INSTITUTO FEDERAL DE MINAS GERAIS - *Campus* Formiga
Curso de Ciência da Computação

RUÍDO ADVERSARIAL COMO MECANISMO DE DEFESA EM CAPTCHAS: AVALIAÇÃO EXPERIMENTAL DO ATAQUE FGSM SOBRE REDES NEURAIS CONVOLUCIONAIS E PROPOSTA DE MODELO CONCEITUAL

Jhon Kennedy Oliveira Silva

Orientador: Roger Santos Ferreira

Formiga - MG

2026

JHON KENNEDY OLIVEIRA SILVA

RUÍDO ADVERSARIAL COMO MECANISMO DE DEFESA EM CAPTCHAS: AVALIAÇÃO EXPERIMENTAL DO ATAQUE FGSM SOBRE REDES NEURAIS CONVOLUCIONAIS E PROPOSTA DE MODELO CONCEITUAL

Monografia do trabalho de conclusão de curso apresentado ao Instituto Federal de Minas Gerais - Campus Formiga, como requisito parcial para a obtenção do título de Bacharel em Ciência da Computação.

Orientador: Roger Santos Ferreira

Formiga - MG

2026

Silva, Jhon Kennedy Oliveira
S586r Ruído adversarial como mecanismo de defesa em CAPTCHAs: avaliação experimental do ataque FGSM sobre Redes Neurais Convolucionais e proposta de modelo conceitual /Jhon Kennedy Oliveira Silva - Formiga : IFMG, 2026.
64 p. :il. color.

Orientador: Roger Santos Ferreira

Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Formiga.

1. CAPTCHA. 2. Ruído adversarial. 3. Redes Neurais Convolucionais. 4. FGSM. 5. Segurança da informação I. Ferreira, Roger Santos. II. Título.

CDD 006,3

Ficha catalográfica elaborada pela Bibliotecária Lívia Renata Santos- CRB/6-2561

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS
Campus Formiga
Diretoria de Ensino
Docência Área Acadêmica de Computação
Rua São Luiz Gonzaga, s/n - Bairro São Luiz - CEP 35570-000 - Formiga - MG
- www.ifmg.edu.br

JHON KENNEDY OLIVEIRA SILVA

RUÍDO ADVERSARIAL COMO MECANISMO DE DEFESA EM CAPTCHAS: AVALIAÇÃO EXPERIMENTAL DO ATAQUE FGSM SOBRE REDES NEURAIS CONVOLUCIONAIS E PROPOSTA DE MODELO CONCEITUAL

Trabalho de Conclusão de Curso apresentado ao Instituto Federal de Minas Gerais - Campus Formiga, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

APROVADA em 11 de novembro de 2026.

BANCA EXAMINADORA

Prof. Roger Santos Ferreira (Orientador)

Prof. Mário Luiz Rodrigues Oliveira (IFMG)

Prof. Everthon Valadão (IFMG)

Formiga, 11 de novembro de 2026.

Documento assinado eletronicamente por **Everthon Valadao dos Santos**, **Professor**, em 11/11/2026, às 18:27, conforme Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Roger Santos Ferreira**, **Professor**, em 11/11/2026, às 18:27, conforme Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Mario Luiz Rodrigues Oliveira**, **Professor**, em 11/11/2026, às 18:28, conforme Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade do documento pode ser conferida no site https://sei.ifmg.edu.br/consultadocs informando o código verificador **2945163** e o código CRC **F0DD590C**.

À minha família e a todos que, concordando ou não, contribuíram de alguma forma para a minha formação.

Agradecimentos

Agradeço primeiramente a Deus, por me conceder saúde, força e discernimento para chegar até aqui e para não desistir nos momentos em que o caminho pareceu longo demais.

Aos meus pais, que nunca mediram esforços para que eu pudesse estudar. Cada conquista minha carrega o trabalho e a renúncia de vocês; esta também é de vocês.

Ao meu orientador, Prof. Roger Santos Ferreira, pela paciência, pelas correções e pela disposição em conduzir este trabalho. Suas observações foram essenciais para que esta pesquisa tomasse a forma que tem hoje.

Aos professores do curso de Ciência da Computação do IFMG — Campus Formiga, pelos conhecimentos compartilhados ao longo de toda a graduação.

E a todos os que estiveram comigo nessa luta, familiares, amigos e colegas de curso, obrigado por cada palavra de incentivo. Nada disso teria sido possível sozinho.

"Podemos enxergar apenas um pouco adiante, mas podemos ver que há muito a ser feito."
(Alan Turing)

Resumo

CAPTCHAs baseados em texto e imagem perderam eficácia diante das Redes Neurais Convolucionais: solucionadores treinados em poucos minutos já resolvem parte considerável dos desafios ainda em uso na *web*. Uma linha de defesa investigada na literatura inverte o problema e usa contra a máquina a fragilidade dela mesma, o ruído adversarial. Os estudos sobre CAPTCHA revisados, porém, não distinguem duas explicações possíveis para a queda de acurácia que esse ruído provoca: a direção da perturbação, calculada sobre o gradiente da função de perda, ou simplesmente a magnitude da alteração imposta aos *pixels*. A resposta define o custo da defesa, já que um CAPTCHA precisa continuar legível para o usuário humano. Este trabalho mediu essa diferença. Uma Revisão Sistemática da Literatura conduzida segundo as diretrizes de Kitchenham e Charters (2007), com o fluxo de seleção relatado no formato PRISMA 2020, reuniu sete estudos primários e delimitou a lacuna. Em seguida, aplicou-se o ataque *Fast Gradient Sign Method* a uma ResNet-20 pré-treinada, sobre as 10.000 imagens de teste do CIFAR-10, em doze níveis de intensidade. Um grupo de controle recebeu ruído aleatório de mesma magnitude L_∞ nas mesmas imagens, e as diferenças foram submetidas ao teste exato de McNemar. A acurácia caiu de 92,59% para 23,48% já em $\varepsilon = 0{,}05$ e chegou a 9,90% em $\varepsilon = 0{,}60$, com $p < 0{,}001$ em todos os níveis. No mesmo $\varepsilon = 0{,}05$, o ruído aleatório manteve 91,53%, e o teste pareado entre os dois grupos confirmou a diferença ($p < 0{,}001$). A degradação decorre, portanto, da direção da perturbação, e não apenas da quantidade de ruído. As métricas SSIM e PSNR situaram em $\varepsilon = 0{,}15$ o limite em que a similaridade estrutural média ainda alcança 0,9278, com PSNR de 30,50 dB. A partir desses resultados, propõe-se um modelo conceitual de arquitetura para geração dinâmica de desafios CAPTCHA reforçados por perturbações adversariais. O experimento foi conduzido sobre imagens naturais e em cenário *white-box*: a validação em desafios CAPTCHA reais e com participantes humanos permanece como trabalho futuro.

Palavras-chave: CAPTCHA; Ruído adversarial; Redes Neurais Convolucionais; FGSM; Segurança da informação.

Abstract

Text- and image-based CAPTCHAs have lost effectiveness against Convolutional Neural Networks: solvers trained in a matter of minutes already break a substantial share of the challenges still in use on the web. One line of defense studied in the literature turns the problem around and uses the networks' own fragility against them, in the form of adversarial noise. The CAPTCHA studies reviewed, however, do not separate two possible explanations for the drop in accuracy that this noise causes: the direction of the perturbation, computed from the gradient of the loss function, or simply the magnitude of the change imposed on the *pixels*. The answer determines the cost of the defense, since a CAPTCHA must remain legible to the human user. This work measured that difference. A systematic literature review following the guidelines of Kitchenham and Charters (2007), with the selection flow reported in the PRISMA 2020 format, gathered seven primary studies and delimited the gap. The *Fast Gradient Sign Method* was then applied to a pre-trained ResNet-20 over the 10,000 test images of CIFAR-10, at twelve perturbation levels. A control group received random noise of the same L_∞ magnitude on the same images, and the differences were assessed with the exact McNemar test. Accuracy fell from 92.59% to 23.48% at $\varepsilon = 0.05$ and reached 9.90% at $\varepsilon = 0.60$, with $p < 0.001$ at every level. At that same $\varepsilon = 0.05$, random noise still yielded 91.53%, and a paired test between the two groups confirmed the difference ($p < 0.001$). The degradation therefore comes from the direction of the perturbation, not merely from the amount of noise. SSIM and PSNR placed at $\varepsilon = 0.15$ the limit at which mean structural similarity still reaches 0.9278, with a PSNR of 30.50 dB. Building on these results, a conceptual architecture is proposed for the dynamic generation of CAPTCHA challenges reinforced by adversarial perturbations. The experiment was conducted on natural images and in a *white-box* setting: validation on real CAPTCHA challenges and with human participants remains future work.

Keywords: CAPTCHA; Adversarial noise; Convolutional Neural Networks; FGSM; Information security.

Lista de ilustrações

Figura 1 – Efeito do ruído adversarial gerado pelo ataque FGSM em $\varepsilon = 0{,}15$ . . . 27
Figura 2 – Fluxograma do processo de geração e validação do desafio 48
Figura 3 – Diagrama de sequência da interação entre usuário, cliente, servidor e modelo substituto 49

Lista de gráficos

Gráfico 1 – Acurácia global: FGSM *versus* ruído aleatório de mesma magnitude . . 42
Gráfico 2 – SSIM e PSNR médios: FGSM e ruído aleatório 44
Gráfico 3 – Matrizes de confusão sobre o subconjunto classificado corretamente em condições limpas 45
Gráfico 4 – Degradação das métricas condicionais por nível de perturbação 63

Lista de tabelas

Tabela 1 – Síntese do processo de seleção segundo o fluxo do PRISMA 31
Tabela 2 – Motivos de exclusão na triagem por título e resumo 32
Tabela 3 – Acurácia global: FGSM *versus* ruído aleatório 41
Tabela 4 – Teste exato de McNemar: pares discordantes e valor-p por nível de perturbação 43
Tabela 5 – Resultados completos por nível de intensidade da perturbação 62

Lista de quadros

Quadro 1 – Questões de pesquisa da RSL 29
Quadro 2 – Expressões de busca aplicadas às bases 30
Quadro 3 – Critérios de exclusão aplicados na triagem 30
Quadro 4 – Artigos incluídos para síntese de estado da arte 33
Quadro 5 – Tabela de contingência do teste de McNemar 36
Quadro 6 – Comparação entre as propostas de CAPTCHA baseadas em defesa contra redes profundas 53

Lista de abreviaturas e siglas

ACM	*Association for Computing Machinery*
BIM	*Basic Iterative Method*
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CAPTCHA	*Completely Automated Public Turing test to tell Computers and Humans Apart*
CCT	*Classifier Cross-task Transferability*
CE	Critério de Exclusão
CIFAR	*Canadian Institute For Advanced Research*
CNN	*Convolutional Neural Network* (Rede Neural Convolucional)
CUDA	*Compute Unified Device Architecture*
FGSM	*Fast Gradient Sign Method*
GAN	*Generative Adversarial Network* (Rede Generativa Adversarial)
GPU	*Graphics Processing Unit*
IEEE	*Institute of Electrical and Electronics Engineers*
I-FGSM	*Iterative Fast Gradient Sign Method*
IFMG	Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais
JPEG	*Joint Photographic Experts Group*
PGD	*Projected Gradient Descent*
PRISMA	*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*
PSNR	*Peak Signal-to-Noise Ratio*
QP	Questão de Pesquisa
RAM	*Random Access Memory*
RSL	Revisão Sistemática da Literatura
SoK	*Systematization of Knowledge*

SSIM	*Structural Similarity Index Measure*
SVM	*Support Vector Machine* (Máquina de Vetores de Suporte)
TCC	Trabalho de Conclusão de Curso
VRAM	*Video Random Access Memory*
YOLO	*You Only Look Once* (família de detectores de objetos; YOLOX é uma de suas variantes)

Lista de símbolos

ε	Intensidade da perturbação adversarial (*epsilon*)
L_∞	Norma do infinito (limite máximo de variação por *pixel*)
p	Valor-p do teste estatístico
α	Nível de significância dos testes estatísticos
b, c	Pares discordantes do teste de McNemar

Sumário

1 INTRODUÇÃO 18
1.1 Contextualização e justificativa 18
1.2 Problema de pesquisa 20
1.3 Objetivos 20
1.3.1 Objetivo geral 20
1.3.2 Objetivos específicos 20
1.4 Abordagem metodológica 21
1.5 Estrutura do trabalho 22

2 FUNDAMENTAÇÃO TEÓRICA 23
2.1 Mecanismos CAPTCHA 23
2.2 Redes Neurais Convolucionais (CNNs) 25
2.3 Ruído Adversarial 26

3 MATERIAIS E MÉTODOS 28
3.1 Caracterização da pesquisa 28
3.2 Procedimento da Revisão Sistemática da Literatura 29
3.2.1 Estratégia de Busca e Seleção 29
3.2.2 Fluxo de Seleção e Estudos Incluídos 30
3.2.3 Síntese dos Resultados da RSL 33
3.3 Hipóteses de pesquisa 35
3.3.1 Teste de McNemar 36
3.4 Materiais utilizados 37
3.5 Procedimento experimental 39
3.5.1 Etapa 1 — Preparação dos dados 39
3.5.2 Etapa 2 — Avaliação da linha de base (*baseline*) 39
3.5.3 Etapa 3 — Aplicação da perturbação 40
3.5.4 Etapa 4 — Avaliação sob ataque (adversarial) 40
3.5.5 Etapa 5 — Análise comparativa e validação 40

4 RESULTADOS E ANÁLISE 41
4.1 Acurácia do modelo em condições limpas (*baseline*) 41
4.2 Impacto do ataque FGSM na acurácia global 41
4.3 Validação estatística: teste de McNemar 43
4.4 Quantificação da imperceptibilidade (SSIM e PSNR) 44
4.5 Análise qualitativa: convergência para classes-atratoras 45

4.6 Síntese e validação da hipótese 45

5 PROPOSTA DE MODELO CONCEITUAL 47
5.1 Visão geral da arquitetura proposta 47
5.1.1 Geração do desafio base 47
5.1.2 Cálculo da perturbação adversarial via modelo substituto 47
5.1.3 Injeção da perturbação e validação de fidelidade e eficácia 50
5.1.4 Validação da resposta humana 50
5.2 Fundamentação e limitações de projeto 51

6 CONSIDERAÇÕES FINAIS 54
6.1 Principais contribuições 54
6.2 Limitações 55
6.3 Trabalhos futuros 56

REFERÊNCIAS 58

APÊNDICES 61

APÊNDICE A – RESULTADOS COMPLETOS DA VARREDURA DE INTENSIDADE 62

APÊNDICE B – DEGRADAÇÃO DAS MÉTRICAS CONDICIONAIS 63

1 Introdução

1.1 Contextualização e justificativa

Um solucionador automatizado treinado por 15 minutos em uma única GPU resolve, na primeira tentativa, cerca de 20% dos CAPTCHAs de texto ainda ativos na *web* (OUSAT et al., 2024). O dado resume o problema tratado aqui: as defesas visuais que sustentaram esses testes por duas décadas deixaram de acompanhar a evolução das Redes Neurais Convolucionais (CNNs), hoje capazes de interpretar imagens distorcidas e ruidosas com precisão próxima da humana (MORADI et al., 2024).

A resposta investigada neste trabalho inverte a lógica do problema. Em vez de tentar esconder a informação do classificador com mais ruído, aplica-se à imagem uma perturbação calculada a partir do gradiente da própria rede: pequena o suficiente para que a pessoa continue lendo o desafio, e direcionada o suficiente para derrubar a máquina.

Os mecanismos CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) surgiram como uma das principais barreiras contra acessos automatizados maliciosos em sistemas digitais. Embora a comunidade de segurança hoje os considere obsoletos como defesa primária, Ousat et al. (2024) identificaram CAPTCHAs de texto ainda em uso em mais de 3.100 *sites*. Cerca de 29% dos CAPTCHAs detectados pelos autores estão em sites de setores críticos, como direito, governo, finanças e saúde, e cerca de metade desses sites recebe mais de 100 mil visitas por mês. Nesses ambientes, os testes continuam sendo mantidos para conter o envio automatizado de *spam* e mitigar ataques de injeção de credenciais (*credential stuffing*) em páginas de login, contato e comentários.

