SCIENTIFIC Publishing Limited

Journal of Intelligent Communication
https://ojs.ukscip.com/index.php/jic

Article

A Zero-Sum Game-Theoretic Analysis for Cost-Aware Backdoor Attacks and Defenses in Deep Learning

Kassem Kallas [1,2,*], **Carine Tannous** [1] **and Hichem Faraoun** [1]

[1] IMT Atlantique, Inserm UMR 1101, 29200 Brest, France

[2] National Institute of Health and Medical Research, Inserm UMR 1101, 29200 Brest, France

[*] Correspondence: kassem.kallas@imt-atlantique.fr

Received: 2 July 2025; **Revised:** 12 August 2025; **Accepted:** 18 August 2025; **Published:** 4 September 2025

Abstract: Backdoor attacks pose a critical and increasingly realistic security threat to deep neural networks (DNNs), enabling adversaries to implant hidden behaviors that remain dormant under normal conditions while preserving high performance on benign data. Although numerous defenses have been proposed, most works treat the interaction between attackers and defenders in isolation, without a principled mechanism to analyze their strategic interplay under realistic resource constraints. This paper introduces BG_{Cost}, a zero-sum game-theoretic framework that formalizes backdoor attack–defense dynamics with explicit cost-aware utility functions. The attacker seeks to maximize Attack Success Rate (ASR) while maintaining Clean Data Accuracy (CDA) above an acceptance threshold to remain stealthy, whereas the defender aims to limit ASR and preserve CDA while minimizing the computational and accuracy costs induced by mitigation. By embedding resource consumption directly into the utilities of both players, BG_{Cost} provides a structured benchmark to study equilibrium strategies across unconstrained, balanced, and high-cost operational regimes. Through numerical simulations, we show that cost-aware game modeling fundamentally alters equilibrium behavior: unconstrained settings drive extreme strategies, costly defenses weaken robustness, costly attacks suppress adversarial impact, and balanced configurations yield deployment-friendly equilibria with low ASR and high CDA. Rather than proposing a new algorithmic defense, BG_{Cost} serves as a decision-theoretic tool that complements existing mechanisms by revealing how cost constraints shape optimal attacker–defender behavior in practice, guiding the design of realistic and resource-efficient protections against backdoor threats.

Keywords: Adversarial Machine Learning; Backdoor Attacks; Backdoor Defenses; Game Theory; Deep Neural Networks; AI Security; Attack-Defense Strategies

1. Introduction

Deep Neural Networks (DNNs) have recently demonstrated outstanding performance across diverse and critical domains, such as computer vision [1], autonomous driving [2], finance [3], healthcare [4], and many other fields [5]. This rapid expansion, however, has increased concerns regarding their security. At each phase of the DNN lifecycle—from data collection and preprocessing to architecture design, training, and deployment—adversaries may exploit weaknesses [6,7] to undermine reliability. A well-known example is adversarial examples, where carefully crafted inputs at test time can cause models to misclassify [8]. Threats are not limited to inference, as attackers may also compromise the training process. Moreover, the substantial computational needs and large-scale data requirements of DNN training, coupled with limited expertise, often drive practitioners to rely on third-party resources such as Machine Learning as a Service (MLaaS) or pre-trained models [9]. While these solutions increase accessibility and

efficiency, they simultaneously reduce user control and introduce new security risks [8].

Backdoor attacks pose a critical threat to the security of DNN. These attacks involve training a model to respond to a specific trigger pattern by giving incorrect outputs [10,11], such that during inference, any input containing the trigger is misclassified accordingly as illustrated in illustrated in **Figure 1**. The attacker's goal is twofold: first, to ensure that the compromised model performs normally on benign samples to maintain high performance and evade detection; and second, to induce misclassification whenever a trigger is present. Backdoor injection can occur at various stages prior to inference [8], including poisoning the training data [12], altering model parameters during training [13], or manipulating the model at deployment time [14]. Moreover, transfer learning introduces vulnerability, as pre-trained models can inherit backdoors when fine-tuned on new tasks [15].

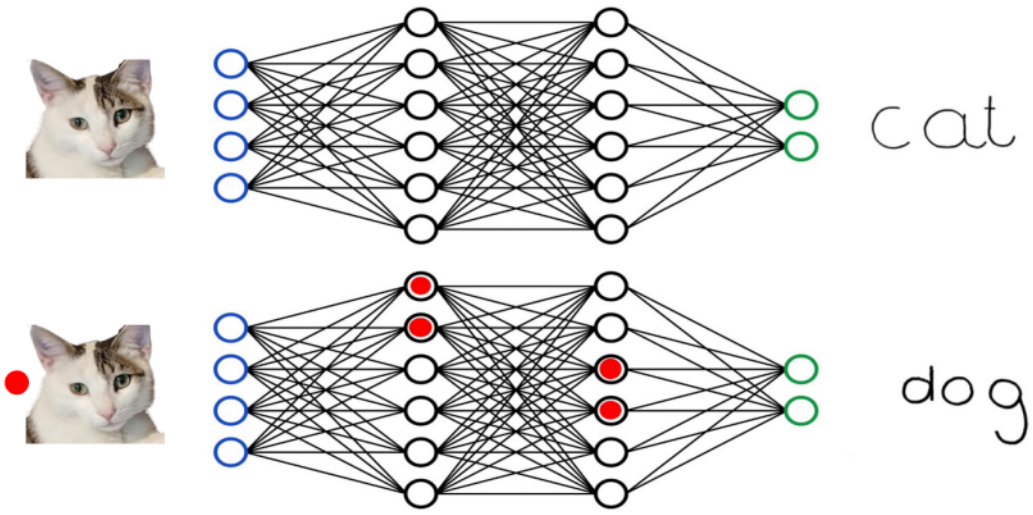


Figure 1. Illustration of backdoor behavior: A clean input is accurately identified as a "cat," but after the addition of a trigger (e.g., a "red circle"), the model incorrectly labels it as a "dog".

A common strategy for launching backdoor attacks is training data poisoning [12,16], where manipulated samples containing a trigger pattern are injected into an otherwise benign dataset, causing the model to learn incorrect associations between the poisoned samples and their target. Depending on the attack design, the labels of poisoned samples may be modified [17] (poison-label attacks) or left intact [16,18] (clean-label attacks), with the latter making detection harder during dataset inspection. Such attacks have been demonstrated across multiple domains [8,19], including NLP [20], audio [21], and computer vision [8,22]. Beyond this distinction, backdoor methods vary widely, spanning class-agnostic and class-specific settings [17], employing diverse trigger types [8], and even adopting properties like transparency [12,23]. Several comprehensive surveys provide systematic overviews of these attacks and their defenses [7,8,10,11,19].

The evolving threat landscape and the ongoing cat-and-mouse game between backdoor attackers and defenders, characterized by the continual emergence of new attacks and defenses [24], motivate this work. Within the specific context of clean-label backdoor attacks on image classification [8,16,18], we ask the following question: Can the interaction between a DNN backdoor attacker and a defender be modeled as a two-player game, with Nash equilibria identified and each player's performance assessed at equilibrium? Related work such as FLGAME [25] has applied game-theoretic methods to federated learning, offering early insights into how attacker–defender dynamics can be formalized.

Our work takes a different direction by focusing on centralized learning and explicitly embedding cost constraints into the game-theoretic framework. This perspective enables a structured analysis of how resource limitations shape adversarial behavior, while retaining the adversarial rigor of a zero-sum formulation. In earlier work [26], Kallas et al. modeled the interaction as a strict zero-sum game, where the attacker's gain equaled the defender's loss. Yet this simplifying assumption overlooks realistic cases where both players incur costs simultaneously—for instance, when defensive measures degrade benign accuracy or when aggressive attacks demand significant computational resources and increase detection risk.

Within this framework, the defender's objective is to maximize Clean Data Accuracy (CDA) while minimizing the Attack Success Rate (ASR), carefully weighing the cost of deploying defensive measures. Conversely, the attacker aims to maximize ASR while ensuring CDA remains above a detection threshold, balancing this goal against the computational cost of injecting and triggering backdoors. By incorporating cost constraints, we refine the strategic interplay between the players, preserving the adversarial nature of the game while introducing more realistic considerations of resource limitations.

Our approach maintains the simplicity of a two-player game and the same set of strategies introduced in our previous work. By embedding cost considerations directly into the utility functions while preserving the zero-sum structure, we examine how these trade-offs influence optimal strategies and game equilibria. Through numerical simulations, we demonstrate the impact of cost-aware decision-making and analyze attacker-defender interactions under realistic resource constraints.

Contributions of this paper include:

1. Introducing a zero-sum game-theoretic framework for DNN backdoor attacks and defenses with cost constraints.
2. Embedding cost constraints symmetrically in attacker and defender utility functions to model computational trade-offs.
3. Analyzing the equilibrium strategies under cost-aware conditions using numerical simulations.

