

UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA
CIÊNCIA DA COMPUTAÇÃO

Mariana dos Santos Dick

**Defesa de Backdoors em Aprendizado de Máquina: Uma Análise do Estado da Arte,
Benchmarks e Aplicações**

Florianópolis
2026

Mariana dos Santos Dick

Defesa de Backdoors em Aprendizado de Máquina: Uma Análise do Estado da Arte, Benchmarks e Aplicações

Trabalho de Conclusão de Curso submetido ao Curso de Graduação em Ciência da Computação do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para obtenção do título de Bacharel em Ciência da Computação.

Orientadora: Larissa Gremelmaier Rosa, Me.

Coorientador: Prof. Jean E. Martina, Dr.

Florianópolis

2026

Mariana dos Santos Dick

**Defesa de Backdoors em Aprendizado de Máquina: Uma Análise do Estado da Arte,
Benchmarks e Aplicações**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Ciência da Computação e aprovado em sua forma final pelo curso de Graduação em Ciência da Computação.

Florianópolis, 03 de julho de 2026.

Prof^a. Lúcia Helena Martins Pacheco, Dr.
Coordenadora do Curso

Banca Examinadora:

Larissa Gremelmaier Rosa, Me.
Orientadora
Universidade Federal de Santa Catarina

Prof. Jean E. Martina, Dr.
Coorientador
Universidade Federal de Santa Catarina

Prof. Ricardo Felipe Custodio, Dr.
Avaliador
Universidade Federal de Santa Catarina

Marcos Tomaszewski, Me.
Avaliador
Universidade Federal de Santa Catarina

À minha família de sangue, de escolha e de quatro patas

AGRADECIMENTOS

Agradeço à minha orientadora, Larissa, por guiar este trabalho com dedicação e cuidado, do primeiro rascunho à versão final. Agradeço também ao meu coorientador, professor Jean, pelas contribuições a esta pesquisa.

À minha família, em especial aos meus pais, Cícero e Maria Porcínia, que abdicaram de muito para que eu pudesse me dedicar integralmente aos estudos. Vocês não estiveram no dia a dia da sala de aula, mas sentaram comigo em cada sala de matrícula e foram os primeiros a saber da aprovação. Até hoje, é com vocês que quero dividir cada conquista.

Aos meus irmãos, meus melhores amigos desde o dia em que nasci. À Duda, por me ensinar a viver com intensidade e alegria, e por fazer da sua casa com a Bruna o meu segundo lar. Ao Juninho, por ouvir todos os meus planos, sonhos e ideias aleatórias, e por sentar comigo até a matemática e os circuitos digitais fazerem sentido.

Ao Ricardo, meu namorado, e à Canjica, minha gata, meu porto seguro. Vocês pouco souberam de máscaras ou gatilhos, mas cuidaram da parte que nenhuma disciplina ensina, a cabeça e o coração de quem escreve. Obrigada por estarem ao meu lado em todos os dias desta caminhada, dos mais difíceis aos mais felizes.

Aos meus amigos Gabi, Novais, Beirão, Ismael e Luiza, pela companhia nas alegrias e pelo apoio nos momentos difíceis. Vocês tornaram tudo mais leve, e pretendo levá-los comigo para a vida. E, em especial, à Lari, pelos mais de cinco anos de amizade muito bem vividos, entre apartamento dividido, viagens, vinhos, fofocas e comida boa. Uma das pessoas com quem posso contar para viver as coisas boas da vida, em todas as fases que ainda vêm.

Por fim, agradeço à Universidade Federal de Santa Catarina e a todos os professores que fizeram parte da minha formação. Seus ensinamentos foram fundamentais para a construção deste trabalho e para que eu chegasse até aqui.

A treacherous weapon is ever a danger to the hand.
(J. R. R. Tolkien, *The Two Towers*, 1954)

RESUMO

Apesar dos avanços e da ampla adoção, o aprendizado de máquina permanece vulnerável a ataques adversariais, sendo os ataques de *backdoor* uma ameaça significativa à integridade dos modelos. Este estudo integra à plataforma de *benchmark* BackdoorBench o NFT (*Augmented Neural Fine-Tuning*), uma técnica recente de defesa de pós-treinamento que a ferramenta ainda não contemplava, e a avalia comparativamente sob um protocolo padronizado. Para fundamentar essa escolha, foi realizada uma Revisão Sistemática da Literatura (RSL) que categorizou as defesas conforme o estágio do ciclo de vida do modelo: pré-treinamento, treinamento, pós-treinamento e tempo de inferência. O NFT foi implementado como um novo módulo na plataforma, seguindo suas convenções, e avaliado em comparação com defesas existentes diante de uma variedade de ataques de *backdoor* no conjunto de dados CIFAR-10. Os resultados indicam que o NFT remove de forma eficaz *backdoors* de gatilho localizado, geométrico e dinâmico (BadNets, WaNet e InputAware), preservando a acurácia em dados limpos, equiparando-se às defesas de melhor desempenho e apresentando maior estabilidade entre execuções. Nesses casos, a avaliação independente em ambiente padronizado confirma o desempenho relatado pelos autores originais do método. Em contrapartida, sua eficácia contra gatilhos aditivos e de mesclagem é limitada, com mitigação apenas parcial do sinal aditivo do SIG e falha diante dos gatilhos de mesclagem (Blended) e de baixa frequência (LF), situações em que defesas baseadas em poda e desaprendizado se mostram superiores. Uma investigação controlada da discrepância entre esses resultados e os relatados na avaliação original do NFT atribuiu tal diferença a fatores do protocolo de avaliação, e não a falhas do método: a receita de treino do ataque, que altera a removibilidade do gatilho para toda a família de ajuste fino, e o orçamento de purificação, sob o qual o NFT reproduz o desempenho originalmente relatado. A análise demonstra que nenhuma defesa é dominante contra todos os tipos de ataque, evidenciando o compromisso entre a eficácia de mitigação e a preservação da utilidade do modelo.

Palavras-chave: Aprendizado de Máquina. Ataques de Backdoor. Benchmark.

ABSTRACT

Despite its advances and widespread adoption, machine learning remains vulnerable to adversarial attacks, including backdoor attacks, which pose a significant threat to model integrity. This study integrates into the BackdoorBench benchmark platform a recent post-training defense technique that the tool did not yet include, NFT (Augmented Neural Fine-Tuning), and evaluates it comparatively under a standardized protocol. To support this choice, a Systematic Literature Review (SLR) was conducted, categorizing defenses by model lifecycle stage: pre-training, training, post-training, and inference time. NFT was implemented as a new module on the platform, following its conventions, and evaluated against existing defenses and a variety of backdoor attacks on the CIFAR-10 dataset. The results indicate that NFT effectively removes backdoors with localized, geometric, and dynamic triggers (BadNets, WaNet, and InputAware), preserving clean accuracy, matching the best-performing defenses, and exhibiting greater stability across runs. In these cases, the independent evaluation in a standardized environment confirms the performance reported by the method’s original authors. In contrast, its effectiveness against additive and blending triggers is limited, with only partial mitigation of the additive SIG signal and failure against the blending (Blended) and low-frequency (LF) triggers, situations in which pruning- and unlearning-based defenses prove superior. A controlled investigation of the discrepancy between these results and those reported in NFT’s original evaluation attributed the difference to factors of the evaluation protocol rather than to flaws in the method: the attack’s training recipe, which changes the trigger’s removability for the entire fine-tuning family, and the purification budget, under which NFT reproduces its originally reported performance. The analysis demonstrates that no defense is dominant against all attack types, highlighting a trade-off between mitigation effectiveness and model utility.

Keywords: Machine Learning. Backdoor Attacks. Benchmark.

LISTA DE FIGURAS

Figura 1 – Fluxo geral de um ataque de <i>backdoor</i> , do envenenamento do conjunto de treinamento à ativação do gatilho na inferência.	18
Figura 2 – Taxonomia das defesas contra <i>backdoor</i> segundo o estágio do ciclo de vida do modelo em que atuam; o reparo de modelo em pós-treinamento, foco deste trabalho, aparece destacado. Adaptado de (WU et al., 2025).	21
Figura 3 – Fluxo de seleção dos estudos primários da revisão sistemática, no formato PRISMA.	28
Figura 4 – Integração do NFT como novo módulo de defesa no fluxo da BackdoorBench.	34
Figura 5 – Exemplos dos gatilhos avaliados sobre imagens do CIFAR-10. Para cada ataque, a linha superior mostra a imagem limpa e a inferior, a versão com o gatilho.	36
Figura 6 – Compromisso entre utilidade (CA, eixo horizontal) e segurança (ASR, eixo vertical), com um painel por ataque e cor e marcador por defesa, em média sobre três sementes. A posição desejável é o canto inferior direito, que combina CA alta e ASR baixa; o NFT aparece destacado com uma estrela, e o ponto sem defesa serve de referência.	39
Figura 7 – Evolução da ASR do NFT ao longo das iterações de ajuste fino, para os ataques avaliados (semente 0).	41
Figura 8 – Ablação do ataque SIG: ASR sem defesa e após o NFT ao variar (a) a amplitude Δ do sinal, com <i>pratio</i> fixo, e (b) a fração de envenenamento, com Δ fixo.	42
Figura 9 – Projeção <i>t-SNE</i> das representações internas (penúltima camada) de amostras limpas e com gatilho, no modelo infectado e após a purificação pelo NFT, para o BadNets (acima) e o Blended (abaixo).	44

LISTA DE TABELAS

Tabela 1 – Síntese dos dados extraídos dos estudos primários selecionados.	28
Tabela 1 – Síntese dos dados extraídos dos estudos primários selecionados (continuação).	29
Tabela 2 – Métricas de avaliação. A seta indica a direção desejável após a defesa.	37
Tabela 3 – Taxa de Sucesso do Ataque (ASR, em %) após cada defesa, em média sobre três sementes ($\pm$ desvio-padrão). Ataques agrupados por natureza do gatilho: localizado, geométrico e dinâmico à esquerda; aditivos e de mesclagem à direita.	38
Tabela 4 – Acurácia em Dados Limpos (CA, em %) após cada defesa, em média sobre três sementes ($\pm$ desvio-padrão).	38
Tabela 5 – Acurácia Robusta (RA, em %) após cada defesa, em média sobre três sementes ($\pm$ desvio-padrão).	39
Tabela 6 – Ablação do ataque SIG: ASR (%) sem defesa e após o NFT, variando isoladamente a amplitude Δ do sinal e a fração de envenenamento (<i>pratio</i>), em uma única semente.	42
Tabela 7 – Taxa de Sucesso do Ataque (ASR, em %) após cada defesa em duas taxas de envenenamento, em uma única semente. Os valores em 10% constam da Tabela 3.	45
Tabela 8 – Taxa de Sucesso do Ataque (ASR, em %) após cada defesa no GTSRB, em média sobre três sementes ($\pm$ desvio-padrão).	45
Tabela 9 – Tempo de execução (em minutos) de cada defesa sobre um mesmo modelo atacado, ordenado do menor ao maior. Os valores são indicativos e dependem do <i>hardware</i> ; o relevante é a ordem relativa entre as defesas.	46
Tabela 10 – Síntese qualitativa do desempenho das defesas por categoria de gatilho, no CIFAR-10.	47

SUMÁRIO

1	INTRODUÇÃO	12
1.1	OBJETIVOS	13
1.1.1	Geral	13
1.1.2	Específicos	13
1.2	ORGANIZAÇÃO DO TRABALHO	14
2	FUNDAMENTAÇÃO TEÓRICA	15
2.1	APRENDIZADO DE MÁQUINA E O PROCESSO DE TREINAMENTO .	15
2.1.1	O Paradigma do Aprendizado Supervisionado	15
2.1.1.1	<i>Generalização, Sobreajuste e Regularização</i>	<i>15</i>
2.2	ATAQUES DE ENVENENAMENTO DE DADOS (DATA POISONING) .	16
2.2.1	O Modelo de Ameaça do Atacante	16
2.2.1.1	<i>Conhecimento do Atacante</i>	<i>16</i>
2.2.1.2	<i>Capacidade do Atacante</i>	<i>17</i>
2.2.2	Categorização dos Ataques por Objetivo	17
2.2.2.1	<i>Ataques de Disponibilidade</i>	<i>17</i>
2.2.2.2	<i>Ataques de Integridade</i>	<i>18</i>
2.2.3	Estratégias de Perturbação	19
2.2.3.1	<i>Label-Flipping</i>	<i>19</i>
2.2.3.2	<i>Ataques Clean-Label</i>	<i>19</i>
2.2.3.3	<i>Tipos de Gatilhos de Backdoor</i>	<i>19</i>
2.3	DEFESAS CONTRA ATAQUES DE BACKDOOR	20
2.3.1	Defesas de Pré-Treinamento (Sanitização de Dados)	21
2.3.1.1	<i>Detecção de Outliers no Espaço de Entrada</i>	<i>21</i>
2.3.1.2	<i>Detecção no Espaço de Características Latentes</i>	<i>22</i>
2.3.2	Defesas Durante o Treinamento (Treinamento Robusto)	22
2.3.2.1	<i>Regularização e Aumento de Dados</i>	<i>22</i>
2.3.2.2	<i>Treinamento Adversarial</i>	<i>22</i>
2.3.3	Defesas de Pós-Treinamento (Reparo de Modelo)	23
2.3.3.1	<i>Modificação da Estrutura do Modelo</i>	<i>23</i>
2.3.3.2	<i>Ajuste Fino dos Pesos (Fine-tuning)</i>	<i>23</i>
2.3.3.3	<i>Desaprendizado Adversarial</i>	<i>24</i>
2.3.3.4	<i>Ajuste Fino por Máscaras Neurais</i>	<i>24</i>
2.3.4	Defesas em Tempo de Inferência	25
2.3.4.1	<i>Detecção de Entradas Suspeitas</i>	<i>25</i>
2.3.4.2	<i>Purificação de Entradas</i>	<i>25</i>
3	REVISÃO SISTEMÁTICA DA LITERATURA	26

3.1	QUESTÕES DE PESQUISA	26
3.2	FONTES DE DADOS E ESTRATÉGIA DE BUSCA	26
3.3	CRITÉRIOS DE INCLUSÃO E EXCLUSÃO	26
3.4	PROCESSO DE SELEÇÃO DOS ESTUDOS	27
3.5	ANÁLISE DO ESTADO DA ARTE E IDENTIFICAÇÃO DA LACUNA DE PESQUISA	29
4	METODOLOGIA	33
4.1	DELIMITAÇÃO DO ESCOPO	33
4.2	PREPARAÇÃO DO AMBIENTE EXPERIMENTAL	33
4.3	INTEGRAÇÃO DO NFT À BACKDOORBENCH	34
4.4	PROTOCOLO DE AVALIAÇÃO EXPERIMENTAL	36
4.4.1	Conjunto de Dados e Modelos Comprometidos	36
4.4.2	Ataques Avaliados	36
4.4.3	Defesas Avaliadas	37
4.4.4	Métricas e Procedimento de Avaliação	37
5	RESULTADOS E DISCUSSÃO	38
5.1	RESULTADOS COMPARATIVOS	38
5.2	DISCUSSÃO DOS RESULTADOS	40
5.2.1	Eficácia do NFT por tipo de ataque	40
5.2.2	Investigação das limitações frente a gatilhos aditivos e de mesclagem	41
5.2.2.1	<i>Mitigação parcial do sinal aditivo (SIG)</i>	41
5.2.2.2	<i>Falha frente ao gatilho de mesclagem (Blended)</i>	43
5.2.2.3	<i>O padrão comum aos gatilhos aditivos</i>	43
5.2.3	Robustez dos achados à taxa de envenenamento	44
5.2.4	Generalização ao GTSRB	45
5.2.5	Custo computacional das defesas	46
5.2.6	Síntese comparativa e implicações práticas	47
6	CONCLUSÃO	49
	DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL GENE- RATIVA	50
	REFERÊNCIAS	51
	APÊNDICE A – GUIA DE EXECUÇÃO DA AVALIAÇÃO EXPERI- MENTAL	55
	APÊNDICE B – ARTIGO EM FORMATO SBC	56

1 INTRODUÇÃO

O sucesso do aprendizado de máquina em diversas aplicações tem sido inerentemente dependente da crescente disponibilidade de poder computacional e de grandes conjuntos de dados para treinamento. Para que um modelo de aprendizado de máquina generalize de forma eficaz, é fundamental que seus dados de treinamento sejam representativos do ambiente real no qual será aplicado (JORDAN; MITCHELL, 2015).

No entanto, em situações práticas, a integridade dos dados frequentemente é comprometida. Muitas aplicações coletam dados de fontes não confiáveis ou sem supervisão humana adequada (FAN et al., 2022), levantando uma questão central sobre a segurança efetiva desses sistemas (BARRENO et al., 2006). A manipulação deliberada dos dados de treinamento representa uma ameaça real, investigada desde os primeiros ataques que visavam subverter aplicações, como filtros de spam (NELSON et al., 2008).

A vulnerabilidade central nesse contexto é conhecida como ataque de *data poisoning*. Nesse ataque, um adversário injeta ou manipula deliberadamente parte dos dados de treinamento para comprometer o aprendizado do modelo (BIGGIO; ROLI, 2018). As consequências de um ataque de *data poisoning* variam conforme o objetivo do atacante, resultando em duas categorias principais: disponibilidade e integridade (CINÀ et al., 2023).

Ataques de disponibilidade têm como objetivo degradar o desempenho geral do modelo, elevando sua taxa de erro para a maioria das entradas legítimas e, essencialmente, tornando-o inutilizável ou não confiável (FAN et al., 2022). Por outro lado, ataques de integridade são mais sutis e mantêm bom desempenho geral, causando falhas específicas e controladas. Um exemplo emblemático é o ataque de *backdoor*, no qual um gatilho inserido nos dados de treinamento induz o modelo a associar esse padrão a resultados incorretos específicos, ativáveis posteriormente em produção (GU et al., 2019).

Essa ameaça não é apenas teórica, sendo identificada como a principal preocupação na implementação prática de aprendizado de máquina em ambientes industriais (GROSSE et al., 2023). Em resposta, a comunidade científica propôs diversas abordagens defensivas, com mais de 70 técnicas diferentes documentadas na literatura recente (CINÀ et al., 2023). Apesar da diversidade de abordagens defensivas propostas, uma análise crítica revela que a eficácia relatada de muitos ataques e defesas é frequentemente consequência de experimentos inconsistentes ou pouco realistas (SCHWARZSCHILD et al., 2021).

Essa dificuldade é intensificada pela escassez de ferramentas práticas e centralizadas para avaliação comparativa. Embora existam bibliotecas de código aberto notáveis, como a Adversarial Robustness Toolbox (ART) e a secml (NICOLAE et al., 2018; PINTOR et al., 2022), elas cobrem apenas uma fração das técnicas existentes. Essa fragmentação deixa desenvolvedores sem um método sistemático para validar seus modelos. Para endereçar essa lacuna, surgiram plataformas de *benchmarking* especializadas, como o BackdoorBench, que oferecem uma implementação integrada de dezenas de algoritmos de ataque e defesa (atualmente 16 ataques e 28 métodos de defesa e detecção) (WU et al., 2024). A plataforma fornece um

protocolo de avaliação padronizado e um conjunto de ferramentas de análise para facilitar comparações justas e reprodutíveis (WU et al., 2025).

Neste contexto, com o objetivo de contribuir para a consolidação e o avanço da área, propomos realizar uma análise aprofundada do estado da arte em defesas contra *backdoor* e expandir a plataforma BackdoorBench. A contribuição prática consiste em implementar nessa plataforma uma técnica de defesa já proposta na literatura, mas ainda não disponível de forma padronizada no *benchmark*, e avaliá-la em comparação com as abordagens existentes. Este estudo busca fornecer uma visão crítica das soluções atuais e disponibilizar um recurso prático que fortaleça a capacidade da comunidade de avaliar e desenvolver sistemas de aprendizado de máquina mais seguros.

1.1 OBJETIVOS

1.1.1 Geral

O objetivo geral deste trabalho é realizar uma análise crítica e experimental do cenário de defesas contra ataques de *backdoor* no aprendizado de máquina. Para isso, o trabalho faz uso da plataforma de *benchmarking* BackdoorBench e a expande, integrando-lhe uma técnica recente de defesa de pós-treinamento ainda não contemplada, e conduz uma análise comparativa de sua eficácia e robustez, frente às soluções já existentes e a um conjunto diverso de ataques, a fim de gerar conhecimento sobre as limitações e vantagens inerentes a essas estratégias.