Segundo Tariq et al. (2023), a eficácia dessa classe de defesa dependia historicamente da suposição de que tarefas cognitivas e perceptuais (como a leitura de caracteres distorcidos ou a identificação de objetos em imagens ruidosas) seriam facilmente resolvidas por humanos, mas permaneceriam inviáveis para sistemas automatizados. Esse pressuposto, contudo, perdeu validade com os avanços da Inteligência Artificial. Para Ye et al. (2018), a popularização de arquiteturas de *Deep Learning* permitiu o desenvolvimento de ataques sofisticados, que alcançam altas taxas de acerto na resolução de CAPTCHAs tradicionais.

Conforme apontado por Weng et al. (2019), modelos modernos conseguem localizar e extrair caracteres mesmo sob distorções severas e fundos complexos. Por outro lado, em experimentos próprios conduzidos em seu estudo de sistematização, Dionysiou e Athanasopoulos (2020) mostraram que tentar aplicar um classificador único direto na imagem bruta, sem tratamento prévio, resulta em acurácias irrisórias de apenas 0,7% com CNN e 0,1% com SVM. Isso indica que as etapas de pré-processamento e segmentação

ainda são vitais para o sucesso do ataque. Ou seja, os CAPTCHAs clássicos falham principalmente porque os atacantes conseguem isolar e limpar as distorções físicas da imagem antes da etapa de reconhecimento.

A fragilidade das defesas tradicionais fica clara na análise da *web* real realizada por Ousat et al. (2024). O estudo constatou que mais de 77% dos CAPTCHAs ativos em produção utilizam apenas esquemas de cores simples ou ruídos de fundo elementares, como linhas, curvas e pontos. Mesmo sob essas proteções convencionais, cerca de 20% dos desafios são quebrados logo na primeira tentativa por solucionadores genéricos (OUSAT et al., 2024). Esses dados mostram que o uso de ruído aleatório ou estático atingiu um limite de eficácia. Essa estagnação justifica a transição para métodos de proteção baseados em perturbações adversariais direcionadas, que atacam diretamente a estrutura das redes neurais.

Uma alternativa promissora para restabelecer essa barreira de segurança explora o ponto fraco estrutural dos classificadores profundos: a alta sensibilidade a perturbações adversariais. Como explicam Goodfellow, Bengio e Courville (2016), o ruído adversarial baseia-se em pequenas alterações matemáticas calculadas a partir do sinal do gradiente da função de perda do modelo, induzindo erros severos de classificação na máquina. Dinh, Tran-Trung e Hoang (2023) complementam que essas perturbações sabotam o algoritmo, mas permanecem imperceptíveis ao olho humano se a intensidade for mantida dentro de limites estritos. Essa estratégia inverte a lógica tradicional de ataque, transformando a fragilidade das próprias redes neurais em uma contramedida ativa.

Como a preparação da imagem é o ponto crítico para o sucesso dos ataques, essa estratégia de defesa torna-se ainda mais eficiente se for desenhada para resistir a esse processo de limpeza. Esse princípio fundamenta a ideia de exemplos adversariais imutáveis (*immutable adversarial examples*), proposta por Osadchy et al. (2017). Nesses casos, a perturbação matemática é calculada para sobreviver a filtros de imagem convencionais, como os gaussianos ou de mediana. Isso garante que o ruído continue ativo e engane o classificador mesmo após o atacante tentar aplicar filtros e tratamentos na imagem.

Essa linha de defesa já tem trabalhos publicados. Dinh, Tran-Trung e Hoang (2023) combinam exemplos adversariais com transferência de estilo neural e relatam queda acentuada na taxa de acerto de solucionadores automáticos; Chen, Jiang e Chen (2020) exploram a estilização de imagens com o mesmo objetivo. Nenhum dos dois, porém, verifica se a degradação observada vem da direção da perturbação ou apenas do quanto a imagem foi alterada. É essa a pergunta que este trabalho responde.

1.2 Problema de pesquisa

Os estudos sobre CAPTCHA analisados na revisão sistemática (Seção 3.2) sustentam que o ruído adversarial degrada o desempenho de classificadores baseados em aprendizado profundo, mas não distinguem duas explicações possíveis para essa degradação: se ela decorre da **direção** da perturbação, calculada a partir do gradiente da função de perda, ou simplesmente da **magnitude** da alteração imposta aos *pixels*. Sem essa distinção, não é possível afirmar que o efeito observado seja específico do ruído adversarial, e não um efeito genérico de degradação da imagem.

Essa lacuna é relevante para o problema de segurança em questão, porque uma defesa baseada em CAPTCHA precisa preservar a legibilidade humana do desafio. Se a queda de acurácia dependesse apenas da magnitude, bastaria degradar mais a imagem, ao custo de torná-la ilegível também para pessoas. Se depender da direção, torna-se possível degradar o classificador mantendo a perturbação em níveis discretos.

Formula-se, assim, a questão que orienta este trabalho: a aplicação de ruído adversarial reduz de forma estatisticamente significativa a acurácia de Redes Neurais Convolucionais, e essa redução decorre da direção informada pelo gradiente ou apenas da magnitude da perturbação? Dessa questão derivam as hipóteses formais enunciadas na Seção 3.3 e submetidas a teste no Capítulo 4.

1.3 Objetivos

1.3.1 Objetivo geral

Avaliar experimentalmente, sob controle estatístico, se a aplicação de ruído adversarial reduz de forma significativa a acurácia de Redes Neurais Convolucionais em tarefas de classificação de imagem, e delimitar a faixa de intensidade em que essa redução ocorre sem degradação visual acentuada, de modo a fundamentar a proposição de um modelo conceitual de CAPTCHA reforçado por perturbações adversariais.

Considera-se o objetivo atendido se as hipóteses nulas H_{0a} e H_{0b} (Seção 3.3) forem rejeitadas ao nível de significância $\alpha = 0{,}05$ e se for possível identificar a faixa de intensidade em que a acurácia sob FGSM cai de forma significativa com SSIM médio $\geq 0{,}90$, critério de fidelidade definido na Seção 3.5.3.

1.3.2 Objetivos específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

a) mapear, por meio de uma Revisão Sistemática da Literatura, as vulnerabilidades documentadas dos mecanismos CAPTCHA frente a solucionadores baseados

em aprendizado profundo, bem como as abordagens de defesa propostas;

b) quantificar a degradação da acurácia de uma Rede Neural Convolucional submetida ao ataque *Fast Gradient Sign Method* em doze níveis de intensidade de perturbação;

c) isolar experimentalmente a causa dessa degradação, comparando o efeito do ruído adversarial ao de um ruído aleatório de mesma magnitude L_∞;

d) verificar a significância estatística das diferenças observadas por meio do teste de McNemar exato (Seção 3.3.1);

e) determinar, com base nas métricas de similaridade estrutural e relação sinal-ruído de pico, o limiar de intensidade até o qual a perturbação preserva a legibilidade visual da imagem;

f) propor um modelo conceitual de arquitetura para a geração dinâmica de desafios CAPTCHA reforçados por ruído adversarial.

1.4 Abordagem metodológica

Para responder à questão formulada, o trabalho parte da fundamentação obtida em uma Revisão Sistemática da Literatura e avalia empiricamente o ataque FGSM (*Fast Gradient Sign Method*), método formalizado por Goodfellow, Shlens e Szegedy (2015), aplicando-o a um cenário diferente do estudo original: uma rede ResNet-20 pré-treinada no conjunto de dados CIFAR-10. Embora os autores originais já sustentassem que a direção da perturbação importa mais do que o ponto específico no espaço (fato demonstrado por eles por meio de experimentos de controle voltados ao treinamento como regularizador), esta pesquisa adota uma abordagem diferente.

Ao desenho experimental, acrescentou-se um grupo de controle submetido a ruído aleatório de mesma magnitude L_∞ diretamente na etapa de inferência. Essa modelagem permite mensurar e comparar, de forma pareada e sobre as mesmas imagens do conjunto de teste, a acurácia sob perturbação direcionada e sob ruído aleatório, submetendo a diferença ao rigor estatístico do teste exato de McNemar. A contribuição desta monografia concentra-se, portanto, na operacionalização dessa validação empírica em uma arquitetura contemporânea e na proposição de um modelo conceitual aplicado a mecanismos CAPTCHA.

Cabe delimitar desde já o alcance do estudo. O experimento é conduzido sobre imagens naturais do CIFAR-10, e não sobre desafios CAPTCHA reais, e assume cenário *white-box*, com acesso completo aos gradientes do modelo atacado. O modelo conceitual apresentado no Capítulo 5 é, por consequência, uma proposta fundamentada nos resultados obtidos, cuja validação em ambiente real permanece como trabalho futuro.

1.5 Estrutura do trabalho

Além desta introdução, o trabalho está organizado em cinco capítulos. O Capítulo 2 reúne a fundamentação teórica, abordando os mecanismos CAPTCHA, as Redes Neurais Convolucionais e o ruído adversarial. O Capítulo 3 descreve os materiais e métodos, incluindo o protocolo da Revisão Sistemática da Literatura, as hipóteses de pesquisa e o procedimento experimental. O Capítulo 4 expõe e analisa os resultados obtidos. O Capítulo 5 apresenta o modelo conceitual de arquitetura proposto a partir desses resultados e discute suas limitações de projeto. Por fim, o Capítulo 6 retoma as hipóteses, sintetiza as contribuições e aponta as limitações e os desdobramentos futuros da pesquisa. Complementam o texto dois apêndices, com os resultados completos da varredura de intensidade e o detalhamento das métricas condicionais.

2 Fundamentação Teórica

2.1 Mecanismos CAPTCHA

Tariq et al. (2023) conceituam que o CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) é um mecanismo de segurança desenvolvido para distinguir usuários humanos de sistemas automatizados, como *bots*.

Estes testes geralmente consistem em desafios que exigem algum tipo de interpretação visual, textual ou comportamental que, tradicionalmente, é mais fácil para uma pessoa realizar do que para um computador, amplamente utilizados em *sites* e sistemas digitais para prevenir ações automatizadas maliciosas, como criação de contas falsas, envio de *spam*, tentativas de acesso indevido e ataques automatizados (TARIQ et al., 2023).

Ao analisar a presença dessas defesas, Ousat et al. (2024) mostram que, embora os CAPTCHAs de texto sejam considerados defasados diante das novas técnicas de inteligência artificial, eles ainda persistem em milhares de plataformas. Esses testes continuam sendo empregados em páginas de login, contato e seções de comentários para conter o envio automatizado de *spam* e mitigar ataques de *credential stuffing*. No entanto, seu funcionamento limita-se a desafios baseados apenas na transcrição de caracteres alfanuméricos distorcidos. Essa fragilidade estrutural os torna altamente vulneráveis a solucionadores modernos, mesmo quando os ataques operam com baixo custo computacional (OUSAT et al., 2024).

Entretanto, com o avanço das técnicas de aprendizado de máquina e, especialmente, das Redes Neurais Convolucionais (CNNs), a capacidade dos sistemas computacionais de interpretar esses desafios aumentou significativamente, tornando necessária a adoção de estratégias que ampliem a resistência dos CAPTCHAs diante de ataques automatizados (MORADI et al., 2024).

Essa vulnerabilidade é agravada pela grande diferença de custos entre defensores e atacantes. Como demonstram Ousat et al. (2024), solucionadores baseados em arquiteturas modernas de detecção de objetos, como o YOLOX, conseguem quebrar esquemas recentes de CAPTCHA com apenas 15 minutos de treinamento em uma única GPU (NVIDIA RTX A5000). Essa facilidade cria um dilema para os projetistas: adicionar ruídos visuais e distorções severas até aumenta a robustez contra os robôs, mas prejudica drasticamente a usabilidade e a experiência do usuário humano (MORADI et al., 2024; OUSAT et al., 2024).

Historicamente, a segurança desses mecanismos baseava-se na suposição de que tarefas perceptuais, como a segmentação de caracteres distorcidos ou a interpretação de

imagens ruidosas, seriam triviais para humanos e inviáveis para máquinas. No entanto, revisões recentes indicam que muitos esquemas tradicionais, sejam eles baseados em texto ou em seleção de imagens, apresentam graves vulnerabilidades (TARIQ et al., 2023).

Conforme discutido por Ye et al. (2018), defesas clássicas baseadas em ruídos complexos de fundo, linhas de oclusão e distorções geométricas tornaram-se insuficientes contra os pipelines modernos de ataque. Esses sistemas de quebra utilizam Redes Generativas Adversariais (GANs) para sintetizar bases de treinamento robustas e fazer o pré-processamento das imagens via redes de tradução, como a *Pix2Pix*. A tarefa final de reconhecimento, contudo, é delegada a Redes Neurais Convolucionais (CNNs) convencionais refinadas por aprendizado por transferência.

Além disso, embora esses solucionadores dispensem rotinas de segmentação projetadas manualmente (*hand-engineered*), o pré-processador realiza uma segmentação implícita ao alargar e padronizar o espaçamento entre caracteres colapsados (YE et al., 2018). Os testes mostram que os mecanismos isolados de sobreposição (*overlapping*) reduzem a acurácia do classificador de 65% para 25,1%, mas não derrubam a taxa de quebra para menos do que o limite de segurança de 1%. Com isso, a experiência do usuário é prejudicada muito antes de o ataque automatizado ser contido (YE et al., 2018).

Como reflexo dessa vulnerabilidade, Noury e Rezaei (2020) demonstram que uma CNN convencional, denominada *Deep-CAPTCHA*, alcança acurácias de 98,94% na quebra de desafios numéricos e 98,31% nos alfanuméricos. O modelo foi treinado com 500.000 amostras sintéticas no caso numérico e 700.000 no alfanumérico e estruturado com saídas Softmax paralelas que realizam o reconhecimento sem a necessidade de uma etapa explícita de segmentação. Esse desempenho prova que tentar impedir a segmentação física dos caracteres, de forma isolada, já não sustenta a segurança. Cabe ressalvar, contudo, que todos os desafios avaliados foram gerados por uma única biblioteca Python, o que limita a generalização do resultado para esquemas em produção.

Além das vulnerabilidades relacionadas à capacidade de reconhecimento visual, falhas de projeto também comprometem a segurança desses mecanismos, uma vez que Hernandez-Castro, Barrero e R-Moreno (2023) citam que, por meio da metodologia BASE-CASS, identificou-se que no CAPTCHA CaptchaStar ocorria vazamento de informações estatísticas laterais (*side-channel*) que permitiam identificar a resposta correta sem a necessidade de resolver efetivamente o desafio cognitivo proposto ao usuário.

Diante desse cenário, torna-se necessário desenvolver mecanismos capazes de acompanhar a evolução das técnicas de reconhecimento automatizado (HERNANDEZ-CASTRO; BARRERO; R-MORENO, 2023), mantendo um equilíbrio crucial entre a segurança do teste e sua usabilidade para humanos. Nesse contexto, estratégias baseadas na aplicação de perturbações adversariais sob medida mostram-se promissoras para degradar significativamente o desempenho de classificadores profundos sem comprometer a legibilidade do

desafio por usuários legítimos (DINH; TRAN-TRUNG; HOANG, 2023; MORADI et al., 2024).

2.2 Redes Neurais Convolucionais (CNNs)

Faceli et al. (2021) apontam que as Redes Neurais Convolucionais (CNNs) são uma arquitetura de aprendizado profundo amplamente utilizada em tarefas de processamento e análise de imagens. Seu desenvolvimento está relacionado à necessidade de criar modelos capazes de identificar automaticamente características relevantes em dados visuais, reduzindo a dependência de técnicas manuais de extração de atributos. Por essa razão, as CNNs desempenham um papel importante em aplicações como reconhecimento de objetos, classificação de imagens, detecção de padrões e sistemas de visão computacional.

A principal característica das CNNs está na utilização de operações de convolução, que permitem analisar diferentes regiões de uma imagem por meio de filtros ou *kernels*. Durante o treinamento, esses filtros aprendem a identificar características relevantes dos dados, como bordas, contornos, texturas e formas. Nas primeiras camadas, geralmente são identificados padrões mais simples, enquanto camadas posteriores combinam essas informações para representar estruturas mais complexas, permitindo ao modelo construir uma representação hierárquica da imagem (GOODFELLOW; BENGIO; COURVILLE, 2016).

Para Goodfellow, Bengio e Courville (2016), além das camadas convolucionais, uma CNN pode apresentar camadas de *pooling*, responsáveis por reduzir a dimensionalidade das representações obtidas e contribuir para a diminuição do custo computacional. Operações como *max pooling* selecionam os valores de maior ativação em determinadas regiões, preservando características consideradas relevantes. Essa redução também pode tornar o modelo menos sensível a pequenas variações na posição dos elementos presentes na imagem, favorecendo sua capacidade de generalização.

