dataset also affects the accuracy achieved by the mixed strategy implemented by the defender. Intuitively, a higher prior probability of a weak adversary g perturbing the dataset, as opposed to a stronger adversary G, results in higher accuracy from the mixed strategy. Conversely, if there is a higher probability that the adversary is stronger, the resulting accuracy from selecting the mixed Bayesian Stackelberg strategy will be lower.

4. Conclusion

In this paper, we develop a Bayesian Stackelberg game in which one of the players, a defender (learner), responds to the actions of other players (adversaries), who are multiple intelligent opponents. The defender directly searches for an optimal, high-rewarding strategy given prior knowledge of the adversaries. Our research demonstrates the effectiveness of the Bayesian Stackelberg equilibrium model in obtaining an optimal mixed strategy when confronted with adversaries, each having multiple attack strategies. Our approach empirically shows that the mixed strategy is the best solution when the defender is unaware of the type of adversaries it may encounter in real-world applications.

The Bayesian Stackelberg game formulation centers on solving the payoff matrices of the defender and adversary strategies for accuracy and classification errors. We derive an optimal mixed strategy by formulating the interaction between the defender and adversaries as a Bayesian Stackelberg game. The solution enables the defender to mix strategies more effectively between CNN models and exhibit increased robustness to targeted and perturbation attacks.

References

[1] M. Gupta and R. K. Dwivedi, "Blockchain- based secure and efficient scheme for medical data," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 5, 6 2023.

[2] X. Sun, H. Wang, J. Li, and J. Pei, "Publishing anonymous survey rating data," *Data Mining and Knowledge Discovery*, vol. 23, pp. 379–406, 11 2011.

[3] J. Yin, M. Tang, J. Cao, and H. Wang, "Apply transfer learning to cybersecurity: Predicting exploitability of vulnerabilities by description," *Knowledge-Based Systems*, vol. 210, 10 2020.

[4] H. Wang, Y. Zhang, and J. Cao, "Ubiquitous computing environments and its usage access control," vol. 152, 01 2006, p. 6.

[5] A. Akan and M. Vural, "Just noticeable difference for machines to generate adversarial images," *arXiv Preprint arXiv*, 2020, accepted. Available upon request.

[6] J. Yin, M. Tang, J. Cao, H. Wang, M. You, and Y. Lin, "Vulnerability exploitation time prediction: an integrated framework for dynamic imbalanced learning," *World Wide Web*, pp. 401–423, 01 2022.

[7] J. Zhang, X. Tao, and H. Wang, "Outlier detection from large distributed databases," *World Wide Web*, vol. 17, 07 2014.

[8] E. Kabir and H. Wang, "Conditional purpose based access control model for privacy protection," vol. 92, 01 2009, pp. 137–144.

[9] A. Tripathi and J. Prakash, "Blockchain enabled interpolation based reversible data hiding mechanism for protecting records," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 5, 5 2023.

[10] X. Sun, H. Wang, J. Li, and Y. Zhang, "Injecting purpose and trust into data anonymisation," *Computers Security*, vol. 30, pp. 332–345, 07 2011.

[11] E. Kabir, "A role-involved purpose-based access control model," *Information Systems Frontiers*, vol. 14, pp. 809–822, 07 2012.

[12] L. Sun, J. Ma, H. Wang, and Y. Zhang, "Cloud service description model: An extension of usdl for cloud services," *IEEE Transactions on Services Computing*, vol. PP, pp. 1–1, 08 2015.

[13] N. Carlini and A. A., "Simple black-box adversarial attacks," *arXiv preprint arXiv*, 2019.

[14] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *Proceedings of the IEEE*, 2018.

[15] P. Anay, T. Zhenyi, L. Shuijing, B. Gautham, and C. Girish, "Robust deep reinforcement learning with adversarial attacks," 2017.

[16] S. Siuly, O. Alcin, E. Kabir, A. Sengur, H. Wang, Y. Zhang, and F. Whittaker, "A new framework for automatic detection of patients with mild cognitive impairment using resting-state eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. PP, pp. 1–1, 07 2020.