This work advances the understanding of adversarial interactions in DNNs, providing a structured approach to identifying optimal strategies for both attackers and defenders while integrating real-world constraints into decision-making. To make our objectives more explicit, the goal of this paper is to investigate how cost constraints alter the dynamics of backdoor attacks and defenses in deep learning. We propose the following hypothesis.

H1. *Cost-aware modeling significantly changes equilibrium strategies compared to unconstrained settings.*

H2. *Balanced cost scenarios yield the most practical trade-off between security and performance.*

H3. *Extreme cost conditions (high attack or high defense costs) reduce the effectiveness of one player while creating vulnerabilities exploitable by the other.*

These hypotheses guide our analysis and are examined through the numerical simulations presented in Section 5.

2. Backdoor Attack

This section introduces the backdoor attack model considered in this work: a targeted, clean-label, data-poisoning attack in an image classification setting. We provide the rationale for this choice, formally define the attack, and describe the adversarial strategies and countermeasures.

2.1. Motivation for Our Threat Model

Backdoor attacks have become a major security threat in deep learning, leading to extensive research on their techniques and countermeasures [7,8,10,11,19]. In this work, we consider the common case of an attacker compromising a supervised learning model trained for image classification, a setting widely investigated in the literature. Early demonstrations, such as BadNets [17], revealed how easily poisoned samples with hidden triggers could corrupt classification tasks, while later studies highlighted their impact on sensitive domains like face recognition and other vision-based applications [7].

The focus here is on targeted backdoor attacks, where the adversary forces the model to consistently misclassify inputs containing a trigger into a specific target class [17]. This differs from untargeted poisoning, which instead seeks to cause arbitrary misclassifications or broadly reduce model performance, similar to Byzantine attacks [8]. Among the many possible strategies, we concentrate on data poisoning-based backdoors, where the training dataset is manipulated to implant malicious behaviors. Such manipulation can occur at different points in the pipeline, including data collection, dataset preparation, or when relying on third-party sources [8,11,17].

Additionally, we assume a clean-label backdoor setting [18], meaning that while the attacker manipulates the

input images, the corresponding class labels remain unchanged. This makes the attack highly stealthy, as conventional dataset validation techniques focus primarily on detecting label inconsistencies, and particularly dangerous, as clean-label backdoors maintain high model performance on benign inputs, which makes them difficult to detect despite embedding harmful behavior. Prior studies [7,8] highlight their effectiveness in compromising deep neural networks while remaining covert.

Our motivation for selecting this threat model lies in its practical relevance to real-world, high-stakes domains such as autonomous driving and biometric authentication [8,10].

2.2. Formalization

A deep neural network (DNN) is defined as a function $\mathcal{F}_\theta$ trained on a dataset $D_{tr} = \{(x_i, y_i)\}_{i=1}^{N_{tr}}$, where $X = \{x_i\}_{i=1}^{N_{tr}}$ represents input images and $Y = \{y_i\}_{i=1}^{N_{tr}} \in C$ denotes class labels. The network parameters θ are optimized by minimizing:

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N_{tr}} \mathcal{L}(\mathcal{F}_\theta(x_i), y_i) \tag{1}$$

The trained model is then evaluated on a test dataset $D_{ts} = \{(x_j, y_j)\}_{j=1}^{N_{ts}}$ and its performance is assessed using Clean Data Accuracy (*CDA*):

$$CDA(\mathcal{F}_\theta, D_{ts}) = \frac{\sum_{j=1}^{N_{ts}} I(x_j, y_j)}{N_{ts}} \tag{2}$$

where $I(x_j, y_j) = 1$ if $\mathcal{F}_\theta(x_j) = y_j$, and 0 otherwise.

A backdoor attack modifies the model such that it misclassifies inputs when a specific trigger x_t is present. The attacker poisons the training data by altering a subset P of m samples from D_{tr}:

$$P = \{(\tilde{x}_i, \tilde{y}_i)\}_{i=1}^{m} \tag{3}$$

$$\tilde{x}_i = (1 - \Delta_{tr}) \times x_i + \Delta_{tr} \times x_t \tag{4}$$

Here, Δ_{tr} represents the backdoor trigger strength, controlling the trigger's visibility and x_t is the trigger. A successful attack increases the Attack Success Rate (ASR) while keeping the model's Clean Data Accuracy (CDA) high to evade detection:

$$ASR(\mathcal{F}_\theta^{po}, D_{ts}^{po}) = \frac{\sum_{j=1}^{|D_{ts}^{po}|} I(\tilde{x}_j, t)}{|D_{ts}^{po}|} \tag{5}$$

$$\begin{aligned} CDA &= (CDA_{cb} + CDA_{cp})/2 \\ ASR &= (ASR_{cb} + ASR_{cp})/2 \end{aligned} \tag{6}$$

We further distinguish two variants of the metrics used throughout the paper. CDA_{cb} denotes clean-data accuracy on the *clean* test set (benign samples), while CDA_{cp} denotes accuracy on the *poisoned* test set (samples containing the trigger) after being cleaned. Similarly, ASR_{cb} and ASR_{cp} are computed on clean and poisoned partitions, respectively. The total *CDA* and *ASR* values are defined as the average of their two components, as shown in Eq. (6), which balances benign and poisoned behavior into a single indicator. This separation and subsequent averaging provide finer insight into how strategies impact both normal performance and attack persistence.

2.3. Summary of Notation

Table 1 summarizes the notation used throughout the paper.

Table 1. Table of Notations.

D_{tr}	Training Dataset
D_{ts}	Testing Dataset
D_{tr}^{po}	Poisoned Training Dataset
D_{ts}^{po}	Poisoned Testing Dataset
$\mathcal{F}_\theta$	Deep Learning Model
$\mathcal{F}_\theta^P$	Backdoored Deep Learning Model
$\mathcal{L}$	Loss Function
C	Number of Classes
CDA	Clean Data Accuracy
ASR	Attack Success Rate
Δ_{tr}	Attack Strength During Training
Δ_{ts}	Attack Strength During Testing
α_{tr}	Fraction of Poisoned Training Samples
α_{def}	Fraction of Samples Processed by Defense
Δ_{def}	Defense Strength
x_t	Backdoor Trigger
BG_{Cost}	Backdoor Game with Cost-Constrained Control
μ_A	Attacker's Utility
μ_D	Defender's Utility
S_A^*	Attacker's Strategy at Equilibrium
S_D^*	Defender's Strategy at Equilibrium
λ_A	Cost Coefficient for Attacker
λ_D	Cost Coefficient for Defender
CA	Attacker's Cost Function
CD	Defender's Cost Function
PA	Attacker's Mixed Strategy Probability Distribution
PD	Defender's Mixed Strategy Probability Distribution
μ_A^*	Attacker's Utility at Equilibrium
μ_D^*	Defender's Utility at Equilibrium
ASR_{cb}	Attack Success Rate on Clean Data
ASR_{cp}	Attack Success Rate on Poisoned Data
CDA_{cb}	Clean Data Accuracy on Clean Data
CDA_{cp}	Clean Data Accuracy on Poisoned Data
$\mathbf{1}[CDA > CDA_{inf}]$	Indicator for CDA Threshold
S_A	Attacker's Strategy Set
S_D	Defender's Strategy Set

2.4. Adversarial Strategies: Attack and Defense

Attacker Strategy. We consider the SIG attack [16], where the backdoor trigger is a sinusoidal or ramp signal embedded into images. The attacker:

- Selects a target class t.
- Applies a structured trigger to a fraction α_{tr} of images.
- Ensures the poisoned dataset D_{tr}^{po} trains the model to associate x_t with class t.

Defender Strategy. We assume a reverse-engineering-based defense that estimates and removes the trigger using:

$$x^{cl} = \frac{\chi_i^{in} - \Delta_{def} \times \hat{x}_t}{1 - \Delta_{def}} \tag{7}$$

Here, $\hat{x}_t$ is the estimated trigger, and Δ_{def} controls the defense strength.

3. Game Theory in a Nutshell

Game theory provides a formal framework for analyzing strategic interactions in which rational decision-makers, or players, select actions to maximize their expected payoffs. Since its introduction by von Neumann and Morgenstern [27], it has been widely applied in domains such as security, economics, and machine learning. The central

premise assumes rational players seeking to optimize their utility, though in practice this can be limited by bounded rationality or other behavioral factors that cause deviations from idealized strategies [28].

3.1. Normal-Form Games

A normal-form game models strategic interactions by defining available strategies and payoffs. A two-player game is given by:

$$G =< S_A, S_D, \mu_A, \mu_D > \tag{8}$$

where:

- S_A and S_D are the strategy sets available to the attacker (A) and defender (D).
- μ_A, μ_D are the utility functions, mapping strategy profiles to numerical payoffs.

In zero-sum games, the gain of one player directly corresponds to the loss of the other, meaning $\mu_D = -\mu_A$. These models are purely adversarial scenarios, such as cybersecurity, where improving an attack's success directly harms the defender's efforts.