O processo de aprendizagem das CNNs ocorre por meio de algoritmos de otimização que ajustam os parâmetros da rede com base nos erros produzidos durante o treinamento. A partir de um conjunto de dados contendo exemplos de entrada e, em tarefas supervisionadas, seus respectivos rótulos, a rede modifica gradualmente seus pesos para melhorar as previsões. O uso da retropropagação do erro (*backpropagation*) permite calcular como cada parâmetro contribuiu para o erro final, possibilitando sua atualização durante o processo de treinamento (GOODFELLOW; BENGIO; COURVILLE, 2016).

Dinh, Tran-Trung e Hoang (2023) observam que é justamente essa capacidade que torna as CNNs eficazes contra CAPTCHAs: a rede aprende, direto dos *pixels*, a separar caracteres, objetos e texturas do fundo, sem que nenhuma regra precise ser programada à mão. Os números da seção anterior dão a dimensão disso — 98,94% de acerto em

CAPTCHAs numéricos sintéticos (NOURY; REZAEI, 2020) e até 92% em esquemas comerciais de texto (YE et al., 2018).

Essa força tem um preço. Ainda segundo Dinh, Tran-Trung e Hoang (2023), as CNNs são sensíveis a alterações mínimas na entrada: uma perturbação calculada a partir do gradiente muda a saída do modelo sem alterar de forma perceptível a imagem. É essa assimetria — muita capacidade de reconhecimento, pouca robustez a pequenas variações direcionadas — que a próxima seção detalha e que este trabalho explora.

2.3 Ruído Adversarial

Segundo Osadchy et al. (2017), o ruído adversarial refere-se a pequenas perturbações introduzidas de forma intencional nos dados de entrada de um modelo de aprendizado de máquina com o objetivo de alterar sua previsão. Em aplicações de visão computacional, essas perturbações podem ser adicionadas a imagens de maneira que, embora sejam pouco perceptíveis ou até imperceptíveis para o ser humano, provoquem alterações significativas na resposta de uma rede neural.

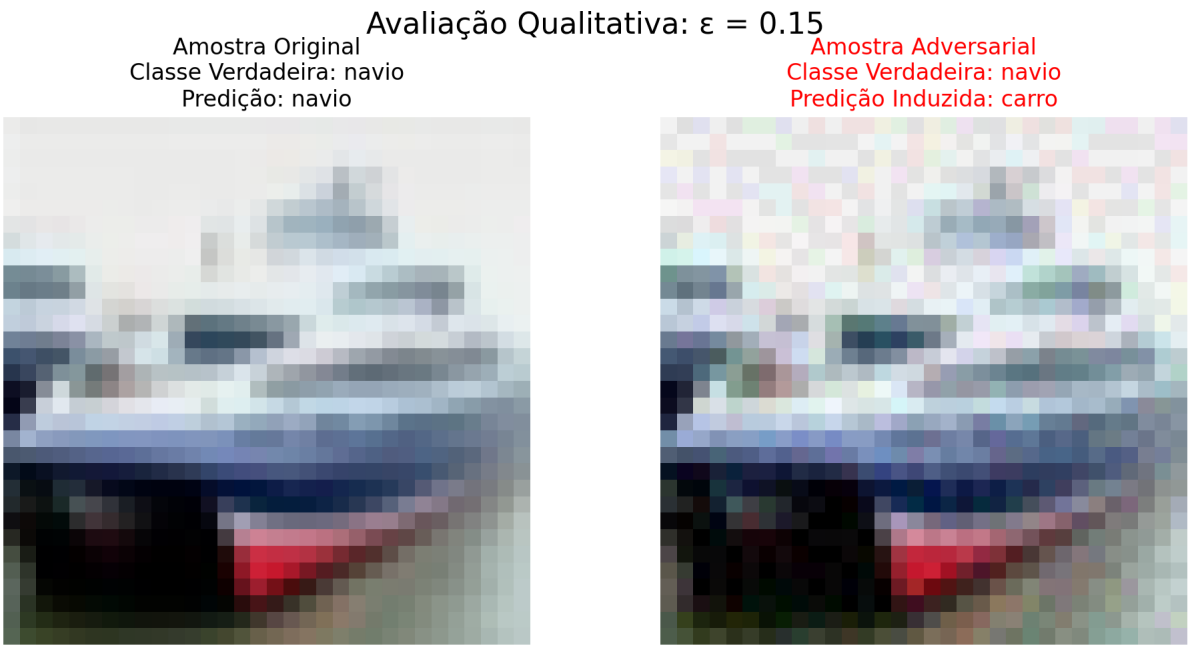
Esse fenômeno evidencia uma limitação importante dos modelos de aprendizado profundo, que podem apresentar elevada precisão em condições normais, mas demonstrar comportamento diferente diante de entradas cuidadosamente modificadas (OSADCHY et al., 2017).

A Figura 1 ilustra esse efeito com um exemplo extraído do próprio experimento desta monografia: uma imagem do conjunto CIFAR-10 corretamente classificada como "navio" passa a ser classificada como "carro" após a aplicação do ataque FGSM em $\varepsilon = 0{,}15$, nível em que o Índice de Similaridade Estrutural médio ainda é de 0,928. Trata-se de um caso ilustrativo, selecionado entre as imagens cuja classificação foi alterada pelo ataque, e não de uma amostra estatisticamente representativa do conjunto.

Dinh, Tran-Trung e Hoang (2023) citam que entre as abordagens utilizadas para a geração de exemplos adversariais estão métodos baseados no gradiente da função de perda do modelo. Essas técnicas utilizam informações obtidas durante o processo de otimização para determinar quais alterações na entrada podem provocar maior impacto sobre a previsão. A intensidade da perturbação pode ser controlada por parâmetros específicos, permitindo estabelecer um equilíbrio entre a alteração necessária para afetar o modelo e a preservação das características originais da imagem.

No contexto das Redes Neurais Convolucionais (CNNs), o ruído adversarial expõe uma limitação estrutural intrínseca dessas arquiteturas. Como apontado por Osadchy et al. (2017), a vulnerabilidade de modelos de aprendizado profundo a exemplos adversariais decorre primordialmente de sua natureza linear por partes (*piecewise linear nature*).

Figura 1 – Efeito do ruído adversarial gerado pelo ataque FGSM em $\varepsilon = 0{,}15$



Fonte: Autor (2026)

Projetadas com ativações e blocos intencionalmente lineares para viabilizar a otimização em larga escala, essas redes tornam-se suscetíveis em espaços de alta dimensão. Sob essa ótica, pequenas perturbações orientadas de forma coerente ao longo da direção do gradiente acumulam-se rapidamente na ativação das camadas subsequentes, alterando drasticamente a classificação final do modelo sem introduzir distorções semânticas perceptíveis ao olho humano (GOODFELLOW; SHLENS; SZEGEDY, 2015; OSADCHY et al., 2017).

Osadchy et al. (2017) foram os primeiros a inverter o uso dessa vulnerabilidade. Em vez de tratá-la como ameaça, propuseram o DeepCAPTCHA: o servidor gera a imagem-desafio, aplica sobre ela uma perturbação adversarial e pede que o usuário identifique o conteúdo original. A rede atacante classifica a imagem de forma errada e com alta confiança; a pessoa continua vendo o mesmo objeto. No protótipo avaliado com 472 participantes, a taxa de acerto humano foi de 86,67% por desafio, contra no máximo 0,7% de chance de um *bot* passar por sorteio.

A ideia depende de três condições, e cada uma delas se torna uma variável experimental nesta pesquisa. A perturbação precisa ser intensa o bastante para enganar o classificador, discreta o bastante para não atrapalhar o usuário, e específica o bastante para que o efeito não se reduza a uma degradação genérica da imagem — caso em que qualquer ruído serviria e o cálculo do gradiente seria dispensável. Medir onde ficam esses limites é o objeto dos Capítulos 3 e 4.

3 Materiais e Métodos

3.1 Caracterização da pesquisa

Esta pesquisa caracteriza-se como aplicada, de natureza exploratória e experimental, com abordagem quantitativa (GIL, 2017). Inicialmente, o estudo valeu-se de uma fase exploratória fundamentada em uma Revisão Sistemática da Literatura (RSL), cujo objetivo foi mapear o estado da arte sobre mecanismos CAPTCHA, identificar suas vulnerabilidades frente às modernas arquiteturas de *Deep Learning* e analisar técnicas promissoras de defesa, com ênfase na geração de exemplos adversariais.

A partir das lacunas teóricas e metodológicas identificadas na RSL, o trabalho evoluiu para uma fase prática e experimental. Dessa forma, a pesquisa metodológica estabelece uma ponte direta entre a teoria levantada na RSL e a prática computacional, culminando na proposição e na avaliação experimental de um cenário de ataque e defesa utilizando Redes Neurais Convolucionais (CNNs) consolidadas na literatura.

Quanto ao delineamento, o experimento funciona como um estudo computacional que aplica o método FGSM, proposto por Goodfellow, Shlens e Szegedy (2015), a um cenário tecnológico diferente. Uma replicação estrita do experimento original seria inviável por conta de diferenças profundas na arquitetura das redes (como o uso da ResNet-20 em vez dos modelos Maxout e classificadores rasos da época) e na própria parametrização do espaço de entrada devido à normalização dos dados.

Nesse contexto, a estratégia de usar as falhas de acurácia de um classificador para avaliar e propor defesas baseia-se na abordagem de Noury e Rezaei (2020). Em seu estudo de vulnerabilidades, os autores analisaram os erros de classificação do próprio solucionador (*Deep-CAPTCHA*) para sugerir proteções estáticas e de baixo custo. Essas recomendações incluíam a escolha de caracteres fáceis de confundir (como os dígitos 3, 7, 8 e 9) e o uso de caracteres em tons mais claros (menor intensidade em escala de cinza), aproximando-os do ruído de fundo.

Esta pesquisa vai além desse modelo ao propor uma contramedida matemática e dinâmica. Em vez de depender de layouts visuais fixos ou de ruídos aleatórios passivos, aplica-se uma perturbação calculada ativamente a partir do gradiente do classificador pelo método FGSM.

Por conta disso, o desenho experimental deste trabalho se afasta do uso de ruído aleatório apenas como regularização no treinamento. O foco aqui está na comparação empírica e sistemática, em tempo de teste, entre a perturbação direcionada por gradiente e um grupo de controle submetido a ruído aleatório de mesma magnitude L_∞ aplicado

sobre as mesmas imagens. Ao analisar o fenômeno sob essas condições controladas e aplicar a validação estatística pareada pelo teste exato de McNemar, torna-se possível isolar o impacto do gradiente e atribuir com segurança as diferenças de acurácia à direção espacial da perturbação, oferecendo um teste independente e rigoroso do método.

3.2 Procedimento da Revisão Sistemática da Literatura

Para mapear o estado da arte e estruturar o problema de pesquisa, realizou-se uma RSL baseada nas diretrizes de Kitchenham e Charters (2007) para Engenharia de Software. Como este Trabalho de Conclusão de Curso (TCC) foi conduzido por um único pesquisador, o método foi adaptado para focar na transparência da busca e na seleção de um *corpus* direcionado ao tema, mantendo o rigor na escolha das fontes.

A estruturação da RSL serviu para responder a cinco questões norteadoras, formuladas para compreender o cenário atual de segurança e a usabilidade das defesas visuais:

Quadro 1 – Questões de pesquisa da RSL

Código	**Questão de pesquisa**
QP1	Quais vulnerabilidades já foram descobertas em mecanismos CAPTCHA?
QP2	Quais tipos de CAPTCHA são mais (ou menos) seguros?
QP3	Quais abordagens recentes foram propostas para melhorar a segurança dos CAPTCHAs?
QP4	Como técnicas de Inteligência Artificial e Aprendizado de Máquina têm influenciado a quebra de CAPTCHAs?
QP5	Quais lacunas persistem na área e ainda precisam ser investigadas?

Fonte: Autor (2026)

3.2.1 Estratégia de Busca e Seleção

A busca bibliográfica ocorreu em março de 2025 nas bases de dados ACM *Digital Library*, IEEE *Xplore* e *ScienceDirect* (Elsevier). Para garantir a relevância dos estudos, o levantamento incluiu apenas artigos publicados em periódicos ou anais de conferências classificados nos estratos A1, A2 ou B1 do sistema Qualis CAPES na área de Computação. As strings de busca utilizadas nas consultas estão detalhadas no Quadro 2.

A triagem dos resultados seguiu os Critérios de Exclusão (CE) listados no Quadro 3. Essa etapa eliminou duplicatas, artigos desalinhados ao escopo de segurança de redes e trabalhos que não apresentavam estudos empíricos primários.

Quadro 2 – Expressões de busca aplicadas às bases

Código	**Expressão de busca**
E1	`captcha AND security`
E2	`captcha AND "machine learning"`
E3	`captcha AND "deep learning"`
E4	`captcha AND attack`
E5	`captcha AND solver`

Fonte: Autor (2026)

Quadro 3 – Critérios de exclusão aplicados na triagem

Código	**Motivo de exclusão**
CE0	Artigo não disponível para leitura do texto completo
CE1	Estudo duplicado
CE2	Não aborda segurança da informação ou controle de acesso
CE3	Não aborda de forma central o mecanismo CAPTCHA
CE4	Literatura cinzenta (teses, dissertações, *blogs* ou relatórios industriais)
CE5	Estudo secundário (revisões de literatura, surveys ou mapeamentos sistemáticos)

Fonte: Autor (2026)

A literatura cinzenta (CE4) foi excluída por não passar por revisão por pares, o que impede assegurar a qualidade metodológica dos resultados. Os estudos secundários (CE5) também foram excluídos, porque resumem estudos primários que podem já estar entre os selecionados; mantê-los faria o mesmo resultado ser contado duas vezes na síntese.

3.2.2 Fluxo de Seleção e Estudos Incluídos

O processo de seleção dos artigos baseou-se no diagrama de fluxo do protocolo PRISMA 2020 (PAGE et al., 2021). Como o PRISMA é uma diretriz de relato, e não de condução, sua aplicação neste estudo serviu para estruturar de forma transparente o fluxo de triagem, enquanto a execução da busca seguiu o método adaptado de Kitchenham e Charters (2007). Por conta disso, a checklist completa de 27 itens da diretriz PRISMA não foi aplicada ou anexada ao manuscrito, o que configura uma escolha de escopo assumida entre as limitações metodológicas deste trabalho.

As mesmas expressões de busca (Quadro 2) foram aplicadas, nas três bases, aos campos de título, resumo e termos de indexação. A diferença de estratégia entre as bases restringiu-se ao uso de filtros de acesso aberto. No IEEE *Xplore* e na *ScienceDirect*, utilizaram-se, respectivamente, as opções *Open Access Only* e *Open access & Open archive*, que limitam os resultados a artigos disponíveis gratuitamente para leitura integral. Esses

Tabela 1 – Síntese do processo de seleção segundo o fluxo do PRISMA

Etapa	Descrição	Quantidade
Identificação	Artigos encontrados nas bases de dados	766
ACM Digital Library	Título, resumo e termos de indexação	450
ScienceDirect	Título, resumo e termos de indexação; filtro *Open access & Open archive*	299
IEEE Xplore	Título, resumo e termos de indexação; filtro *Open Access Only*	17
Deduplicação	Registros repetidos removidos via gerenciador de referências	1
Triagem	Registros avaliados por título e resumo	765
	Registros excluídos na triagem inicial	731
Elegibilidade	Artigos selecionados para leitura completa do texto	34
	Artigos excluídos após leitura completa (motivos de escopo/surveys)	27
Incluídos	**Estudos selecionados para a síntese final**	**7**

Fonte: Autor (2026)

filtros antecipam, já na etapa de busca, o critério de exclusão CE0 (artigo não disponível para leitura do texto completo). Na ACM *Digital Library*, não foi localizado filtro equivalente na interface de busca; nessa base, os registros sem texto completo acessível foram eliminados na triagem, pela aplicação do mesmo critério CE0. Essa escolha segue a orientação de Kitchenham e Charters (2007), segundo a qual as decisões tomadas na estratégia de busca devem ser registradas e justificadas. A restrição ao acesso aberto explica, ao menos em parte, as diferenças no volume de registros recuperados entre as bases e constitui uma limitação desta revisão, pois estudos relevantes publicados sem acesso aberto podem não ter sido identificados.