1.1.2 Específicos

- Realizar uma revisão sistemática da literatura sobre ataques de *backdoor* e suas defesas, para mapear o estado da arte e identificar uma técnica de defesa relevante ainda ausente dos *benchmarks* de referência.
- Analisar a arquitetura e o protocolo de avaliação da BackdoorBench, para compreender seus módulos e viabilizar a integração da nova defesa.
- Implementar a defesa selecionada, o NFT (*Augmented Neural Fine-Tuning*) (KARIM et al., 2024), como um novo módulo da BackdoorBench, para incorporá-la à plataforma de forma fiel e reprodutível.
- Executar experimentos controlados para avaliar a eficácia da defesa contra um conjunto diverso de ataques de *backdoor*, de gatilhos localizados, geométricos, dinâmicos e de mesclagem ou aditivos, e compará-la à das defesas já existentes.
- Analisar criticamente os resultados, discutir os compromissos entre a eficácia de mitigação e a preservação da utilidade do modelo e apontar direções para pesquisas futuras.

1.2 ORGANIZAÇÃO DO TRABALHO

O restante deste trabalho está estruturado da seguinte forma: o Capítulo 2 apresenta a fundamentação teórica, detalhando conceitos essenciais sobre aprendizado de máquina, a taxonomia dos ataques de envenenamento de dados e, com foco especial, os diferentes tipos de ataques e defesas de *backdoor*. O Capítulo 3 apresenta a Revisão Sistemática da Literatura, descrevendo seu protocolo, sintetizando os trabalhos correlatos e identificando, com base na lacuna no estado da arte, a defesa a ser integrada. O Capítulo 4 detalha a metodologia experimental, incluindo a delimitação do escopo, a preparação do ambiente e o protocolo de avaliação das defesas. O Capítulo 5 apresenta uma comparação dos resultados obtidos com a defesa integrada em relação às abordagens existentes, além de discutir os compromissos identificados entre a mitigação dos ataques e a preservação da utilidade do modelo. Por fim, o Capítulo 6 sumariza as contribuições, aponta as limitações e delinea direções para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 APRENDIZADO DE MÁQUINA E O PROCESSO DE TREINAMENTO

2.1.1 O Paradigma do Aprendizado Supervisionado

O aprendizado de máquina compreende um conjunto de métodos que permitem aos sistemas computacionais melhorar seu desempenho em determinada tarefa por meio da análise de dados (BISHOP, 2007). Entre as diversas abordagens existentes, o aprendizado supervisionado destaca-se como um dos paradigmas mais fundamentais e amplamente utilizados, especialmente em tarefas relacionadas ao reconhecimento de padrões (GOODFELLOW; BENGIO; COURVILLE, 2016).

No aprendizado supervisionado, o objetivo do sistema é aprender uma função de mapeamento $Y = F(Z, W)$, onde Z representa uma entrada (por exemplo, uma imagem), W corresponde ao conjunto de parâmetros ajustáveis do modelo (como pesos de uma rede neural), e Y é a saída prevista pelo sistema (por exemplo, uma classe associada à imagem). O treinamento ocorre a partir de um conjunto de treinamento, composto por pares de exemplos rotulados $\{(Z^p, D^p)\}_{p=1}^P$, onde D^p é a saída correta esperada para a entrada Z^p (LECUN et al., 1998).

O objetivo do treinamento é encontrar a configuração ideal dos parâmetros W que minimize uma função de perda, responsável por quantificar a discrepância entre a saída produzida pelo modelo $F(Z, W)$ e a saída desejada D (LECUN et al., 1998). Para problemas complexos, como reconhecimento de objetos em imagens de alta resolução, são empregados modelos com alta capacidade de aprendizado, como Redes Neurais Convolucionais (CNNs), que possuem milhões de parâmetros treináveis para capturar representações ricas e hierárquicas dos dados (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

2.1.1.1 Generalização, Sobreajuste e Regularização

O principal objetivo de um modelo de aprendizado de máquina não é obter desempenho perfeito sobre os dados vistos durante o treinamento, mas sim generalizar o conhecimento adquirido para novos dados não vistos anteriormente (GOODFELLOW; BENGIO; COURVILLE, 2016). Para avaliar essa capacidade, o conjunto total de dados é tradicionalmente dividido em três subconjuntos: o de treinamento, usado diretamente no ajuste dos parâmetros; o de validação, que guia decisões de desenvolvimento, como a escolha de hiperparâmetros, sem participar do ajuste dos pesos; e o de teste, reservado exclusivamente à avaliação final e intocado ao longo do processo, oferecendo uma medida imparcial da generalização (BISHOP, 2007; GOODFELLOW; BENGIO; COURVILLE, 2016).

O principal obstáculo a essa generalização é o sobreajuste (*overfitting*): modelos de alta capacidade podem memorizar particularidades e até ruídos estatísticos do conjunto de treinamento, apresentando desempenho excepcional no treino, porém significativamente inferior

no teste (GOODFELLOW; BENGIO; COURVILLE, 2016; LECUN et al., 1998). Sua mitigação combina arquiteturas adequadas ao tipo de dado, como as próprias CNNs, que reduzem o número de parâmetros livres por meio de campos receptivos locais e compartilhamento de pesos, e técnicas de regularização que controlam a complexidade do modelo, como termos de penalidade adicionados à função de perda e o *dropout*, que desativa neurônios aleatoriamente durante o treinamento (KRIZHEVSKY; SUTSKEVER; HINTON, 2012; LECUN et al., 1998).

2.2 ATAQUES DE ENVENENAMENTO DE DADOS (DATA POISONING)

Os ataques de envenenamento de dados consistem na manipulação proposital do conjunto de treinamento por um adversário, com o objetivo de corromper o processo de aprendizado e comprometer o comportamento do modelo durante a inferência (GOLDBLUM et al., 2022). A eficácia da ameaça reside na capacidade do atacante de inserir amostras maliciosas que, mesmo constituindo uma fração mínima do volume de dados, são suficientes para distorcer a fronteira de decisão aprendida pelo modelo (SHAFABI et al., 2018). Para uma compreensão e defesa eficazes, a análise desta ameaça é classicamente estruturada a partir de um modelo de ameaça que formaliza o comportamento do adversário (BIGGIO; ROLI, 2018).

2.2.1 O Modelo de Ameaça do Atacante

O modelo de ameaça estrutura a avaliação de vulnerabilidades a partir de três eixos principais: o objetivo do ataque, o conhecimento que o adversário possui sobre o sistema e sua capacidade de manipulação dos dados (BIGGIO; ROLI, 2018). O primeiro eixo, o objetivo, define o que o atacante pretende alcançar e é classicamente dividido em duas categorias, violações de integridade e violações de disponibilidade (BARRENO et al., 2006), que estruturam a Subseção 2.2.2, onde são detalhadas junto dos ataques que as concretizam. Os dois demais eixos são descritos a seguir.

2.2.1.1 Conhecimento do Atacante

A efetividade e a estratégia dos ataques de envenenamento de dados variam conforme o nível de conhecimento que o atacante possui sobre o sistema, classicamente dividido em três cenários (BIGGIO; ROLI, 2018; CINÀ et al., 2023). No *white-box*, o adversário conhece o algoritmo, a arquitetura, os parâmetros treinados e o conjunto de treinamento; embora menos comum na prática, o cenário é útil para avaliar a robustez sob condições extremas. No *black-box*, ele apenas envia entradas ao modelo e observa as saídas, guiando ataques indiretos por meio de modelos substitutos. O cenário intermediário, *gray-box*, é o mais realista: o adversário dispõe de conhecimento parcial, como a arquitetura e o tipo de dados processados, mas desconhece os pesos treinados e os dados originais.

2.2.1.2 Capacidade do Atacante

A capacidade do atacante define seu nível de influência sobre o modelo. Essa capacidade é tradicionalmente dividida em duas categorias principais: exploratória e causativa (BARRENO et al., 2006; BIGGIO; ROLI, 2018). Os ataques exploratórios, conhecidos como ataques de evasão, ocorrem exclusivamente na fase de inferência, com manipulação pontual das entradas para enganar modelos já treinados.

Este trabalho foca especificamente nos ataques causativos, ou ataques de envenenamento, nos quais o adversário interfere diretamente na fase de treinamento, manipulando os dados que serão utilizados para ajustar os parâmetros do modelo (BARRENO et al., 2006; BIGGIO; ROLI, 2018). Essa influência pode se manifestar de diversas formas:

- **Injeção de dados em fontes abertas:** O adversário insere amostras maliciosas em fontes públicas ou pouco controladas, esperando que essas amostras sejam coletadas inadvertidamente para o treinamento do modelo. Nesse caso, o atacante não controla diretamente o processo de treinamento, apenas a composição parcial dos dados (CINÀ et al., 2023; SHAFahi et al., 2018).
- **Controle do processo de treinamento (*outsourcing*):** Em situações em que o treinamento é terceirizado para serviços externos, como em *Machine Learning as a Service* (MLaaS), o provedor malicioso pode manipular tanto os dados quanto os procedimentos de treinamento (por exemplo, ajustar hiperparâmetros) para gerar modelos comprometidos (BadNets). Esses modelos se comportam normalmente durante a validação, escondendo o *backdoor*, até serem ativados em produção (GU et al., 2019).

2.2.2 Categorização dos Ataques por Objetivo

2.2.2.1 Ataques de Disponibilidade

Os ataques de disponibilidade, também denominados indiscriminados, visam comprometer amplamente a funcionalidade do modelo, maximizando a perda média no conjunto de teste e degradando sua acurácia geral, em vez de mirar instâncias específicas (GOLDBLUM et al., 2022; FAN et al., 2022). Um exemplo clássico é a tentativa de inutilizar um filtro de spam, induzindo falsos positivos em excesso até que o administrador seja obrigado a desativá-lo (NELSON et al., 2008). A estratégia típica insere amostras maliciosas no treinamento para distorcer a fronteira de decisão e elevar os erros sobre dados legítimos (CINÀ et al., 2023). Esses ataques, estudados inicialmente em modelos clássicos como as *Support Vector Machines*, podem ser formulados como problemas de otimização que geram iterativamente amostras envenenadas para maximizar o erro (BIGGIO; NELSON; LASKOV, 2012).

2.2.2.2 Ataques de Integridade

Ao contrário dos ataques de disponibilidade, os ataques de integridade são projetados para serem discretos e difíceis de identificar. O adversário busca inserir falhas de classificação específicas sem afetar o desempenho geral do modelo em dados legítimos (BIGGIO; ROLI, 2018; BARRENO et al., 2006), englobando principalmente os ataques direcionados e os ataques de *backdoor* (GOLDBLUM et al., 2022). Um exemplo clássico é o sistema de reconhecimento facial que passa a identificar erroneamente uma pessoa específica quando esta utiliza um determinado acessório, como óculos (CHEN et al., 2017).

Nos ataques direcionados, o objetivo é induzir o modelo a errar em uma instância específica previamente definida, sem degradar sua acurácia global. Uma estratégia recorrente nesse cenário é o ataque de *clean-label*, em que as amostras maliciosas aparentam ter rótulos corretos e pertencer legitimamente à classe base. Contudo, são manipuladas de forma sutil para que suas representações internas se aproximem da amostra-alvo. Durante o treinamento, o modelo desloca sua fronteira de decisão para acomodar essas amostras, o que leva à classificação incorreta da instância-alvo (SHAFahi et al., 2018).

Os ataques de *backdoor*, ou ataques do tipo cavalo de Troia, são uma forma eficaz de ataque de integridade. A técnica consiste em introduzir uma vulnerabilidade latente no modelo durante o treinamento, geralmente por meio de dados envenenados que incluem um padrão secreto chamado de gatilho (*trigger*) (GOLDBLUM et al., 2022; GU et al., 2019). Esse gatilho pode ser um adesivo visual, uma frase específica ou um padrão de pixels. Quando presente, ele ativa o comportamento malicioso aprendido: o modelo classifica a entrada envenenada como pertencente à classe designada pelo atacante, mesmo que visualmente ela não pertença àquela classe (CINÀ et al., 2023).

Durante a validação, como os dados utilizados não contêm o gatilho, o modelo aparenta funcionar normalmente, o que dificulta a detecção da ameaça (GU et al., 2019). Avanços recentes, como o Hidden Trigger Backdoor Attack, mostraram que é possível realizar esse tipo de ataque mesmo sem que o gatilho esteja visível nos dados de treinamento, mantendo a vulnerabilidade completamente oculta até sua ativação (SAHA; SUBRAMANYA; PIRSIYAVASH, 2020). A Figura 1 resume o fluxo geral de um ataque de *backdoor*, do envenenamento do treino à ativação na inferência.

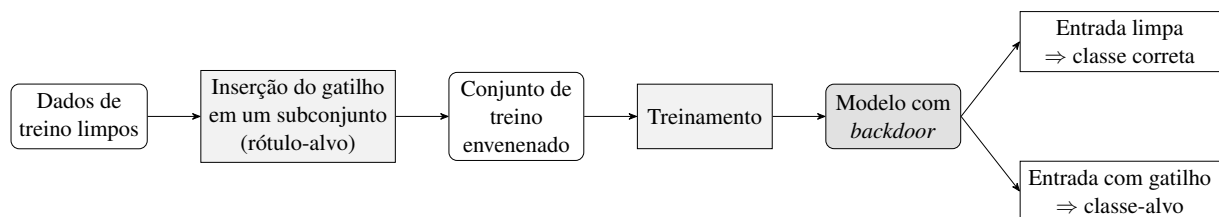


Figura 1 – Fluxo geral de um ataque de *backdoor*, do envenenamento do conjunto de treinamento à ativação do gatilho na inferência.

2.2.3 Estratégias de Perturbação

A geração de amostras de envenenamento eficazes pode ser formulada como um problema de otimização, no qual o adversário busca encontrar a menor perturbação possível que ainda assim seja capaz de induzir o comportamento malicioso desejado pelo ataque, seja ele de disponibilidade ou de integridade (GOLDBLUM et al., 2022). As estratégias de perturbação diferem quanto à sua complexidade, visibilidade e impacto no desempenho do modelo.

2.2.3.1 *Label-Flipping*

A técnica de *label-flipping* (troca de rótulo) é uma das formas mais simples e intuitivas de envenenamento de dados. Nela, o atacante seleciona amostras legítimas de um conjunto de dados e altera seus rótulos para classes incorretas (GOLDBLUM et al., 2022). Por exemplo, uma imagem de um gato pode ser rotulada como "cachorro". Durante o treinamento, essas amostras induzem o modelo a aprender uma fronteira de decisão incorreta. Apesar de sua simplicidade, essa técnica pode gerar ruído facilmente perceptível e é passível de detecção por inspeção manual. Os primeiros estudos sobre ataques de envenenamento, especialmente contra SVMs, já utilizavam essa abordagem, mostrando que mesmo rótulos trocados de forma estratégica podem maximizar o erro do modelo (BIGGIO; NELSON; LASKOV, 2012).

2.2.3.2 *Ataques Clean-Label*

Para superar as limitações da técnica anterior, surgiram os ataques *clean-label*, que são consideravelmente mais sofisticados. Nestes, o adversário mantém os rótulos corretos, mas aplica pequenas perturbações, muitas vezes imperceptíveis, a amostras da classe base. O objetivo é fazer com que a representação interna dessas amostras se aproxime da representação da amostra-alvo que se deseja atacar. Durante o retreinamento, o modelo ajusta sua fronteira de decisão de forma a acomodar os venenos, o que leva à classificação incorreta da amostra-alvo, mesmo sem que ela tenha sido modificada (SHAFahi et al., 2018). Como os rótulos são mantidos corretos e as alterações visuais são mínimas, essa técnica torna os venenos extremamente difíceis de serem detectados (GOLDBLUM et al., 2022; SHAFahi et al., 2018).

2.2.3.3 *Tipos de Gatilhos de Backdoor*

Em ataques de *backdoor*, o envenenamento de dados ocorre pela inserção de um padrão específico, o gatilho, em um subconjunto de amostras de treinamento. O objetivo é que o modelo aprenda uma associação maliciosa entre a presença do gatilho e uma classe-alvo escolhida pelo atacante. Uma vez treinado, o modelo se comportará normalmente com dados legítimos, mas classificará qualquer entrada que contenha o gatilho na classe-alvo, efetivamente ativando o

backdoor. A natureza e a complexidade desses gatilhos variam significativamente, podendo ser categorizados da seguinte forma:

- **Gatilhos de *patch*:** Correspondem à forma canônica de gatilho, popularizada pelo ataque BadNets (GU et al., 2019). Consistem na inserção de um padrão visual localizado, como um pequeno quadrado ou logotipo, em uma área específica da imagem. A forte dependência das Redes Neurais Convolucionais (CNNs) em características locais faz com que o *patch* se torne um indicador altamente preditivo para a classe-alvo (GOLDBLUM et al., 2022).
- **Gatilhos globais (aditivos ou de mesclagem):** Diferentemente dos *patches*, estes gatilhos são aplicados por meio de transformações aditivas ou de mesclagem que afetam a imagem inteira. Exemplos incluem a adição de um padrão de ruído de baixa amplitude ou a alteração sutil de propriedades como brilho e saturação. Por serem distribuídas por toda a imagem, essas modificações tendem a ser menos perceptíveis, aumentando o desafio da detecção (GOLDBLUM et al., 2022).
- **Gatilhos geométricos:** Também afetam a imagem inteira, mas, em vez de adicionar ou mesclar um padrão, alteram a geometria da própria imagem por meio de deformações suaves e imperceptíveis, como o empenamento espacial empregado pelo WaNet (NGUYEN; TRAN, 2021). Por não introduzirem conteúdo novo, costumam ser tratados como uma classe distinta dos gatilhos aditivos.
- **Gatilhos dinâmicos:** Em vez de um padrão fixo, empregam gatilhos que variam de amostra para amostra, em padrão e localização, o que torna ineficaz a detecção baseada em assinaturas fixas (SALEM et al., 2022). O ataque InputAware leva essa ideia ao extremo ao gerar, por meio de uma rede auxiliar, um gatilho específico para cada entrada (NGUYEN; TRAN, 2020).
- **Gatilhos semânticos:** Representam a classe mais avançada de gatilhos, pois não inserem padrões artificiais, mas exploram características semânticas já existentes e plausíveis nos dados, como associar uma característica natural, por exemplo “carros de cor verde”, à classe “pássaro” (BAGDASARYAN et al., 2020). A detecção é extremamente complexa, pois o gatilho é indistinguível de uma amostra benigna.

2.3 DEFESAS CONTRA ATAQUES DE BACKDOOR

Em resposta à diversidade e sofisticação dos ataques de envenenamento de dados, uma vasta gama de estratégias defensivas foi desenvolvida. Para uma análise sistemática, essas abordagens são frequentemente organizadas conforme o estágio do ciclo de vida do modelo de aprendizado de máquina em que atuam (CINÀ et al., 2023; WU et al., 2025). Seguindo uma taxonomia proeminente na literatura recente, as defesas são divididas em quatro categorias

principais: (1) pré-treinamento, (2) durante o treinamento, (3) pós-treinamento e (4) em tempo de inferência (WU et al., 2025). A Figura 2 organiza essas categorias ao longo do ciclo de vida do modelo e situa, em cada uma, exemplos de defesas, incluindo o NFT estudado neste trabalho.

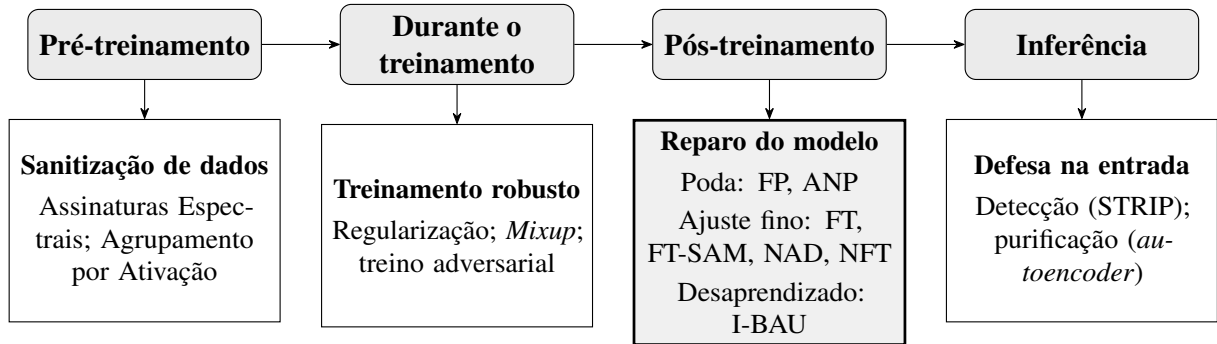


Figura 2 – Taxonomia das defesas contra *backdoor* segundo o estágio do ciclo de vida do modelo em que atuam; o reparo de modelo em pós-treinamento, foco deste trabalho, aparece destacado. Adaptado de (WU et al., 2025).

