Faculdade de Tecnologia de Americana "Ministro Ralph Biasi"

Curso Superior de Tecnologia em Segurança da Informação

Brayan Yuki Cazelle

Bruno Vêneri Ferreira

DETECÇÃO DE DEEPFAKES EM REDES SOCIAIS POR INTELIGÊNCIA ARTIFICIAL:

ANÁLISE COMPARATIVA DE TÉCNICAS E IMPACTO DA COMPRESSÃO

Americana, SP

2026

Brayan Yuki Cazelle
Bruno Vêneri Ferreira

DETECÇÃO DE DEEPFAKES EM REDES SOCIAIS POR INTELIGENCIA ARTIFIAL: ANÁLISE COMPARATIVA DE TÉCNICAS E IMPACTOS DA COMPRESSÃO

Trabalho de graduação apresentado como exigência parcial para obtenção do título de Tecnólogo em Segurança da Informação pelo Centro Paula Souza – Faculdade de Tecnologia de Americana – Ministro Ralph Biasi.
Área de concentração: Segurança da Informação

Americana, 10 de Junho de 2026

Banca Examinadora:

Maria Cristina Aranda
Doutora
Fatec Americana – Ministro Ralph Biasi

Selma Guedes Cavalcante Marcante
Especialista
Fatec Americana – Ministro Ralph Biasi

Maxwel Vitorino da Silva
Mestre
Fatec Americana – Ministro Ralph Biasi

DETECÇÃO DE DEEPFAKES EM REDES SOCIAIS POR INTELIGÊNCIA ARTIFICIAL: ANÁLISE COMPARATIVA DE TÉCNICAS E IMPACTO DA COMPRESSÃO

Brayan Yuki Cazelle, FATEC Americana Ministro Ralph Biasi,
brayan.cazelle@aluno.cps.sp.gov.br

Bruno Vêneri Ferreira, FATEC Americana Ministro Ralph Biasi,
bruno.ferreira103@aluno.cps.sp.gov.br

Orientadora: Prof.ª Maria Cristina Aranda, FATEC Americana Ministro Ralph Biasi,
maria.aranda@cps.sp.gov.br

Resumo

O avanço dos modelos de inteligência artificial generativa viabilizou a produção de conteúdos sintéticos altamente realistas, conhecidos como *deepfakes*, que se disseminam rapidamente em plataformas como Instagram, TikTok, YouTube e X. Este artigo analisa as principais técnicas baseadas em Inteligência Artificial (IA) para detecção de *deepfakes* em redes sociais, avaliando eficácia, limitações e vulnerabilidades em condições reais de uso. A pesquisa adota abordagem de revisão bibliográfica comparativa de natureza qualitativa, com análise quantitativa indireta baseada em resultados reportados na literatura científica indexada em IEEE Xplore, ACM Digital Library e Google Scholar (2018–2024). Foram analisadas quatro abordagens: Redes Neurais Convolucionais (CNNs), Transformadores de Visão (ViT), *rPPG* e análise espectral, avaliadas por métricas como Acurácia, Precisão, *Recall*, F1-*Score* e AUC-ROC sobre os conjuntos de dados FaceForensics++, DFDC e Celeb-DF. Os resultados mostram que nenhuma técnica isolada supera as demais em todos os cenários avaliados, sendo a compressão de vídeo aplicada pelas plataformas o principal fator de degradação dos detectores. Como complemento à revisão bibliográfica, foi conduzido um experimento de verificação empírica que quantificou a degradação de sinais visuais sob compressão progressiva, corroborando os achados da literatura. A principal contribuição deste trabalho é consolidar comparativamente técnicas recentes considerando especificamente o impacto da compressão e do ambiente de redes sociais sobre o desempenho dos modelos. Conclui-se que abordagens híbridas e multimodais com generalização entre domínios representam o caminho mais promissor para a detecção eficaz de *deepfakes* em ambientes digitais públicos.

Palavras-chave: *deepfakes*; detecção por inteligência artificial; redes sociais; segurança da informação; aprendizado profundo.