3.2. Equilibrium Concepts

A Nash equilibrium [29,30] represents a situation where no player can improve their outcome by changing their strategy unilaterally, assuming the other player's strategy remains fixed. Formally, it satisfies:

$$\begin{aligned} \mu_A(S_A^*, S_D^*) \geq \mu_A(S_A, S_D^*) \ \ \forall S_A \in S_A \\ \mu_D(S_A^*, S_D^*) \geq \mu_D(S_A^*, S_D) \ \ \forall S_D \in S_D \end{aligned} \tag{9}$$

In zero-sum games, the equilibrium corresponds to a saddle point, where neither player can improve their outcome.

There are two primary types of equilibria: - Pure strategy Nash equilibrium: Players choose a single action deterministically. - Mixed strategy Nash equilibrium: Players randomize over actions, balancing unpredictability and optimality.

3.3. Solving Normal-Form Games

A game is dominance solvable [30] if dominated strategies can be iteratively eliminated. More generally, Nash equilibria in two-player zero-sum games satisfy:

$$\max_{s_A} \min_{s_D} u_A(s_A, s_D) = \min_{s_D} \max_{s_A} u_A(s_A, s_D) \tag{10}$$

These can be computed using linear programming or specialized algorithms like Lemke-Howson for bimatrix games.

For mixed strategies, where players randomize over available actions, the expected utility is:

$$\begin{aligned} U_A(P_A, P_D) = \sum_{s_A, s_D} P_A(s_A) u_A(s_A, s_D) P_D(s_D) \\ U_D(P_A, P_D) = \sum_{s_A, s_D} P_A(s_A) u_D(s_A, s_D) P_D(s_D) \end{aligned} \tag{11}$$

Mixed-strategy equilibria ensure optimality when pure strategies lack a stable solution.

3.4. Implications for Adversarial Machine Learning

Game theory provides a structured approach for modeling the interactions between attacker and defender in adversarial machine learning. By formulating backdoor attacks and defenses as a zero-sum game, we can derive optimal strategies for both sides (players). Introducing cost constraints further enhances the realism of the model by incorporating practical trade-offs between attack strength and computational resources.

This game-theoretic foundation underpins the Backdoor Game with Cost-Constrained Control (BG_{Cost}), extending previous work [26] to incorporate practical computational constraints.

4. Zero-Sum Cost-Constrained Backdoor Game

4.1. Overview

We frame the interaction between a backdoor attacker and a DNN defender as a two-player zero-sum game with cost constraints, where both players are assumed to act rationally with full knowledge of the game's structure, but without certainty about the other player's selected strategy. In contrast to earlier formulations that emphasized only adversarial objectives, our model explicitly incorporates *cost-aware decision-making*, requiring both attacker and defender to weigh resource limitations when optimizing their strategies.

Within this framework, the defender's objective is to preserve clean data accuracy (CDA) while suppressing the attack success rate (ASR), yet must do so under the burden of computational overhead and the risk of accuracy degradation caused by defensive measures. Conversely, the attacker seeks to maximize ASR while keeping CDA above a rejection threshold to remain stealthy, and simultaneously minimize the computational expense of poisoning data or deploying triggers. These competing trade-offs mirror real-world scenarios, where stronger attacks or defenses inevitably demand greater resources and carry higher risks.

Following the different scenarios presented by Kallas et al. [26], we adopt the Backdoor Game with Maximum Control (BG Max) as our starting framework and extend it to the Backdoor Game with Cost-Constrained Control (BG_{Cost}). This new formulation captures real-world limitations where both attackers and defenders must optimize not only for effectiveness but also for computational efficiency. By embedding cost considerations directly into the utility functions, our model better reflects the strategic balancing acts encountered in practical adversarial environments.

Despite introducing cost constraints, the game remains zero-sum because:

- The fundamental adversarial structure is preserved, where an increase in ASR corresponds to a proportional decrease in CDA.
- Costs are symmetrically embedded into both players' utility functions, ensuring that constraints do not introduce asymmetry.
- Optimal strategies at equilibrium remain dictated by competitive interactions, with both players adapting to resource limitations while maintaining opposing objectives.

Thus, the BG_{Cost} framework extends the previous maximum control scenario while embedding practical constraints, offering a structured and realistic model for analyzing attack-defense interactions under computational and operational limitations.

4.2. Formal Definition of the Game

The Backdoor Game with Cost-Constrained Control (BG_{Cost}) is formulated as a two-player zero-sum game, where the attacker (A) and the defender (D) engage in a strategic interaction defined as follows:

$$BG_{Cost} =< S_A, S_D, \mu_A, \mu_D > \tag{12}$$

where:

- S_A and S_D are the strategy spaces of the attacker and defender, respectively.
- μ_A: $S_A \times S_D \rightarrow \mathbb{R}$ is the attacker's utility function.
- μ_D : $S_A \times S_D \rightarrow \mathbb{R}$ is the defender's utility function, where $\mu_D = -\mu_A$, ensuring the zero-sum property.

The strategy spaces are defined as:

$$S_A = (\alpha_{tr}, \Delta_{tr}, \Delta_{ts}) \in [0,1]^3, S_D = \Delta_{def} \in [0,1] \tag{13}$$

The attacker's strategies include adjusting the poisoning ratio α_{tr} during training and controlling the trigger strengths Δ_{tr} and Δ_{ts}. The defender, in response, can adjust their defense strength Δ_{def}.

4.3. Constructing a Utility Function

To integrate cost constraints while preserving the zero-sum structure, we define cost-constrained utility functions for both players.

Attacker's Utility:

$$\mu_A = ASR \times \mathbf{1}[CDA > CDA_{inf}] - \lambda_A C_A(\alpha_{tr}, \Delta_{tr}, \Delta_{ts}) - \lambda_D C_D(\Delta_{def}) \tag{14}$$

where:

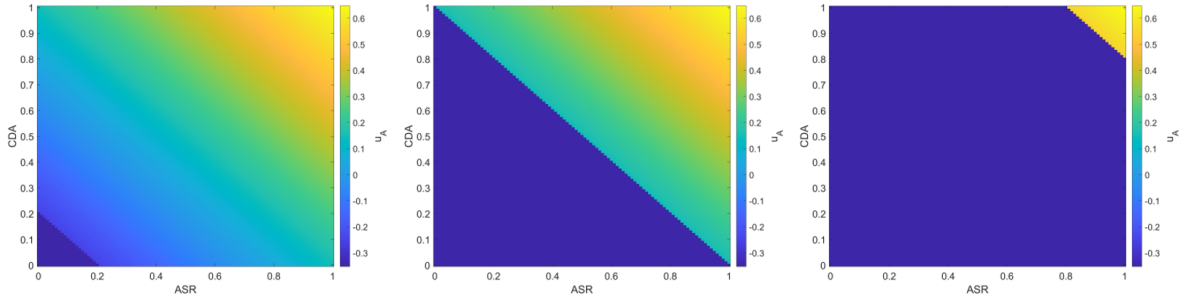
- ASR is the attack success rate.
- $\mathbf{1}[CDA > CDA_{\text{inf}}]$ ensures that the attacker's success is counted only if the defender does not reject the model.
- $C_A(\alpha_{tr}, \Delta_{tr}, \Delta_{ts}) = \alpha_{tr}\ (\Delta_{tr}^2 + \Delta_{ts}^2)$ represents the attacker's computational cost, increasing quadratically with trigger strengths and linearly with poisoning ratio.
- $C_D\ (\Delta_{def}) = \log(1 + \Delta_{def})$ represents the defender's cost, increasing logarithmically as the defense intensity grows.
- λ_A and λ_D are weight parameters that control the influence of cost constraints on the attack and defense, respectively.

Defender's Utility:

$$\mu_D = -\mu_A \tag{15}$$

It ensures that the game remains strictly zero-sum.

Figure 2 shows the attacker's utility under the balanced BG_{Cost} setup with $\lambda_A = \lambda_D = 1.0$. Lower thresholds (e.g., $CDA_{\text{inf}} = 0.1$) leave the attacker with a broad region of positive utility, while higher thresholds (e.g., $CDA_{\text{inf}} = 0.9$) restrict this region considerably. This shows how selecting a proper acceptance criteria CDA_{inf} by the defender can directly constrain the attacker's effective strategy space.



(**a**) $CDA_{\text{inf}} = 0.1$. (**b**) $CDA_{\text{inf}} = 0.5$. (**c**) $CDA_{\text{inf}} = 0.9$.

Figure 2. Balanced BG_{Cost} utilities for different CDA_{inf} under $\lambda_A = \lambda_D = 1.0$.

4.4. Cost Constraints and Game Dynamics

The introduction of cost constraints preserves the zero-sum structure of the game while compelling both players to strategically allocate their resources. In our formulation, the coefficients λ_A and λ_D weight decision-theoretic trade-offs inside the utilities: λ_A captures the attacker's budget pressure (e.g., how costly it is to increase poisoning ratio or trigger strength while staying stealthy), and λ_D captures the defender's budget pressure (e.g., how costly it is

to raise defense intensity without harming benign accuracy). Importantly, these costs are *not* direct measurements of runtime or memory; rather, they are design weights that shape optimal strategies under resource awareness within the game.