2.3.1 Defesas de Pré-Treinamento (Sanitização de Dados)

A primeira linha de defesa contra ataques de envenenamento atua diretamente sobre o conjunto de dados antes do início do treinamento do modelo. Conhecidas como técnicas de sanitização de dados (*data sanitization*), seu objetivo é identificar e remover amostras potencialmente maliciosas de um conjunto de dados não confiável (CINÀ et al., 2023). A premissa central é que, para um ataque ser eficaz, as amostras de veneno devem ser suficientemente distintas das amostras limpas para influenciar o aprendizado, o que, por sua vez, pode torná-las detectáveis como anomalias ou *outliers* (CINÀ et al., 2023).

2.3.1.1 Detecção de Outliers no Espaço de Entrada

A detecção de *outliers* no espaço de entrada é uma das abordagens mais diretas para a sanitização de dados, com raízes no campo da estatística robusta (GOLDBLUM et al., 2022). Essas defesas operam identificando e removendo pontos que se afastam excessivamente do centro de sua respectiva classe. Técnicas como a L2, que descarta os pontos mais distantes do centroide da classe, e a Slab, que projeta os dados em uma linha que conecta os centroides de duas classes antes de remover os pontos mais afastados, exemplificam essa estratégia (FAN et al., 2022).

Apesar da sua simplicidade, a eficácia dessa abordagem é limitada, especialmente em domínios de alta dimensão. Em conjuntos de dados complexos como os de imagens, a alta variância natural dos dados pode mascarar o sinal das amostras envenenadas, tornando a separação entre dados limpos e maliciosos ineficaz (TRAN; LI; MADRY, 2018).

Adicionalmente, a detecção de *outliers* no espaço de entrada é vulnerável a ataques adaptativos, nos quais um adversário cria amostras de veneno intencionalmente projetadas

para não parecerem anomalias (FAN et al., 2022; GOLDBLUM et al., 2022). Na prática, foi demonstrado que ataques de envenenamento cuidadosamente elaborados podem gerar amostras que não são identificadas por esses detectores, mesmo com limiares de detecção generosos (CHEN et al., 2017).

2.3.1.2 *Detecção no Espaço de Características Latentes*

Devido às limitações da análise no espaço de entrada, uma alternativa mais robusta é analisar os dados no espaço de características latentes, ou seja, nas representações internas aprendidas por uma rede neural (GOLDBLUM et al., 2022; CINÀ et al., 2023). A intuição é que, nesse espaço, os sinais necessários para a classificação são amplificados, tornando a diferença entre amostras limpas e contaminadas mais pronunciada (TRAN; LI; MADRY, 2018; GOLDBLUM et al., 2022). Em ataques de *backdoor*, por exemplo, o gatilho tende a gerar ativações neurais distintas que podem ser identificadas (TRAN; LI; MADRY, 2018). Duas técnicas proeminentes que exploram essa propriedade são as Assinaturas Espectrais (TRAN; LI; MADRY, 2018) e o Agrupamento por Ativação (CINÀ et al., 2023).

2.3.2 **Defesas Durante o Treinamento (Treinamento Robusto)**

Em vez de sanitizar os dados previamente, uma segunda categoria de defesas modifica o próprio processo de aprendizado para torná-lo inerentemente mais robusto à presença de amostras envenenadas (CINÀ et al., 2023; WU et al., 2025). O objetivo central dessas abordagens, conhecidas como treinamento robusto, é limitar a influência que amostras maliciosas individuais podem exercer sobre os parâmetros finais do modelo.

2.3.2.1 *Regularização e Aumento de Dados*

Uma estratégia primária de robustez é o uso de regularização, que penaliza a complexidade excessiva do modelo e o incentiva a aprender funções de decisão mais simples e suaves. Essa suavidade torna o modelo menos propenso a se ajustar a *outliers* e, consequentemente, mais resistente ao envenenamento (GOLDBLUM et al., 2022). De forma complementar, técnicas de aumento de dados como o *Mixup* e o *CutMix*, que criam novas amostras combinando partes de imagens e rótulos existentes, também promovem essa suavização e dificultam a memorização de gatilhos de *backdoor* específicos (FAN et al., 2022; GOLDBLUM et al., 2022).

2.3.2.2 *Treinamento Adversarial*

Outra abordagem relevante é o treinamento adversarial. Embora tenha sido originalmente projetado para defender contra ataques de evasão, que ocorrem no momento da inferência, ele força o modelo a aprender características mais robustas e generalizáveis ao ser treinado com

exemplos adversariais. Como efeito secundário, pode conferir resistência contra certos ataques de envenenamento, pois o modelo se torna menos sensível a pequenas perturbações nos dados, incluindo os gatilhos de *backdoor* mais sutis (GOLDBLUM et al., 2022).

2.3.3 Defesas de Pós-Treinamento (Reparo de Modelo)

Atuando sobre um modelo já treinado e potencialmente comprometido, esta categoria de defesas busca detectar a presença de um *backdoor* e, em seguida, repará-lo ou mitigá-lo. Essas estratégias são particularmente úteis em cenários onde não há acesso ao conjunto de dados de treinamento original, ou apenas a uma pequena porção de dados limpos (WU et al., 2025). As abordagens distribuem-se nas frentes descritas a seguir, com destaque para as defesas que integram a comparação experimental deste trabalho:

2.3.3.1 Modificação da Estrutura do Modelo

Esta abordagem, chamada sanitização de modelo, identifica e remove componentes da rede responsáveis pelo comportamento do *backdoor* (CINÀ et al., 2023). A técnica mais comum é a poda de neurônios (*neuron pruning*), que elimina neurônios com base em métricas de contribuição ao ataque, como padrões de ativação anômalos ou alta sensibilidade a perturbações adversariais (WU et al., 2025; GOLDBLUM et al., 2022). Duas defesas dessa categoria são avaliadas neste trabalho.

A Fine-Pruning (FP) (LIU; DOLAN-GAVITT; GARG, 2018) observa que neurônios explorados pelo *backdoor* tendem a permanecer inativos diante de entradas legítimas: a técnica registra a ativação média dos neurônios da última camada convolucional em dados limpos, poda iterativamente os menos ativos e, em seguida, aplica ajuste fino para recuperar a acurácia e enfraquecer o ataque residual.

Já a Adversarial Neuron Pruning (ANP) (WU; WANG, 2021) parte da observação de que neurônios associados ao *backdoor* são mais sensíveis a perturbações adversariais aplicadas aos pesos da rede: sob essas perturbações, a defesa otimiza uma máscara contínua por neurônio, usando um pequeno conjunto de dados limpos, e remove os neurônios que a máscara identifica como mais implicados no comportamento malicioso.

2.3.3.2 Ajuste Fino dos Pesos (Fine-tuning)

Esta abordagem ajusta os pesos do modelo comprometido utilizando um pequeno conjunto de dados limpos. O objetivo consiste em eliminar a associação maliciosa entre o gatilho e o rótulo-alvo, substituindo o conhecimento indesejado por informações benignas (GOLDBLUM et al., 2022). Na forma tradicional, o ajuste fino (FT) reotimiza todos os pesos da rede com base nos dados limpos, promovendo o esquecimento progressivo da associação maliciosa, que deixa de ser reforçada durante o processo.

Aprimoramentos dessa abordagem visam superar limitações conhecidas do procedimento. O método FT-SAM (ZHU et al., 2023) baseia-se na observação de que neurônios associados ao *backdoor* frequentemente apresentam pesos de norma elevada, pouco afetados pelo ajuste fino convencional. Para mitigar esse problema, o otimizador é substituído pela minimização sensível à curvatura (*Sharpness-Aware Minimization*). Em cada iteração, uma maximização interna perturba os pesos na vizinhança do ponto atual antes do passo de descida, conduzindo a mínimos mais planos e reduzindo a norma dos neurônios comprometidos.

O método Neural Attention Distillation (NAD) (LI et al., 2021) utiliza a destilação, em que uma cópia do modelo comprometido, ajustada com dados limpos, atua como professora. O modelo original é então reajustado com um termo adicional que alinha seus mapas de atenção intermediários aos da professora, realinhando as regiões da imagem utilizadas pela rede e eliminando a atenção associada ao gatilho. Uma estratégia relacionada, Neural Cleanse (NC) (WANG et al., 2019), primeiramente reconstrói o gatilho por meio de otimização e, em seguida, o neutraliza reajustando o modelo para desaprender a associação com o rótulo-alvo.

2.3.3.3 Desaprendizado Adversarial

Sem a necessidade de podar a rede ou restringir o ajuste a camadas específicas, o I-BAU (*Implicit Backdoor Adversarial Unlearning*) (ZENG et al., 2022) formula a purificação como um problema minimax. Um passo interno sintetiza, a partir dos dados limpos, perturbações adversariais universais que simulam o efeito de possíveis gatilhos. Em seguida, um passo externo reajusta o modelo para desaprender a resposta a essas perturbações. A resolução do problema por meio de hipergradientes implícitos elimina a necessidade de reconstruir o gatilho para cada classe, tornando o procedimento computacionalmente eficiente.

2.3.3.4 Ajuste Fino por Máscaras Neurais

Uma abordagem recente de ajuste fino busca mitigar o risco de degradação da acurácia ao restringir os parâmetros atualizados. Em vez de reotimizar todos os pesos da rede, o NFT (*Augmented Neural Fine-Tuning*) (KARIM et al., 2024) aprende um conjunto reduzido de máscaras escalares, uma por canal, que reescalam as camadas de normalização por lote (*Batch Normalization*), mantendo os demais parâmetros do modelo congelados.

O método parte da premissa de que a associação maliciosa se manifesta de forma concentrada em um subconjunto de canais, de modo que reescalá-los, com o auxílio de um pequeno conjunto de dados limpos, é suficiente para suprimir o *backdoor* e preservar a função benigna (KARIM et al., 2024). Ao ajustar apenas essas máscaras, o método se diferencia das demais técnicas de ajuste fino pela parcimônia no conjunto de parâmetros otimizados. O custo de execução efetivo é analisado na avaliação experimental. O mecanismo detalhado do método, acompanhado do pseudocódigo correspondente, é apresentado na Capítulo 4, juntamente com a descrição de sua integração à plataforma.

2.3.4 Defesas em Tempo de Inferência

A última linha de defesa atua durante a fase de inferência, quando o modelo já está em produção e sendo utilizado para fazer previsões (WU et al., 2025). Diferentemente das categorias anteriores, o objetivo aqui não é reparar o modelo, mas sim neutralizar o ataque no momento da sua execução, impedindo que uma entrada contendo um gatilho malicioso ative o *backdoor*. Essas técnicas podem ser divididas em duas abordagens principais:

2.3.4.1 Detecção de Entradas Suspeitas

Esta estratégia analisa cada entrada antes de passá-la ao modelo, funcionando como um filtro. Se uma entrada é classificada como suspeita de conter um gatilho, ela pode ser rejeitada ou sinalizada para análise posterior. A defesa STRIP (GAO et al., 2019), por exemplo, mede a aleatoriedade (ou entropia) das previsões do modelo ao sobrepor a imagem de entrada com vários padrões. Uma baixa entropia nas previsões é um forte indicador de que o modelo está reagindo a um gatilho de *backdoor*, e a entrada é considerada suspeita.

2.3.4.2 Purificação de Entradas

Uma segunda abordagem, em vez de apenas rejeitar a entrada, busca ativamente purificá-la. O objetivo é pré-processar a amostra para remover ou neutralizar o gatilho malicioso, permitindo que o modelo original faça uma classificação correta sobre a versão limpa da entrada (WU et al., 2025). Uma técnica comum para isso é o uso de um autoencoder treinado em dados limpos. Ao passar uma imagem com *backdoor* pelo autoencoder, o processo de reconstrução tende a eliminar perturbações anômalas, como os gatilhos, que não fazem parte da distribuição natural dos dados, limpando a entrada antes da classificação final (GOLDBLUM et al., 2022).

3 REVISÃO SISTEMÁTICA DA LITERATURA

Para garantir uma análise abrangente e imparcial do conhecimento existente, foi conduzida uma Revisão Sistemática da Literatura (RSL), seguindo as diretrizes de Kitchenham (KITCHENHAM et al., 2004). As etapas do protocolo de revisão são descritas a seguir.

3.1 QUESTÕES DE PESQUISA

Para guiar a Revisão Sistemática da Literatura e delimitar o escopo da pesquisa, foram formuladas as seguintes questões de pesquisa (QPs):

- **QP1:** Como a literatura atual classifica as estratégias de defesa contra ataques de *backdoor*, e quais técnicas podem ser consideradas o estado da arte em cada categoria?
- **QP2:** Quais são os protocolos de avaliação e as métricas estabelecidos na literatura para medir a eficácia e os compromissos (por exemplo, entre robustez e acurácia) de defesas contra backdoors?
- **QP3:** Dentre os *benchmarks* de segurança em ML existentes, qual oferece a maior cobertura de ataques e defesas de *backdoor* e qual lacuna de implementação (por exemplo, uma defesa promissora ainda não integrada) ele apresenta?

3.2 FONTES DE DADOS E ESTRATÉGIA DE BUSCA

A busca pelos estudos primários foi realizada nas seguintes bases de dados digitais, amplamente reconhecidas: IEEE Xplore, ACM Digital Library e Springer. Para maximizar a cobertura da literatura relevante e alinhar a busca aos objetivos específicos deste trabalho, foi construída a seguinte *string* de busca genérica, que foi adaptada conforme a sintaxe específica de cada base de dados pesquisada.

String de Busca

```
("backdoor attack" OR "trojan attack")
AND ("image classification" OR "CIFAR-10" OR "ImageNet" OR "vision")
AND ("defense" OR "mitigation")
AND ("benchmark" OR "BackdoorBench")
```

3.3 CRITÉRIOS DE INCLUSÃO E EXCLUSÃO

Para garantir a relevância e a qualidade dos estudos selecionados, foram definidos os seguintes critérios:

- **CrITÉrios de Inclusão:**

- CI1.** Artigos publicados em conferências e periódicos de alta relevância, revisados por pares.
- CI2.** Estudos que abordam explicitamente ataques de *backdoor* (ou *trojan attack*) contra modelos de aprendizado de máquina supervisionado.
- CI3.** Estudos que propõem, avaliam, comparam ou fazem a análise de técnicas de defesa contra ataques de *backdoor*.
- CI4.** Artigos publicados a partir de 2017, considerando o crescimento do interesse em ataques de *backdoor* com o avanço do *deep learning*.

- **CrITÉrios de Exclusão:**

- CE1.** Artigos cujo foco principal seja outros tipos de ataque, como ataques de evasão ou ataques de privacidade.
- CE2.** Estudos que não tratem de dados no domínio de imagem, dado que este é o foco da maioria dos *benchmarks* e ataques canônicos.
- CE3.** Artigos que não estejam disponíveis em texto completo ou escritos no idioma inglês.
- CE4.** Trabalhos secundários, como outras revisões sistemáticas ou *surveys*.

3.4 PROCESSO DE SELEÇÃO DOS ESTUDOS

Para a identificação dos estudos primários, a busca foi realizada nas bases de dados digitais IEEE Xplore, ACM Digital Library e Springer. A pesquisa na base ACM Digital Library, contudo, retornou um único resultado, que foi descartado por não atender aos critérios de inclusão; dessa forma, os estudos considerados foram provenientes das bases IEEE Xplore e Springer. Seguindo o protocolo de Kitchenham (KITCHENHAM et al., 2004), realizou-se uma triagem inicial por título e resumo, seguida da leitura integral dos artigos pré-selecionados, extraíndo-se de cada estudo aprovado dados-chave como a técnica de defesa, a disponibilidade de código e os *benchmarks* utilizados. O processo de seleção, desde a busca inicial até a inclusão final dos estudos primários, está ilustrado na Figura 3.

O processo resultou em 106 artigos, sem duplicatas entre as bases, dos quais 12 foram pré-selecionados para leitura integral após a aplicação dos critérios de inclusão e exclusão. Desses, 2 foram classificados como trabalhos secundários (*surveys*) e, embora utilizados para a fundamentação teórica, foram excluídos da análise final, conforme o critério de exclusão CE4. O estudo sobre modelos de linguagem foi mantido por sua alta relevância, apesar de tangenciar o critério CE2, para fornecer uma perspectiva mais ampla. Assim, 10 artigos foram definidos como o conjunto final de estudos primários para a extração de dados. Complementarmente à busca sistemática, realizou-se ainda uma busca dirigida na literatura recente, voltada especificamente à

seleção da defesa a ser integrada à plataforma, cujo resultado é retomado mais adiante, na análise da lacuna.

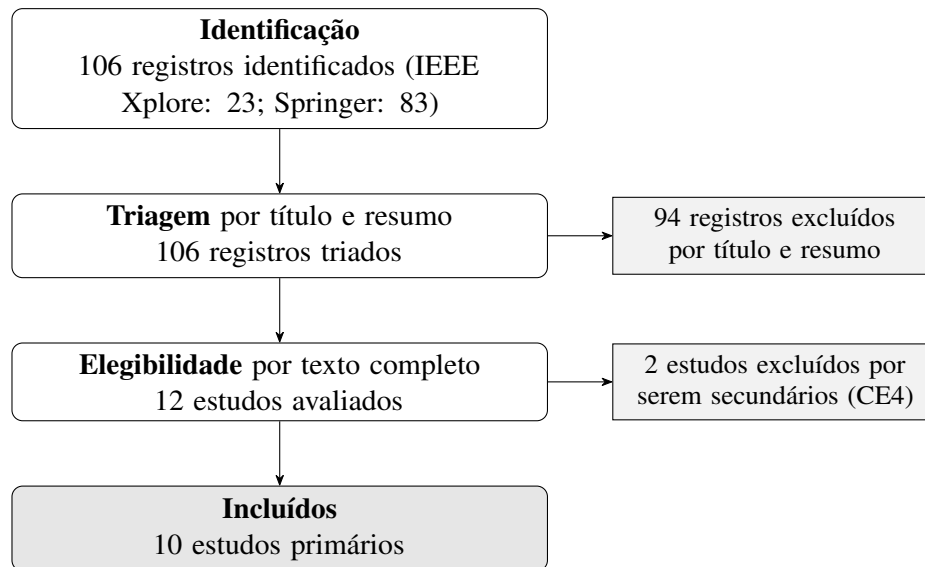


Figura 3 – Fluxo de seleção dos estudos primários da revisão sistemática, no formato PRISMA.

A Tabela 1 apresenta a síntese dos dados extraídos dos 10 estudos primários selecionados. Para cada trabalho, foram identificadas suas principais contribuições, a disponibilidade pública do código-fonte e os conjuntos de dados ou *benchmarks* utilizados em seus experimentos, contextualizando o domínio de avaliação de cada proposta.

Tabela 1 – Síntese dos dados extraídos dos estudos primários selecionados.

Estudo	Técnica Proposta	Disponibilidade de Código	Benchmarks / Conjuntos de Dados Utilizados
(WU et al., 2025)	Plataforma de <i>benchmark</i> unificada para ataques e defesas de <i>backdoor</i> (BackdoorBench).	Sim	Múltiplos conjuntos de dados de visão computacional.
(PANG et al., 2022)	Plataforma de avaliação holística para backdoors neurais (TrojanZoo).	Sim	Múltiplos conjuntos de dados de visão computacional.
(YAN et al., 2022)	Defesa para Aprendizado Federado baseada em mecanismo de dupla atenção (FDDAM).	Não	Conjuntos de dados de classificação de imagem.

Tabela 1 – Síntese dos dados extraídos dos estudos primários selecionados (continuação).

Estudo	Técnica Proposta	Disponibilidade de Código	Benchmarks / Conjuntos de Dados Utilizados
(LI et al., 2024)	Defesa de pós-treinamento contra backdoors condicionais à quantização (EFRAP).	Sim	Múltiplos conjuntos de dados de <i>benchmark</i> .
(ZHANG et al., 2023)	Defesa baseada em aprendizado de representação desconfundida (CBD).	Sim	Múltiplos conjuntos de dados de <i>benchmark</i> .
(ZHAO et al., 2024)	Ataque <i>clean-label</i> (Cbat) e defesa baseada em sanitização de dados (CbatD) para modelos de linguagem.	Não	Conjuntos de dados de classificação de texto.
(HASAN et al., 2024)	Defesa para Aprendizado Federado que detecta e remove clientes maliciosos por consenso (GuardFL).	Não	CIFAR-10, MNIST, FMNIST.
(SALEM et al., 2022)	Classe de ataques de <i>backdoor</i> com gatilhos dinâmicos (Random Backdoor, BaN, c-BaN).	Não	MNIST, CelebA, CIFAR-10.
(DOAN et al., 2021)	Framework de ataque que aprende o gatilho de forma otimizada e furtiva (LIRA).	Sim	MNIST, CIFAR-10, GTSRB, T-ImageNet.
(SUN et al., 2023)	Ataque de <i>backdoor</i> invisível com gatilhos dinâmicos para a tarefa de reidentificação de pessoas (DT-IBA).	Sim	Conjuntos de dados de <i>benchmark</i> de Person Re-ID.