Abstract

The advancement of generative artificial intelligence models has enabled the production of highly realistic synthetic content, known as *deepfakes*, which spread

rapidly across platforms such as Instagram, TikTok, YouTube, and X. This article analyzes the main AI-based techniques for *deepfake* detection on social media, focusing on effectiveness, limitations, and vulnerabilities under real-world conditions. The research adopts a comparative bibliographic review of qualitative nature, supported by indirect quantitative analysis based on results reported in the scientific literature indexed in IEEE Xplore, ACM Digital Library, and Google Scholar (2018–2024). Four main approaches were analyzed: Convolutional Neural Networks (CNNs), Vision Transformers (ViT), *rPPG*, and spectral analysis, evaluated through Accuracy, Precision, Recall, F1-Score, and AUC-ROC metrics on the FaceForensics++, DFDC, and Celeb-DF datasets. Findings indicate that no single technique outperforms the others across all evaluated scenarios, with video compression applied by social media platforms being the primary factor degrading detector performance. As a complement to the literature review, an empirical verification experiment was conducted to quantify the degradation of visual signals under progressive compression, corroborating the findings reported in prior work. The main contribution of this work is a comparative consolidation of recent detection techniques, with specific focus on how compression and the social media environment affect model performance. It is concluded that hybrid and multimodal approaches combined with cross-domain generalization strategies represent the most promising path for effective *deepfake* detection in public digital environments.

Keywords: *deepfakes*; artificial intelligence detection; social media; information security; deep learning.

1 INTRODUÇÃO

A expansão acelerada das redes sociais transformou a forma como informações visuais e audiovisuais são produzidas, consumidas e compartilhadas em escala global. Nesse mesmo movimento, o avanço dos modelos de Inteligência Artificial (IA) generativa, em especial as Redes Adversariais Generativas (GANs, do inglês *Generative Adversarial Networks*), propostas por Goodfellow *et al.* (2014), e os modelos de difusão, viabilizou a criação de conteúdos sintéticos altamente realistas, conhecidos como *deepfakes* (Rössler *et al.*, 2019). Tais conteúdos simulam rostos, vozes e expressões humanas com elevado grau de verossimilhança, dificultando a distinção entre material legítimo e manipulado.

No ambiente das redes sociais, a disseminação de *deepfakes* ocorre de forma rápida e massiva, potencializando riscos associados à desinformação, manipulação de opinião pública, fraudes digitais e ataques à reputação de figuras públicas. Plataformas como Instagram, Facebook, TikTok, X (antigo Twitter) e YouTube são vetores frequentes desse tipo de conteúdo, especialmente quando direcionado a instituições públicas, agentes governamentais e processos eleitorais (Dolhansky *et al.*, 2020). O impacto econômico do fenômeno é igualmente expressivo: segundo projeções do Deloitte Center for Financial Services (2024), fraudes impulsionadas por

IA generativa devem escalar de US$ 12,3 bilhões em 2023 para US$ 40 bilhões nos Estados Unidos até 2027, evidenciando a urgência de respostas técnicas estruturadas.

Pesquisas recentes apontam avanços em técnicas de detecção automática por meio de Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*), Transformadores de Visão (ViT, do inglês *Vision Transformers*), análise de sinais fisiológicos remotos (*rPPG*, do inglês *remote Photoplethysmography*) e abordagens baseadas no domínio da frequência (Verdoliva, 2020; Haliassos *et al.*, 2022). Persistem, porém, desafios relevantes quanto à generalização dos modelos diante de diferentes conjuntos de dados e à eficácia em conteúdos submetidos a algoritmos de compressão típicos das plataformas sociais, que degradam artefatos visuais essenciais para a identificação de manipulações (Sun *et al.*, 2023).

Diante desse cenário, este trabalho apresenta como problema de pesquisa: as técnicas baseadas em IA podem ser aplicadas de forma eficaz para a detecção de *deepfakes* em conteúdos veiculados em redes sociais de acesso público?