A busca inicial retornou 766 registros, que apresentaram pouca sobreposição devido à especificidade dos filtros de exportação de cada base. Após a triagem por leitura do texto completo, restaram sete artigos primários para responder às questões norteadoras da pesquisa, conforme listado no Quadro 4.

Após a remoção de 1 duplicata, restaram 765 registros para a triagem por título e resumo, e 731 deles foram excluídos. A Tabela 2 detalha os motivos. A maior parte

das exclusões (647 registros, cerca de 88,5%) decorreu de falta de aderência temática: 421 registros não tratavam o mecanismo CAPTCHA de forma central (CE3), incluídos três sumários de edição recuperados pela busca, e 226 não abordavam segurança da informação (CE2). Esse volume reflete a amplitude das expressões de busca (Quadro 2), que, ao combinar o termo *captcha* com termos genéricos como *security* e *attack*, recuperaram estudos que mencionam o mecanismo apenas de passagem. Outros 62 registros foram excluídos por indisponibilidade do texto completo (CE0), a maioria na ACM Digital Library, base em que não havia filtro de acesso aberto; 17 por serem estudos secundários (CE5); 4 por se tratarem de literatura cinzenta (CE4); e 1, um estudo sobre esquemas de CAPTCHA manuscrito em árabe para dispositivos móveis, foi excluído sem que o motivo fosse registrado no instrumento de seleção.

Tabela 2 – Motivos de exclusão na triagem por título e resumo

Código	**Motivo**	**Registros**
CE0	Texto completo não disponível	62
CE2	Não aborda segurança da informação	226
CE3	Não aborda de forma central o mecanismo CAPTCHA	421
CE4	Literatura cinzenta	4
CE5	Estudo secundário	17
—	Excluído sem motivo registrado	1
Total excluído na triagem		**731**

Fonte: Autor (2026)

Os 34 artigos restantes foram lidos na íntegra, e 27 foram excluídos por não responderem diretamente às questões de pesquisa do Quadro 1 ou por se enquadrarem no critério CE5. O motivo individual de cada exclusão nessa etapa não foi registrado no instrumento de seleção, o que limita a rastreabilidade da revisão.

Um dos estudos incluídos, Dionysiou e Athanasopoulos (2020), configura-se como sistematização de conhecimento (*Systematization of Knowledge* – SoK) e, por isso, enquadra-se formalmente no critério de exclusão CE5. Sua inclusão constitui uma exceção deliberada: além de classificar estudos anteriores, os autores conduzem experimentos próprios, construindo e avaliando classificadores baseados exclusivamente em aprendizado de máquina sobre um esquema de CAPTCHA textual (DIONYSIOU; ATHANASOPOULOS, 2020). São esses resultados empíricos primários, e não a parte de revisão do artigo, que este trabalho utiliza na síntese a seguir.

Além dos sete estudos incluídos, o texto recorre a fontes complementares, que não integram o corpus final da RSL e não foram usadas para responder às questões do Quadro 1. Elas se dividem em dois grupos. O primeiro reúne obras de referência metodológica e técnica, como Goodfellow, Shlens e Szegedy (2015), He et al. (2016), Wang et al. (2004) e Madry et al. (2018), que fundamentam o experimento. O segundo reúne trabalhos sobre CAPTCHA que chegaram à etapa de leitura completa da RSL, mas não integraram o

Quadro 4 – Artigos incluídos para síntese de estado da arte

Estudo	**Título**
Osadchy et al. (2017)	No Bot Expects the DeepCAPTCHA! Introducing Immutable Adversarial Examples, with Applications to CAPTCHA Generation
Ye et al. (2018)	Yet Another Text CAPTCHA Solver: A Generative Adversarial Network Based Approach
Weng et al. (2019)	Towards Understanding the Security of Modern Image CAPTCHAs and Underground CAPTCHA-Solving Services
Dionysiou e Athanasopoulos (2020)	SoK: Machine vs. Machine – A Systematic Classification of Automated Machine Learning-Based CAPTCHA Solvers
Hernandez-Castro, Barrero e R-Moreno (2023)	Breaking CaptchaStar Using the BASECASS Methodology
Dinh et al. (2024)	A Comprehensive Analysis of Cognitive CAPTCHAs Through Eye Tracking
Moradi et al. (2024)	Image CAPTCHAs: When Deep Learning Breaks the Mold

Fonte: Autor (2026)

corpus final por não responderem diretamente às questões do Quadro 1, entre eles Ousat et al. (2024) e Wang et al. (2020). Somam-se a esse grupo dois *preprints* publicados no repositório arXiv e localizados por busca manual complementar, Tariq et al. (2023) e Noury e Rezaei (2020), que se enquadrariam como literatura cinzenta segundo o critério CE4 e, por isso, são empregados apenas como fontes de contextualização, sem peso na síntese da revisão. Também chegaram à leitura completa sem integrar o corpus final os trabalhos de Dinh, Tran-Trung e Hoang (2023) e Chen, Jiang e Chen (2020), retomados na comparação do Quadro 6. Esse procedimento segue a orientação de Kitchenham e Charters (2007), segundo a qual a busca em bases digitais não é suficiente por si só e deve ser complementada por outras fontes, inclusive pela consulta manual a listas de referências de estudos relevantes.

3.2.3 Síntese dos Resultados da RSL

A síntese do *corpus* selecionado é apresentada a seguir, organizada por questão de pesquisa (Quadro 1), e dá base ao desenho experimental deste trabalho.

QP1 – Vulnerabilidades já descobertas. A maioria das vulnerabilidades encontradas não vem da sofisticação técnica de ataques customizados, mas de falhas de projeto dos próprios testes. Como demonstram Hernandez-Castro, Barrero e R-Moreno (2023), foi possível quebrar o CAPTCHA *CaptchaStar* por meio da metodologia *BASECASS* sem desenvolver solucionadores complexos. Para isso, os autores usaram apenas métricas

simples de aleatoriedade do teste *ENT* (como entropia e média aritmética), classificadores comuns de prateleira (*off-the-shelf*) e parametrização padrão. Esse cenário confirma que o vazamento de informações por canais laterais (*side-channel*) inviabiliza a segurança do teste (HERNANDEZ-CASTRO; BARRERO; R-MORENO, 2023). Esse resultado mostra um limite da defesa estudada neste trabalho. O ruído adversarial engana o classificador do atacante, mas não corrige falhas no projeto do desafio: se o gerador deixar padrões estatísticos exploráveis, o atacante resolve o CAPTCHA sem precisar classificar a imagem. Por isso, o ruído adversarial só funciona como uma proteção a mais em um desafio que já seja bem projetado.

QP2 – Tipos de CAPTCHA mais e menos seguros. Os CAPTCHAs textuais tradicionais perderam eficácia diante das novas técnicas de visão computacional, que quebram múltiplos esquemas em tempos inferiores a 50 milissegundos (YE et al., 2018). A inclusão do estudo de Weng et al. (2019) amplia esse cenário, por ser o único trabalho primário do *corpus* que cobre de forma sistemática a segurança de CAPTCHAs de imagem modernos, como os de seleção, deslizamento e clique. Sem essa perspectiva, a análise do estado da arte ficaria restrita a desafios puramente textuais, impedindo uma comparação direta com o domínio do *DeepCAPTCHA* (OSADCHY et al., 2017), cuja proteção também é aplicada no plano bidimensional de imagens naturais de objetos.

QP3 – Abordagens recentes para melhorar a segurança. As defesas mais promissoras exploram as limitações estruturais e matemáticas das próprias redes neurais. Essas abordagens usam exemplos adversariais para derrubar a acurácia de classificadores automatizados sem prejudicar a leitura por humanos legítimos (OSADCHY et al., 2017). Como a preparação da imagem é o ponto crítico para o sucesso dos ataques, essa contramedida ganha força ao ser projetada de forma imutável: a perturbação é calculada para sobreviver a filtros convencionais (como gaussianos ou de mediana), garantindo que o ruído adversarial continue ativo mesmo após tentativas de limpeza (OSADCHY et al., 2017). Essa estratégia de proteção ativa ganha força ao se observar que a recomendação parte dos próprios autores dos ataques. Em seu trabalho sobre a quebra de CAPTCHAs de imagem, Weng et al. (2019) incluem, entre as contramedidas propostas, a geração de perturbações adversariais imperceptíveis, com a vantagem adicional de mitigar também os ataques baseados em serviços *on-line* de visão computacional. Os autores ilustram essa defesa com uma perturbação que descrevem como gerada pelo FGSM, com nível de ruído $\varepsilon = 0{,}07$ e 40 passos de iteração (WENG et al., 2019). Como o FGSM original é um método de passo único (GOODFELLOW; SHLENS; SZEGEDY, 2015), a configuração com 40 iterações corresponde, na prática, à sua variante iterativa (BIM, também chamada I-FGSM), próxima do PGD descrito por Madry et al. (2018).

QP4 – Influência da Inteligência Artificial e do Aprendizado de Máquina na quebra. O uso de *pipelines* que combinam Redes Generativas Adversariais (GANs)

e aprendizado por transferência permite remover ruídos de fundo de forma sistemática, viabilizando a quebra de múltiplos esquemas textuais (YE et al., 2018). O aprendizado de máquina, porém, não basta sozinho: nos experimentos próprios de Dionysiou e Athanasopoulos (2020), a aplicação direta de um classificador único sobre a imagem bruta, sem tratamento prévio, resultou em acurácias de apenas 0,7% com CNN e 0,1% com SVM sobre o esquema textual mais difícil do conjunto avaliado. Isso indica que o pré-processamento e a segmentação continuam sendo etapas essenciais para o ataque funcionar: os CAPTCHAs clássicos falham porque os atacantes limpam as distorções físicas antes do reconhecimento (DIONYSIOU; ATHANASOPOULOS, 2020).

QP5 – Lacunas que persistem. Os estudos do *corpus* sustentam a viabilidade teórica do ruído adversarial como mecanismo de defesa ativo, proposto por Osadchy et al. (2017) e recomendado por Weng et al. (2019), mas carecem de testes práticos que isolem com rigor a causa da degradação que ele provoca no momento da inferência. Essa lacuna empírica é o objeto de estudo do experimento detalhado nas próximas seções. Os trabalhos de Dinh, Tran-Trung e Hoang (2023) e Chen, Jiang e Chen (2020), que seguem a mesma linha de defesa mas não integram o *corpus* da RSL, conforme explicado na Seção 3.2, também não fazem essa verificação, como observado no Capítulo 1.

3.3 Hipóteses de pesquisa

Com base nas evidências levantadas pela literatura, as quais apontam que o ruído adversarial ataca falhas estruturais baseadas em gradiente nos classificadores de Inteligência Artificial, este trabalho propõe-se a validar experimentalmente o impacto dessa técnica. Para isso, definem-se dois pares de hipóteses estatísticas, correspondentes às duas partes da questão de pesquisa formulada na Seção 1.2.

O primeiro par verifica se o ruído adversarial reduz a acurácia do classificador:

a) $\mathbf{H_{0a}}$: a aplicação de ruído adversarial (FGSM) sobre imagens naturais não altera significativamente a acurácia de Redes Neurais Convolucionais (CNNs) em tarefas de classificação visual;

b) $\mathbf{H_{1a}}$: a aplicação de ruído adversarial (FGSM) sobre imagens naturais reduz significativamente a acurácia de Redes Neurais Convolucionais (CNNs) em tarefas de classificação visual.

O segundo par verifica se essa redução decorre da direção da perturbação, e não apenas de sua magnitude:

a) $\mathbf{H_{0b}}$: a acurácia sob FGSM é igual à acurácia sob ruído aleatório de mesma magnitude, aplicado às mesmas imagens;

b) $\mathbf{H_{1b}}$: a acurácia sob FGSM é menor que a acurácia sob ruído aleatório de mesma magnitude, aplicado às mesmas imagens.

Os dois pares são avaliados, em cada nível de ε, pelo teste exato de McNemar, adequado à comparação de classificações pareadas sobre as mesmas amostras. No primeiro par, comparam-se as predições de cada imagem na condição limpa e sob FGSM (teste bilateral); no segundo, as predições de cada imagem sob FGSM e sob ruído aleatório (teste unilateral, coerente com o sentido de H_{1b}). Adota-se o nível de significância $\alpha = 0{,}05$ em ambos os testes. Como o conjunto de teste tem $n = 10.000$ imagens, valores-p muito pequenos são esperados em praticamente qualquer comparação; por isso, além do valor-p, são reportados os pares discordantes de cada teste (b e c), que expressam o tamanho do efeito.

Caso H_{0a} seja rejeitada, conclui-se que a perturbação reduz a acurácia do classificador; caso H_{0b} também seja rejeitada, conclui-se que essa redução não se explica apenas pela magnitude da alteração imposta aos *pixels*, mas pela direção informada pelo gradiente. A extrapolação teórica desses resultados para o domínio dos mecanismos CAPTCHA é detalhada no Capítulo 5.

3.3.1 Teste de McNemar

O teste de McNemar (MCNEMAR, 1947) verifica se duas proporções medidas sobre as mesmas unidades diferem de forma significativa. Ele se aplica a dados pareados e dicotômicos: cada unidade é observada em duas condições e, em cada uma, o resultado assume um de dois valores. Neste trabalho, a unidade é a imagem do conjunto de teste, e o resultado é o acerto ou o erro do classificador.

Em cada comparação, as 10.000 imagens são distribuídas em uma tabela de contingência 2×2 (Quadro 5). A casela a reúne as imagens classificadas corretamente nas duas condições; b, as corretas na primeira condição e incorretas na segunda; c, o inverso; e d, as incorretas nas duas. As caselas a e d não informam sobre a diferença entre as condições, pois nelas a imagem teve o mesmo resultado nas duas. O teste usa apenas os pares discordantes b e c.

Quadro 5 – Tabela de contingência do teste de McNemar

	Condição 2: acerto	**Condição 2: erro**
Condição 1: acerto	a	b
Condição 1: erro	c	d

Fonte: Adaptado de McNemar (1947).

Sob a hipótese nula de que as duas condições têm a mesma taxa de acerto, uma imagem que muda de resultado tem a mesma probabilidade de mudar em um sentido

ou no outro. Assim, dado o total de pares discordantes $n = b + c$, o número b segue uma distribuição binomial de parâmetros n e 0,5. O valor-p é a probabilidade, sob essa distribuição, de se observar um desequilíbrio entre b e c igual ou maior que o observado. Adota-se aqui a forma exata do teste, calculada diretamente pela distribuição binomial, em vez da aproximação pela estatística $\chi^2 = (b-c)^2/(b+c)$, que depende de n suficientemente grande.

O McNemar foi escolhido por causa da forma como o experimento foi montado. As duas condições são aplicadas às mesmas imagens, então as observações não são independentes, e testes para proporções independentes, como o teste z de duas proporções, não levariam esse pareamento em conta. Dietterich (1998) analisou testes estatísticos para comparar classificadores e concluiu que, quando o modelo é avaliado uma única vez sobre um mesmo conjunto de teste, o teste de McNemar é o único com taxa de erro tipo I aceitável. Esse é o caso deste trabalho, em que um único modelo pré-treinado é avaliado sobre o mesmo conjunto de teste nas duas condições.

Como exemplo, na comparação entre a condição limpa e o FGSM com $\varepsilon = 0{,}05$, 6.912 imagens passaram de acerto a erro (b) e apenas 1 passou de erro a acerto (c). Sob a hipótese nula, esses dois números deveriam ser próximos; um desequilíbrio dessa ordem produz valor-p inferior a 10^{-300}.

3.4 Materiais utilizados

Conjunto de dados (*dataset*): o conjunto público de imagens CIFAR-10 foi utilizado como ambiente experimental controlado para validar o efeito do ruído adversarial em modelos convolucionais. Essa escolha metodológica permite isolar o impacto da perturbação sobre o algoritmo antes de qualquer extrapolação para sistemas CAPTCHA reais, avaliando a degradação do desempenho da rede sem a interferência de variáveis externas, texturas ou ruídos nativos do CAPTCHA. Além disso, o CIFAR-10 é o conjunto de referência da literatura de robustez adversarial, em que se define o orçamento convencional de 8/255, o que permite comparar os valores de ε deste trabalho com os de outros estudos. O *dataset* é composto por 60.000 imagens coloridas de 32×32 pixels, das quais foram utilizadas as 10.000 da partição de teste para inferência.