3.5 ANÁLISE DO ESTADO DA ARTE E IDENTIFICAÇÃO DA LACUNA DE PESQUISA

A análise dos trabalhos selecionados revela um campo de pesquisa dinâmico, caracterizado por uma evolução contínua tanto dos ataques de *backdoor* quanto das defesas correspondentes. Para sistematizar esse cenário, estudos recentes, como o da plataforma BackdoorBench, propõem uma classificação das estratégias defensivas de acordo com o estágio do ciclo de vida do modelo em que atuam: pré-treinamento, voltado à detecção de amostras envenenadas; durante

o treinamento, com foco na inibição da injeção do *backdoor*; pós-treinamento, destinado à detecção ou remoção do *backdoor* em modelos já treinados; e em tempo de inferência, visando impedir a ativação do *backdoor* durante a produção (WU et al., 2025).

No estágio de pré-treinamento, o foco está na sanitização dos dados. Abordagens como a CbatD identificam e removem amostras envenenadas ao localizar aquelas com as menores perdas de treinamento e, posteriormente, calcular a relevância de suas características para confirmar a contaminação (ZHAO et al., 2024). Durante o treinamento, o objetivo é aumentar a robustez do processo de otimização. Estratégias como a FDDAM utilizam um mecanismo de dupla atenção, tanto em nível de canal quanto espacial, para induzir o modelo a ignorar os gatilhos e concentrar-se no conteúdo semântico da imagem (YAN et al., 2022). Outra abordagem, a CBD, emprega aprendizado de representação deconfundida para separar os efeitos causais dos dados dos efeitos espúrios introduzidos pelo gatilho (ZHANG et al., 2023).

Em cenários nos quais o modelo já foi comprometido, as defesas de pós-treinamento propõem mecanismos de reparo. Um exemplo é a EFRAP, que neutraliza o *backdoor* ao adotar uma estratégia de arredondamento não convencional durante a quantização, mantendo a acurácia do modelo (LI et al., 2024). A diversidade dessas abordagens evidencia uma especialização crescente, com defesas projetadas para contextos complexos, como o Aprendizado Federado. Nesse contexto, a GuardFL detecta e remove clientes maliciosos por meio de um mecanismo de consenso majoritário (HASAN et al., 2024), além da já mencionada quantização de modelos (LI et al., 2024).

Observa-se um padrão relevante nessa organização: as categorias não recebem atenção uniforme, e a maior parte das defesas propostas pertence ao pós-treinamento, como constata a própria BackdoorBench (WU et al., 2025). Esse desbalanceamento se explica por alguns fatores. O primeiro é o alinhamento das defesas de pós-treinamento ao modelo de ameaça considerado mais realista: em cenários de terceirização do treinamento ou uso de modelos pré-treinados, o defensor geralmente possui apenas o modelo já comprometido e um pequeno conjunto de dados limpos, o que corresponde exatamente ao pressuposto dessas técnicas (KARIM et al., 2024). A sanitização de dados e o treinamento robusto, em contraste, exigem controle sobre os dados ou sobre o próprio processo de treinamento (CINÀ et al., 2023).

Além disso, as defesas de pré-treinamento baseadas em detecção de anomalias apresentam limitações em espaços de alta dimensão (TRAN; LI; MADRY, 2018) e são vulneráveis a ataques adaptativos (CHEN et al., 2017), o que diminui sua atratividade. Ademais, o protocolo de avaliação do pós-treinamento, que envolve aplicar a defesa a um modelo comprometido fixo e medir a variação da Taxa de Sucesso do Ataque (ASR) e da Acurácia em Dados Limpos (CA), é direto e padronizável, o que favorece a rápida iteração e a incorporação em *benchmarks* (PANG et al., 2022). Esses fatores ajudam a explicar a proeminência das técnicas de purificação por ajuste fino, categoria à qual pertence o NFT.

O aumento na sofisticação dos ataques e defesas impõe à avaliação rigorosa um desafio central no campo. Para enfrentar essa complexidade, a comunidade vem desenvolvendo plataformas de *benchmark*, como a TrojanZoo, que propõe uma avaliação holística e prática de

backdoors neurais (PANG et al., 2022), e, com escopo ainda mais amplo, a BackdoorBench (WU et al., 2025). Essas ferramentas têm como objetivo estabelecer protocolos padronizados de avaliação, viabilizando a mensuração justa do progresso na área (WU et al., 2025; PANG et al., 2022). Além disso, tais plataformas permitem quantificar o equilíbrio entre a eficácia dos ataques, mensurada pela ASR, e a manutenção da utilidade do modelo, refletida na CA (PANG et al., 2022).

Ataques como o LIRA são projetados para maximizar a ASR ao mesmo tempo em que produzem gatilhos imperceptíveis, preservando o desempenho do modelo em dados limpos (DOAN et al., 2021). De modo análogo, ataques dinâmicos propostos por Salem et al. atingem altas taxas de sucesso com impacto desprezível na utilidade do modelo (SALEM et al., 2022). Em contraste, defesas como a EFRAP têm como prioridade mitigar o *backdoor* (reduzindo a ASR) ao mesmo tempo que mantêm a acurácia em dados limpos, o que ressalta a necessidade de *benchmarks* consistentes para avaliar esse compromisso (LI et al., 2024).

A literatura aponta a BackdoorBench como a plataforma de *benchmark* mais abrangente atualmente, reconhecimento que é reforçado por outras iniciativas da área, como a TrojanZoo (WU et al., 2025; PANG et al., 2022). No entanto, o rápido surgimento de novas ameaças frequentemente supera a capacidade dessas plataformas de se atualizarem. Entre as ameaças recentes, destacam-se os gatilhos dinâmicos, que empregam padrões e localizações variáveis para cada amostra envenenada, dificultando a detecção baseada em assinaturas fixas (SALEM et al., 2022), além de ataques como o LIRA, que otimizam o gatilho para garantir tanto eficácia quanto imperceptibilidade (DOAN et al., 2021).

Adicionalmente, ataques direcionados a domínios específicos vêm introduzindo novas vulnerabilidades: por exemplo, na reidentificação de pessoas, ataques de *backdoor* invisíveis com gatilhos dinâmicos podem associar a identidade de um alvo a qualquer indivíduo que utilize determinado item, como uma mochila (SUN et al., 2023); já na quantização de modelos, ataques podem ser programados para se manifestarem apenas após a otimização do modelo para produção, permanecendo latentes durante a validação (LI et al., 2024). Isso evidencia uma lacuna contínua entre o desenvolvimento dessas técnicas e sua avaliação sistemática e padronizada.

A partir da análise do estado da arte, este trabalho delimita seu foco. A decisão de partida foi expandir a BackdoorBench, identificada na revisão como a plataforma mais abrangente, integrando-lhe uma defesa de pós-treinamento ainda ausente. Isso impõe um critério: a defesa precisa ser compatível com o modelo de ameaça da plataforma, baseado em ataques de ativação direta. O pós-treinamento mostrou-se a categoria mais promissora para essa busca, por concentrar técnicas de purificação por ajuste fino diretamente comparáveis às já implementadas.

Sob esse critério, a EFRAP surgiu como primeira opção, por ser recente e ter código disponível, mas atua sobre *backdoors* condicionados à quantização, que só se manifestam após a otimização do modelo para produção (LI et al., 2024), fora do escopo da plataforma. Uma busca direcionada levou então ao NFT (*Augmented Neural Fine-Tuning*) (KARIM et al., 2024), uma defesa de ajuste fino que dispensa a busca adversarial de métodos como a ANP (WU; WANG, 2021) e a I-BAU (ZENG et al., 2022), oferecendo uma alternativa mais simples e diretamente

comparável às defesas já presentes na plataforma.

Em síntese, a revisão aborda e responde às questões de pesquisa propostas. Sobre a QP1, a literatura classifica as defesas conforme o estágio do ciclo de vida do modelo em que atuam (pré-treinamento, durante o treinamento, pós-treinamento e inferência), sendo o pós-treinamento o foco das técnicas mais inovadoras, especialmente as de purificação via ajuste fino. Em relação à QP2, o protocolo vigente consiste em aplicar a defesa ao modelo comprometido e quantificar o compromisso entre ASR e CA, conforme padronizado pelas plataformas de *benchmark*. Quanto à QP3, a BackdoorBench destaca-se como a plataforma com maior abrangência de ataques e defesas, enquanto a lacuna identificada, corroborada pela busca complementar, refere-se à ausência do NFT, uma defesa de pós-treinamento alinhada ao escopo da plataforma.

4 METODOLOGIA

Este capítulo apresenta a metodologia experimental utilizada para avaliar, no *benchmark* BackdoorBench, a defesa selecionada com base na revisão da literatura (Capítulo 3). São detalhados a delimitação do escopo do estudo, a preparação do ambiente experimental e o protocolo de avaliação que orienta a comparação entre as defesas.

4.1 DELIMITAÇÃO DO ESCOPO

O foco deste estudo recai sobre ataques de *backdoor* devido a dois fatores principais: sua elevada relevância como ameaça e a viabilidade de simulação em ambiente experimental. Ataques de *backdoor* comprometem a integridade dos modelos de maneira furtiva, pois um sistema infectado pode operar normalmente até ser ativado por um gatilho secreto (GU et al., 2019; SAHA; SUBRAMANYA; PIRSIYAVASH, 2020). Essa característica torna tais ataques uma preocupação central tanto para a indústria (GROSSE et al., 2023) quanto para a comunidade acadêmica (WU et al., 2025).

Sob a perspectiva da viabilidade, a simulação de ataques canônicos, como o BadNets, constitui um procedimento consolidado (GU et al., 2019). A disponibilidade de plataformas como o BackdoorBench, que oferecem implementações prontas, possibilita direcionar o esforço de pesquisa para a avaliação das defesas, em vez da implementação dos ataques, tornando o escopo do projeto mais gerenciável e robusto (WU et al., 2025).

4.2 PREPARAÇÃO DO AMBIENTE EXPERIMENTAL

A etapa experimental foi conduzida utilizando a plataforma BackdoorBench (WU et al., 2025), selecionada na revisão sistemática por integrar, em um único arcabouço, implementações padronizadas de diversos ataques e defesas de *backdoor*, bem como um protocolo de avaliação uniforme. Essa padronização possibilita uma comparação justa entre técnicas, ao assegurar que ataques, defesas e métricas sejam computados de maneira idêntica para todos os métodos avaliados.

Os experimentos foram executados em uma GPU NVIDIA RTX 5070 Ti, utilizando PyTorch 2.11.0 e torchvision 0.26.0 sobre CUDA 12.8. Por se tratar de *hardware* recente, foi necessário adotar essa versão do PyTorch, mais nova que a originalmente empregada pela BackdoorBench, o que exigiu ajustes pontuais de compatibilidade no código da plataforma, sem alteração da lógica dos algoritmos. Toda a matriz experimental foi conduzida em ambiente controlado, com sementes fixadas em cada execução, a fim de garantir a reprodutibilidade dos resultados.

4.3 INTEGRAÇÃO DO NFT À BACKDOORBENCH

A BackdoorBench estrutura ataques e defesas como módulos independentes que compartilham um fluxo de trabalho unificado. Cada defesa de pós-treinamento recebe como entrada o resultado de um ataque, ou seja, o modelo comprometido e os conjuntos de dados correspondentes, aplica seu procedimento de reparo a um pequeno subconjunto de dados limpos e é avaliada pelas mesmas rotinas de medição da plataforma. A integração de uma nova defesa requer implementar um módulo compatível com essa interface e fornecer um arquivo de configuração com seus hiperparâmetros, o que assegura que a técnica adicionada seja executada e avaliada de maneira idêntica às já existentes, condição essencial para uma comparação justa.

Seguindo essa convenção, o NFT foi implementado como um novo módulo de defesa que estende a classe base da plataforma e reaproveita suas rotinas de carregamento do modelo comprometido, de avaliação e de salvamento do modelo reparado. No fluxo da BackdoorBench, ele ocupa a mesma posição das defesas já existentes, como ilustra a Figura 4, antes de detalharmos o mecanismo interno do método.

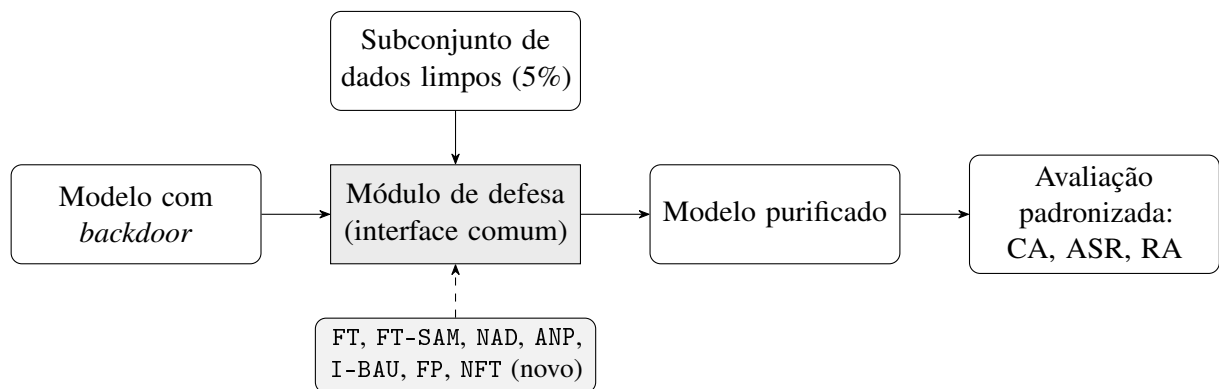


Figura 4 – Integração do NFT como novo módulo de defesa no fluxo da BackdoorBench.

Estabelecido o encaixe, resta explicar o que o módulo faz internamente. Uma rede neural processa a imagem em camadas, e entre elas estão as camadas de normalização, que ajustam a intensidade com que cada canal de informação segue adiante na rede. Cada canal tem um fator de escala, denotado γ , que regula essa intensidade. O NFT parte da hipótese de que o *backdoor* se concentra em alguns canais, e de que atenuá-los basta para neutralizá-lo, sem reajustar a rede inteira. Esses canais não são identificados de forma explícita: o método ajusta todas as máscaras simultaneamente para que o modelo continue classificando bem os dados limpos e, como o gatilho não está presente nesses dados, os canais que servem apenas ao *backdoor* podem ser atenuados sem prejudicar a tarefa benigna.

Para isso, o método associa a cada canal uma máscara m , um fator ajustável entre 0 e 1 que multiplica a escala original, resultando na escala efetiva $\gamma \odot m$: quando a máscara vale 1, o canal permanece intacto; quanto mais próxima de 0, mais o canal é atenuado. Apenas essas máscaras são otimizadas durante a defesa, com um pequeno conjunto de dados limpos, enquanto

todos os demais pesos da rede permanecem congelados.

O ajuste das máscaras combina três componentes, que reaparecem no laço do Algoritmo 1. O *MixUp* é uma técnica de aumento de dados que cria imagens intermediárias misturando pares de exemplos limpos, dificultando a reativação do gatilho e dando nome ao método (*augmented*). O regularizador penaliza o afastamento das máscaras em relação a 1, isto é, mede o quanto cada máscara se distanciou da posição neutra ($\|\mathbf{1} - m\|_1$) e desencoraja mudanças desnecessárias, concentrando as alterações nos canais comprometidos. A projeção, por fim, limita cada máscara ao intervalo $[\mu(l), 1]$, impedindo que um canal seja zerado por completo, com um piso $\mu(l)$ que depende da profundidade l da camada. Ao final, cada máscara é fundida de volta na escala da normalização ($\gamma \leftarrow \gamma \odot m$), e a rede recupera a arquitetura original, sem custo adicional de inferência.

Algoritmo 1: Augmented Neural Fine-Tuning (NFT)

Entrada: modelo com *backdoor* f_θ ; subconjunto de dados limpos D_c com N_c amostras por classe; número de iterações T ; minilotes por iteração B ; taxa de aprendizado η ; coeficiente η_c ; parâmetros da projeção α e β .

Saída: modelo purificado $f_{\theta'}$.

- 1: substituir cada camada de normalização por uma versão mascarada, de escala efetiva $\gamma \odot m$; inicializar $m \leftarrow \mathbf{1}$
- 2: congelar todos os parâmetros do modelo, exceto as máscaras m
- 3: **para** $t = 1, \dots, T$ **faça**
- 4: **para cada** um dos B minilotes de 128 amostras reamostradas de D_c **faça**
- 5: amostrar $\lambda \sim \text{Beta}(1, 1)$ e formar $\tilde{x}$ por *MixUp*
- 6: $\mathcal{L} \leftarrow \mathcal{L}_{\text{CE}}^{\text{MixUp}}(f(\tilde{x})) + \frac{\eta_c}{N_c} \|\mathbf{1} - m\|_1$
- 7: atualizar as máscaras m por um passo de gradiente sobre $\mathcal{L}$, com taxa η
- 8: **para cada** camada mascarada l **faça**
- 9: $m_l \leftarrow \text{clip}(m_l, \mu(l), 1)$, com $\mu(l) = \alpha e^{-\beta l}$
- 10: fundir as máscaras na escala, $\gamma \leftarrow \gamma \odot m$, restaurando a normalização padrão
- 11: **retorna** $f_{\theta'}$

A implementação foi desenvolvida em conformidade com o código de referência disponibilizado pelos autores do NFT¹, reproduzindo seus elementos centrais: o comportamento das camadas de normalização mascaradas, o orçamento de iterações com reamostragem e a projeção das máscaras a cada passo. Como verificação empírica de fidelidade, observou-se que, contra o BadNets, o módulo reduz a taxa de sucesso do ataque a 2,3%, valor próximo ao 1,74% relatado pelos autores na avaliação original, o que indica que a integração preserva o comportamento esperado do método.

¹ Disponível em <https://github.com/nazmul-karim170/NFT-Augmented-Backdoor-Purification>.

4.4 PROTOCOLO DE AVALIAÇÃO EXPERIMENTAL

4.4.1 Conjunto de Dados e Modelos Comprometidos

A avaliação utilizou o conjunto de dados CIFAR-10 (KRIZHEVSKY, 2009) e a arquitetura PreActResNet18 (HE et al., 2016), combinação amplamente adotada na literatura de *backdoor* e padrão na BackdoorBench. Cada modelo comprometido foi obtido treinando-se a rede sobre o conjunto envenenado, com taxa de envenenamento de 10% e classe-alvo fixa, seguindo o protocolo padrão da plataforma. Esses modelos comprometidos constituem o ponto de partida comum sobre o qual todas as defesas são posteriormente aplicadas.

4.4.2 Ataques Avaliados

Para abordar diferentes tipos de gatilhos, foram selecionados seis ataques, conforme ilustrado na Figura 5: BadNets (GU et al., 2019), que insere um *patch* localizado; WaNet (NGUYEN; TRAN, 2021), baseado em deformação geométrica imperceptível; InputAware (NGUYEN; TRAN, 2020), cujo gatilho é dinâmico e dependente da entrada, gerado por uma rede específica para cada amostra; Blended (CHEN et al., 2017), que sobrepõe um padrão global à imagem inteira; SIG (BARNI; KALLAS; TONDI, 2019), um ataque *clean-label* que adiciona um sinal senoidal global à imagem; e LF (*Low-Frequency*) (ZENG et al., 2021), que aplica um gatilho global concentrado nas baixas frequências do espectro da imagem.