[17] J.-Y. Li, Z.-H. Zhan, H. Wang, and J. Zhang, "Data-driven evolutionary algorithm with perturbation-based ensemble surrogates," *IEEE Transactions on Cybernetics*, vol. PP, pp. 1–13, 08 2020.

[18] Y.-F. Ge, H. Wang, E. Bertino, Z.-H. Zhan, J. Cao, Y. Zhang, and J. Zhang, "Evolutionary dynamic database partitioning optimization for privacy and utility," *IEEE Transactions on Dependable and Secure Computing*, pp. 1–17, 2023.

[19] C. Wang, B. Sun, K.-J. Du, J.-Y. Li, Z.-H. Zhan, S.-W. Jeon, H. Wang, and J. Zhang, "A novel evolutionary algorithm with column and sub-block local search for sudoku puzzles," *IEEE Transactions on Games*, vol. PP, pp. 1–11, 01 2023.

[20] E. Kabir, A. Mahmood, H. Wang, and A. Mustafa, "Microaggregation sorting framework for k-anonymity statistical disclosure control in cloud computing," *IEEE Transactions on Cloud Computing*, vol. PP, pp. 408–417, 08 2020.

[21] H. Wang, Y. Zhang, J. Cao, and V. Varadharajan, "Achieving secure and flexible m-services through tickets," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 33, pp. 697 – 708, 12 2003.

[22] L. Dritsoula and P. Loiseau, "A game-theoretic analysis of adversarial classification," *IEEE Transactions on Information Forensics and Security*, 2017.

[23] J.-Q. Yang, Q.-T. Yang, K.-J. Du, C.-H. Chen, H. Wang, S.-W. Jeon, J. Zhang, and Z.-H. Zhan, "Bi-directional feature fixation-based particle swarm optimization for large-scale feature selection," *IEEE Transactions on Big Data*, vol. PP, pp. 1–14, 01 2022.

[24] J. Yin, M. Tang, J. Cao, M. You, H. Wang, and M. Alazab, "Knowledge-driven cybersecurity intelligence: Software vulnerability coexploitation behavior discovery," *IEEE Transactions on Industrial Informatics*, vol. PP, pp. 1–9, 01 2022.

[25] J.-Y. Li, K.-J. Du, Z.-H. Zhan, H. Wang, and J. Zhang, "Distributed differential evolution with adaptive resource allocation," *IEEE transactions on cybernetics*, vol. PP, 03 2022.

[26] R. Sarki, K. Ahmed, H. Wang, Y. Zhang, and K. Wang, "Convolutional neural network for multi-class classification of diabetic eye disease," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 9, no. 4, 12 2021.

[27] E. Wong and L. Rice, "Fast is better than free: Revisiting adversarial training," 2020.

[28] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and V. A, "Towards deep learning models resistant to adversarial attacks," 2018.

[29] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:6706414

[30] K. Grosse, D. Pfaff, and M. Smith, "The limitations of model uncertainty in adversarial settings," 2018.

[31] A. Anish, C. Nicholas, and W. David, "Maximum efficiency and output of class-f power amplifiers," *International Conference on Machine Learning (ICML)*, p. 1802.00420, 2018.

[32] T. Fiez, B. Chasnov, and L. Ratliff, "Implicit learning dynamics in stackelberg games: equilibria characterization, convergence analysis, and empirical study," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20. JMLR.org, 2020.

[33] P. Paruchuri, J. P. Pearce, J. Marecki, M. Tambe, F. Ordonez, and S. Kraus, "Playing games for security: an efficient exact algorithm for solving bayesian stackelberg games," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, ser. AAMAS '08. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems, 2008, p. 895–902.

[34] Y.-F. Ge, W.-J. Yu, J. Cao, H. Wang, Z.-H. Zhan, Y. Zhang, and J. Zhang, "Distributed memetic algorithm for outsourced database fragmentation," *IEEE Transactions on Cybernetics*, vol. PP, pp. 1–14, 11 2020.