Modelo CNN: optou-se pela utilização de um modelo previamente treinado (ResNet-20) fornecido pelo repositório *chenyaofo/pytorch-cifar-models*. A ResNet-20 é a versão da arquitetura que o próprio artigo original da ResNet propôs para o CIFAR-10 (HE et al., 2016). A decisão de utilizar uma rede com pesos consolidados na literatura visa isolar as variáveis relacionadas ao treinamento, permitindo que a principal variável experimental seja o ataque FGSM. A acurácia do modelo sobre o conjunto de teste limpo, sem qualquer perturbação, foi verificada experimentalmente antes da aplicação do ataque, servindo como

linha de base (*baseline*) real de comparação (HE et al., 2016; KRIZHEVSKY, 2009).

Ferramentas e ambiente de desenvolvimento:

a) **Hardware**: processador AMD Ryzen 5 5600G (3,9 GHz), 16 GB de RAM e GPU NVIDIA GeForce RTX 3060 (12 GB de VRAM), sistema operacional Windows 11 Pro 64 *bits*;

b) **Software**: Python (v3.11.9), PyTorch (v2.5.1+cu124) e Torchvision (v0.20.1+cu124) para manipulação de tensores e cálculo do gradiente via CUDA (v12.4). As métricas de classificação (*accuracy*, *precision*, *recall* e *F1-score*) foram calculadas com *scikit-learn*; a significância estatística das diferenças observadas foi avaliada por meio do teste de McNemar (SciPy); a similaridade perceptual entre imagens originais e adversariais foi quantificada com as métricas SSIM e PSNR (*scikit-image*). Pandas e Matplotlib foram utilizados para estruturação tabular e geração de gráficos e matrizes de confusão. O código-fonte do experimento, os dados brutos e as figuras geradas estão publicamente disponíveis em <https://github.com/jhonkennedy06/tcc-ruido-adversarial-captcha>.

Técnica de ruído adversarial: o algoritmo FGSM (*Fast Gradient Sign Method*) foi escolhido por ser um dos ataques adversariais mais reconhecidos e clássicos da literatura, servindo como *baseline* estruturado para avaliar o impacto de perturbações direcionadas pelo gradiente. O FGSM altera cada componente da imagem (*pixel* e canal) em exatamente $\pm\varepsilon$, conforme o sinal do gradiente: $\eta = \varepsilon\,\mathrm{sign}(\nabla_x J(\theta, x, y))$ (GOODFELLOW; SHLENS; SZEGEDY, 2015).

Como grupo de controle, um segundo conjunto de imagens foi gerado com ruído aleatório de sinal: cada componente recebe, de forma independente e com igual probabilidade, o valor $+\varepsilon$ ou $-\varepsilon$ (ruído de Rademacher). O controle reproduz, assim, a mesma estrutura da perturbação do FGSM, e a única diferença entre os dois grupos é a origem do sinal: sorteado no controle e calculado a partir do gradiente no FGSM. Por construção, as duas perturbações têm a mesma norma L_∞, igual a ε, e a mesma norma L_2, igual a $\varepsilon\sqrt{d}$, em que d é o número de componentes da imagem. Um ruído uniforme no intervalo $[-\varepsilon, \varepsilon]$ teria a mesma norma L_∞, mas norma L_2 esperada de apenas $\varepsilon\sqrt{d/3}$ (cerca de 58% da do FGSM), pois a maior parte de seus componentes ficaria abaixo de ε em módulo; a comparação seria, então, desfavorável ao controle. Com magnitudes idênticas nas duas normas, uma eventual diferença de acurácia entre os grupos não pode ser atribuída à quantidade de alteração imposta aos *pixels*, mas à direção da perturbação.

Nota: o parâmetro ε foi aplicado no espaço de *pixels* normalizado (após a normalização por média e desvio-padrão do canal), e não diretamente no espaço de *pixel* bruto $[0, 1]$, convenção que deve ser considerada na comparação com valores de ε reportados por outros trabalhos da literatura. Como a normalização

divide cada canal pelo seu desvio-padrão ($\sigma \approx 0{,}20$), uma variação ε no espaço normalizado equivale a $\varepsilon \times \sigma$ no espaço de pixel $[0, 1]$. Assim, $\varepsilon = 0{,}05$ corresponde a $0{,}05 \times 0{,}20 \approx 0{,}010$, ou cerca de 2,6/255. Chama-se orçamento de perturbação o valor máximo de ε admitido em um ataque; na literatura de robustez adversarial em CIFAR-10, o valor convencional é 8/255. O valor $\varepsilon = 0{,}05$ fica, portanto, cerca de três vezes abaixo desse orçamento, o que reforça a magnitude do efeito observado. Pela mesma conversão, o limite $\varepsilon = 0{,}15$ adotado neste trabalho corresponde a cerca de 7,7/255, próximo do orçamento convencional.

3.5 Procedimento experimental

Para avaliar o impacto do ruído adversarial na eficácia das Redes Neurais Convolucionais, o experimento computacional foi conduzido em um ambiente acelerado por hardware (GPU via CUDA v12.4), utilizando os *frameworks* PyTorch (v2.5.1+cu124) e Torchvision (v0.20.1+cu124). O procedimento seguiu um *pipeline* rigoroso de cinco etapas, detalhado a seguir.

3.5.1 Etapa 1 — Preparação dos dados

A avaliação foi conduzida utilizando o conjunto de dados CIFAR-10, amplamente adotado na literatura de visão computacional. Foram selecionadas 10.000 imagens de teste do *dataset* original. Para garantir a reprodutibilidade do experimento, a semente aleatória (*seed*) foi fixada no valor 42. O modelo alvo, isto é, o modelo atacado no experimento, escolhido para a classificação foi a arquitetura ResNet-20, previamente treinada para reconhecer as categorias de imagens contidas no CIFAR-10.

3.5.2 Etapa 2 — Avaliação da linha de base (*baseline*)

Antes da introdução de qualquer perturbação, o modelo CNN foi executado sobre as 10.000 imagens originais (sem ruído) para estabelecer a linha de base de acurácia global. Além das métricas globais, cujo denominador são as 10.000 imagens, foram definidas métricas condicionais (acurácia, *precision*, *recall* e *F1-score* em média macro), calculadas exclusivamente sobre o subconjunto de imagens que o classificador identificou corretamente na condição limpa. Por construção, essas métricas valem 100% em $\varepsilon = 0$; esse nível funciona apenas como âncora de normalização para monitorar a deterioração provocada pelas perturbações, e não como medida do desempenho nativo do modelo.

3.5.3 Etapa 3 — Aplicação da perturbação

O conjunto original de imagens foi submetido ao ataque *Fast Gradient Sign Method* (FGSM) em doze níveis de intensidade (ε), de 0,05 a 0,60, em incrementos de 0,05. A faixa cobre desde um valor abaixo do orçamento convencional de 8/255 da literatura de CIFAR-10 ($\varepsilon = 0{,}05$ corresponde a cerca de 2,6/255, conforme a Nota da Seção 3.4) até intensidades em que a perturbação já é visível e a acurácia fica perto do acerto ao acaso, de 10%. O passo de 0,05, no entanto, deixa sem pontos intermediários o trecho entre $\varepsilon = 0$ e $\varepsilon = 0{,}05$, justamente onde ocorre a maior parte da queda de acurácia. Em cada nível, as mesmas imagens receberam também o ruído aleatório de controle descrito na Seção 3.4, com o mesmo valor de ε. Após a soma da perturbação, cada imagem resultante, tanto do FGSM quanto do controle, foi recortada (*clipping*) ao intervalo válido de *pixel*, isto é, ao intervalo $[0, 1]$ expresso no espaço normalizado, de modo que nenhuma imagem contivesse valores fora da faixa representável.

Para acompanhar a degradação visual, calcularam-se o Índice de Similaridade Estrutural (SSIM) e a Relação Sinal-Ruído de Pico (PSNR) entre cada imagem original e suas versões perturbadas, sobre as 10.000 imagens do conjunto de teste e para os dois grupos (FGSM e controle). As métricas foram computadas no espaço de *pixel* $[0, 1]$ com a biblioteca *scikit-image*, em sua configuração padrão para o SSIM: janela deslizante uniforme de 7×7 *pixels*, com os canais de cor avaliados separadamente e depois promediados. Adotou-se como critério de fidelidade aceitável o SSIM médio $\geq 0{,}90$.

3.5.4 Etapa 4 — Avaliação sob ataque (adversarial)

Em cada nível de ε, o mesmo modelo ResNet-20 classificou as imagens perturbadas pelo FGSM e as imagens do grupo de controle. Registraram-se, para cada grupo, a acurácia global e, para o FGSM, as métricas condicionais descritas na Etapa 2 e as matrizes de confusão.

3.5.5 Etapa 5 — Análise comparativa e validação

Os resultados de cada nível foram comparados com a linha de base e entre os dois grupos. As hipóteses da Seção 3.3 foram testadas pelo teste exato de McNemar, ao nível de significância $\alpha = 0{,}05$: H_{0a} pela comparação pareada entre as predições na condição limpa e sob FGSM, e H_{0b} pela comparação pareada entre as predições sob FGSM e sob ruído aleatório. Para cada teste e cada nível de ε, registraram-se os pares discordantes (b e c) e o valor-p. Os resultados são apresentados no Capítulo 4.

4 Resultados e Análise

4.1 Acurácia do modelo em condições limpas (*baseline*)

O modelo ResNet-20 pré-treinado atingiu acurácia de 92,59% sobre o conjunto de teste do CIFAR-10 sem qualquer perturbação, valor consistente com o desempenho característico dessa arquitetura reportado pelo repositório de origem (*chenyaofo/pytorch-cifar-models*). Esse valor constitui a linha de base real de comparação para todas as análises subsequentes. Das 10.000 imagens, 9.259 foram classificadas corretamente nessa condição; esse subconjunto é o denominador das métricas condicionais apresentadas a seguir e no Apêndice B.

4.2 Impacto do ataque FGSM na acurácia global

A Tabela 3 e o Gráfico 1 apresentam a degradação da acurácia global sob ataque FGSM, comparada ao grupo de controle de ruído aleatório de mesma magnitude L_∞. A varredura completa contemplou doze níveis de intensidade, de $\varepsilon = 0{,}05$ a $\varepsilon = 0{,}60$, em incrementos de 0,05, além da condição sem ruído ($\varepsilon = 0$); a tabela reproduz os níveis mais representativos, e a série completa consta do Apêndice A.

Tabela 3 – Acurácia global: FGSM *versus* ruído aleatório

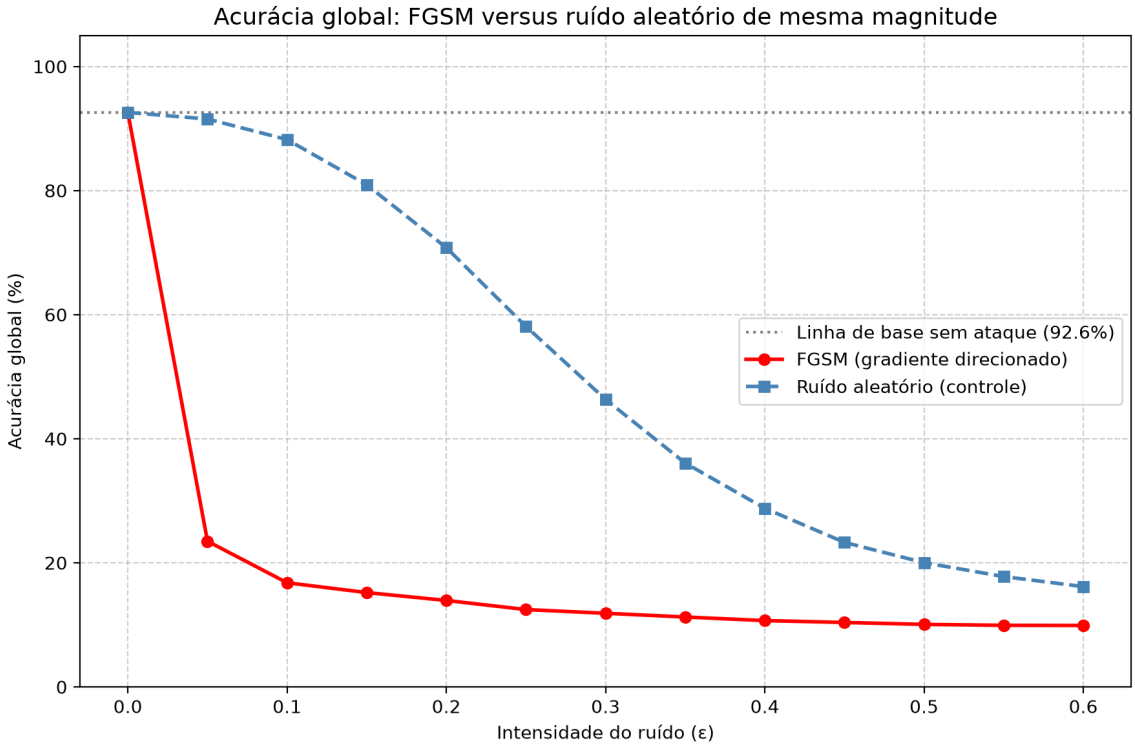
ε	Acurácia global FGSM (%)	Acurácia global ruído aleatório (%)	Redução FGSM (p.p.)
0,00	92,59	92,59	0,00
0,05	23,48	91,53	69,11
0,10	16,78	88,21	75,81
0,20	13,93	70,79	78,66
0,40	10,68	28,78	81,91
0,60	9,90	16,16	82,69

Fonte: Autor (2026)

Os resultados evidenciam uma queda abrupta de acurácia já na menor intensidade testada: com $\varepsilon = 0{,}05$, uma perturbação de magnitude sutil, a acurácia despenca de 92,59% para 23,48%, uma redução de 69,11 pontos percentuais. Em contraste, o grupo de controle com ruído aleatório de magnitude idêntica mantém acurácia de 91,53% no mesmo nível de perturbação, praticamente inalterado em relação ao *baseline*.

Essa discrepância, confirmada estatisticamente na Seção 4.3, é a evidência empírica central deste trabalho: com a mesma magnitude, a perturbação direcionada pelo gradiente

Gráfico 1 – Acurácia global: FGSM *versus* ruído aleatório de mesma magnitude



Fonte: Autor (2026)

da função de perda causa uma degradação muito maior que a aleatória, e, na faixa de baixa intensidade, praticamente toda a queda de acurácia se deve à direção da perturbação, e não à sua magnitude. Isso confirma experimentalmente a premissa teórica discutida na Seção 2.3, de que o ruído adversarial explora uma falha estrutural da rede, e não uma simples sensibilidade genérica a ruído.

Em $\varepsilon = 0{,}60$ (perturbação máxima testada), a acurácia sob FGSM converge para 9,90%, próxima do nível de acerto aleatório para um problema de 10 classes (10%), indicando colapso quase completo da capacidade de classificação do modelo.

As métricas condicionais acompanham esse colapso: em $\varepsilon = 0{,}60$, a acurácia condicional cai para 9,95%, a *precision* para 12,0%, o *recall* para 10,1% e o *F1-score* para 4,74% (Apêndice B). Em intensidades elevadas, ambos os ruídos degradam o classificador — em $\varepsilon = 0{,}60$, o ruído aleatório reduz a acurácia a 16,16% e o FGSM a 9,90% —, mas nesse regime a perturbação já é visualmente evidente (Seção 4.4), o que a inviabiliza como mecanismo de CAPTCHA. É, portanto, na faixa de baixa intensidade, onde a imagem permanece legível, que a diferença entre ruído direcionado e ruído aleatório se manifesta com clareza.

4.3 Validação estatística: teste de McNemar

Para testar formalmente os dois pares de hipóteses definidos na Seção 3.3, aplicou-se o teste exato de McNemar, com $\alpha = 0{,}05$, sobre as classificações pareadas de cada imagem em cada nível de ε. A Tabela 4 apresenta, para os dois testes, os pares discordantes (b e c) e o valor-p.