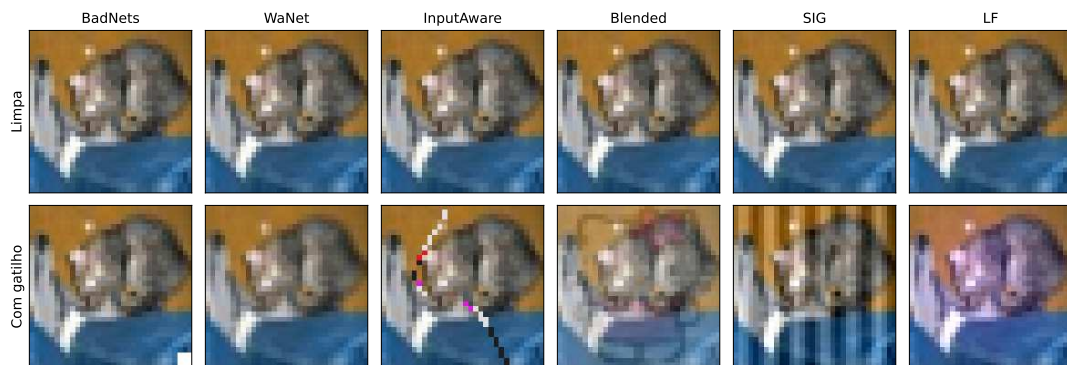


Figura 5 – Exemplos dos gatilhos avaliados sobre imagens do CIFAR-10. Para cada ataque, a linha superior mostra a imagem limpa e a inferior, a versão com o gatilho.

Para o SIG, adotou-se a configuração padrão da BackdoorBench, que implementa a formulação aditiva original de Barni et al. (BARNI; KALLAS; TONDI, 2019): um sinal senoidal de amplitude elevada ($\Delta = 40$, em escala de 0 a 255) somado à imagem inteira. Por ser um ataque *clean-label*, apenas amostras já pertencentes à classe-alvo são envenenadas, sem troca de rótulo; com taxa de 10% sobre o conjunto completo, isso abrange a totalidade da classe-alvo. Essa parametrização difere da empregada na avaliação original do NFT (KARIM et al., 2024),

que usa um sinal mais fraco, sobreposto por mesclagem ($\alpha = 0,2$), e envenena uma fração menor das amostras (80% da classe-alvo). A distinção será retomada na discussão dos resultados.

4.4.3 Defesas Avaliadas

A defesa proposta, o NFT, foi comparada a seis defesas de pós-treinamento já implementadas na BackdoorBench (Figura 2): ajuste fino tradicional (FT), ajuste fino com minimização sensível à curvatura (FT-SAM) (ZHU et al., 2023), destilação por atenção neural (NAD) (LI et al., 2021), poda adversarial de neurônios (ANP) (WU; WANG, 2021), desaprendizado adversarial implícito (I-BAU) (ZENG et al., 2022) e poda com ajuste fino (FP) (LIU; DOLAN-GAVITT; GARG, 2018). Os mecanismos dessas defesas foram detalhados na Seção 2.3. O NFT foi executado conforme a configuração recomendada pelos autores, enquanto as defesas de comparação utilizaram os hiperparâmetros padrão da plataforma.

À exceção da fração de dados limpos, o NFT reproduz fielmente a configuração do código de referência dos autores. Três controles asseguraram a equidade da comparação entre todas as defesas: a aplicação aos mesmos modelos comprometidos, gerados com a mesma semente; o uso da mesma fração de dados limpos, 5% do conjunto de treinamento; e a execução no mesmo ambiente computacional. O segundo desses controles é justamente a alteração mencionada: a avaliação original do NFT emprega uma fração menor de dados limpos, e igualá-la à das demais defesas garante que todas operem sob o mesmo orçamento de dados.

4.4.4 Métricas e Procedimento de Avaliação

A avaliação emprega três métricas, todas expressas em valores percentuais e resumidas na Tabela 2. Em conjunto, elas capturam tanto a utilidade preservada do modelo quanto a eficácia da defesa: a CA mede o quanto o modelo continua acertando em dados limpos, a ASR mede o quanto o ataque ainda funciona, e a RA vai além da simples neutralização do gatilho, verificando se a predição correta foi de fato recuperada. Para mitigar o efeito da aleatoriedade do treinamento, toda a matriz de ataques e defesas foi executada com três sementes independentes, reportando-se a média e o desvio-padrão.

Tabela 2 – Métricas de avaliação. A seta indica a direção desejável após a defesa.

Métrica	O que mede	Desejável
Acurácia em Dados Limpos (CA)	Fração de amostras limpas classificadas corretamente	↑
Taxa de Sucesso do Ataque (ASR)	Fração de amostras com gatilho classificadas na classe-alvo	↓
Acurácia Robusta (RA)	Fração de amostras com gatilho classificadas no rótulo verdadeiro	↑

5 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados da avaliação experimental da defesa NFT, integrada à plataforma BackdoorBench, em comparação com defesas de pós-treinamento já disponíveis na ferramenta, conforme o protocolo descrito no Capítulo 4. Inicialmente, são apresentados os resultados comparativos referentes a um conjunto diversificado de ataques de *backdoor*. Posteriormente, os resultados são discutidos, com ênfase no equilíbrio entre a eficácia de mitigação e a preservação da utilidade do modelo.

5.1 RESULTADOS COMPARATIVOS

A Tabela 3 apresenta a ASR de cada defesa contra cada ataque, a Tabela 4 apresenta a CA correspondente e a Tabela 5, a RA. Os valores correspondem à média sobre três execuções independentes (sementes 0, 1 e 2), acompanhada do respectivo desvio-padrão. Nas tabelas de ASR e de RA, o melhor valor por coluna é destacado em negrito.

Tabela 3 – Taxa de Sucesso do Ataque (ASR, em %) após cada defesa, em média sobre três sementes ($\pm$ desvio-padrão). Ataques agrupados por natureza do gatilho: localizado, geométrico e dinâmico à esquerda; aditivos e de mesclagem à direita.

Defesa	BadNets	WaNet	InputAware	Blended	SIG	LF
Sem defesa	94,5 $\pm$ 0,0	95,6 $\pm$ 1,1	93,9 $\pm$ 3,8	99,8 $\pm$ 0,1	97,5 $\pm$ 0,6	99,1 $\pm$ 0,2
NFT	2,3 $\pm$ 1,1	1,4 $\pm$ 0,6	1,9 $\pm$ 0,4	83,5 $\pm$ 3,8	23,5 $\pm$ 2,6	77,6 $\pm$ 6,4
FT	1,3 $\pm$ 0,2	1,3 $\pm$ 0,5	26,6 $\pm$ 24,3	92,8 $\pm$ 4,3	4,3 $\pm$ 1,4	78,2 $\pm$ 1,1
FT-SAM	1,6 $\pm$ 0,4	1,0 $\pm$ 0,2	2,5 $\pm$ 0,7	14,8 $\pm$ 4,1	0,4 $\pm$ 0,1	5,9 $\pm$ 1,2
NAD	1,6 $\pm$ 0,2	1,4 $\pm$ 0,6	32,8 $\pm$ 27,2	73,0 $\pm$ 7,6	3,6 $\pm$ 0,5	53,8 $\pm$ 25,0
ANP	0,0 $\pm$ 0,0	0,4 $\pm$ 0,3	0,3 $\pm$ 0,3	3,1 $\pm$ 1,6	2,2 $\pm$ 2,1	0,4 $\pm$ 0,3
I-BAU	1,2 $\pm$ 0,6	18,6 $\pm$ 11,0	19,5 $\pm$ 26,0	11,7 $\pm$ 5,7	2,1 $\pm$ 1,6	15,0 $\pm$ 8,5
FP	0,9 $\pm$ 0,1	5,8 $\pm$ 3,2	15,1 $\pm$ 3,0	19,1 $\pm$ 8,0	2,1 $\pm$ 1,5	37,0 $\pm$ 9,8

Tabela 4 – Acurácia em Dados Limpos (CA, em %) após cada defesa, em média sobre três sementes ($\pm$ desvio-padrão).

Defesa	BadNets	WaNet	InputAware	Blended	SIG	LF
Sem defesa	91,5 $\pm$ 0,1	90,9 $\pm$ 0,3	91,3 $\pm$ 0,1	93,4 $\pm$ 0,1	84,6 $\pm$ 0,1	93,3 $\pm$ 0,0
NFT	90,6 $\pm$ 0,3	91,9 $\pm$ 0,3	91,6 $\pm$ 0,3	92,0 $\pm$ 0,0	89,3 $\pm$ 0,2	91,7 $\pm$ 0,1
FT	90,5 $\pm$ 0,1	93,2 $\pm$ 0,3	93,1 $\pm$ 0,1	92,8 $\pm$ 0,1	90,7 $\pm$ 0,2	92,6 $\pm$ 0,3
FT-SAM	91,6 $\pm$ 0,1	93,3 $\pm$ 0,1	93,3 $\pm$ 0,4	92,7 $\pm$ 0,1	91,3 $\pm$ 0,2	92,5 $\pm$ 0,0
NAD	89,4 $\pm$ 0,2	93,1 $\pm$ 0,3	93,1 $\pm$ 0,1	91,9 $\pm$ 0,2	89,5 $\pm$ 0,1	92,2 $\pm$ 0,2
ANP	85,5 $\pm$ 0,2	87,7 $\pm$ 3,7	84,8 $\pm$ 1,1	85,7 $\pm$ 0,9	80,4 $\pm$ 0,4	86,7 $\pm$ 1,1
I-BAU	87,8 $\pm$ 1,6	90,6 $\pm$ 1,3	90,8 $\pm$ 0,5	89,7 $\pm$ 0,6	86,4 $\pm$ 0,3	88,2 $\pm$ 1,3
FP	91,6 $\pm$ 0,2	93,1 $\pm$ 0,1	93,2 $\pm$ 0,2	92,5 $\pm$ 0,2	90,5 $\pm$ 0,1	92,4 $\pm$ 0,3

Tabela 5 – Acurácia Robusta (RA, em %) após cada defesa, em média sobre três sementes ($\pm$ desvio-padrão).

Defesa	BadNets	WaNet	InputAware	Blended	SIG	LF
Sem defesa	5,3 $\pm$ 0,0	4,1 $\pm$ 1,1	5,9 $\pm$ 3,5	0,2 $\pm$ 0,1	2,5 $\pm$ 0,6	0,8 $\pm$ 0,2
NFT	88,2 $\pm$ 1,2	88,1 $\pm$ 1,9	87,8 $\pm$ 0,4	15,0 $\pm$ 3,3	45,6 $\pm$ 1,5	20,3 $\pm$ 5,8
FT	89,9 $\pm$ 0,2	89,4 $\pm$ 1,2	68,2 $\pm$ 22,3	6,4 $\pm$ 3,6	64,7 $\pm$ 1,4	20,1 $\pm$ 1,2
FT-SAM	91,0 $\pm$ 0,2	90,3 $\pm$ 1,2	89,5 $\pm$ 0,8	61,2 $\pm$ 2,3	63,3 $\pm$ 4,9	82,6 $\pm$ 0,9
NAD	88,8 $\pm$ 0,4	89,0 $\pm$ 0,9	62,2 $\pm$ 24,7	22,9 $\pm$ 6,0	56,0 $\pm$ 3,1	41,7 $\pm$ 21,7
ANP	89,7 $\pm$ 0,9	88,4 $\pm$ 1,5	81,9 $\pm$ 1,8	50,5 $\pm$ 8,7	54,0 $\pm$ 5,3	83,2 $\pm$ 2,5
I-BAU	87,7 $\pm$ 1,9	71,3 $\pm$ 12,2	68,4 $\pm$ 20,8	50,7 $\pm$ 3,6	53,2 $\pm$ 1,2	69,4 $\pm$ 7,3
FP	91,3 $\pm$ 0,2	84,3 $\pm$ 4,8	79,0 $\pm$ 2,4	51,9 $\pm$ 4,4	53,2 $\pm$ 3,2	52,1 $\pm$ 5,8

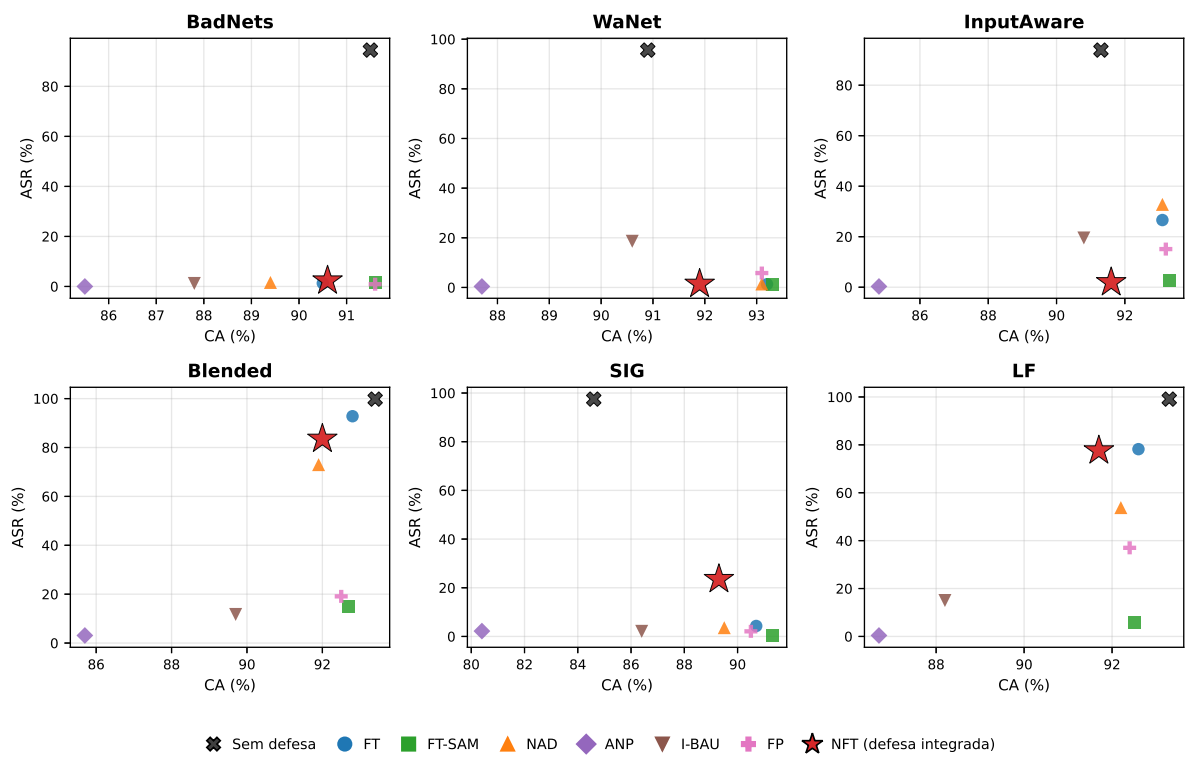


Figura 6 – Compromisso entre utilidade (CA, eixo horizontal) e segurança (ASR, eixo vertical), com um painel por ataque e cor e marcador por defesa, em média sobre três sementes. A posição desejável é o canto inferior direito, que combina CA alta e ASR baixa; o NFT aparece destacado com uma estrela, e o ponto sem defesa serve de referência.

De modo geral, todas as defesas mantêm níveis razoáveis de acurácia em dados limpos, entre 80% e 93%. Por isso, a principal diferença entre elas está na capacidade de reduzir a ASR, e nenhuma apresenta o melhor desempenho contra todos os ataques ao mesmo tempo, como mostra a Figura 6. A Tabela 5 acrescenta uma informação importante: a RA, fração de amostras com gatilho que voltam a ser classificadas no rótulo correto. Nos ataques mais bem mitigados (BadNets e WaNet), a RA é alta, superando 84% para quase todas as defesas, sinal de que as amostras envenenadas recuperam sua classe verdadeira. Já nos gatilhos aditivo e de mesclagem

(Blended e SIG), a RA permanece baixa sempre que a ASR se mantém alta, o que evidencia a dificuldade geral das defesas nesses cenários.

As taxas de sucesso dos ataques sem defesa coincidem com as reportadas na avaliação de referência da BackdoorBench (WU et al., 2025) para o CIFAR-10 e a arquitetura PreActResNet18, o que confirma que a reprodução dos ataques é fiel. Vale notar que o estudo da BackdoorBench também relata que defesas de ajuste fino, como FT e NAD, falham contra o Blended, o mesmo padrão observado aqui. Os comportamentos discutidos a seguir são, portanto, propriedades do *benchmark*, válidas para além deste trabalho, e não efeitos do setup adotado.

5.2 DISCUSSÃO DOS RESULTADOS

5.2.1 Eficácia do NFT por tipo de ataque

O NFT demonstra alta eficácia contra ataques de gatilho localizado e geométrico. Contra o BadNets e o WaNet, reduz a ASR de cerca de 95% para perto de 2%, mantendo a CA em torno de 91%, desempenho comparável ao das principais defesas de ajuste fino, como FT, FT-SAM e NAD. O método também se destaca no InputAware, cujo gatilho, embora dinâmico e dependente da entrada, permanece de natureza localizada: a ASR cai para 1,9%, com a CA preservada, igualando as melhores defesas e ficando atrás apenas da ANP, que tem CA menor. Nesse ataque, o NFT ainda se mostrou bem mais estável entre as sementes do que as outras defesas de ajuste fino, que oscilaram fortemente. Essa estabilidade reforça a robustez do método diante de gatilhos estruturados, mesmo quando variam de uma amostra para outra.

Contra o SIG, um gatilho aditivo sob regime *clean-label*, o NFT alcançou apenas mitigação parcial: a ASR caiu de cerca de 98% para 23,5%, com baixa variação entre as sementes. Embora a redução seja expressiva ante o modelo sem defesa, o valor permanece bem acima do das outras defesas de ajuste fino, que neutralizam o ataque quase por completo, como FT (4,3%) e FT-SAM (0,4%). A trajetória de otimização estabiliza nesse patamar ao longo de todo o ajuste (Figura 7), o que aponta para uma limitação do mecanismo de máscaras diante desse tipo de gatilho, e não uma simples falta de convergência, questão retomada na subseção seguinte. A acurácia robusta correspondente, de cerca de 46%, indica que parte das amostras com gatilho deixa de ir para a classe-alvo, mas não retorna completamente ao rótulo verdadeiro.

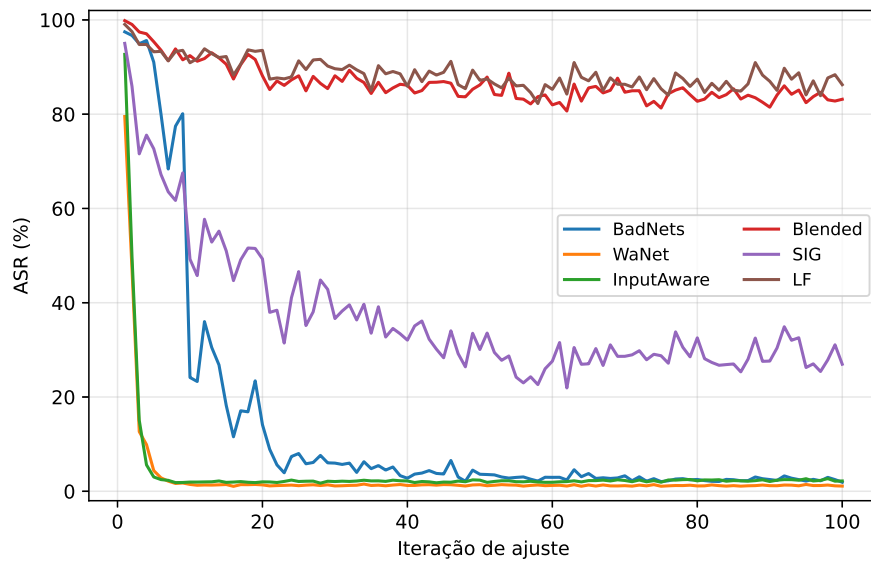


Figura 7 – Evolução da ASR do NFT ao longo das iterações de ajuste fino, para os ataques avaliados (semente 0).

5.2.2 Investigação das limitações frente a gatilhos aditivos e de mesclagem

Os resultados comparativos revelaram um limite claro do NFT: ele neutraliza bem os gatilhos localizados e geométricos, mas perde eficácia diante do SIG, do Blended e do LF. O que aproxima esses três ataques é a forma de inserir o gatilho: em vez de um adesivo num canto da imagem, todos o espalham por todos os pixels, somando, mesclando ou alterando frequências, de modo que ele se dilui nas características legítimas da cena. Esta seção examina em detalhe os casos do SIG e do Blended e, em seguida, mostra que o LF reproduz o mesmo comportamento, evidenciando que a limitação é comum a toda essa categoria.

5.2.2.1 Mitigação parcial do sinal aditivo (SIG)

Antes dos números, convém recuperar o que torna o SIG um caso difícil. Ao contrário de um adesivo localizado, como o do BadNets, seu gatilho é um sinal senoidal somado à imagem inteira, como uma leve textura ondulada distribuída por todos os pixels. Por estar diluído, e não concentrado em um canto, ele se confunde com as características legítimas da imagem, o que antecipa a dificuldade de uma defesa que atua canal a canal para removê-lo.