Key cost drivers reflected in the utilities include:

- **Computational effort (attack side):** Higher poisoning rates and stronger (Δ_{tr}, Δ_{ts}) imply greater effort/risk for the attacker, which is penalized via CA($\cdot$) and scaled by λ_A.
- **Training/processing effort (defense side):** Stronger defenses (larger Δ_{def} and coverage α_{def}) are penalized via CD ($\cdot$) and scaled by λ_D , capturing the budgetary impact of more intensive mitigation.
- **Energy/operational footprint:** Both intense attacks and defenses imply higher operational burden; the cost terms encourage budget-aware choices when such burden matters.

Our utilities are evaluated using the same training/evaluation pipeline as the underlying model; BG_{Cost} introduces no additional asymptotic time/space complexity beyond sweeping the strategy grid. Empirical profiling of wall-clock time or peak memory is therefore orthogonal to the game definition and can vary with hardware and implementation. Inference latency and memory are not explicitly modeled here; instead, **Table 2** provides representative settings for (λ_A, λ_D) that let practitioners encode their own operational constraints. By embedding these weights in a structured zero-sum framework, BG_{Cost} offers a resource-aware perspective on adversarial interactions, guiding strategy selection without prescribing a specific runtime or memory budget.

Table 2. Cost Constraint Configurations and Their Effects.

Scenario	λ_A (Attack Cost)	λ_D (Defense Cost)	Expected Behavior
Unconstrained	0.1–0.5	0.1–0.5	Minimal cost impact, aggressive strategies.
Balanced	0.5–1.5	0.5–1.5	Trade-off between performance and cost.
Costly Attacks	1.0–3.0	0.1–0.5	Attackers favor low-cost triggers.
Costly Defenses	0.1–0.5	1.0–3.0	Defenders avoid expensive defenses.
High Constraints	2.0–5.0	2.0–5.0	Both players optimize for low-cost strategies.

5. Simulation Results and Discussion

The experimental framework used in this study is presented in this section. Beginning with the dataset and model architecture, followed by a formal definition of the game setup. We then analyze the resulting utility matrices and discuss the equilibrium strategies that emerge. A summary of the key findings obtained from the simulations is presented at the end.

5.1. Dataset and Models

To explore strategic interactions in backdoor attacks within our game-theoretic framework, we rely on the MNIST dataset in our simulations, as its controlled environment makes it ideal for clearly evaluating both attack and defense strategies.

Our experimental setup employs a shallow convolutional neural network (CNN) architecture composed of four main components: a first convolutional layer with 64 filters followed by max pooling, a second convolutional layer with 128 filters and max pooling, a fully connected layer with 256 neurons, and an output layer of 10 neurons. Both convolutional layers use kernels of size 5 with ReLU activations. In each game scenario and for each strategy profile, the CNN is trained for 100 epochs with a batch size of 64. Following training, utilities are derived from the observed clean data accuracy (CDA) and attack success rate (ASR), with each value contributing to the entries of the utility matrix. For reference, the baseline test accuracy of the benign model $\mathcal{F}_\theta$—in the absence of any attack—reaches 99.07%.

5.2. Game Setup

As explained in Section 4, the BFcost framework models the interactions between attackers and defenders while accounting for the cost of their actions. The attacker controls the parameters Δ_{tr} and Δ_{ts}, which define the trigger strength during training and testing, respectively. The defender, in turn, controls Δ_{def}, determining the in-

tensity of defensive countermeasures. Unlike previous game setups, BG_{Cost} explicitly integrates the cost associated with each player's actions, influencing their strategy selection and optimal choices.

Table 3 presents the key parameters in the strategy set for BG_{Cost}, along with their respective value ranges, which define the space in which both players optimize their strategies while balancing performance and resource costs. To ensure a consistent evaluation of how varying cost constraints influence equilibrium strategies and game outcomes, we analyze different cost scenarios by setting λ_A and λ_D to the average values of their respective intervals for each scenario, as shown in **Table 2**. Additionally, for each game instance, we quantize the attacker's and defender's overlay powers to ensure a structured evaluation of strategic choices. The maximum overlay power Δ_{tr} is empirically selected based on the highest achievable ASR in the absence of defense. This enables precise analysis of how cost constraints shape decision-making in adversarial settings, revealing optimal strategies under various budget conditions.

Table 3. Parameters and Value Ranges in BG_{Cost} Strategy Set.

Player	Parameter	Range
Attacker	Δ_{tr}	{0.01,⋯, 0.09} ∪ {0. 1,⋯, 0.5}
Attacker	Δ_{ts}	{0.01,⋯, 0.09} ∪ {0. 1,⋯, 0.5}
Attacker	α_{tr}	{0.05, 0. 1,⋯, 0.9, 1.0}
Defender	Δ_{def}	{0.01,⋯, 0.09} ∪ {0. 1,⋯, 0.5}
Defender	α_{def}	{0.05, 0. 1,⋯, 0.9, 1.0}

5.3. Analysis of the Utility Matrices

Unconstrained Game: In the Unconstrained Game scenario (**Figure 3**), where the cost parameters are set to λ_A = 0.3 and λ_D = 0.3, both players operate with minimal cost constraints, allowing for greater strategic flexibility. The utility matrix reveals distinct patterns, with dark blue zones appearing at the bottom and in the rightmost columns where α_{def} > 0.8, indicating low utility for the attacker and a strong advantage for the defender. Column-wise analysis shows that as Δ_{def} increases, the defender's utility improves, particularly at Δ_{def} = 0.5, explaining the vertical dark blue lines that highlight the likelihood of high Δ_{def} values being chosen at equilibrium. Conversely, yellow zones, which signify high attacker utility, are concentrated in rows corresponding to medium-to-high values of α_{tr} and Δ_{tr}, suggesting that the attacker prefers these strategies in pure or mixed strategy equilibria.

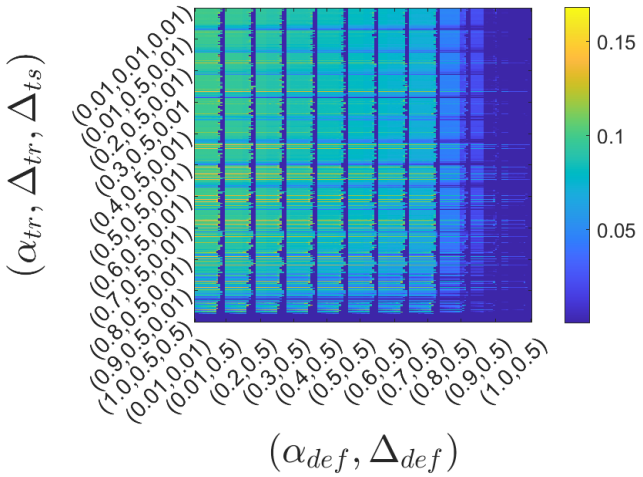


Figure 3. BG_{Cost} Unconstrained: μ_A and sine trigger.

This equilibrium behavior happens because the low-cost setting encourages more aggressive play from both sides, attacker and defender. The attacker can increase α_{tr} and Δ_{tr} without significant penalties, leading to stronger backdoor strategies. Meanwhile, the defender responds by leveraging high Δ_{def}, attempting to neutralize the attack while minimizing damage to benign samples. As a result, both sides adopt strategies, making intense interactions more likely at equilibrium.

Balanced Game: In the Balanced Game scenario (**Figure 4**), where the cost parameters are set to λ_A = 1.0 and λ_D = 1.0, the introduction of cost constraints leads to a significant expansion of the dark blue zones in the utility matrix. This shows that both players prefer to balance performance and cost, making them less likely to choose extreme strategies. Compared to the unconstrained game (**Figure 3**), the attacker's aggressive strategies are considerably reduced, as shown by the reduced presence of yellow areas, which correspond to high utility for the attacker. This suggests that the attacker now favors lower values of α_{tr} and Δ_{tr}, ensuring that the backdoor remains effective while minimizing costs.

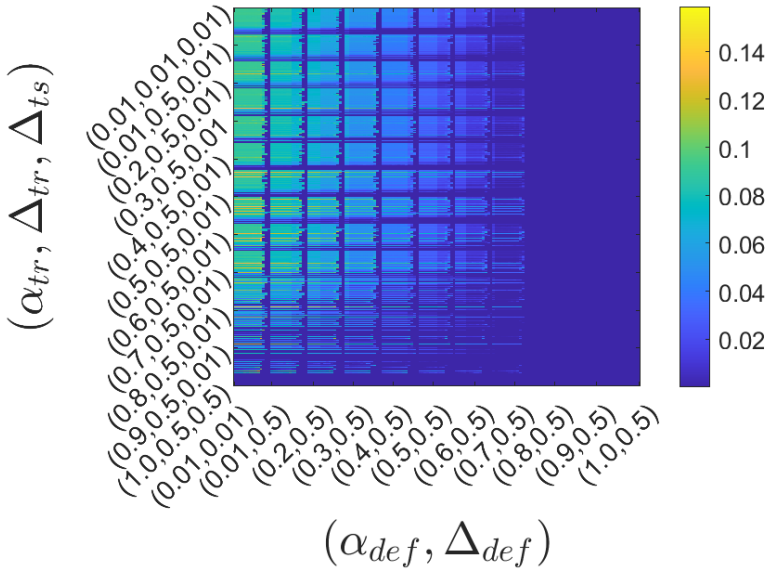


Figure 4. BG_{Cost} Balanced: μ_A and sine trigger.