Tabela 4 – Teste exato de McNemar: pares discordantes e valor-p por nível de perturbação

	H_{0a}: limpo × FGSM			**H_{0b}: FGSM × aleatório**			
ε	b	c	p	b	c	$(b-c)/n$ (p.p.)	p
0,05	6912	1	$< 10^{-300}$	6805	0	68,05	$< 10^{-300}$
0,10	7589	8	$< 10^{-300}$	7150	7	71,43	$< 10^{-300}$
0,15	7763	24	$< 10^{-300}$	6599	27	65,72	$< 10^{-300}$
0,20	7898	32	$< 10^{-300}$	5772	86	56,86	$< 10^{-300}$
0,25	8057	44	$< 10^{-300}$	4702	138	45,64	$< 10^{-300}$
0,30	8118	45	$< 10^{-300}$	3670	223	34,47	$< 10^{-300}$
0,35	8185	52	$< 10^{-300}$	2726	245	24,81	$< 10^{-300}$
0,40	8241	50	$< 10^{-300}$	2088	278	18,10	$< 10^{-300}$
0,45	8277	56	$< 10^{-300}$	1611	320	12,91	$5{,}8 \times 10^{-207}$
0,50	8308	56	$< 10^{-300}$	1335	340	9,95	$2{,}0 \times 10^{-139}$
0,55	8329	62	$< 10^{-300}$	1140	354	7,86	$1{,}2 \times 10^{-96}$
0,60	8338	69	$< 10^{-300}$	998	372	6,26	$1{,}3 \times 10^{-66}$

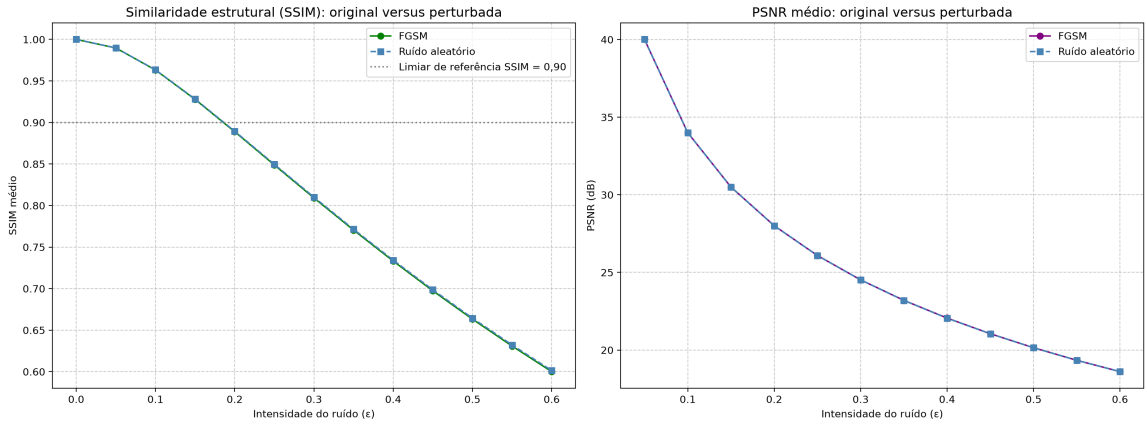
Fonte: Autor (2026)

Nota: H_{0a}: b = imagens corretas na condição limpa e incorretas sob FGSM; c = o inverso. H_{0b}: b = imagens corretas sob ruído aleatório e incorretas sob FGSM; c = o inverso. $n = 10.000$.

No primeiro teste, H_{0a} foi rejeitada em todos os níveis avaliados, com $p < 0{,}001$. Em $\varepsilon = 0{,}05$, por exemplo, 6.912 imagens classificadas corretamente na condição limpa passaram a ser classificadas incorretamente sob FGSM, contra apenas uma no sentido inverso. Em diversos níveis, o valor-p exato sofreu *underflow* numérico e é reportado como $p < 10^{-300}$. A queda de acurácia sob ruído adversarial, portanto, não pode ser atribuída ao acaso.

No segundo teste, H_{0b} foi rejeitada em todos os níveis avaliados, com $p < 0{,}001$. O tamanho do efeito aparece nos pares discordantes: em $\varepsilon = 0{,}05$, 6.805 imagens foram classificadas corretamente sob ruído aleatório e incorretamente sob FGSM, contra nenhuma no sentido inverso, uma diferença líquida de 68,05 pontos percentuais em favor do controle. Essa diferença diminui à medida que ε cresce e o próprio ruído aleatório passa a degradar o classificador, chegando a 6,26 pontos percentuais em $\varepsilon = 0{,}60$. Com a mesma magnitude de perturbação, portanto, o FGSM reduz a acurácia de forma significativamente maior que o ruído aleatório.

Gráfico 2 – SSIM e PSNR médios: FGSM e ruído aleatório



Fonte: Autor (2026)

4.4 Quantificação da imperceptibilidade (SSIM e PSNR)

Para validar empiricamente a alegação teórica de que o ruído adversarial é "visualmente imperceptível" (Seção 2.3), calculou-se a similaridade estrutural (SSIM) e a relação sinal-ruído de pico (PSNR) entre as imagens originais e suas versões perturbadas, para cada nível de ε, sobre as 10.000 imagens do conjunto de teste e para os dois grupos (FGSM e ruído aleatório).

Os testes mostram que a similaridade estrutural diminui de forma contínua com o aumento da intensidade do ruído. Para manter o SSIM médio acima de 0,90, valor adotado aqui como referência para garantir a integridade da imagem, a intensidade máxima viável foi de $\varepsilon = 0{,}15$. Nesse ponto, o SSIM médio registrou 0,9278 e o PSNR médio ficou em 30,50 dB.

A partir de $\varepsilon = 0{,}20$, o índice médio cai para 0,8889 e as distorções se tornam visivelmente mais severas, chegando a um SSIM de apenas 0,6001 no limite máximo de $\varepsilon = 0{,}60$. Esses dados definem com precisão a faixa segura para aplicar o método FGSM dentro da arquitetura de defesa proposta.

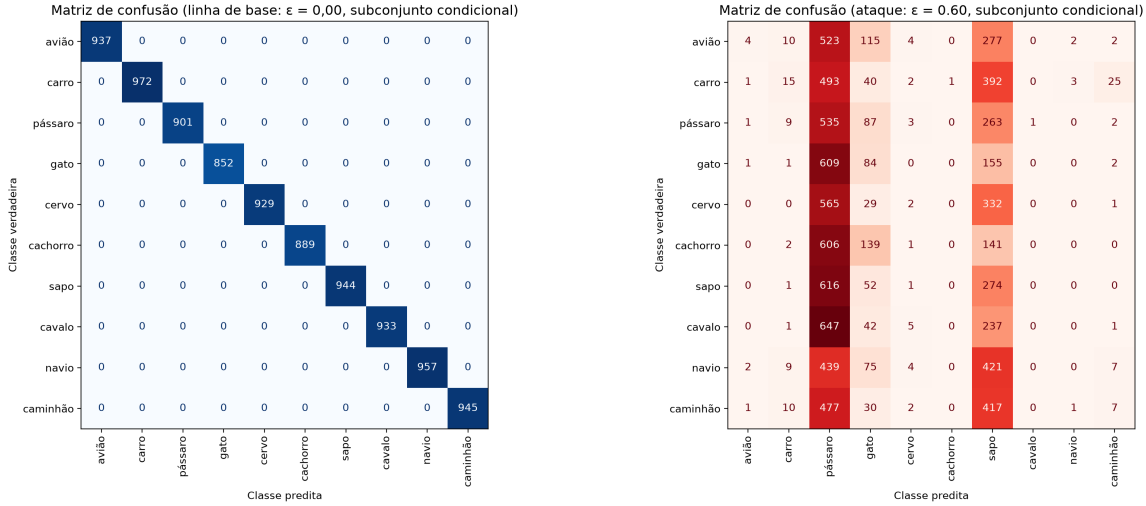
Os valores de SSIM e PSNR do ruído aleatório são praticamente idênticos (diferença máxima de 0,0013 no SSIM e de 0,01 dB no PSNR) aos do FGSM em todos os níveis de ε (Apêndice A), o que era esperado, já que as duas perturbações alteram cada componente da imagem em exatamente $\pm\varepsilon$. Os dois grupos impõem, portanto, a mesma degradação visual medida pelas métricas, mas produzem efeitos muito diferentes sobre o classificador (Seções 4.2 e 4.3).

Na prática, isso fixa o teto de operação a ser adotado pelo modelo conceitual apresentado no Capítulo 5: acima de $\varepsilon \approx 0{,}15$, a perturbação deixa de atender ao critério de fidelidade adotado nesta pesquisa.

Cabe, contudo, uma ressalva metodológica. SSIM e PSNR são métricas objetivas de similaridade entre sinais e funcionam como aproximação da percepção humana, não como medida direta dela. A afirmação de imperceptibilidade adotada neste trabalho é, portanto, uma afirmação métrica: sua confirmação perceptual exigiria um experimento com participantes humanos, não conduzido nesta pesquisa.

4.5 Análise qualitativa: convergência para classes-atratoras

Gráfico 3 – Matrizes de confusão sobre o subconjunto classificado corretamente em condições limpas



Fonte: Autor (2026)

A matriz de confusão em $\varepsilon = 0{,}60$ revela um padrão que a acurácia isolada não mostra: em vez de os erros se espalharem pelas dez classes, o modelo passa a classificar a maioria das imagens perturbadas como "pássaro" ou "sapo", independentemente da classe verdadeira. O ataque não confunde o classificador ao acaso; empurra as predições para regiões concentradas do espaço de decisão. Este trabalho não investigou a causa do fenômeno. A explicação mais direta seria a de que essas duas classes ocupam regiões amplas na fronteira aprendida por este modelo específico, hipótese compatível com o argumento de linearidade de Goodfellow, Shlens e Szegedy (2015), mas que exigiria análise adicional para ser confirmada.

4.6 Síntese e validação da hipótese

Os resultados deste capítulo sustentam duas conclusões. A primeira: o ruído adversarial reduz a acurácia da ResNet-20 de forma que não se explica pelo acaso (H_{0a} rejeitada nos doze níveis avaliados, $p < 0{,}001$). A segunda, que é a contribuição própria deste trabalho: essa redução não se explica apenas pela magnitude da perturbação. Sobre

as mesmas imagens e com a mesma magnitude nas normas L_∞ e L_2, o FGSM reduz a acurácia significativamente mais que o ruído aleatório (H_{0b} rejeitada nos doze níveis, $p < 0{,}001$), o que indica que a degradação decorre da direção informada pelo gradiente. Em $\varepsilon = 0{,}05$, a diferença entre os dois grupos é de 68,05 pontos percentuais.

Esse resultado torna defensável, no plano conceitual, o desenho de CAPTCHA proposto a seguir, no Capítulo 5. Não o valida: o experimento não empregou desafios CAPTCHA, nem solucionadores reais, nem participantes humanos.

5 Proposta de Modelo Conceitual

A partir da fundamentação teórica apresentada e das lacunas identificadas na literatura, que evidenciam a ineficácia das defesas visuais tradicionais frente ao *Deep Learning*, bem como dos resultados experimentais apresentados no Capítulo 4, propõe-se um modelo conceitual de arquitetura para a geração dinâmica de desafios CAPTCHA reforçados por ruído adversarial.

5.1 Visão geral da arquitetura proposta

O modelo conceitual é estruturado em quatro etapas, executadas no momento em que o usuário solicita acesso ao sistema protegido.

A Figura 2 apresenta essa visão geral, e a Figura 3 detalha a troca de mensagens entre os componentes.

5.1.1 Geração do desafio base

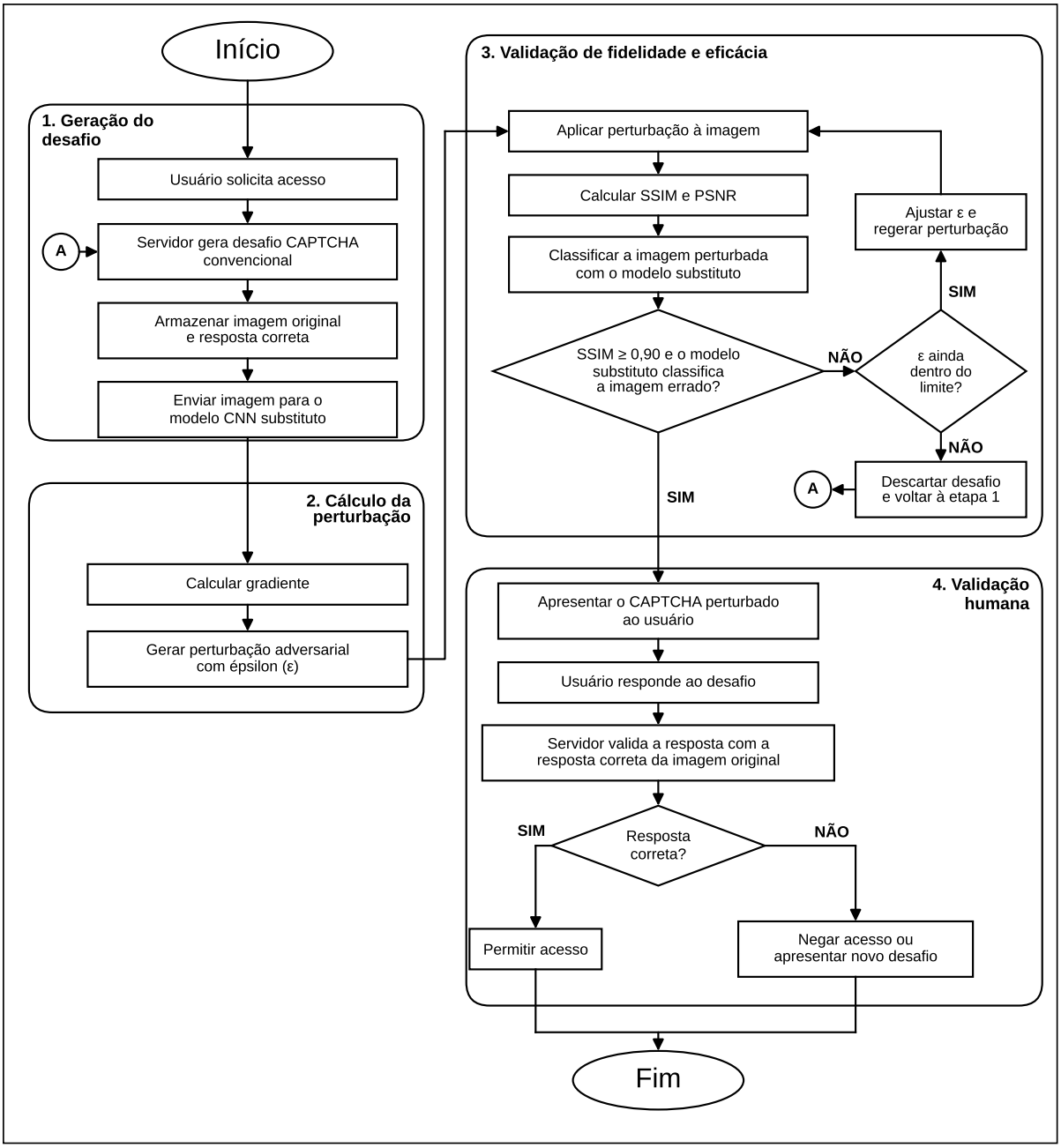
Inicialmente, o servidor gera uma imagem-desafio convencional, que pode envolver texto distorcido, seleção de objetos ou classificação semântica de imagens. Essa geração segue os padrões visuais e estruturais já estabelecidos na literatura de design de CAPTCHA, garantindo que o formato do desafio seja familiar e intuitivo para o usuário.

5.1.2 Cálculo da perturbação adversarial via modelo substituto

Como o servidor não sabe qual classificador um eventual *bot* usará, este trabalho propõe calcular a perturbação sobre um modelo CNN de referência mantido pelo próprio serviço (modelo substituto ou *surrogate model*). Essa é uma decisão de projeto desta proposta. O *DeepCAPTCHA* original, proposto por Osadchy et al. (2017), adota um modelo de ameaça diferente (*white-box*), no qual assume que o atacante conhece integralmente a arquitetura e os parâmetros da rede geradora. A segurança do esquema original baseia-se na indisponibilidade da imagem-fonte e na imutabilidade do ruído, e não no segredo do classificador de referência.

Essa etapa assume a propriedade de transferibilidade dos exemplos adversariais: perturbações geradas contra uma rede de referência tendem a enganar também outras arquiteturas. Goodfellow, Shlens e Szegedy (2015) descrevem o fenômeno, mas o tratam como tendência, e não como garantia — em seus experimentos, a concordância de erro entre modelos distintos variou de 16,0% a 54,6%, conforme a arquitetura comparada.