Esse é o pano de fundo do desempenho apenas parcial do NFT aqui: a ASR cai de forma incompleta, ao passo que a avaliação original do método (KARIM et al., 2024) relata cerca de 0,1%. Como os hiperparâmetros da defesa são equivalentes nos dois cenários, a diferença não está na defesa, mas em como o ataque foi montado. O SIG da BackdoorBench usa a formulação aditiva de Barni et al., com amplitude $\Delta = 40$ (Seção 4.4), enquanto a avaliação original emprega um sinal mais fraco e menos envenenamento. A ablação a seguir isola qual desses fatores pesa mais, variando um parâmetro do ataque por vez (Tabela 6 e Figura 8).

Tabela 6 – Ablação do ataque SIG: ASR (%) sem defesa e após o NFT, variando isoladamente a amplitude Δ do sinal e a fração de envenenamento (*pratio*), em uma única semente.

Δ	<i>pratio</i>	ASR sem defesa	ASR após NFT
40	0,10	98,2	27,0
20	0,10	85,1	23,0
10	0,10	68,2	8,5
40	0,05	96,4	83,4
40	0,02	91,2	81,6

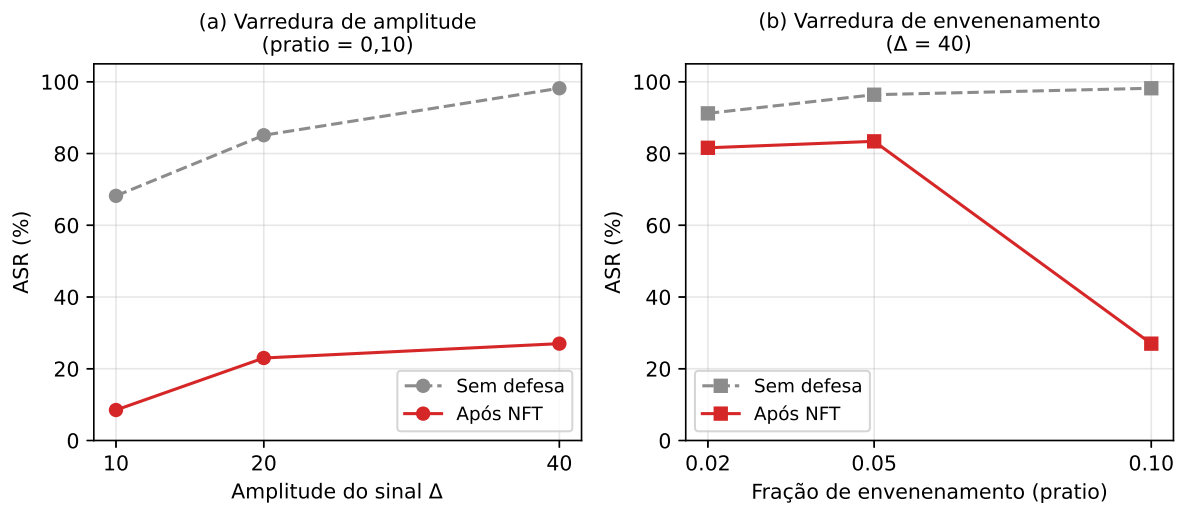


Figura 8 – Ablação do ataque SIG: ASR sem defesa e após o NFT ao variar (a) a amplitude Δ do sinal, com *pratio* fixo, e (b) a fração de envenenamento, com Δ fixo.

O primeiro achado é direto: quanto mais forte o sinal, mais difícil removê-lo. Com o envenenamento fixo, a ASR após o NFT cai de 27,0% em $\Delta = 40$ para 8,5% em $\Delta = 10$, e isso explica por que o método se sai bem diante do sinal mais brando da avaliação original. Parte dessa recuperação vem da própria atenuação do ataque com sinais fracos, já que, sem defesa, a ASR também cai; ainda assim, em termos relativos, a purificação se mostra mais completa quando o sinal é menor.

O segundo achado contraria a intuição: reduzir o envenenamento de 100% para 20% das amostras da classe-alvo piorou o resultado do NFT, de 27,0% para 81,6%. A explicação remete ao modo como o método age. As máscaras do NFT funcionam como botões que reescalam um canal de cada vez; saturar a classe-alvo de gatilho concentra o backdoor em poucos canais, justamente os que esses botões conseguem desligar. Com menos envenenamento, o sinal se espalha de forma difusa por muitos canais e foge ao alcance de um mecanismo que atua sobre cada canal de modo isolado.

Esse contraste delimita o alcance do método. O ajuste fino completo (FT) permanece eficaz no mesmo cenário (4,3%) porque reaprende todos os pesos, e não apenas algumas escalas;

o NFT, sem essa liberdade, não suprime um sinal aditivo de alta amplitude entrelaçado às características legítimas da classe. A limitação, portanto, não está na execução, mas no grau de liberdade que o método se permite.

5.2.2.2 *Falha frente ao gatilho de mesclagem (Blended)*

O caso do Blended é mais grave que o do SIG. Enquanto a avaliação original do NFT relata remoção quase completa do ataque, com ASR de cerca de 0,3%, aqui o método mantém a ASR em torno de 83% durante todo o ajuste, sem qualquer redução (Figura 7). A pergunta que organiza esta subseção é por que a mesma defesa fracassa de forma tão completa, e a resposta, como se verá, está menos no NFT do que na forma como o próprio ataque foi treinado.

A pista decisiva é que a dificuldade não é exclusiva do NFT, mas de toda a família de ajuste fino. O ajuste fino completo (FT) mantém a ASR em cerca de 93% e a NAD em 73%, enquanto a FT-SAM e as defesas de poda e desaprendizado, como a ANP e a I-BAU, reduzem a ASR de forma substancial. O dado que fecha o argumento vem da comparação com o estudo original (KARIM et al., 2024): lá, tanto a FT quanto a NAD removem o Blended com eficácia, ao contrário daqui. O que torna o gatilho removível ou não, portanto, é o modo como o ataque foi treinado, e isso vale para a família de ajuste fino inteira.

Reproduzindo a receita de treino do ataque empregada no estudo original, em especial o agendamento da taxa de aprendizado, o gatilho voltou a ser removível por ajuste fino, e a ASR do NFT caiu para cerca de 35%. O resíduo que ainda separa esse valor dos 0,3% relatados não é estrutural, mas uma questão de medição: ao longo das iterações, o NFT chega a levar a ASR abaixo de 5% nas primeiras passagens, atingindo o patamar do estudo original, para depois subir e oscilar em torno de 40%. Como se reporta a última iteração, registra-se a ponta alta dessa oscilação, e não o fundo onde os dois números coincidem.

5.2.2.3 *O padrão comum aos gatilhos aditivos*

O LF, terceiro gatilho de natureza aditiva avaliado, confirma o padrão já visto no SIG e no Blended. O NFT manteve a ASR em 77,6%, no mesmo nível do ajuste fino completo (FT, 78,2%), enquanto a FT-SAM, a ANP e a I-BAU mitigaram o ataque. A recorrência do mesmo comportamento em três gatilhos distintos, a mesclagem do Blended, o sinal aditivo do SIG e a perturbação de baixa frequência do LF, indica que a limitação não é particularidade de um ataque, mas inerente a essa categoria de gatilho. Vale contrastá-la com o gatilho geométrico do WaNet, que também afeta a imagem inteira e ainda assim é removido com sucesso: o que determina a falha não é a abrangência espacial do gatilho, mas sua natureza aditiva ou de mesclagem.

Essa diferença também aparece no espaço de características aprendido pelo modelo, isto é, na forma como a rede organiza internamente as imagens antes de classificá-las. A Figura 9 usa a técnica *t-SNE* para projetar essas representações em duas dimensões, comparando amostras limpas e com gatilho, antes e depois da purificação. No BadNets, as amostras com gatilho

formam agrupamentos separados no modelo infectado e voltam a se reunir aos agrupamentos legítimos após o NFT, sinal de que recuperaram suas características de classe. No Blended, em contraste, esse agrupamento permanece praticamente inalterado após a purificação, confirmando de forma visual que o *backdoor* não foi removido.

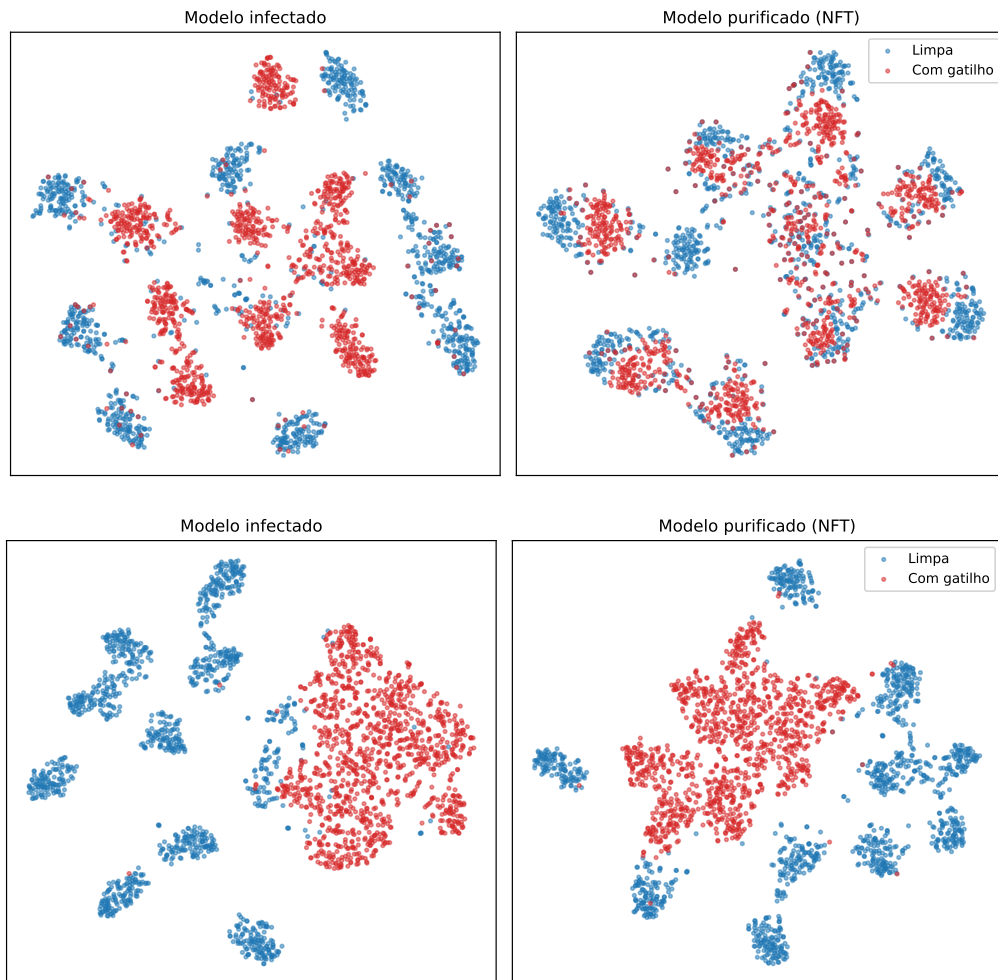


Figura 9 – Projeção t -SNE das representações internas (penúltima camada) de amostras limpas e com gatilho, no modelo infectado e após a purificação pelo NFT, para o BadNets (acima) e o Blended (abaixo).

5.2.3 Robustez dos achados à taxa de envenenamento

A taxa de envenenamento, a fração de dados que o atacante corrompe, é uma escolha dele, não do defensor. Por isso, uma defesa só é confiável se suas conclusões valem para diferentes intensidades de ataque, e não apenas para o valor da matriz principal. Replicando a avaliação a 5% e 20%, o perfil do NFT mostrou-se essencialmente invariante (Tabela 7): contra gatilhos localizados e geométricos, a ASR permaneceu baixa nas três taxas (BadNets entre 4,8% e 5,8%; WaNet em torno de 1%), enquanto contra o Blended manteve-se ineficaz em todas elas, entre 73% e 83%. A intensidade do ataque, portanto, não altera a conclusão qualitativa sobre quais gatilhos o método consegue neutralizar e quais resistem a ele.

Tabela 7 – Taxa de Sucesso do Ataque (ASR, em %) após cada defesa em duas taxas de envenenamento, em uma única semente. Os valores em 10% constam da Tabela 3.

Defesa	<i>pratio</i> = 5%				<i>pratio</i> = 20%			
	BadNets	WaNet	Blended	SIG	BadNets	WaNet	Blended	SIG
Sem defesa	92,0	94,5	99,1	96,4	96,7	94,1	99,9	—
NFT	5,8	0,9	73,3	83,4	4,8	1,1	79,6	—
FT	5,1	34,1	97,3	84,7	1,3	0,6	3,8	—
FT-SAM	1,4	0,8	17,3	0,9	1,3	0,9	3,4	—
NAD	1,8	46,8	96,9	78,8	2,0	0,6	2,8	—
ANP	0,0	0,3	21,9	11,2	0,0	0,7	1,3	—
I-BAU	0,7	14,2	13,5	1,4	1,4	1,3	16,4	—
FP	1,5	3,7	30,9	30,9	0,8	0,9	12,8	—

5.2.4 Generalização ao GTSRB

Um achado obtido em um único conjunto de dados pode refletir peculiaridades dele, e não uma propriedade geral do método. Para testar se as conclusões do CIFAR-10 se sustentam, replicou-se a avaliação no GTSRB, um conjunto de placas de trânsito com 43 classes, mantendo a arquitetura, a taxa de envenenamento de 10% e três sementes. Cinco dos seis ataques foram avaliados; o SIG ficou de fora porque, sendo *clean-label* restrito à classe-alvo, a 10% exigiria mais amostras do que a classe possui. A Tabela 8 apresenta a ASR após cada defesa; a acurácia em dados limpos é comentada na discussão a seguir.

Tabela 8 – Taxa de Sucesso do Ataque (ASR, em %) após cada defesa no GTSRB, em média sobre três sementes ($\pm$ desvio-padrão).

Defesa	BadNets	WaNet	InputAware	Blended	LF
Sem defesa	94,9 $\pm$ 0,2	97,2 $\pm$ 1,2	92,6 $\pm$ 5,7	100,0 $\pm$ 0,0	99,4 $\pm$ 0,0
NFT	43,3 $\pm$ 28,1	0,0 $\pm$ 0,0	17,0 $\pm$ 18,2	94,6 $\pm$ 2,5	84,9 $\pm$ 9,7
FT	33,7 $\pm$ 15,4	5,9 $\pm$ 3,1	67,3 $\pm$ 15,5	99,9 $\pm$ 0,0	65,4 $\pm$ 27,9
FT-SAM	1,9 $\pm$ 1,2	0,0 $\pm$ 0,0	17,7 $\pm$ 21,9	15,8 $\pm$ 6,1	0,2 $\pm$ 0,2
NAD	28,0 $\pm$ 15,8	4,0 $\pm$ 1,2	41,7 $\pm$ 37,9	98,7 $\pm$ 0,9	70,9 $\pm$ 31,1
ANP	0,0 $\pm$ 0,0	0,0 $\pm$ 0,0	0,0 $\pm$ 0,0	11,3 $\pm$ 8,1	0,0 $\pm$ 0,0
I-BAU	0,1 $\pm$ 0,1	0,0 $\pm$ 0,0	0,0 $\pm$ 0,0	51,6 $\pm$ 36,6	11,7 $\pm$ 9,3
FP	0,0 $\pm$ 0,0	0,4 $\pm$ 0,5	1,7 $\pm$ 1,8	66,0 $\pm$ 16,1	26,7 $\pm$ 22,4

O achado central do CIFAR-10 se reproduz: o NFT falha contra os gatilhos aditivos e de mesclagem (ASR de 94,6% no Blended e 84,9% no LF), mas purifica de forma eficaz e estável o gatilho geométrico do WaNet, e a dificuldade segue atingindo toda a família de ajuste fino, não só o NFT. A acurácia em dados limpos ficou entre 89% e 99% para todas as defesas, confirmando que a purificação preserva a utilidade do modelo também aqui.

Surge, porém, uma diferença em relação ao CIFAR-10. Contra o BadNets e o InputAware, que o NFT resolvia com folga no conjunto anterior, todas as defesas de ajuste fino passaram a oscilar fortemente entre as sementes, enquanto a FT-SAM e as defesas de poda e

desaprendizado seguiram estáveis. Parte dessa instabilidade vem da escassez de dados limpos por classe: com 43 classes e distribuição desbalanceada, sobram poucas amostras por classe para o ajuste fino reaprender o comportamento benigno. Esse fator explica a variação entre execuções, mas não o nível de falha que resta, cuja causa permanece em aberto.

Em conjunto, o GTSRB corrobora o achado central e acrescenta uma ressalva. A fragilidade do NFT frente a gatilhos aditivos e de mesclagem reaparece sem alteração, mas a purificação por ajuste fino contra gatilhos localizados e dinâmicos mostrou-se sensível ao conjunto de dados, ao passo que poda, desaprendizado e ajuste fino sensível à curvatura se mantiveram robustos.

5.2.5 Custo computacional das defesas

Até aqui, as defesas foram comparadas pela eficácia: o quanto reduzem a ASR e preservam a acurácia. Mas a escolha de uma defesa na prática também depende do custo, e aqui o NFT merece atenção, pois costuma-se associar à dispensa da busca adversarial um ganho de eficiência que, como se verá, não se confirma no tempo de execução. A Tabela 9 reúne o tempo de cada defesa sobre um mesmo modelo atacado, medido no BadNets com o CIFAR-10, taxa de envenenamento de 5% e mesma GPU, ordenado do menor ao maior.

Tabela 9 – Tempo de execução (em minutos) de cada defesa sobre um mesmo modelo atacado, ordenado do menor ao maior. Os valores são indicativos e dependem do *hardware*; o relevante é a ordem relativa entre as defesas.

Defesa	Tempo total (min)
I-BAU	0,8
NAD	1,0
FT	3,1
FT-SAM	5,3
ANP	5,5
NFT	8,3
FP	23,3

Embora cada iteração do NFT seja rápida, seu orçamento elevado de otimização o coloca entre as defesas mais caras (8,3 minutos), à frente de toda a família de ajuste fino. Esse custo não é compensado por convergência precoce: a purificação costuma exigir o orçamento completo, pois contra alguns ataques a ASR só estabiliza ao fim das iterações. Interromper o ajuste mais cedo, quando a purificação já ocorreu, exigiria saber de antemão o ponto certo de parada, o que pressupõe um acesso à ASR sobre dados com gatilho que o defensor em geral não tem.

Convém, então, qualificar a relação entre custo e busca adversarial. Embora essa etapa seja cara por iteração, ela não determina o custo total, que depende sobretudo do número de iterações: a defesa mais rápida do conjunto é a I-BAU, que realiza busca adversarial, e a mais lenta é a FP, que não a realiza. O NFT, embora prescindida dessa busca, é mais lento que a FT, a NAD e a própria I-BAU, por conta de seu orçamento elevado de iterações. Sua vantagem prática,

portanto, não está no tempo de execução, mas na simplicidade conceitual e na preservação da arquitetura original.

5.2.6 Síntese comparativa e implicações práticas

Reunindo as duas dimensões avaliadas, eficácia e custo, os resultados reforçam que nenhuma defesa domina todas as outras. A Tabela 10 resume o desempenho de cada defesa por categoria de gatilho, classificando-o como eficaz, parcial ou falho conforme a ASR média após a defesa, e acrescenta a preservação da acurácia em dados limpos e o custo de execução. O quadro torna visível o padrão central do trabalho: a divisão entre os gatilhos que o ajuste fino remove e os que resistem a ele.

Tabela 10 – Síntese qualitativa do desempenho das defesas por categoria de gatilho, no CIFAR-10.

Defesa	Localiz./Geom.	Dinâmico	Adit./Mescl.	CA	Custo
FT	Eficaz	Parcial	Falha	Alta	Baixo
FT-SAM	Eficaz	Eficaz	Eficaz	Alta	Médio
NAD	Eficaz	Parcial	Parcial	Alta	Baixo
ANP	Eficaz	Eficaz	Eficaz	Baixa	Médio
I-BAU	Eficaz	Parcial	Eficaz	Média	Baixo
FP	Eficaz	Parcial	Parcial	Alta	Alto
NFT	Eficaz	Eficaz	Falha	Alta	Alto

Eficácia (ASR média após defesa):	Eficaz $\leq 10\%$	Parcial 10–50%	Falha $> 50\%$
CA (perda ante sem defesa):	Alta < 2 pts	Média 2–5 pts	Baixa > 5 pts
Custo (Tabela 9):	Baixo ≤ 3 min	Médio 3–6 min	Alto > 6 min

A leitura por linhas e colunas confirma duas conclusões já discutidas. Nenhuma defesa é eficaz em todas as colunas sem comprometer a CA: a ANP mitiga todas as categorias de gatilho, mas é a que mais sacrifica a acurácia em dados limpos. E a coluna dos gatilhos aditivos e de mesclagem é a que mais separa as defesas: enquanto poda e desaprendizado a enfrentam bem, o ajuste fino tem desempenho irregular ali, com a FT e o NFT falhando e as demais oscilando entre parcial e eficaz. Várias defesas, além disso, variaram bastante entre as sementes, o que reforça a importância de reportar médias; nesse aspecto, o NFT destacou-se pela estabilidade comparativamente alta.