On the defender's side, the strategy becomes more calculated yet remains proactive. The defender exhibits higher α_{def} and Δ_{def} values in comparison to the attacker's relatively conservative strategy, as evidenced by the dark blue regions concentrated at higher defense levels. This suggests that while the defender does not fully maximize their defensive efforts due to cost considerations, they still adopt relatively strong countermeasures to maintain control over ASR. The equilibrium in this setting reflects a more structured strategic balance, where both players avoid extreme moves and instead aim for cost-efficient yet effective strategies that align with their objectives without overspending resources.

Costly Attacks Game: In the Costly Attacks scenario (**Figure 5**), where attack costs are high (λ_A = 2.0) and defense costs are low (λ_D = 0.3), the attacker is forced to adopt low-cost strategies, meaning minimal values for α_{tr} , Δ_{tr}, and Δ_{ts} to reduce resource expenditure. This causes the attacker to adopt weaker strategies, which are less effective but more sustainable under the cost constraints. Meanwhile, the defender, benefiting from a low-cost defense environment, takes advantage of the attacker's reduced aggression by deploying stronger countermeasures, reflected in higher values of α_{def} and Δ_{def}. This strategic imbalance is evident in the expanded yellow zones in the utility matrix, which indicate that the attacker's most viable strategies now lie in low-value profiles, while the defender freely strengthens its defenses without significant cost.

Costly Defenses Game: Conversely, in the Costly Defenses scenario (**Figure 6**), where λ_A = 0.3 and λ_D = 2.0, the roles are reversed. The defender, now facing significant cost constraints, must limit the intensity of its defensive actions, resulting in lower values of α_{def} and Δ_{def}. The weakened defense creates an opening for the attacker, who, while still favoring low-cost strategies, faces less resistance and thus maintains a minimal but persistent level of effectiveness. This contrast between costly attacks and costly defenses underscores the fundamental impact of resource constraints on strategic decision-making: when attacks are costly, the attacker weakens, allowing the defender to dominate, whereas when defenses are costly, the defender is forced to scale back, giving the attacker more room to operate.

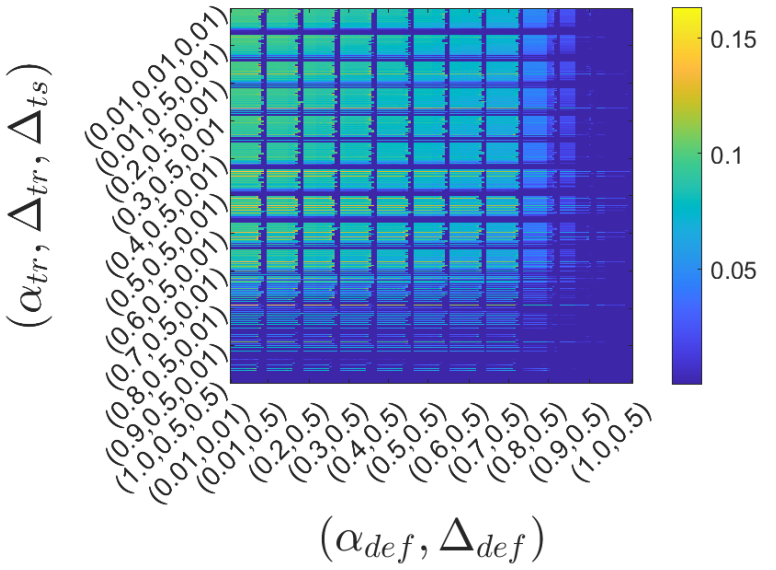


Figure 5. BG_{Cost} Costly Attacks: μ_A and sine trigger.

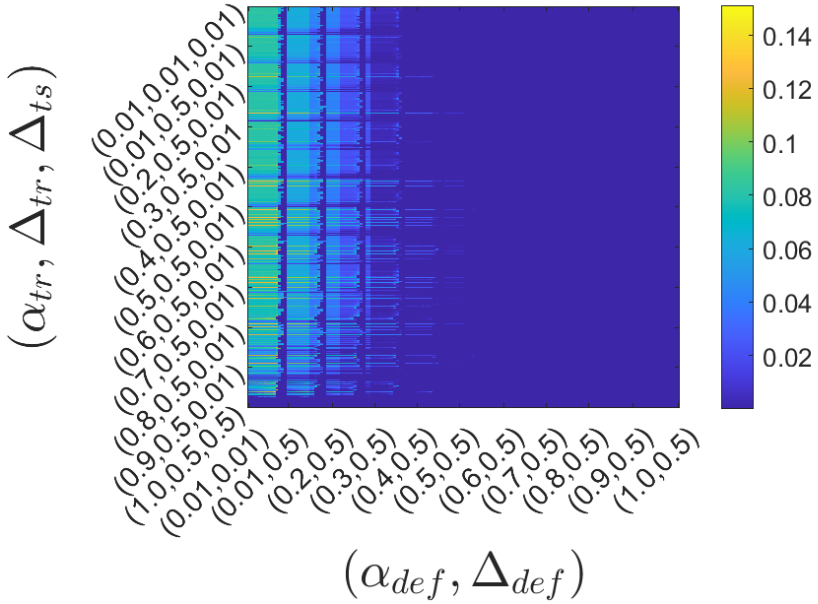


Figure 6. BG_{Cost} Costly Defenses: μ_A and sine trigger.

High Constraints Game: Finally, in the High Constraints scenario (**Figure 7**), both the attacker and defender face significant cost restrictions, with $\lambda_A = 2.0$ and $\lambda_D = 2.0$, forcing them to prioritize low-cost strategies to maximize their respective utilities, as indicated in Equations 14 and 15. This cost-driven limitation explains the prevalence of dark blue zones in the utility matrix, indicating low attacker utility and relatively higher defender utility. Since high-cost strategies are no longer viable, both players adopt a conservative approach, with the attacker favoring lower values of α_{tr} , Δ_{tr}, and Δ_{ts} , while the defender reduces its defense intensity by decreasing either α_{def}, Δ_{def}, or both. The result is a fragile equilibrium, where neither the attacker nor the defender can assert significant influence, leading to weak attacks met by equally weak defenses—a situation that minimizes extreme outcomes but compromises overall model robustness.

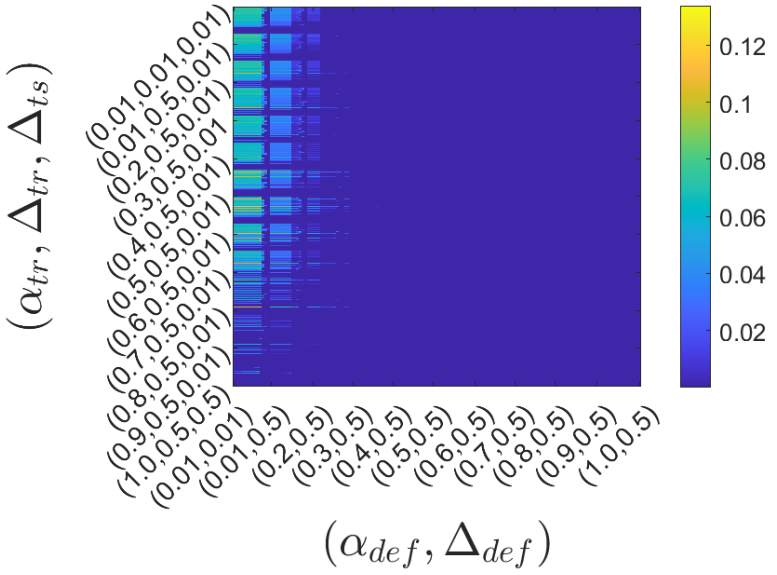
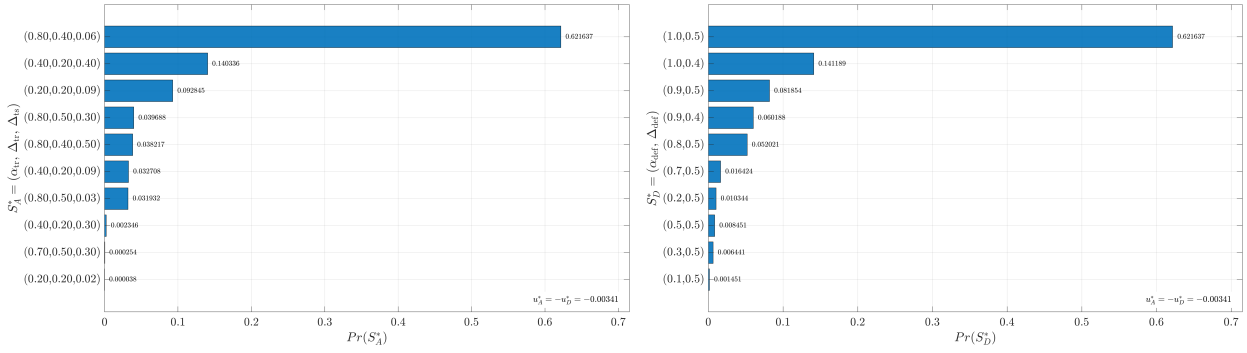


Figure 7. BG_{Cost} High Constraints: μ_A and sine trigger.