[35] F. Liu, X. Zhou, J. Cao, Z. Wang, W. Tianben, H. Wang, and Y. Zhang, "Anomaly detection in quasi-periodic time series based on automatic data segmentation and attentional lstm-cnn," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 08 2020.

[36] H. Wang, J. Cao, and Y. Zhang, "A flexible payment scheme and its role-based access control," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, pp. 425–436, 04 2005.

[37] J. Shu, X. Jia, K. YANG, and H. Wang, "Privacy-preserving task recommendation services for crowdsourcing," *IEEE Transactions on Services Computing*, vol. PP, pp. 1–1, 01 2018.

[38] Y. Zhang, Y. Shen, H. Wang, J. Yong, and X. Jiang, "On secure wireless communications for iot under eavesdropper collusion," *IEEE Transactions on Automation Science and Engineering*, vol. 13, pp. 1–13, 12 2015.

[39] K. Cheng, L. Wang, Y. Shen, H. Wang, Y. Wang, X. Jiang, and H. Zhong, "Secure k-nn query on encrypted cloud data with multiple keys," *IEEE Transactions on Big Data*, vol. PP, pp. 1–1, 05 2017.

[40] H. Wang, Y. Zhang, and J. Cao, "Effective collaboration with information sharing in virtual universities," *IEEE Trans. Knowl. Data Eng.*, vol. 21, pp. 840–853, 06 2009.

[41] R. Singh, S. Subramani, J. Du, Y. Zhang, H. Wang, Y. Miao, and K. Ahmed, "Antisocial behavior identification from twitter feeds using traditional machine learning algorithms and deep learning," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 4, p. e17, May 2023. [Online]. Available: https://publications.eai.eu/index.php/sis/article/view/3184

[42] J. Yin, M. Tang, J. Cao, M. You, H. Wang, and M. Alazab, "Knowledge-driven cybersecurity intelligence: Software vulnerability co-exploitation behaviour discovery," *IEEE Transactions on Industrial Informatics*, 2022.

[43] S. Siuly, Alçin, H. Wang, Y. Li, and P. Wen, "Exploring rhythms and channels-based eeg biomarkers for early detection of alzheimer's disease," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. PP, pp. 1–15, 04 2024.

[44] J. Zhang, H. Li, X. Liu, Y. Luo, F. Chen, and H. Wang, "On efficient and robust anonymization for privacy protection on massive streaming categorical information," *IEEE Transactions on Dependable and Secure Computing*, vol. PP, pp. 1–1, 09 2015.

[45] Y.-F. Ge, M. Orlowska, J. Cao, H. Wang, and Y. Zhang, "Mdde: multitasking distributed differential evolution for privacy-preserving database fragmentation," *The VLDB Journal*, vol. 31, pp. 1–19, 01 2022.

[46] Z.-J. Wang, Z.-H. Zhan, Y. Lin, W.-J. Yu, H. Wang, S. Kwong, and J. Zhang, "Automatic niching differential evolution with contour prediction approach for multimodal optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. PP, pp. 1–1, 04 2019.

[47] Y. Zhang, Y. Gong, Y. Gao, H. Wang, and J. Zhang, "Parameter-free voronoi neighborhood for evolutionary multimodal optimization," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 335–349, 2020.

[48] Z. Zhang, L. Teng, M. Zhou, and H. Wang, "Enhanced branch-and-bound framework for a class of sequencing problems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, pp. 1–11, 05 2019.

[49] T. Huang, Y.-J. Gong, S. Kwong, H. Wang, and J. Zhang, "A niching memetic algorithm for multi-solution traveling salesman problem," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 3, pp. 508–522, 2019.

[50] J. Bose and G. Gidel, "Adversarial example games," *Proc NeurIPS*, 2020.

[51] W. Shi, W.-n. Chen, S. Kwong, J. Zhang, H. Wang, G. Tianlong, H. Yuan, and J. Zhang, "A coevolutionary estimation of distribution algorithm for group insurance portfolio," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. PP, pp. 1–15, 07 2021.