Do ponto de vista prático, esses resultados sugerem uma escolha de defesa guiada pelo tipo de ameaça esperada, que costuma se dividir em dois grandes cenários. No primeiro, o gatilho é visível e localizado, como um adesivo colado em uma placa de trânsito, um logotipo sobreposto ou uma marca no canto da imagem, à maneira do BadNets, ou ainda uma deformação geométrica sutil, como a do WaNet. Contra esse tipo de ameaça, o NFT é uma opção atraente, pois iguala o desempenho das melhores defesas com boa estabilidade entre execuções, sem recorrer a podas ou buscas adversariais.

No segundo cenário, o gatilho se dilui pela imagem inteira como uma perturbação tênue e quase imperceptível, somada ou mesclada aos pixels, como no Blended e no SIG. Aqui o NFT

perde eficácia, e tornam-se preferíveis a poda de neurônios da ANP, que remove diretamente as unidades comprometidas da rede, o desaprendizado adversarial da I-BAU, que treina o modelo a esquecer a resposta ao gatilho, e o ajuste fino sensível à curvatura da FT-SAM, que foi a defesa de desempenho mais consistente ao longo dos seis ataques.

Em síntese, o NFT integra-se de forma competitiva às defesas da BackdoorBench, destacando-se nos gatilhos localizados e geométricos e oferecendo uma alternativa conceitualmente simples, enquanto suas fragilidades frente ao SIG e ao Blended delimitam os cenários em que estratégias complementares são preferíveis. O conjunto dos resultados evidencia o compromisso central da área: nenhuma defesa concilia, ao mesmo tempo, eficácia ampla de mitigação, preservação da utilidade e baixo custo.

6 CONCLUSÃO

Este trabalho investigou os ataques de *backdoor* em aprendizado de máquina, com o objetivo de analisar o estado da arte em defesas e oferecer uma contribuição prática e experimental à área. Uma Revisão Sistemática da Literatura mapeou as estratégias de defesa segundo o estágio do ciclo de vida do modelo e evidenciou que o pós-treinamento é uma das categorias mais ativas, com novas técnicas de purificação surgindo mais rápido do que as plataformas de *benchmark* conseguem incorporá-las.

Diante dessa lacuna, integrou-se à BackdoorBench uma técnica recente ainda ausente, o NFT, implementada como um novo módulo segundo as convenções da plataforma e avaliada sob um protocolo padronizado. A avaliação mostrou um método competitivo, eficaz contra gatilhos localizados, geométricos e dinâmicos e estável entre execuções, mas frágil contra gatilhos aditivos e de mesclagem, uma limitação que ele divide com toda a família de ajuste fino. Nenhuma defesa se mostrou superior em todos os cenários, e a replicação em um segundo conjunto de dados confirmou esse quadro, acrescentando que a confiabilidade do ajuste fino depende da quantidade de dados limpos disponível por classe.

Uma parte importante do trabalho veio de investigar um resultado divergente. Sob o protocolo da plataforma, o NFT falhou contra um ataque que sua avaliação original removia quase por completo, e a reconstrução controlada dessa diferença mostrou que a causa não era uma falha do método, mas escolhas do protocolo de avaliação: a forma como o ataque foi treinado e o orçamento concedido à defesa. Esse achado mostra que o desempenho relatado de uma defesa pode não se repetir sob outro protocolo, o que pede cautela ao comparar técnicas avaliadas em condições diferentes e reforça o valor de benchmarks padronizados.

O trabalho tem limitações que apontam caminhos para sua continuação. A avaliação restringiu-se a uma única arquitetura e a conjuntos de classificação de imagens, e a comparação ateu-se a defesas de pós-treinamento e a ataques de ativação direta. Superar essas fronteiras é a principal direção de trabalho futuro: estender a avaliação a outras arquiteturas, domínios e ataques, incluindo ameaças que só se manifestam sob condições específicas. Além disso, o aumento gradual da taxa de sucesso do ataque ao longo das iterações sugere investigar critérios de parada antecipada para o NFT. Por fim, o módulo desenvolvido foi disponibilizado publicamente, com a intenção de propô-lo para integração à BackdoorBench e ampliar a cobertura da plataforma.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL GENERATIVA

Em conformidade com a Resolução Normativa nº 217/2025/CUn, com a Portaria nº 2.664/2026 do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e com o Código de Conduta para Autores em Publicações da Sociedade Brasileira de Computação (SBC), a autora declara o uso de ferramentas de Inteligência Artificial Generativa (IAG) na elaboração deste trabalho.

Foram empregados os modelos Claude Opus 4.8 e Claude Fable 5 (Anthropic), acessados por meio da ferramenta Claude Code, que apoiaram a implementação, a revisão e a documentação do módulo de defesa integrado à plataforma BackdoorBench, bem como a revisão de trechos do texto deste trabalho.

Todo o conteúdo gerado ou sugerido pela ferramenta foi revisado criticamente pela autora antes de sua incorporação ao trabalho. A autora assume a responsabilidade integral pela precisão técnica, pela originalidade e pela correção do conteúdo final apresentado, incluindo eventuais imprecisões decorrentes do uso dessas tecnologias. As ferramentas de IAG não são autoras deste trabalho, papel que cabe exclusivamente à autora identificada na folha de rosto.

REFERÊNCIAS

- BAGDASARYAN, E. et al. How to backdoor federated learning. In: **The 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)**. [S.l.]: PMLR, 2020. (Proceedings of Machine Learning Research), p. 2938–2948.
- BARNI, M.; KALLAS, K.; TONDI, B. A new backdoor attack in CNNs by training set corruption without label poisoning. In: IEEE. **IEEE International Conference on Image Processing (ICIP)**. [S.l.], 2019. p. 101–105.
- BARRENO, M. et al. Can machine learning be secure? In: **Proceedings of the 2006 ACM Symposium on Information, computer and communications security**. [S.l.: s.n.], 2006. p. 16–25.
- BIGGIO, B.; NELSON, B.; LASKOV, P. Poisoning attacks against support vector machines. In: **Proceedings of the 29th International Conference on Machine Learning (ICML)**. [S.l.]: Omnipress, 2012.
- BIGGIO, B.; ROLI, F. Wild patterns: Ten years after the rise of adversarial machine learning. **Pattern Recognition**, Elsevier, v. 84, p. 317–331, 2018.
- BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2007. (Information Science and Statistics). ISBN 9780387310732.
- CHEN, X. et al. Targeted backdoor attacks on deep learning systems using data poisoning. **arXiv preprint arXiv:1712.05526**, 2017.
- CINÀ, A. E. et al. Wild patterns reloaded: A survey of machine learning security against training data poisoning. **ACM Computing Surveys**, ACM New York, NY, v. 55, n. 13s, p. 1–39, 2023.
- DOAN, K. et al. Lira: Learnable, imperceptible and robust backdoor attacks. In: **Proceedings of the IEEE/CVF international conference on computer vision**. [S.l.: s.n.], 2021. p. 11966–11976.
- FAN, J. et al. A survey on data poisoning attacks and defenses. In: IEEE. **2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)**. [S.l.], 2022. p. 48–55.
- GAO, Y. et al. STRIP: A defence against trojan attacks on deep neural networks. In: **Proceedings of the 35th Annual Computer Security Applications Conference (ACSAC)**. [S.l.]: ACM, 2019. p. 113–125.
- GOLDBLUM, M. et al. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 45, n. 2, p. 1563–1580, 2022.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Cambridge, MA: MIT Press, 2016. <http://www.deeplearningbook.org>.
- GROSSE, K. et al. Machine learning security in industry: A quantitative survey. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 18, p. 1749–1762, 2023.
- GU, T. et al. BadNets: Evaluating backdooring attacks on deep neural networks. **IEEE Access**, IEEE, v. 7, p. 47230–47244, 2019.

- HASAN, M. H. et al. Guardfl: Detection and removal of malicious clients through majority consensus. In: IEEE. **2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)**. [S.l.], 2024. p. 1112–1117.
- HE, K. et al. Identity mappings in deep residual networks. In: **Computer Vision – ECCV 2016**. [S.l.]: Springer, 2016. (Lecture Notes in Computer Science), p. 630–645.
- JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015.
- KARIM, N. et al. Augmented neural fine-tuning for efficient backdoor purification. In: **Computer Vision – ECCV 2024**. [S.l.]: Springer, 2024. (Lecture Notes in Computer Science), p. 401–418.
- KITCHENHAM, B. et al. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, n. 2004, p. 1–26, 2004.
- KRIZHEVSKY, A. **Learning Multiple Layers of Features from Tiny Images**. [S.l.], 2009.
- KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. **Advances in neural information processing systems**, v. 25, 2012.
- LECUN, Y. et al. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998.
- LI, B. et al. Nearest is not dearest: Towards practical defense against quantization-conditioned backdoor attacks. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2024. p. 24523–24533.
- LI, Y. et al. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In: **9th International Conference on Learning Representations (ICLR)**. [S.l.]: OpenReview.net, 2021.
- LIU, K.; DOLAN-GAVITT, B.; GARG, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In: **Research in Attacks, Intrusions, and Defenses (RAID)**. [S.l.]: Springer, 2018. (Lecture Notes in Computer Science), p. 273–294.
- NELSON, B. et al. Exploiting machine learning to subvert your spam filter. **LEET**, v. 8, n. 1-9, p. 16–17, 2008.
- NGUYEN, T. A.; TRAN, A. T. Input-aware dynamic backdoor attack. In: **Advances in Neural Information Processing Systems 33 (NeurIPS)**. [S.l.: s.n.], 2020.
- NGUYEN, T. A.; TRAN, A. T. WaNet – imperceptible warping-based backdoor attack. In: **International Conference on Learning Representations (ICLR)**. [S.l.: s.n.], 2021.
- NICOLAE, M.-I. et al. Adversarial robustness toolbox v1. 0.0. **arXiv preprint arXiv:1807.01069**, 2018.
- PANG, R. et al. Trojanzoo: Towards unified, holistic, and practical evaluation of neural backdoors. In: **2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)**. [S.l.: s.n.], 2022. p. 684–702.

- PINTOR, M. et al. secml: Secure and explainable machine learning in python. **SoftwareX**, Elsevier, v. 18, p. 101095, 2022.
- SAHA, A.; SUBRAMANYA, A.; PIRSIYAVASH, H. Hidden trigger backdoor attacks. In: **Proceedings of the AAAI conference on artificial intelligence**. [S.l.: s.n.], 2020. v. 34, n. 07, p. 11957–11965.
- SALEM, A. et al. Dynamic backdoor attacks against machine learning models. In: IEEE. **2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)**. [S.l.], 2022. p. 703–718.
- SCHWARZSCHILD, A. et al. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In: PMLR. **International Conference on Machine Learning**. [S.l.], 2021. p. 9389–9398.
- SHAFARI, A. et al. Poison frogs! targeted clean-label poisoning attacks on neural networks. **Advances in neural information processing systems**, v. 31, 2018.
- SUN, W. et al. Invisible backdoor attack with dynamic triggers against person re-identification. **IEEE Transactions on Information Forensics and Security**, IEEE, v. 19, p. 307–319, 2023.
- TRAN, B.; LI, J.; MADRY, A. Spectral signatures in backdoor attacks. In: **Advances in Neural Information Processing Systems 31 (NeurIPS)**. [S.l.: s.n.], 2018. p. 8011–8021.
- WANG, B. et al. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In: **2019 IEEE Symposium on Security and Privacy (SP)**. [S.l.: s.n.], 2019. p. 707–723.
- WU, B. et al. **BackdoorBench: a comprehensive benchmark of backdoor learning (repositório de código)**. 2024. <https://github.com/SCLBD/BackdoorBench>.
- WU, B. et al. Backdoorbench: A comprehensive benchmark and analysis of backdoor learning. **International Journal of Computer Vision**, Springer, p. 1–88, 2025.
- WU, D.; WANG, Y. Adversarial neuron pruning purifies backdoored deep models. In: **Advances in Neural Information Processing Systems 34 (NeurIPS)**. [S.l.: s.n.], 2021. p. 16913–16925.
- YAN, J. et al. A federated learning backdoor attack defense method based on dual attention mechanism. In: IEEE. **2022 IEEE Eighth International Conference on Big Data Computing Service and Applications (BigDataService)**. [S.l.], 2022. p. 143–148.
- ZENG, Y. et al. Adversarial unlearning of backdoors via implicit hypergradient. In: **International Conference on Learning Representations (ICLR)**. [S.l.: s.n.], 2022.
- ZENG, Y. et al. Rethinking the backdoor attacks' triggers: A frequency perspective. In: **IEEE/CVF International Conference on Computer Vision (ICCV)**. [S.l.]: IEEE, 2021. p. 16453–16461.
- ZHANG, Z. et al. Backdoor defense via deconfounded representation learning. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2023. p. 12228–12238.
- ZHAO, S. et al. Exploring clean label backdoor attacks and defense in language models. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, IEEE, 2024.

ZHU, M. et al. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In: **IEEE/CVF International Conference on Computer Vision (ICCV)**. [S.l.]: IEEE, 2023. p. 4443–4454.

APÊNDICE A – GUIA DE EXECUÇÃO DA AVALIAÇÃO EXPERIMENTAL

Este apêndice indica como reproduzir os experimentos descritos neste trabalho. A defesa NFT integrada à plataforma BackdoorBench, composta pelo módulo `defense/nft.py` e pela configuração `config/defense/nft/config.yaml`, encontra-se disponível publicamente em: <https://github.com/dickmariana/BackdoorBench/>.

A instalação do ambiente, o download dos conjuntos de dados e o treinamento dos ataques seguem o procedimento oficial da plataforma, documentado no `README.md` do próprio repositório. Em suma, treina-se cada ataque a partir de seu módulo em `attack/`, o que salva o modelo comprometido em `record/<nome>/attack_result.pt`.

Sobre esse resultado, a defesa NFT é executada com:

```
python ./defense/nft.py \  
  --result_file <nome> \  
  --yaml_path ./config/defense/nft/config.yaml \  
  --dataset cifar10
```

Os hiperparâmetros seguem a configuração recomendada pelos autores do método, documentada no próprio arquivo de configuração (Capítulo 4). Ao final, o modelo purificado e o arquivo `nft_df.csv`, com a trajetória de CA, ASR e RA por iteração, são salvos em `record/<nome>/defense/nft/`.

As defesas de comparação (FT, FT-SAM, NAD, ANP, I-BAU e FP) já integram a plataforma e são executadas de forma análoga, conforme o `README.md`. Para reproduzir a matriz completa do Capítulo 5, o procedimento é repetido para cada combinação de ataque e defesa com taxa de envenenamento de 10% (`-pratio 0.1`) nas sementes 0, 1 e 2 (`-random_seed`), reportando-se média e desvio-padrão.

APÊNDICE B – ARTIGO EM FORMATO SBC

Em atendimento à exigência da disciplina de Projeto II, este apêndice reproduz o artigo científico elaborado a partir deste trabalho de conclusão de curso, redigido segundo o *template* da Sociedade Brasileira de Computação (SBC). O artigo condensa a motivação, a fundamentação, a revisão sistemática, a integração do NFT à plataforma e os resultados apresentados nos capítulos anteriores.

Defesa de Backdoors em Aprendizado de Máquina: Uma Análise do Estado da Arte, Benchmarks e Aplicações

Mariana dos Santos Dick¹, Larissa Gremelmaier Rosa¹, Jean Everson Martina¹

¹Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis – SC – Brasil

mariana.dick@grad.ufsc.br, {larissa.rosa, jean.martina}@ufsc.br

Abstract. *Machine learning remains vulnerable to backdoor attacks, in which a trigger planted in the training data induces controlled misclassification at inference time. This work integrates NFT (Augmented Neural Fine-Tuning, ECCV 2024), a post-training defense not yet available in the BackdoorBench platform, and evaluates it against six defenses and six attacks under a standardized protocol. NFT effectively removes localized, geometric, and dynamic triggers while preserving clean accuracy, but fails against additive and blending triggers, a limitation shared by the fine-tuning family. A controlled investigation attributes the divergence from the method’s original evaluation to protocol factors, notably the attack’s training recipe, rather than to flaws in the method.*

Resumo. *O aprendizado de máquina permanece vulnerável a ataques de backdoor, nos quais um gatilho inserido nos dados de treinamento induz erros controlados na inferência. Este trabalho integra à plataforma BackdoorBench o NFT (Augmented Neural Fine-Tuning, ECCV 2024), uma defesa de pós-treinamento ainda ausente da ferramenta, e o avalia frente a seis defesas e seis ataques sob um protocolo padronizado. O NFT remove com eficácia gatilhos localizados, geométricos e dinâmicos, preservando a acurácia em dados limpos, mas falha contra gatilhos aditivos e de mesclagem, limitação compartilhada pela família de ajuste fino. Uma investigação controlada atribui a divergência em relação à avaliação original do método a fatores do protocolo, em especial a receita de treino do ataque, e não a falhas do método.*

1. Introdução

O sucesso do aprendizado de máquina depende de dados de treinamento representativos do ambiente real de aplicação. Em situações práticas, porém, muitas aplicações coletam dados de fontes não confiáveis ou sem supervisão adequada [Fan et al. 2022], o que expõe os modelos a ataques de envenenamento de dados (*data poisoning*), nos quais um adversário injeta ou manipula deliberadamente parte do conjunto de treinamento [Biggio and Roli 2018]. Entre as variantes de maior impacto está o ataque de *backdoor*: um gatilho inserido nos dados de treinamento induz o modelo a associar esse padrão a saídas incorretas específicas, ativáveis posteriormente em produção [Gu et al. 2019]. Como o modelo comprometido opera normalmente na ausência do gatilho, a ameaça é furtiva, sendo apontada como a principal preocupação na implantação industrial de aprendizado de máquina [Grosse et al. 2023].

Em resposta, a comunidade propôs dezenas de técnicas defensivas [Cinà et al. 2023]. A eficácia relatada, contudo, é frequentemente consequência de

experimentos inconsistentes ou pouco realistas [Schwarzschild et al. 2021], o que motivou o surgimento de plataformas de *benchmark* como a *TrojanZoo* [Pang et al. 2022] e a *BackdoorBench* [Wu et al. 2025], que padronizam ataques, defesas e métricas para permitir comparações justas e reproduzíveis. Ainda assim, novas técnicas surgem mais rápido do que as plataformas conseguem incorporá-las, criando uma lacuna entre o estado da arte e sua avaliação sistemática.

Este trabalho contribui para reduzir essa lacuna em duas frentes. Primeiro, uma Revisão Sistemática da Literatura (RSL) mapeia as defesas contra *backdoor* segundo o estágio do ciclo de vida do modelo em que atuam e identifica, no pós-treinamento, uma técnica recente ainda ausente da *BackdoorBench*: o *NFT* (*Augmented Neural Fine-Tuning*) [Karim et al. 2024]. Segundo, o NFT é implementado como um novo módulo da plataforma, seguindo suas convenções, e avaliado frente a seis defesas de pós-treinamento e seis ataques representativos, com análise do compromisso entre mitigação e utilidade do modelo, da robustez à taxa de envenenamento, da generalização a um segundo conjunto de dados e do custo computacional.

O restante do artigo organiza-se da seguinte forma. A Seção 2 revisa a fundamentação e os trabalhos relacionados. A Seção 3 descreve a integração do NFT e o protocolo experimental. A Seção 4 apresenta e discute os resultados e a Seção 5 conclui o artigo.

2. Fundamentação e Trabalhos Relacionados

Ataques de *backdoor*. No ataque de *backdoor*, o adversário envenena uma fração dos dados de treinamento com um gatilho associado a uma classe-alvo. O modelo aprende a correlação espúria e passa a classificar na classe-alvo qualquer entrada que contenha o gatilho, mantendo desempenho normal em dados limpos [Gu et al. 2019]. Os gatilhos variam de *patches* localizados, como no *BadNets* [Gu et al. 2019], a padrões globais mesclados à imagem inteira, como no *Blended* [Chen et al. 2017], sinais senoidais aditivos em regime *clean-label*, como no *SIG* [Barni et al. 2019], deformações geométricas imperceptíveis, como no *WaNet* [Nguyen and Tran 2021], gatilhos dinâmicos dependentes da entrada, como no *InputAware* [Nguyen and Tran 2020], e perturbações concentradas nas baixas frequências, como no *LF* [Zeng et al. 2021].