5.4. Analysis of Equilibrium Strategies

5.4.1. Unconstrained Game

With minimal cost penalties ($\lambda_A = \lambda_D = 0.3$), both attacker and defender can pursue aggressive strategies. The attacker distributes probabilities across medium–high poisoning ratios (α_{tr}, Δ_{tr}) and adapts the test-time trigger strength Δ_{ts} to trade stealth for reliability. Notably, as shown in **Figure 8**, the strategy (0.8, 0.4, 0.06) carries the largest probability mass (0.621637), suggesting a preference for stealthy yet effective configurations. The defender responds with strong countermeasures, concentrating on (1.0, 0.5) with the same high probability (0.621637). This equilibrium suppresses ASR but does so at the expense of benign performance: CDA = 0.892 with CDA_{cb} = 0.893 and CDA_{cp} = 0.892 (**Table 4**). While attacks are partially neutralized, the drop in CDA makes deployment unattractive since defensive intensity harms model reliability.



(a) Attacker **(b)** Defender

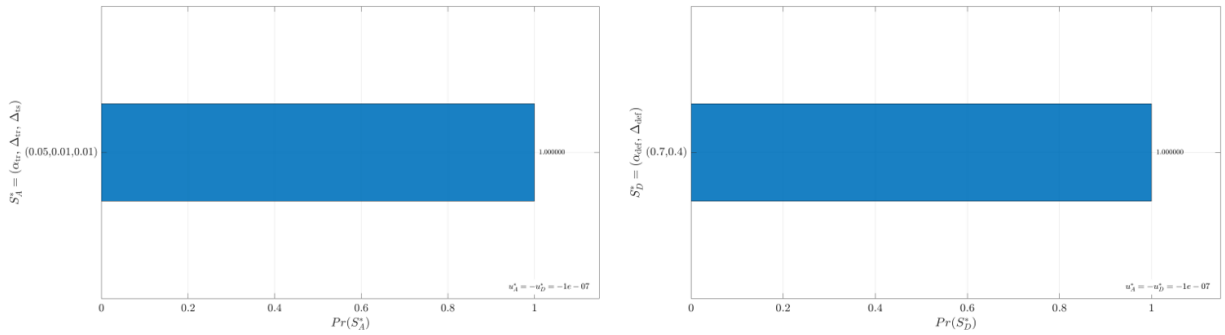
Figure 8. Mixed Strategy Equilibrium for Unconstrained BG_{Cost} with Sin Trigger.

5.4.2. Balanced Game

Under symmetric constraints ($\lambda_A = \lambda_D = 1.0$), players converge to cost-efficient pure strategies. The attacker consistently adopts (0.05, 0.01, 0.01), while the defender chooses (0.7, 0.4), both with probability 1.0, as illustrated in **Figure 9**. This equilibrium is favorable: the ASR remains low (0.109) while CDA is preserved at 0.937, outperforming all other cases. The outcome aligns with the heatmap analysis and highlights this configuration as the most deployment-friendly, achieving strong mitigation without excessive accuracy loss.

Table 4. Performance at the Equilibrium for Different BG_{Cost} Cases.

Metric	Unconstr.	Balanced	Costly Att.	Costly Def.	High Const.
ASR_{cb}	0.134	0.101	0.170	0.102	0.086
ASR_{cp}	0.130	0.117	0.101	0.123	0.112
CDA_{cb}	0.893	0.960	0.795	0.844	0.786
CDA_{cp}	0.892	0.914	0.870	0.798	0.807
ASR	0.132	0.109	0.135	0.113	0.099
CDA	0.892	0.937	0.832	0.821	0.796

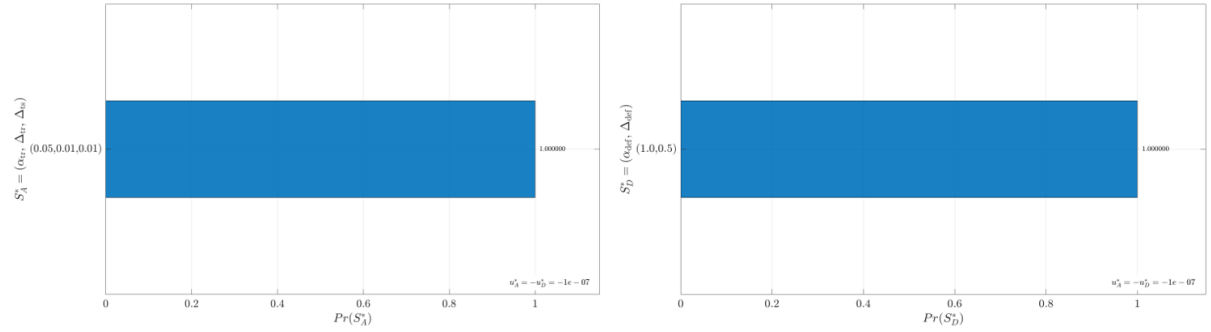


(**a**) Attacker (**b**) Defender

Figure 9. Mixed Strategy Equilibrium for Balanced BG_{Cost} with Sin Trigger.

5.4.3. Costly Attacks Game

When attacks are expensive (λ_A high) and defenses remain affordable (λ_D low), the attacker is forced to retreat to the weakest configuration (0.05, 0.01, 0.01) with probability 1.0, as shown in **Figure 10**. The defender seizes this opportunity, escalating to (1.0, 0.5) with probability 1.0 (see **Figure 10**). This dynamic reduces ASR to 0.135, but the aggressive defense significantly degrades CDA, which falls to 0.832 overall (CDA_{cb} = 0.795). The result demonstrates that excessive defense can backfire by harming benign accuracy; more calibrated defenses would strike a better balance between robustness and usability.

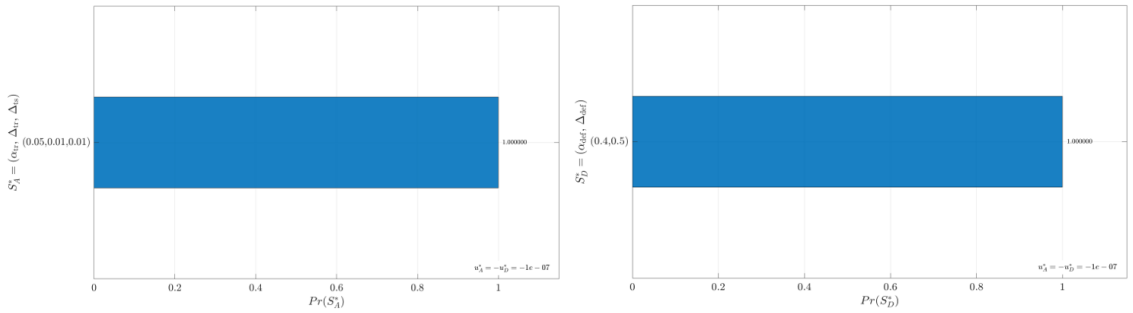


(**a**) Attacker (**b**) Defender

Figure 10. Mixed Strategy Equilibrium for Costly Attacks BG_{Cost} with Sin Trigger.

5.4.4. Costly Defenses Game

Here, the defender faces high costs (λ_D large), restricting their options. As presented in **Figure 11**, the equilibrium settles on (0.4, 0.5) with probability 1.0, a moderate defense that avoids over-expenditure. The attacker, unchanged from the costly attack case, sticks to (0.05, 0.01, 0.01). The outcome is modest: ASR = 0.113 and CDA = 0.821. While not catastrophic, this case reflects the limitations imposed by expensive defenses: the system avoids collapse but cannot achieve strong guarantees, emphasizing the need for lightweight and adaptive defense strategies.