O objetivo geral consiste em analisar as técnicas de IA aplicadas à detecção de *deepfakes* no contexto das redes sociais, avaliando sua eficácia, limitações e vulnerabilidades por meio de revisão bibliográfica qualitativa. Como objetivos específicos, busca-se: (i) compreender a relação entre segurança da informação e inteligência artificial no ambiente das redes sociais; (ii) analisar os principais métodos de criação de *deepfakes* e identificar vulnerabilidades em conteúdos de interesse público disseminados em plataformas sociais; (iii) comparar o desempenho das principais técnicas de detecção baseadas em IA segundo métricas padronizadas como Acurácia, Precisão, *Recall*, F1-*Score* e AUC-ROC (Área sob a Curva ROC, do inglês *Area Under the Receiver Operating Characteristic Curve*); e (iv) identificar os principais conjuntos de dados de referência utilizados na avaliação dessas técnicas, como FaceForensics++, DFDC e Celeb-DF.

A relevância do estudo fundamenta-se no crescimento do uso de *deepfakes* em campanhas de desinformação, fraudes digitais e manipulação política, especialmente em ambientes públicos como redes sociais (Li *et al.*, 2020). Ferramentas capazes de gerar rostos e vozes sintéticas estão hoje disponíveis gratuitamente, o que abaixou significativamente a barreira técnica para produção de conteúdo manipulado com fins

maliciosos. Esse cenário afeta instituições públicas, processos democráticos e a confiança coletiva nas informações digitais. O desenvolvimento de técnicas baseadas em IA para detecção de *deepfakes* tornou-se, assim, requisito da cibersegurança moderna e da proteção da integridade informacional (Brundage *et al.*, 2018).

Para tanto, são investigadas diferentes abordagens propostas na literatura recente, considerando arquitetura dos modelos, métricas de avaliação, conjuntos de dados de referência e desempenho em cenários de compressão de vídeo, com vistas a mapear o estado da arte e identificar lacunas de aprimoramento tecnológico.

O artigo está organizado em cinco seções. A seção 2 apresenta o referencial teórico, abrangendo a relação entre Segurança da Informação e IA nas redes sociais, as modalidades de *deepfakes*, as técnicas de detecção baseadas em IA, os conjuntos de dados de referência e as métricas de avaliação. A seção 3 descreve os materiais e métodos adotados, incluindo um estudo empírico complementar de verificação da degradação por compressão. A seção 4 discute os resultados com análise comparativa das técnicas, os desafios técnicos persistentes e os dados obtidos no estudo empírico. A seção 5 reúne as considerações finais e os direcionamentos para trabalhos futuros.

2 REFERENCIAL TEÓRICO

Para compreender como as técnicas de detecção de *deepfakes* baseadas em IA operam, que riscos representam e quais são seus limites no ambiente das redes sociais, esta seção percorre os conceitos de base necessários: a relação entre Segurança da Informação e IA, as modalidades de criação de *deepfakes*, as técnicas de detecção por IA, os conjuntos de dados de referência e as métricas de avaliação empregadas na literatura.

2.1 Segurança da Informação, integridade digital e Inteligência Artificial nas redes sociais

A Segurança da Informação tem como propósito garantir os pilares de confidencialidade, integridade e disponibilidade dos dados, conforme Whitman e Mattord (2018). No contexto das redes sociais, esses princípios enfrentam desafios adicionais: o volume massivo de conteúdo compartilhado e a baixa rastreabilidade das fontes tornam a verificação de autenticidade uma tarefa técnica e socialmente

complexa. A integridade informacional, em particular, é diretamente ameaçada pela circulação de mídias sintéticas, que corrompem a confiabilidade do conteúdo sem que o receptor disponha de mecanismos imediatos de verificação.

Schneier (2015) aponta que a forte interconectividade dos sistemas modernos amplia a superfície de ataque, especialmente quando plataformas sociais atuam como intermediárias da informação pública. Nesse ambiente, a IA assume um papel dual: tanto instrumentaliza a proteção contra ameaças quanto pode ser explorada como vetor de risco. Anderson e Moore (2006) demonstram que sistemas baseados em IA automatizam a detecção de padrões anômalos e reduzem o tempo de resposta a incidentes; Russell e Norvig (2021), por outro lado, ressaltam que esses sistemas se tornam vulneráveis quando treinados com dados enviesados ou maliciosamente manipulados, o que expõe uma fragilidade estrutural de qualquer modelo de aprendizado de máquina dependente da qualidade dos dados de treinamento.