Defesas por estágio do ciclo de vida. A literatura recente organiza as defesas conforme o estágio em que atuam: pré-treinamento (sanitização dos dados), durante o treinamento (treinamento robusto), pós-treinamento (purificação do modelo comprometido) e tempo de inferência [Wu et al. 2025, Cinà et al. 2023]. A RSL conduzida neste trabalho, seguindo o protocolo de Kitchenham [Kitchenham et al. 2004] sobre as bases IEEE Xplore, ACM DL e Springer, constatou que o pós-treinamento concentra as técnicas mais ativas. Esse desbalanceamento reflete o modelo de ameaça considerado mais realista: em cenários de terceirização do treinamento, o defensor possui apenas o modelo já comprometido e um pequeno conjunto de dados limpos [Karim et al. 2024]. Entre as defesas dessa categoria destacam-se o ajuste fino tradicional (FT), a destilação por atenção neural (NAD) [Li et al. 2021], o ajuste fino com minimização sensível à curvatura (FT-SAM) [Zhu et al. 2023], a poda adversarial de neurônios (ANP) [Wu and Wang 2021], o desaprendizado adversarial implícito (I-BAU) [Zeng et al. 2022] e a poda com ajuste fino (FP) [Liu et al. 2018].

O NFT. O *Augmented Neural Fine-Tuning* [Karim et al. 2024] parte da hipótese de que o *backdoor* se concentra em um subconjunto de canais da rede. Em vez de reajustar os pesos, o método associa a cada canal das camadas de normalização uma máscara $m \in [0, 1]$ que multiplica o fator de escala original ($\gamma \odot m$) e otimiza apenas essas máscaras sobre um pequeno conjunto limpo. O treino combina *MixUp*, que dispensa a síntese de gatilho e a busca adversarial de métodos como ANP e I-BAU, um regularizador ℓ_1 que penaliza o afastamento das máscaras da posição neutra e uma projeção que impõe a cada máscara um piso dependente da profundidade da camada. Ao final, as máscaras são fundidas de volta na escala de normalização, preservando a arquitetura original sem custo adicional de inferência.

3. Metodologia

Integração à plataforma. A *BackdoorBench* estrutura ataques e defesas como módulos independentes com um fluxo unificado: cada defesa de pós-treinamento recebe o modelo comprometido e os dados do ataque, aplica seu procedimento de reparo com um pequeno subconjunto limpo e é avaliada pelas mesmas rotinas de medição. O NFT foi implementado como um novo módulo que estende a classe base da plataforma e reaproveita suas rotinas de carregamento, avaliação e salvamento, em conformidade com o código de referência dos autores¹: comportamento das camadas de normalização mascaradas, orçamento de iterações com reamostragem e projeção das máscaras a cada passo. Como verificação de fidelidade, contra o *BadNets* o módulo reduz a taxa de sucesso do ataque a 2,3%, próximo do 1,74% relatado na avaliação original. O módulo foi disponibilizado publicamente².

Protocolo. A avaliação utilizou o CIFAR-10 [Krizhevsky 2009] com a arquitetura PreActResNet18 [He et al. 2016], taxa de envenenamento de 10% e classe-alvo fixa, seguindo o padrão da plataforma. Foram avaliados seis ataques cobrindo diferentes naturezas de gatilho (*BadNets*, *WaNet*, *InputAware*, *Blended*, *SIG* e *LF*) e seis defesas de comparação (FT, FT-SAM, NAD, ANP, I-BAU e FP), executadas com os hiperparâmetros padrão da plataforma; o NFT usou a configuração recomendada por seus autores. Três controles asseguraram a equidade: os mesmos modelos comprometidos, a mesma fração de dados limpos (5%) e o mesmo ambiente computacional. As métricas são a acurácia em dados limpos (CA), a taxa de sucesso do ataque (ASR) e a acurácia robusta (RA, fração de amostras com gatilho classificadas no rótulo verdadeiro). Toda a matriz foi executada com três sementes independentes, reportando-se média e desvio-padrão.

4. Resultados e Discussão

A Tabela 1 apresenta a ASR de cada defesa contra cada ataque no CIFAR-10. Todas as defesas mantêm a CA entre 80% e 93% (o NFT entre 89% e 92%), de modo que a principal diferença entre elas está na capacidade de reduzir a ASR — e nenhuma apresenta o melhor desempenho contra todos os ataques simultaneamente, como ilustra a Figura 1. As taxas de sucesso sem defesa coincidem com as da avaliação de referência da plataforma [Wu et al. 2025], confirmando a fidelidade da reprodução.

Eficácia por tipo de gatilho. O NFT demonstra alta eficácia contra gatilhos localizados e geométricos: contra o *BadNets* e o *WaNet*, reduz a ASR de cerca de 95% para

¹<https://github.com/nazmul-karim170/NFT-Augmented-Backdoor-Purification>

²<https://github.com/dickmariana/BackdoorBench/tree/nft-defense>

Tabela 1. ASR (%) após cada defesa no CIFAR-10, média sobre três sementes ($\pm$ desvio-padrão). Melhor valor por coluna em negrito.

Defesa	BadNets	WaNet	InputAware	Blended	SIG	LF
Sem defesa	94,5 $\pm$ 0,0	95,6 $\pm$ 1,1	93,9 $\pm$ 3,8	99,8 $\pm$ 0,1	97,5 $\pm$ 0,6	99,1 $\pm$ 0,2
NFT	2,3 $\pm$ 1,1	1,4 $\pm$ 0,6	1,9 $\pm$ 0,4	83,5 $\pm$ 3,8	23,5 $\pm$ 2,6	77,6 $\pm$ 6,4
FT	1,3 $\pm$ 0,2	1,3 $\pm$ 0,5	26,6 $\pm$ 24,3	92,8 $\pm$ 4,3	4,3 $\pm$ 1,4	78,2 $\pm$ 1,1
FT-SAM	1,6 $\pm$ 0,4	1,0 $\pm$ 0,2	2,5 $\pm$ 0,7	14,8 $\pm$ 4,1	0,4 $\pm$ 0,1	5,9 $\pm$ 1,2
NAD	1,6 $\pm$ 0,2	1,4 $\pm$ 0,6	32,8 $\pm$ 27,2	73,0 $\pm$ 7,6	3,6 $\pm$ 0,5	53,8 $\pm$ 25,0
ANP	0,0 $\pm$ 0,0	0,4 $\pm$ 0,3	0,3 $\pm$ 0,3	3,1 $\pm$ 1,6	2,2 $\pm$ 2,1	0,4 $\pm$ 0,3
I-BAU	1,2 $\pm$ 0,6	18,6 $\pm$ 11,0	19,5 $\pm$ 26,0	11,7 $\pm$ 5,7	2,1 $\pm$ 1,6	15,0 $\pm$ 8,5
FP	0,9 $\pm$ 0,1	5,8 $\pm$ 3,2	15,1 $\pm$ 3,0	19,1 $\pm$ 8,0	2,1 $\pm$ 1,5	37,0 $\pm$ 9,8

perto de 2%, mantendo a CA em torno de 91%, desempenho comparável às melhores defesas de ajuste fino. No *InputAware*, cujo gatilho é dinâmico mas de natureza localizada, a ASR cai para 1,9% com estabilidade muito superior à das demais defesas de ajuste fino, que oscilaram fortemente entre sementes (FT: 26,6 $\pm$ 24,3; NAD: 32,8 $\pm$ 27,2). Em contraste, contra os gatilhos aditivos e de mesclagem (*Blended*, *SIG* e *LF*) o NFT perde eficácia, padrão que a RA confirma: nesses ataques, as amostras com gatilho não retornam ao rótulo verdadeiro. A comparação com o *WaNet*, que também afeta a imagem inteira e ainda assim é removido, indica que o fator determinante não é a abrangência espacial do gatilho, mas sua natureza aditiva ou de mesclagem.

Investigação do SIG. Contra o *SIG*, o NFT alcançou mitigação apenas parcial (23,5%), enquanto sua avaliação original relata cerca de 0,1%. Como os hiperparâmetros da defesa são equivalentes nos dois cenários, a diferença está na montagem do ataque: a *BackdoorBench* usa a formulação aditiva original com amplitude $\Delta = 40$ [Barni et al. 2019], enquanto a avaliação original do NFT emprega um sinal mais fraco e menos envenenamento. A ablação da Tabela 2 isola os dois fatores. Quanto mais forte o sinal, mais difícil removê-lo (ASR pós-NFT de 27,0% em $\Delta = 40$ para 8,5% em $\Delta = 10$), o que explica o bom desempenho do método diante do sinal brando da avaliação original. Já a redução do envenenamento piora o resultado (de 27,0% para 81,6%): saturar a classe-alvo concentra o *backdoor* em poucos canais, justamente os que as máscaras conseguem atenuar; com menos envenenamento, o sinal se difunde por muitos canais e escapa a um mecanismo que atua canal a canal.

Tabela 2. Ablação do ataque SIG: ASR (%) sem defesa e após o NFT, variando isoladamente a amplitude Δ e a fração de envenenamento (*pratio*), em uma única semente.

Δ	<i>pratio</i>	ASR sem defesa	ASR após NFT
40	0,10	98,2	27,0
20	0,10	85,1	23,0
10	0,10	68,2	8,5
40	0,05	96,4	83,4
40	0,02	91,2	81,6

Investigação do Blended. O caso do *Blended* é mais grave: o NFT mantém a ASR em torno de 83% durante todo o ajuste, enquanto a avaliação original relata 0,3%. A pista

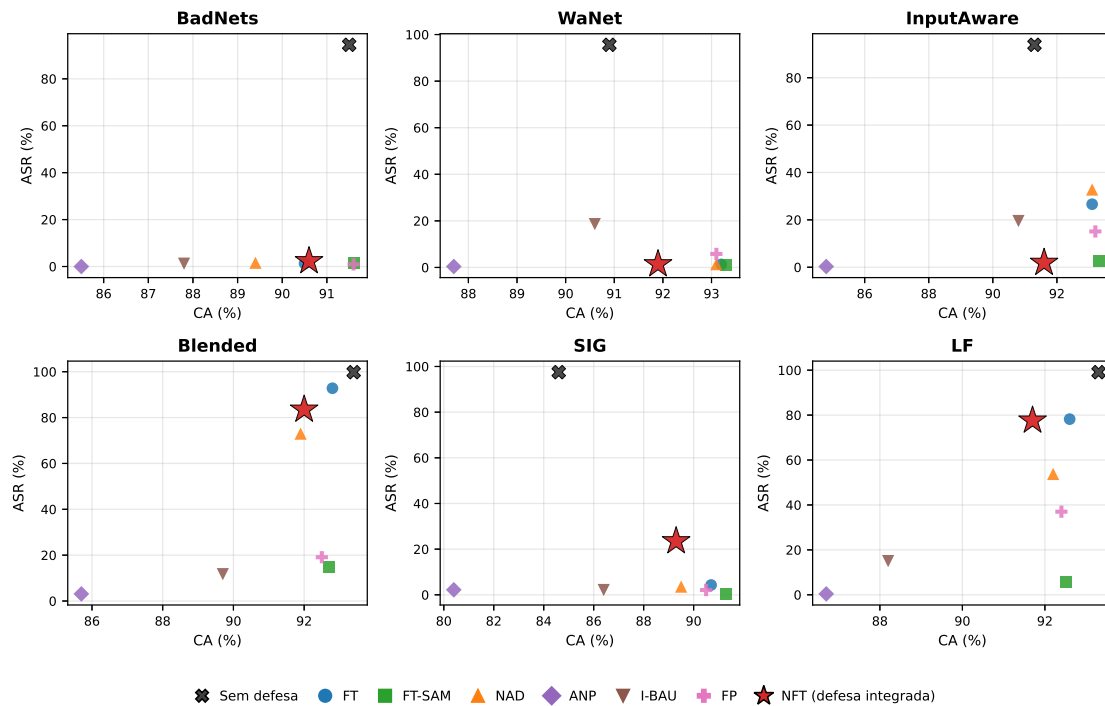


Figura 1. Compromisso entre utilidade (CA) e segurança (ASR), com um painel por ataque e cor e marcador por defesa, em média sobre três sementes. A posição desejável é o canto inferior direito; o NFT aparece destacado com uma estrela.

decisiva é que a dificuldade atinge toda a família de ajuste fino (FT: 93%; NAD: 73%), ao passo que, no estudo original, FT e NAD também removem o ataque. O que torna o gatilho removível ou não, portanto, é a receita de treino do ataque. Reproduzindo a receita da avaliação original na plataforma, em especial o agendamento da taxa de aprendizado, o gatilho voltou a ser removível por ajuste fino e a ASR do NFT caiu para cerca de 35%; ao longo das iterações, o método chega a atingir o patamar do estudo original (abaixo de 5%), oscilando em seguida — o resíduo é uma questão de medição do ponto final, não estrutural. Esse achado mostra que o desempenho relatado de uma defesa pode não se repetir sob outro protocolo, o que pede cautela ao comparar técnicas avaliadas em condições diferentes e reforça o valor de *benchmarks* padronizados.

Robustez e generalização. Replicando a avaliação com taxas de envenenamento de 5% e 20%, o perfil do NFT mostrou-se invariante: ASR baixa contra gatilhos localizados e geométricos nas três taxas e ineficácia contra o *Blended* em todas elas. No GTSRB (43 classes, mesma arquitetura e protocolo), o achado central se reproduz: o NFT falha contra os gatilhos aditivos e de mesclagem (94,6% no *Blended*, 84,9% no *LF*) e purifica com eficácia e estabilidade o gatilho geométrico do *WaNet* (0,0%). Surgiu, porém, uma ressalva: contra o *BadNets* e o *InputAware*, toda a família de ajuste fino passou a oscilar fortemente entre sementes, o que se explica em parte pela escassez de dados limpos por classe, enquanto FT-SAM, poda e desaprendizado se mantiveram robustos.

Custo computacional. Medido sobre um mesmo modelo atacado e mesma GPU, o NFT ficou entre as defesas mais caras (8,3 minutos, contra 0,8 da I-BAU, 3,1 do FT e 23,3 do FP). A dispensa da busca adversarial não se traduz em ganho de tempo, pois o

custo total depende sobretudo do orçamento de iterações, que no NFT é elevado e costuma ser exigido por completo. Sua vantagem prática está na simplicidade conceitual e na preservação da arquitetura original, não no tempo de execução.

Síntese. A Tabela 3 resume o desempenho por categoria de gatilho. Nenhuma defesa é eficaz em todas as categorias sem comprometer a CA: a ANP mitiga todas, mas é a que mais sacrifica a acurácia limpa. Para gatilhos visíveis e localizados ou deformações geométricas, o NFT é uma opção atraente, igualando as melhores defesas com estabilidade superior. Para gatilhos diluídos pela imagem (aditivos ou de mesclagem), tornam-se preferíveis a ANP, a I-BAU e a FT-SAM — esta última, a defesa mais consistente ao longo dos seis ataques.

Tabela 3. Síntese qualitativa do desempenho das defesas por categoria de gatilho, no CIFAR-10. Eficácia pela ASR média após a defesa (Eficaz $\leq 10\%$; Parcial 10–50%; Falha $> 50\%$); CA pela perda ante o modelo sem defesa (Alta < 2 pts; Média 2–5 pts; Baixa > 5 pts); custo pelo tempo de execução (Baixo ≤ 3 min; Médio 3–6 min; Alto > 6 min).

Defesa	Localiz./Geom.	Dinâmico	Adit./Mescl.	CA	Custo
FT	Eficaz	Parcial	Falha	Alta	Baixo
FT-SAM	Eficaz	Eficaz	Eficaz	Alta	Médio
NAD	Eficaz	Parcial	Parcial	Alta	Baixo
ANP	Eficaz	Eficaz	Eficaz	Baixa	Médio
I-BAU	Eficaz	Parcial	Eficaz	Média	Baixo
FP	Eficaz	Parcial	Parcial	Alta	Alto
NFT	Eficaz	Eficaz	Falha	Alta	Alto

5. Conclusão

Este trabalho integrou à *BackdoorBench* uma defesa de pós-treinamento recente ainda ausente da plataforma, o NFT, e a avaliou sob um protocolo padronizado frente a seis defesas e seis ataques. O método mostrou-se competitivo e estável contra gatilhos localizados, geométricos e dinâmicos, mas frágil contra gatilhos aditivos e de mesclagem, limitação que divide com toda a família de ajuste fino. A investigação controlada de um resultado divergente mostrou que a causa não era uma falha do método, mas escolhas do protocolo de avaliação, em especial a receita de treino do ataque — evidência de que comparações entre técnicas avaliadas em condições diferentes exigem cautela e de que *benchmarks* padronizados são essenciais para a área.

A avaliação restringiu-se a uma arquitetura, a conjuntos de classificação de imagens e a defesas de pós-treinamento contra ataques de ativação direta. Trabalhos futuros incluem estender a análise a outras arquiteturas, domínios e ataques, e investigar critérios de parada antecipada para o NFT. O módulo desenvolvido foi disponibilizado publicamente, com a intenção de propô-lo para integração à *BackdoorBench* e ampliar a cobertura da plataforma.

Referências

Barni, M., Kallas, K., and Tondi, B. (2019). A new backdoor attack in CNNs by training set corruption without label poisoning. In *IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE.

- Biggio, B. and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331.
- Chen, X., Liu, C., Li, B., Lu, K., and Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.
- Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., and Roli, F. (2023). Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, 55(13s):1–39.
- Fan, J., Yan, Q., Li, M., Qu, G., and Xiao, Y. (2022). A survey on data poisoning attacks and defenses. In *2022 7th IEEE International Conference on Data Science in Cyberspace (DSC)*, pages 48–55. IEEE.
- Grosse, K., Bieringer, L., Besold, T. R., Biggio, B., and Krombholz, K. (2023). Machine learning security in industry: A quantitative survey. *IEEE Transactions on Information Forensics and Security*, 18:1749–1762.
- Gu, T., Liu, K., Dolan-Gavitt, B., and Garg, S. (2019). BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, Lecture Notes in Computer Science, pages 630–645. Springer.
- Karim, N., Al Arafat, A., Khalid, U., Guo, Z., and Rahnavard, N. (2024). Augmented neural fine-tuning for efficient backdoor purification. In *Computer Vision – ECCV 2024*, Lecture Notes in Computer Science, pages 401–418. Springer.
- Kitchenham, B. et al. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.
- Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.
- Li, Y., Lyu, X., Koren, N., Lyu, L., Li, B., and Ma, X. (2021). Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *9th International Conference on Learning Representations (ICLR)*. OpenReview.net.
- Liu, K., Dolan-Gavitt, B., and Garg, S. (2018). Fine-pruning: Defending against backdooring attacks on deep neural networks. In *Research in Attacks, Intrusions, and Defenses (RAID)*, Lecture Notes in Computer Science, pages 273–294. Springer.
- Nguyen, T. A. and Tran, A. T. (2020). Input-aware dynamic backdoor attack. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*.
- Nguyen, T. A. and Tran, A. T. (2021). WaNet – imperceptible warping-based backdoor attack. In *International Conference on Learning Representations (ICLR)*.
- Pang, R., Zhang, Z., Gao, X., Xi, Z., Ji, S., Cheng, P., Luo, X., and Wang, T. (2022). TrojanZoo: Towards unified, holistic, and practical evaluation of neural backdoors. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 684–702.

- Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., and Goldstein, T. (2021). Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR.
- Wu, B., Chen, H., Zhang, M., Zhu, Z., Wei, S., Yuan, D., Zhu, M., Wang, R., Liu, L., and Shen, C. (2025). Backdoorbench: A comprehensive benchmark and analysis of backdoor learning. *International Journal of Computer Vision*, pages 1–88.
- Wu, D. and Wang, Y. (2021). Adversarial neuron pruning purifies backdoored deep models. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pages 16913–16925.
- Zeng, Y., Chen, S., Park, W., Mao, Z. M., Jin, M., and Jia, R. (2022). Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations (ICLR)*.
- Zeng, Y., Park, W., Mao, Z. M., and Jia, R. (2021). Rethinking the backdoor attacks’ triggers: A frequency perspective. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16453–16461. IEEE.
- Zhu, M., Wei, S., Shen, L., Fan, Y., and Wu, B. (2023). Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4443–4454. IEEE.