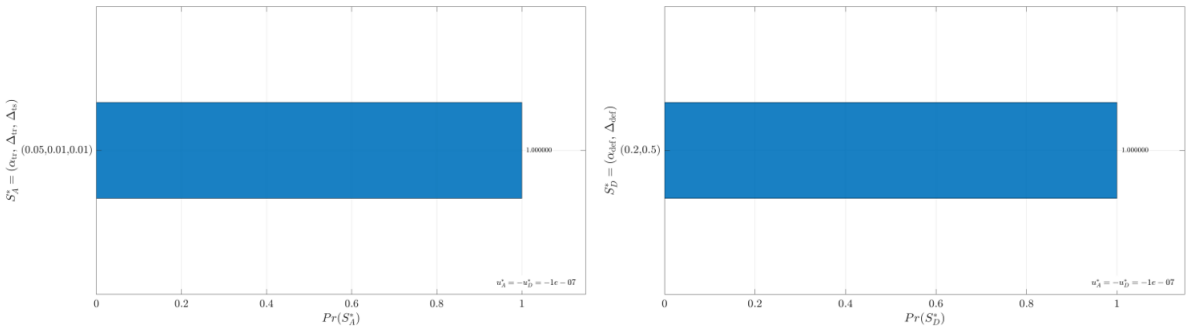


(**a**) Attacker (**b**) Defender

Figure 11. Mixed Strategy Equilibrium for Costly Defenses BG_{Cost} with Sin Trigger.

5.4.5. High Constraints Game

Finally, under strong cost pressure on both sides ($\lambda_A = \lambda_D$ large), neither player can afford high-intensity strategies. Both settle on minimal-effort moves: the attacker uses (0.05, 0.01, 0.01), while the defender employs (0.2, 0.5), each with probability 1.0, as can be seen in **Figure12**. This equilibrium results in the lowest ASR (0.099) but also the weakest CDA (0.796). The low investment produces a fragile equilibrium: attacks are limited, but so is defense, leaving the model simultaneously underperforming and underprotected (**Figure 13**, **Table 4**). This configuration is the least attractive for deployment, as it sacrifices both robustness and accuracy.



(**a**) Attacker (**b**) Defender

Figure 12. Mixed Strategy Equilibrium for High Constraints BG_{Cost} with Sin Trigger.

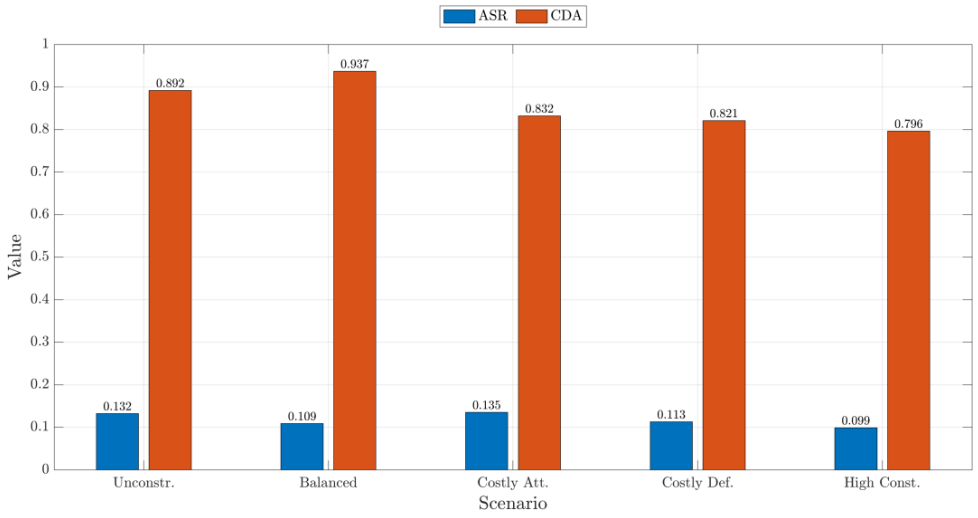


Figure 13. ASR and *CDA* at the Equilibrium for Different BG_{Cost} Cases.

5.5. Summary of Findings

Practical Guidelines & Limitations

- **Preferred operating point:** The Balanced scenario delivers the most favorable ASR–CDA trade-off under cost awareness.
- **Defense calibration:** Avoid over-defending in Unconstrained/Costly-Attacks settings—track CDA_{cb} and set guardrails on Δ_{def} .
- **Budget-aware choices:** In Costly-Defenses/High-Constraints, prioritize lightweight screening, acceptance thresholds (CDA_{inf}), and post-deployment monitoring.
- **Stealthy triggers:** Expect low-power (Δ_{tr}, Δ_{ts}); incorporate trigger-aware audits and anomaly checks around target classes.
- **Limitations:** Results shown for MNIST and a single model family; extension to richer datasets/models and non–zero-sum variants is left to future work.

To consolidate these insights, we provide in **Table 5** and the accompanying summary box a compact overview of the main findings, deployment-oriented guidelines, and limitations. These items synthesize the trends observed in **Figure 13** and the per-scenario equilibria, allowing readers to quickly grasp the strategic implications of each cost setting.

Table 5. Quick-reference synthesis of findings, guidelines, and limitations across BG_{Cost} scenarios.

Scenario	Key Findings	Practical Guidelines	Limitations/Caveats
Unconstrained	Aggressive play on both sides; defender tends to push high Δ_{def}; CDA degradation noticeable while ASR remains non-negligible.	Avoid deploying with weak cost controls; if used, cap defense strength to preserve CDA and rely on monitoring/rollback.	Sensitive to over-defense (CDA drops); cost-agnostic tuning may not generalize to resource-limited settings.
Balanced	Best trade-off: low ASR with high CDA; pure strategies emerge with modest α_{tr} , Δ_{tr} and moderate ($\alpha_{def}, \Delta_{def}$).	Recommended default for deployment; prioritize moderate defense and verification against stealthy, low-power triggers.	Still scenario-specific; assumes reliable estimation of λ_A, λ_D and a stable operating domain.
Costly Attacks	Attacker backs off to low-cost triggers; defender can afford stronger countermeasures; CDA declines if defense is too aggressive.	Exploit attacker's cost pressure; prefer calibrated (not maximal) defenses to avoid unnecessary CDA loss.	Over-defending can erode benign accuracy; watch CDA_{cb} to prevent unacceptable quality drops.
Costly Defenses	Defender scales back; attacker retains minimal but persistent effectiveness; ASR modest, CDA also modest.	Use lightweight, selective defenses (e.g., pre-filtering, targeted inspection) and trigger-aware QA gates.	Budget-bound defenses risk residual backdoors; continuous monitoring is needed to catch low-power attacks.
High Constraints	Both sides conservative; ASR low, but CDA also lowest among settings; fragile equilibrium, limited robustness.	Avoid for production if possible; if unavoidable, enforce strict acceptance thresholds and fail-safe policies.	Under-investment by both players leaves the model brittle, with limited headroom to react to shifts.

Our results show that cost constraints fundamentally shape attacker and defender behavior in deep learning security. When both sides must account for resource limitations, their strategies shift markedly. Equilibrium analysis suggests that defenders should prioritize moderate-cost defenses—overly aggressive defenses harm model performance, while minimal ones let attacks persist. Attackers facing high costs tend to scale back, favoring low-cost, stealthy strategies, which makes subtle anomaly detection a promising countermeasure. In contrast, unconstrained settings encourage stronger attacks and aggressive defenses, but at the expense of degrading CDA.

Among all scenarios, the Balanced configuration emerges as the most practical for deployment. It minimizes ASR while preserving strong model performance, without incurring excessive resource costs. By contrast, the costly defenses scenario highlights how resource limitations impair the defender's ability to apply robust countermeasures, leaving the model vulnerable to persistent but less aggressive attacks. The costly attacks scenario weakens the attacker but may slightly degrade CDA, while the high constraints case is the least desirable: severe restrictions on both players lead to weak attacks met with equally weak defenses, producing an ineffective equilibrium. Overall, the Balanced scenario stands out as the most viable for real-world deployment, achieving an optimal compromise between security, model accuracy, and computational feasibility.

6. Conclusions

This paper presented BG_{Cost}, a game-theoretic framework for modeling the interaction between an attacker and a defender in deep learning backdoor attacks, explicitly incorporating the limitations imposed by cost constraints. By framing the problem as a two-player zero-sum game—where the attacker's gain directly corresponds to the defender's loss—we analyzed how strategic decisions evolve under varying cost scenarios and demonstrated how both players adapt their behaviors when accounting for resource limitations. The framework defines utility functions that integrate Clean Data Accuracy (CDA) and Attack Success Rate (ASR), enabling a structured evaluation of equilibrium strategies. Through numerical simulations, we showed how cost constraints shape optimal decision-making, underscoring the pivotal role of resource-aware trade-offs in adversarial interactions.

A key insight from our study is that cost constraints fundamentally alter the attack-defense dynamics. Unconstrained settings lead to aggressive strategies from both players, while cost-balanced scenarios encourage a more defensive equilibrium. High attack costs discourage attackers from engaging in strong backdoor strategies, giving defenders an advantage, whereas high defense costs weaken the defender's ability to neutralize attacks effectively. The high constraints scenario resulted in a fragile equilibrium where neither player could act optimally due to extreme cost restrictions. Across all settings, the balanced scenario emerged as the most practical for real-world deployment, offering the best trade-off between security and model performance. These findings underscore the need for cost-aware modeling in adversarial machine learning and highlight the importance of flexible defense strategies that can adapt to maintain security without exceeding computational limits.