Figura 2 – Fluxograma do processo de geração e validação do desafio

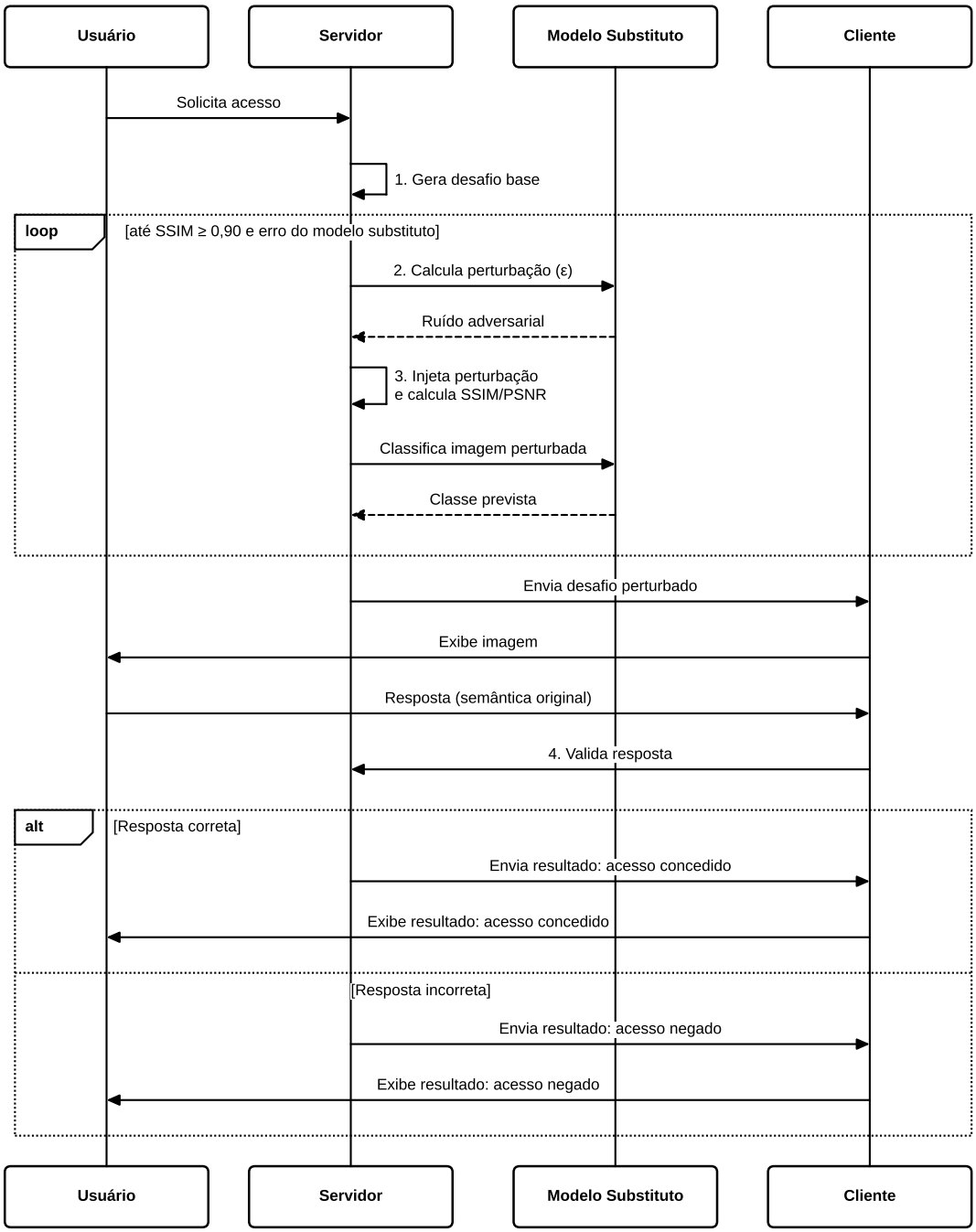


Fonte: Autor (2026)

Dinh, Tran-Trung e Hoang (2023) apoiam-se na mesma propriedade em sua proposta de CAPTCHA estilizado, sem isolá-la nem quantificá-la como variável. Este trabalho também não a mede: a transferibilidade é aqui um pressuposto de projeto, e sua verificação empírica está entre as recomendações do Capítulo 6.

A escolha do fator de perturbação (ε) foi calibrada para respeitar o critério de fidelidade estrutural determinado nos testes deste estudo (Seção 4.4). Adotou-se o limiar de SSIM $\geq 0{,}90$ como limite de qualidade operacional aceitável para o sistema. Essa definição baseia-se estritamente na análise de similaridade de sinais, e não em um limite perceptual

Figura 3 – Diagrama de sequência da interação entre usuário, cliente, servidor e modelo substituto



Fonte: Autor (2026)

absoluto da literatura, já que o índice SSIM (WANG et al., 2004) funciona em uma escala contínua sem faixas normativas fixas. Nos experimentos realizados, o limite máximo que atende a essa restrição é de $\varepsilon = 0{,}15$, ponto no qual o SSIM médio registrou o valor de 0,9278.

5.1.3 Injeção da perturbação e validação de fidelidade e eficácia

Antes de exibir o desafio ao usuário, o sistema aplica o ruído e verifica de forma automática, por meio de métricas como SSIM e PSNR, se a imagem gerada permanece dentro do limite de fidelidade estrutural definido. Essa fase funciona como um controle de qualidade em tempo real, garantindo que a perturbação matemática não comprometa a legibilidade do CAPTCHA com distorções excessivas. O índice SSIM (WANG et al., 2004) funciona neste caso como um indicador objetivo de preservação do sinal, e não como uma garantia perceptual de que os ruídos de alta frequência sejam invisíveis ao olho humano.

Além da fidelidade, o sistema verifica se a perturbação cumpre o seu papel. A imagem perturbada é classificada pelo modelo substituto, e o desafio só segue para o usuário se duas condições forem atendidas ao mesmo tempo: o SSIM da imagem é igual ou superior a 0,90 e o modelo substituto erra a classificação. Sem a segunda condição, uma imagem em que o FGSM não alterou a decisão da rede seria entregue ao cliente sem proteção efetiva contra robôs. Se uma das condições falhar, o valor de ε é ajustado e a perturbação é gerada de novo, sempre dentro do limite de fidelidade definido na Seção 5.1.2. Se nenhum valor de ε dentro desse limite fizer o modelo substituto errar, o desafio é descartado e o servidor gera uma nova imagem-base, o que impede que o processo se repita indefinidamente. Essa verificação segue a lógica do algoritmo de geração de ruído imutável de Osadchy et al. (2017), que filtra a imagem perturbada, testa se a rede ainda a reconhece e, se reconhecer, aumenta o ruído, sem ultrapassar o nível que permite o reconhecimento humano. O erro do modelo substituto é condição necessária, mas não suficiente: ele não garante que o classificador do atacante também erre, o que depende da transferibilidade discutida na Seção 5.1.2.

5.1.4 Validação da resposta humana

O desafio perturbado é, por fim, apresentado na interface do cliente. Osadchy et al. (2017) afirmam que, diferentemente de um CAPTCHA convencional, a resposta esperada pelo servidor corresponde à percepção semântica humana da imagem original (não perturbada), e não à classificação incorreta que uma CNN atacante produziria com alta confiança. O mecanismo explora a divergência entre o que a pessoa vê e o que a rede computa: resolve o desafio quem enxerga a imagem original, e não a classificação induzida pela perturbação (OSADCHY et al., 2017). Vale registrar que essa é uma hipótese de projeto. O modelo aqui descrito não foi implementado nem submetido a usuários, e a verificação dessa premissa está entre os trabalhos futuros.

5.2 Fundamentação e limitações de projeto

Essa proposta arquitetural parte do conceito de exemplos adversariais imutáveis (*immutable adversarial examples*), formalizado por Osadchy et al. (2017). Os autores tratam o ruído adversarial como um mecanismo de defesa ativo e persistente, projetado para resistir a técnicas de pré-processamento e filtragem, como filtros de mediana ou gaussianos, que normalmente conseguiriam remover ruídos visuais convencionais. Essa resistência, porém, não é uma propriedade do FGSM padrão usado na Etapa 2 do modelo proposto. Os próprios autores mostram que o ruído adversarial simples é insuficiente para construir um CAPTCHA seguro, justamente por poder ser removido ou atenuado por filtragem antes da classificação, e por isso propõem um algoritmo específico para tornar o ruído imutável. Neste trabalho, a resistência a filtragem não foi avaliada (Quadro 6). A imutabilidade deve ser entendida, portanto, como um requisito evolutivo do módulo gerador de ruído, a ser incorporado em implementações futuras, e não como uma garantia da arquitetura na forma aqui descrita.

A vantagem das perturbações direcionadas sobre o ruído estático aparece nos próprios experimentos de Weng et al. (2019). Contramedidas passivas produzem ganhos parciais: inserir uma região vazia enganosa em CAPTCHAs de deslizamento reduz pela metade a taxa de sucesso do ataque, e ruído aleatório de fundo dificulta a recuperação da imagem-fonte. Nenhuma delas, porém, leva a taxa de quebra a níveis aceitáveis. É essa insuficiência que justifica perturbações calculadas sobre o gradiente do classificador, e não apenas sobre a aparência da imagem (WENG et al., 2019).

A principal limitação de projeto desta arquitetura é a dependência do cenário de caixa-branca (*white-box*) exigido pelo algoritmo FGSM. O método precisa de acesso direto aos gradientes do classificador para calcular a perturbação. No modelo proposto, esse cálculo é feito sobre o modelo substituto mantido pelo próprio servidor, e não sobre o solucionador real que o atacante utiliza.

Por outro lado, a eficácia do ruído adversarial em cenários práticos de caixa-preta (*black-box*), nos quais o servidor de defesa opera sem conhecer a arquitetura do classificador hostil, baseia-se na propriedade de transferibilidade dos exemplos adversariais. Esse fenômeno, descrito por Goodfellow, Shlens e Szegedy (2015), indica que as perturbações geradas a partir de um modelo interno de referência tendem a enganar redes neurais com arquiteturas distintas. Essa característica diminui o problema do acesso restrito aos gradientes do atacante, embora não elimine totalmente a incerteza sobre a taxa real de sucesso da defesa em condições reais de operação.

O desenvolvimento dessas contramedidas ganha importância diante da alta eficiência dos novos métodos de ataque. Como demonstram Wang et al. (2020), técnicas baseadas em aprendizado por transferência (*transfer learning*) conseguem quebrar CAPTCHAs

tradicionais com facilidade, exigindo apenas 500 amostras reais rotuladas por esquema (1.000 nos esquemas em chinês) para o ajuste fino do modelo. O baixo custo e a facilidade para criar esses solucionadores modernos mostram que é preciso validar, de forma prática e urgente, defesas baseadas em perturbações ativas.

Como a taxa exata de sucesso dessa generalização do ruído não foi avaliada exaustivamente neste trabalho, ela é apontada como a recomendação central para trabalhos futuros (Capítulo 6). Uma mitigação conceitual viável para elevar essa taxa de transferibilidade seria o uso de um comitê (*ensemble*) de diferentes modelos substitutos durante a etapa de geração do ruído, ou ainda a adoção de métodos de perturbação otimizados e mais robustos à transferência, como o PGD, abordagem esta que fica reservada para investigação futura (MADRY et al., 2018).

Para além da robustez matemática, o sucesso prático de uma defesa depende do custo que ela impõe ao usuário legítimo. Dinh et al. (2024) mediram esse custo com rastreamento ocular em 29 participantes: quanto maior a exigência cognitiva do desafio, mais longas as fixações, mais rápidas as sacadas e maior a dilatação pupilar. Entre os cinco modelos de CAPTCHA cognitivo testados, o tempo médio de resolução variou de 9,3 a 19,1 segundos, e a taxa de acerto humano, de 44,8% a 93,1%.

O rastreamento ocular exige equipamento dedicado e não se aplica à *web* aberta. A extensão sugerida aqui é uma proposta deste trabalho, e não uma recomendação dos autores citados: aproximar esse custo cognitivo por métricas que o próprio navegador já captura, como o tempo até a resposta, e usar essa leitura para ajustar ε dentro da faixa validada no Capítulo 4. Trata-se de uma hipótese de projeto, não avaliada nesta pesquisa.

Para situar a proposta em relação às abordagens já publicadas, o Quadro 6 compara as três propostas publicadas na literatura e a deste trabalho quanto ao mecanismo de defesa adotado, à resistência a técnicas de filtragem, à forma de validação e às limitações declaradas.

Quadro 6 – Comparação entre as propostas de CAPTCHA baseadas em defesa contra redes profundas

Proposta	**Mecanismo de defesa**	**Resistência a filtragem**	**Validação**	**Limitação**
DeepCAPTCHA Osadchy et al. (2017)	Ruído adversarial imutável	Projetada explicitamente (resiste a filtro de mediana e gaussiano)	Protótipo com análise de segurança e usabilidade (472 participantes)	Exige reciclagem da rede sob treinamento adversarial do atacante
StyleCAPTCHA Chen, Jiang e Chen (2020)	Estilização por transferência de estilo neural + troca adversarial de estilos guiada pela CCT	Não discutida	Experimental contra redes profundas (sem testes com humanos)	Depende de repertório finito de estilos; usabilidade humana apenas estimada por hipótese (95%)
Adversarial + *style transfer* Dinh, Tran-Trung e Hoang (2023)	FGSM combinado a transferência de estilo neural	Não avaliada	Experimental em CAPTCHAs de texto e imagem	Não avalia transferibilidade entre diferentes arquiteturas
Este trabalho	FGSM sobre modelo substituto com calibração por SSIM	Não avaliada	Modelo conceitual; experimento em CIFAR-10	Cenário *white-box*; sem teste em CAPTCHA real

Fonte: Autor (2026)

6 Considerações Finais

O experimento respondeu à pergunta formulada no Capítulo 1. Com $\varepsilon = 0{,}05$ — perturbação equivalente a cerca de 2,6/255 no espaço de *pixel*, três vezes menor que o orçamento convencionalmente adotado na literatura de robustez em CIFAR-10 — a acurácia da ResNet-20 caiu de 92,59% para 23,48%. Nas mesmas imagens, o ruído aleatório de idêntica magnitude L_∞ manteve 91,53%. O teste exato de McNemar rejeitou, com $p < 0{,}001$, tanto H_{0a}, de que o FGSM não altera a acurácia, quanto H_{0b}, de que FGSM e ruído aleatório de mesma magnitude produzem a mesma acurácia, em todos os níveis avaliados.

Esses números indicam que, com a mesma quantidade de ruído, o que mais derruba o classificador é a direção em que ele é aplicado. A distinção importa porque define o custo da defesa. Se o efeito dependesse da magnitude, proteger o desafio significaria degradar a imagem até torná-la ilegível também para o usuário. Como depende da direção, a perturbação pode ficar em níveis discretos: até $\varepsilon = 0{,}15$ o SSIM médio permanece em 0,9278, com PSNR de 30,50 dB.

Cabe ser explícito sobre o que este trabalho não demonstrou. O experimento usou imagens naturais do CIFAR-10, de 32×32 *pixels*, e não desafios CAPTCHA. Nenhum solucionador real foi atacado, nenhum usuário humano foi consultado e a transferibilidade da perturbação entre arquiteturas não foi medida. O que os resultados sustentam é a plausibilidade do mecanismo descrito no Capítulo 5 — não sua eficácia como defesa em produção. Aquele modelo é uma proposta derivada destes dados, e sua validação continua em aberto.

6.1 Principais contribuições

Este trabalho não propõe um método adversarial novo nem reproduz de forma estrita o experimento de Goodfellow, Shlens e Szegedy (2015): aplica o FGSM, tal como formulado pelos autores, a uma arquitetura e a um conjunto de dados diferentes dos originais. A própria ideia de que a direção da perturbação importa mais do que sua magnitude também já está em Goodfellow, Shlens e Szegedy (2015), cuja comparação com ruído aleatório se dá no contexto do treinamento. A contribuição deste trabalho é verificar essa ideia em tempo de inferência, sob controle explícito: sobre as mesmas imagens, o FGSM foi comparado a um ruído aleatório com a mesma magnitude nas normas L_∞ e L_2, e a diferença entre os dois grupos foi submetida a teste estatístico pareado (H_{0b}, McNemar exato, $p < 0{,}001$), com os pares discordantes reportados como medida do tamanho do efeito. Trata-se, portanto, de uma contribuição de verificação empírica controlada, e não

de uma extensão do método.