[52] A. Alvi, S. Siuly, and H. Wang, "A long short-term memory based framework for early detection of mild cognitive impairment from eeg signals," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. PP, pp. 1–14, 01 2022.

[53] M. N. A. Tawhid, S. Siuly, K. Wang, and H. Wang, "Automatic and efficient framework for identifying

multiple neurological disorders from eeg signals," *IEEE Transactions on Technology and Society*, vol. PP, pp. 1–1, 03 2023.

[54] W.-L. Liu, Y.-J. Gong, W.-n. Chen, Z. Liu, H. Wang, and J. Zhang, "Coordinated charging scheduling of electric vehicles: A mixed-variable differential evolution approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, pp. 1–16, 10 2019.

[55] A. S. Chivukula and X. Yang, "Game theoretical adversarial deep learning with variational adversaries," *in IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 8, pp. pp. 3568–3581, 2021.

[56] Z.-G. Chen, Z.-H. Zhan, H. Wang, and J. Zhang, "Distributed individuals for multiple peaks: A novel differential evolution for multimodal optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. PP, pp. 1–1, 10 2019.

[57] T. Huang, Y.-J. Gong, W.-n. Chen, H. Wang, and J. Zhang, "A probabilistic niching evolutionary computation framework based on binary space partitioning," *IEEE Transactions on Cybernetics*, vol. PP, pp. 1–14, 03 2020.

[58] S. Siuly, S. Khare, V. Bajaj, H. Wang, and Y. Zhang, "A computerized method for automatic detection of schizophrenia using eeg signals," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 1, p. 1, 09 2020.

[59] Y. Zhang, Y. Shen, H. Wang, Y. Zhang, and X. Jiang, "On secure wireless communications for service oriented computing," *IEEE Transactions on Services Computing*, vol. PP, pp. 1–1, 09 2015.

[60] Y. Wang, Y. Shen, H. Wang, J. Cao, and X. Jiang, "Mtmr: Ensuring mapreduce computation integrity with merkle tree-based verifications," *IEEE Transactions on Big Data*, vol. 4, no. 3, pp. 418–431, 2016.

[61] M. Peng, Q. Xie, H. Wang, Y. Zhang, and G. Tian, "Bayesian sparse topical coding," *IEEE Transactions on Knowledge and Data Engineering*, vol. PP, pp. 1–1, 06 2018.

[62] S. Supriya, S. Siuly, H. Wang, and Y. Zhang, "Eeg sleep stages analysis and classification based on weighed complex network features," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. PP, pp. 1–11, 11 2018.

[63] M. Peng, J. Zhu, H. Wang, X. Li, Y. Zhang, X. Zhang, and G. Tian, "Mining event-oriented topics in microblog stream with unsupervised multi-view hierarchical embedding," *ACM Transactions on Knowledge Discovery from Data*, vol. 12, pp. 1–26, 04 2018.

[64] Y.-F. Ge, E. Bertino, H. Wang, J. Cao, and Y. Zhang, "Distributed cooperative coevolution of data publishing privacy and transparency," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, 08 2023.

[65] M. Peng, W. Gao, H. Wang, Y. Zhang, J. Huang, Q. Xie, G. Hu, and G. Tian, "Parallelization of massive textstream compression based on compressed sensing," *ACM Transactions on Information Systems*, vol. 36, pp. 1–18, 08 2017.

[66] J. Ma, L. Sun, H. Wang, Y. Zhang, and U. Aickelin, "Supervised anomaly detection in uncertain pseudoperiodic data streams," *ACM Transactions on Internet Technology*, vol. 16, pp. 1–20, 01 2016.

[67] H. Wang, X. Jiang, and G. Kambourakis, "Special issue on security, privacy and trust in network-based big data," *Information Sciences*, vol. 318, pp. 48–50, 2015.

[68] M. Enamul Kabir, H. Wang, and E. Bertino, "A conditional purpose-based access control model with dynamic roles," *Expert Systems with Applications*, vol. 38, no. 3, pp. 1482–1489, 2011.