It is important to emphasize that BG_{Cost} is not proposed as a new algorithmic defense but rather as a benchmarking and analysis framework. Its contribution lies in providing a structured game-theoretic perspective where different backdoor attacks and defenses can be embedded as player strategies and systematically analyzed under cost constraints. In this way, BG_{Cost} complements—rather than replaces—established defenses by offering a decision-theoretic lens through which their trade-offs can be compared.

In deployment scenarios, the cost-aware perspective reflects realistic trade-offs across industry applications. For example, cloud-based medical AI systems may tolerate moderate computational overhead to ensure robustness against backdoor risks, as reliability is critical in healthcare. Conversely, autonomous driving platforms require defenses that minimize latency and computational burden, prioritizing real-time responsiveness while still mitigating adversarial threats. Our results suggest that balanced defenses, instead of extreme strategies, offer the most practical compromise between performance, security, and resource consumption, making them especially suitable for real-world deployment where both robustness and efficiency are essential.

Future work could expand on this framework in several directions. First, evaluating the model on more complex datasets such as ImageNet and domain-specific benchmarks will help assess its robustness across different applications. Second, exploring non-zero-sum game formulations could provide deeper insights into scenarios where both attackers and defenders incur losses, such as in federated learning environments, going beyond the current strictly zero-sum assumption. Third, incorporating multiple attackers and/or defenders would introduce richer dynamics and collective behaviors, potentially requiring more advanced Bayesian or cooperative game models. Another promising avenue is dynamic adaptation, where players adjust their strategies over time based on observed behaviors, leading to more realistic sequential or reinforcement learning-based approaches. Further investigations into information asymmetry, where one player has more information than the other, could also quantify the impact of strategic uncertainty, offering insights into how varying levels of awareness influence optimal decision-making. Finally, real-world deployment of cost-aware defenses should consider adaptive cost mechanisms, where defenders allocate resources dynamically in response to real-time attack detection, ensuring an optimal balance between security and computational efficiency. Together, these directions will advance adversarial learning research and contribute to more effective, resource-efficient defenses against backdoor attacks.

Author Contributions

Conceptualization, K.K.; methodology, K.K.; software, K.K., C.T., and H.F.; validation, K.K., C.T., and H.F.; formal analysis, K.K.; investigation, K.K., C.T., and H.F.; resources, K.K.; data curation, K.K., C.T., and H.F.; writing—original draft preparation, K.K.; writing—review and editing, K.K., C.T., and H.F.; visualization, K.K.; supervision, K.K.; project administration, K.K. All authors have read and agreed to the published version of the manuscript.

Funding

This work was partially supported by the industrial chair CYBAILE, partly funded by a European and Brittany region grant (reference FEDER-AIDEN-76568) and SSF-ML-DH.

Institutional Review Board Statement

Not applicable. This study did not involve humans, animals, or any data requiring ethical approval.

Informed Consent Statement

Not applicable. This study did not involve human participants.

Data Availability Statement

No new datasets were generated for the purposes of this study. All experiments were conducted using publicly available benchmark data. In particular, the MNIST dataset used in the simulations can be accessed at: http://yann.lecun.com/exdb/mnist/. The utility matrices and equilibrium results reported in this paper are fully reproducible from the methodology and parameter settings described in the manuscript. No proprietary, personal, or restricted data were used.

Acknowledgment

The authors would like to thank Gouenou Coatrieux for his valuable support in the acquisition of funding for this work.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Voulodimos, A.; Doulamis, N.; Doulamis, A.; et al. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 1–13.
2. Grigorescu, S.; Trasnea, B.; Cocias, T.; et al. A Survey of Deep Learning Techniques for Autonomous Driving. *J. Field Robot.* **2020**, *37*, 362–386.
3. Heaton, J.B.; Polson, N.G.; Witte, J.H. Deep Learning for Finance: Deep Portfolios. *Appl. Stoch. Models Bus. Ind.* **2017**, *33*, 3–12.
4. Miotto, R.; Wang, F.; Wang, S.; et al. Deep Learning for Healthcare: Review, Opportunities and Challenges. *Brief. Bioinform.* **2018**, *19*, 1236–1246.
5. Shinde, P.P.; Shah, S. A Review of Machine Learning and Deep Learning Applications. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–6.
6. Gürel, N.M.; Qi, X.; Rimanic, L.; et al. Knowledge Enhanced Machine Learning Pipeline Against Diverse Adversarial Attacks. *Proc. Mach. Learn. Res.* **2021**, *139*, 3976–3987.
7. Le Roux, Q.; Bourbao, E.; Teglia, Y.; et al. A Comprehensive Survey on Backdoor Attacks and Their Defenses in Face Recognition Systems. *IEEE Access* **2024**, *12*, 47433–47468. [CrossRef]
8. Wu, B.; Zhu, Z.; Liu, L.; et al. Attacks in Adversarial Machine Learning: A Systematic Survey from the Life-Cycle Perspective. *arXiv Preprint* **2023**, arXiv:2302.09457.
9. Al-Jarrah, O.Y.; Yoo, P.D.; Muhaidat, S.; et al. Efficient Machine Learning for Big Data: A Review. *Big Data Res.* **2015**, *2*, 87–93.
10. Gao, Y.; Doan, B.G.; Zhang, Z.; et al. Backdoor Attacks and Countermeasures on Deep Learning: A Comprehensive Review. *arXiv Preprint* **2020**, arXiv:2007.10760.
11. Li, Y.; Jiang, Y.; Li, Z.; et al. Backdoor Learning: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *35*, 5–22.
12. Chen, X.; Liu, C.; Li, B.; et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv Preprint* **2017**, arXiv:1712.05526.
13. Schwarzschild, A.; Goldblum, M.; Gupta, A.; et al. Just How Toxic Is Data Poisoning? A Unified Benchmark

for Backdoor and Data Poisoning Attacks. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021.
14. Zheng, T.; Lan, H.; Li, B. Be Careful with PyPI Packages: You May Unconsciously Spread Backdoor Model Weights. In Proceedings of the 6th MLSys Conference, Miami, FL, USA, 4–8 June 2023.
15. Kurita, K.; Michel, P.; Neubig, G. Weight Poisoning Attacks on Pre-Trained Models. *arXiv Preprint* **2020**, arXiv:2004.06660.
16. Barni, M.; Kallas, K.; Tondi, B. A New Backdoor Attack in CNNs by Training Set Corruption Without Label Poisoning. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1–6.
17. Gu, T.; Liu, K.; Dolan-Gavitt, B.; et al. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access* **2019**, *7*, 47230–47244.
18. Turner, A.; Tsipras, D.; Madry, A. Label-Consistent Backdoor Attacks. *arXiv Preprint* **2019**, arXiv:1912.02771.
19. Wu, B.; Wei, S.; Zhu, M.; et al. Defenses in Adversarial Machine Learning: A Survey. *arXiv Preprint* **2023**, arXiv:2312.08890.
20. Sheng, X.; Han, Z.; Li, P.; et al. A Survey on Backdoor Attack and Defense in Natural Language Processing. *arXiv Preprint* **2022**, arXiv:2211.11958.
21. Kong, Y.; Zhang, J. Adversarial Audio: A New Information Hiding Method and Backdoor for DNN-Based Speech Recognition Models. *arXiv Preprint* **2019**, arXiv:1904.03829.
22. Bhalerao, A.; Kallas, K.; Tondi, B.; et al. Luminance-Based Video Backdoor Attack Against Anti-Spoofing Rebroadcast Detection. In Proceedings of the 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), Kuala Lumpur, Malaysia, 27–29 September 2019; pp. 1–6.
23. Li, Y.; Li, Y.; Wu, B.; et al. Invisible Backdoor Attack with Sample-Specific Triggers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 16463–16472.
24. Wang, B.; Cao, X.; Jia, J.; et al. On Certifying Robustness Against Backdoor Attacks via Randomized Smoothing. *arXiv Preprint* **2020**, arXiv:2002.11750.
25. Jia, J.; Yuan, Z.; Sahabandu, D.; et al. FLGAME: A Game-theoretic Defense against Backdoor Attacks in Federated Learning. In *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, New Orleans, LA, USA, 10–16 December 2023. Available online: https://openreview.net/forum?id=TwCGI3rVddj
26. Kallas, K.; Le Roux, Q.; Hamidouche, W.; et al. Strategic Safeguarding: A Game Theoretic Approach for Analyzing Attacker–Defender Behavior in DNN Backdoors. *EURASIP J. Inf. Secur.* **2024**, *2024*, 32.
27. von Neumann, J.; Morgenstern, O. *Theory of Games and Economic Behavior*; Princeton University Press: Princeton, NJ, USA, 2007.
28. Burguillo, J.C. Game Theory. In *Self-Organizing Coalitions for Managing Complexity*; Springer: Cham, Switzerland, 2018; 29, pp. 101–135. [CrossRef]
29. Nash, J. Equilibrium Points in N-Person Games. *Proc. Natl. Acad. Sci. USA* **1950**, *36*, 48–49.
30. Osborne, M.J.; Rubinstein, A. *A Course in Game Theory*; MIT Press: Cambridge, MA, USA, 1994.