Além disso, a medição da imperceptibilidade visual por meio das métricas SSIM e PSNR permitiu estabelecer um limite prático concreto ($\varepsilon \leq 0{,}15$) para o uso dessas perturbações. Pelo critério métrico adotado, esse limite indica uma faixa em que a imagem preserva sua estrutura e ainda degrada fortemente o classificador, fornecendo a base técnica necessária para o modelo conceitual apresentado no Capítulo 5.

6.2 Limitações

Como limitações, destacam-se:

a) o experimento foi conduzido sobre imagens naturais do CIFAR-10, e não sobre CAPTCHAs reais, o que restringe a generalização direta dos resultados;

b) o ataque implementado assume cenário *white-box*, com acesso completo aos gradientes do modelo atacado, cenário que pode não se sustentar diante de solucionadores automatizados que utilizem arquiteturas distintas da usada para gerar o ruído — risco já demonstrado na prática por ataques de transferência de aprendizado contra CAPTCHAs textuais;

c) a eficácia do ruído adversarial foi testada sob o cenário de uma única tentativa de quebra por parte do solucionador (*single-shot attack*). Como alertam Ousat et al. (2024), análises baseadas em uma tentativa isolada tendem a superestimar a força da defesa, já que ataques automatizados reais operam com múltiplas tentativas consecutivas. Na prática, taxas de sucesso moderadas de 63% em 5 tentativas podem subir para 89,85% quando o atacante realiza até 50 tentativas;

d) apenas o método de passo único FGSM foi avaliado nos testes práticos. Métodos iterativos como *BIM*, *DeepFool* (discutidos na literatura de CAPTCHAs estilizados (DINH; TRAN-TRUNG; HOANG, 2023)) e *PGD* (MADRY et al., 2018) não foram contemplados;

e) a imperceptibilidade das perturbações foi avaliada apenas por métricas objetivas (SSIM e PSNR), sem testes com usuários reais;

f) o SSIM foi calculado com a configuração padrão da biblioteca *scikit-image*, com janela uniforme de 7×7 *pixels*, e não com a janela gaussiana de 11×11 *pixels* ($\sigma = 1{,}5$) adotada por Wang et al. (2004); por isso, os valores obtidos não são diretamente comparáveis aos reportados com a configuração original. Além disso, as imagens são pequenas (32×32 *pixels*): a janela cobre cerca de um quinto da largura da imagem, e a biblioteca descarta da média as três linhas e colunas de cada borda, de modo que o índice médio reflete apenas a região central de 26×26 *pixels*. O SSIM foi validado originalmente por Wang et al.

(2004) em 344 imagens de 768×512 *pixels* comprimidas com JPEG e JPEG2000. A sensibilidade da métrica diante de ruídos adversariais de alta frequência em baixa resolução exige, portanto, cautela, e o índice não deve ser tomado como um limite absoluto da percepção humana;

g) a robustez da defesa se limita ao bloqueio de ataques automatizados por aprendizado de máquina. Como apontam Weng et al. (2019), os autores identificaram 152 serviços clandestinos de resolução de CAPTCHA, com cerca de 3.000 trabalhadores cada e receita diária estimada em torno de US$ 1 milhão por serviço. Por isso, sistemas focados apenas na distinção entre computadores e humanos continuam vulneráveis a ataques de retransmissão humana (*relay attacks*), uma limitação intrínseca de segurança comum a todos os testes de Turing modernos;

h) na RSL, o motivo individual de exclusão dos 27 artigos descartados após leitura completa não foi registrado, assim como o de um registro excluído na triagem por título e resumo, o que limita a rastreabilidade da seleção;

i) a varredura de ε usou passo uniforme de 0,05, de modo que a região entre $\varepsilon = 0$ e $\varepsilon = 0{,}05$, onde a acurácia sob FGSM cai de 92,59% para 23,48%, não foi detalhada, e o ponto em que a queda se inicia não pode ser localizado com precisão.

6.3 Trabalhos futuros

Para trabalhos futuros, recomenda-se:

a) a validação do modelo conceitual proposto em bases de CAPTCHA reais, contemplando tanto a taxa de bloqueio de solucionadores automatizados quanto a usabilidade para usuários humanos;

b) a investigação da transferibilidade do ataque entre arquiteturas distintas de CNN, avaliando a robustez do modelo conceitual em cenários *black-box*;

c) a comparação com métodos adversariais iterativos, como BIM, *DeepFool* (DINH; TRAN-TRUNG; HOANG, 2023) e PGD (MADRY et al., 2018), ampliando a validade externa das conclusões aqui apresentadas;

d) a avaliação da robustez da defesa sob ataques iterativos e regimes de múltiplas tentativas, uma vez que análises baseadas em uma única tentativa tendem a superestimar a força da defesa (OUSAT et al., 2024);

e) uma varredura mais fina de ε na faixa $0 < \varepsilon \leq 0{,}05$, para localizar o ponto em que a acurácia sob FGSM começa a cair de forma significativa;

f) a incorporação de mecanismos de imutabilidade ao gerador de ruído, nos moldes do proposto por Osadchy et al. (2017), e a avaliação da resistência da perturbação a filtros de mediana e gaussianos.

Referências

CHEN, H.; JIANG, B.; CHEN, H. StyleCAPTCHA: CAPTCHA based on stylized images to defend against deep networks. In: *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference (FODS '20)*. New York: ACM, 2020. p. 161–170. DOI: 10.1145/3412815.3416895. Citado 4 vezes nas páginas 19, 33, 35 e 53.

DIETTERICH, T. G. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, v. 10, n. 7, p. 1895–1923, 1998. DOI: 10.1162/089976698300017197. Citado na página 37.

DINH, N. et al. A comprehensive analysis of cognitive CAPTCHAs through eye tracking. *IEEE Access*, v. 12, p. 47190–47209, 2024. DOI: 10.1109/ACCESS.2024.3373542. Citado 2 vezes nas páginas 33 e 52.

DINH, N.; TRAN-TRUNG, K.; HOANG, V. T. Augment CAPTCHA security using adversarial examples with neural style transfer. *IEEE Access*, v. 11, p. 83553–83561, 2023. DOI: 10.1109/ACCESS.2023.3298442. Citado 9 vezes nas páginas 19, 25, 26, 33, 35, 48, 53, 55 e 56.

DIONYSIOU, A.; ATHANASOPOULOS, E. SoK: machine vs. machine — a systematic classification of automated machine learning-based CAPTCHA solvers. *Computers & Security*, v. 97, p. 101947, 2020. DOI: 10.1016/j.cose.2020.101947. Citado 4 vezes nas páginas 18, 32, 33 e 35.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. 2. ed. Rio de Janeiro: LTC, 2021. Citado na página 25.

GIL, A. C. *Como elaborar projetos de pesquisa*. 6. ed. São Paulo: Atlas, 2017. Citado na página 28.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. Cambridge: MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>. Citado 2 vezes nas páginas 19 e 25.

GOODFELLOW, I. J.; SHLENS, J.; SZEGEDY, C. Explaining and harnessing adversarial examples. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego: [s.n.], 2015. ArXiv:1412.6572. Disponível em: <https://arxiv.org/abs/1412.6572>. Citado 10 vezes nas páginas 21, 27, 28, 32, 34, 38, 45, 47, 51 e 54.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016. p. 770–778. DOI: 10.1109/CVPR.2016.90. Citado 3 vezes nas páginas 32, 37 e 38.

HERNANDEZ-CASTRO, C.; BARRERO, D. F.; R-MORENO, M. D. Breaking CaptchaStar using the BASECASS methodology. *ACM Transactions on Internet Technology*, v. 23, n. 1, p. 1–12, 2023. Artigo 5. DOI: 10.1145/3546867. Citado 3 vezes nas páginas 24, 33 e 34.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing systematic literature reviews in software engineering*. Keele, 2007. Versão 2.3. EBSE Technical Report. Disponível em: <https://www.elsevier.com/___data/promis_misc/525444systematicreviewsguide.pdf>. Citado 4 vezes nas páginas 29, 30, 31 e 33.

KRIZHEVSKY, A. *Learning multiple layers of features from tiny images*. Toronto, 2009. Relatório técnico. Disponível em: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf>. Citado na página 38.

MADRY, A. et al. Towards deep learning models resistant to adversarial attacks. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver: [s.n.], 2018. ArXiv:1706.06083. Disponível em: <https://arxiv.org/abs/1706.06083>. Citado 5 vezes nas páginas 32, 34, 52, 55 e 56.

MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, v. 12, n. 2, p. 153–157, 1947. DOI: 10.1007/BF02295996. Citado na página 36.

MORADI, M. et al. Image CAPTCHAs: when deep learning breaks the mold. *IEEE Access*, v. 12, p. 112211–112231, 2024. DOI: 10.1109/ACCESS.2024.3442976. Citado 4 vezes nas páginas 18, 23, 25 e 33.

NOURY, Z.; REZAEI, M. Deep-CAPTCHA: a deep learning based CAPTCHA solver for vulnerability assessment. *arXiv preprint arXiv:2006.08296*, 2020. Disponível em: <https://arxiv.org/abs/2006.08296>. Citado 4 vezes nas páginas 24, 26, 28 e 33.

OSADCHY, M. et al. No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation. *IEEE Transactions on Information Forensics and Security*, v. 12, n. 11, p. 2640–2653, 2017. DOI: 10.1109/TIFS.2017.2718479. Citado 11 vezes nas páginas 19, 26, 27, 33, 34, 35, 47, 50, 51, 53 e 57.

OUSAT, B. et al. The matter of CAPTCHAs: an analysis of a brittle security feature on the modern web. In: *Proceedings of the ACM Web Conference 2024 (WWW '24)*. New York: ACM, 2024. p. 1835–1846. Singapore. DOI: 10.1145/3589334.3645619. Citado 6 vezes nas páginas 18, 19, 23, 33, 55 e 56.

PAGE, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, v. 372, p. n71, 2021. DOI: 10.1136/bmj.n71. Citado na página 30.

TARIQ, N. et al. CAPTCHA types and breaking techniques: design issues, challenges, and future research directions. *arXiv preprint arXiv:2307.10239*, 2023. 46 p. Disponível em: <https://arxiv.org/abs/2307.10239>. Citado 4 vezes nas páginas 18, 23, 24 e 33.

WANG, P. et al. Simple and easy: transfer learning-based attacks to text CAPTCHA. *IEEE Access*, v. 8, p. 59044–59058, 2020. DOI: 10.1109/ACCESS.2020.2982945. Citado 2 vezes nas páginas 33 e 51.

WANG, Z. et al. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, v. 13, n. 4, p. 600–612, 2004. DOI: 10.1109/TIP.2003.819861. Citado 5 vezes nas páginas 32, 49, 50, 55 e 56.

WENG, H. et al. Towards understanding the security of modern image CAPTCHAs and underground CAPTCHA-solving services. *Big Data Mining and Analytics*, v. 2, n. 2, p. 118–144, 2019. DOI: 10.26599/BDMA.2019.9020001. Citado 6 vezes nas páginas 18, 33, 34, 35, 51 e 56.

YE, G. et al. Yet another text CAPTCHA solver: a generative adversarial network based approach. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18)*. New York: ACM, 2018. p. 332–348. Toronto, Canada. DOI: 10.1145/3243734.3243754. Citado 6 vezes nas páginas 18, 24, 26, 33, 34 e 35.

Apêndices

APÊNDICE A – Resultados completos da varredura de intensidade

A Tabela 5 apresenta os resultados dos doze níveis de perturbação avaliados, além da condição sem ruído no experimento, dos quais o Capítulo 4 discute um recorte. As colunas de acurácia global têm como denominador as 10.000 imagens do conjunto de teste; as colunas condicionais, o subconjunto classificado corretamente em condições limpas. Os valores de SSIM e PSNR, calculados sobre as 10.000 imagens, referem-se à comparação entre a imagem original e suas versões perturbadas pelo FGSM e pelo ruído aleatório. A coluna McNemar refere-se a H_{0a}; o teste de H_{0b} é apresentado na Tabela 4.

Tabela 5 – Resultados completos por nível de intensidade da perturbação

	Acurácia global (%)		**Condicional (%)**		**FGSM**		**Aleatório**		
ε	FGSM	Aleatório	Acurácia	F1	SSIM	PSNR (dB)	SSIM	PSNR (dB)	McNemar (H_{0a})
0,00	92,59	92,59	100,00	100,00	1,0000	—	1,0000	—	—
0,05	23,48	91,53	25,35	25,93	0,9897	40,01	0,9898	40,01	$< 10^{-300}$
0,10	16,78	88,21	18,04	18,60	0,9632	34,00	0,9634	34,00	$< 10^{-300}$
0,15	15,20	80,92	16,16	16,52	0,9278	30,50	0,9283	30,50	$< 10^{-300}$
0,20	13,93	70,79	14,70	14,24	0,8889	28,01	0,8895	28,01	$< 10^{-300}$
0,25	12,46	58,10	12,98	11,44	0,8489	26,09	0,8497	26,09	$< 10^{-300}$
0,30	11,86	46,33	12,32	9,95	0,8092	24,52	0,8101	24,52	$< 10^{-300}$
0,35	11,26	36,07	11,60	8,59	0,7705	23,20	0,7715	23,20	$< 10^{-300}$
0,40	10,68	28,78	10,99	7,33	0,7332	22,06	0,7343	22,05	$< 10^{-300}$
0,45	10,38	23,29	10,61	6,36	0,6974	21,05	0,6986	21,05	$< 10^{-300}$
0,50	10,07	20,02	10,27	5,63	0,6633	20,15	0,6645	20,15	$< 10^{-300}$
0,55	9,92	17,78	10,04	5,13	0,6309	19,34	0,6322	19,34	$< 10^{-300}$
0,60	9,90	16,16	9,95	4,74	0,6001	18,61	0,6014	18,61	$< 10^{-300}$

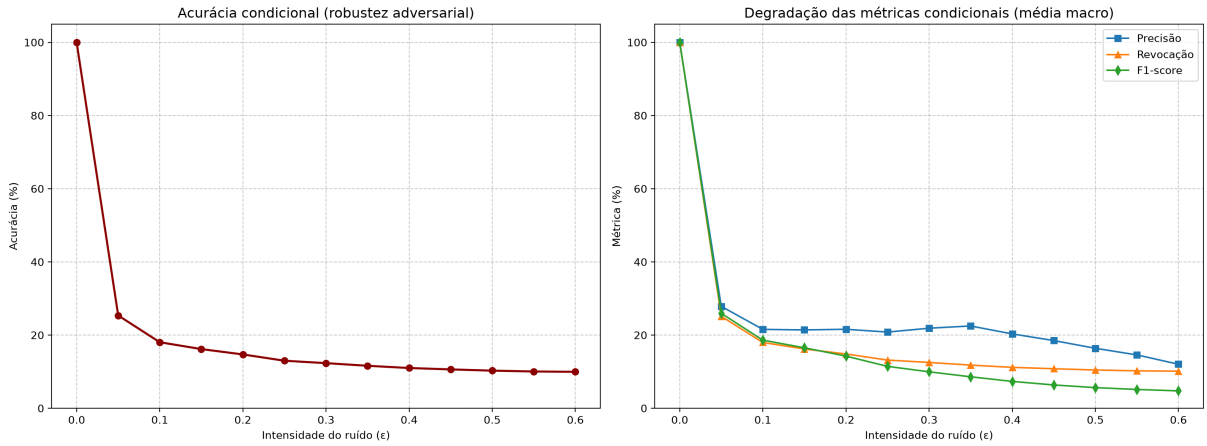
Fonte: Autor (2026)

Os dados brutos que originaram esta tabela, bem como o código que os produziu, estão disponíveis no repositório indicado na Seção 3.4.

APÊNDICE B – Degradação das métricas condicionais

O Gráfico 4 detalha o comportamento das métricas condicionais ao longo de toda a varredura. À esquerda, a acurácia sobre o subconjunto originalmente classificado de forma correta; à direita, precisão, revocação e *F1-score* em média macro. A leitura conjunta evidencia que a precisão se degrada mais lentamente que a revocação, comportamento esperado quando os erros se concentram em poucas classes-atratoras, conforme discutido na Seção 4.5.

Gráfico 4 – Degradação das métricas condicionais por nível de perturbação



Fonte: Autor (2026)