No contexto específico das redes sociais, a IA ocupa posição central na curadoria algorítmica de conteúdo, o que pode ampliar a visibilidade de *deepfakes* por meio de mecanismos de recomendação automática. Brundage *et al.* (2018) alertam que o uso malicioso de IA permite escalar operações de desinformação de forma automatizada. Os algoritmos de recomendação, ao priorizarem conteúdo com alto engajamento, tendem a amplificar materiais sensacionalistas e emocionalmente provocativos, características que *deepfakes* de cunho político ou difamatório frequentemente apresentam.

Esse fenômeno conecta-se ao que Zuboff (2019) denomina *capitalismo de vigilância*: um modelo econômico no qual as experiências e os comportamentos dos usuários são coletados em larga escala e convertidos em dados comportamentais comercializáveis. Nesse modelo, a lógica das plataformas sociais não é neutra; ela prioriza o engajamento como métrica central, criando um ambiente estruturalmente favorável à propagação de conteúdos que gerem reação imediata, incluindo *deepfakes* criados com fins manipulativos. O problema da desinformação por mídia sintética, portanto, não é exclusivamente técnico: possui dimensões econômicas e sistêmicas que ampliam sua escala e resistência.

Outro aspecto relevante para a Segurança da Informação é que a disseminação de *deepfakes* não depende da perfeição técnica do conteúdo. Chesney e Citron (2019)

observam que material sintético não precisa ser impecável para causar dano: basta ser convincente o suficiente para circular amplamente antes de qualquer verificação. O intervalo entre publicação e detecção torna-se, assim, um vetor de vulnerabilidade informacional que demanda soluções automatizadas capazes de operar em tempo compatível com a velocidade de propagação das redes sociais. Ataques adversariais, engenharia social e mecanismos de *trust* digital comprometido são consequências diretas desse cenário, reforçando a necessidade de integrar a detecção de *deepfakes* às estratégias mais amplas de Segurança da Informação, que incluem autenticidade digital, rastreabilidade forense e proteção da integridade informacional.

2.2 Deepfakes no ambiente das redes sociais

O conceito de *deepfakes* está diretamente associado ao avanço das Redes Adversariais Generativas (GANs, do inglês *Generative Adversarial Networks*), propostas por Goodfellow *et al.* (2014), que tornaram possível a geração de mídias sintéticas altamente realistas. A arquitetura GAN opera por meio de dois componentes em competição contínua: uma rede geradora, responsável por produzir conteúdo sintético, e uma rede discriminadora, que tenta distinguir o material fabricado do autêntico. Esse processo iterativo força a rede geradora a aprimorar progressivamente a qualidade do conteúdo produzido, resultando em mídias cada vez mais difíceis de identificar. Nas redes sociais, essas tecnologias passaram a ser exploradas para criar vídeos falsos de figuras públicas, líderes políticos e representantes governamentais.

A literatura técnica identifica ao menos quatro modalidades principais de *deepfakes* visuais (Verdoliva, 2020). A primeira, e mais difundida, é a troca de rosto (*face swap*), na qual o rosto de uma pessoa é substituído pelo de outra, preservando as expressões faciais originais. A segunda é a sincronização labial (*lip sync*), que modifica apenas a região da boca para compatibilizar os movimentos labiais com um áudio distinto, frequentemente usada para fabricar declarações falsas atribuídas a autoridades. A terceira modalidade, conhecida como *puppet master* ou reencenação facial, transfere expressões, movimentos oculares e gestos de uma pessoa ao rosto de outra. A quarta é a síntese facial completa, que gera rostos inteiramente fictícios, sem correspondência com nenhum indivíduo real. Cada uma dessas técnicas produz artefatos visuais distintos, o que implica abordagens de detecção diferentes, tornando a tarefa dos detectores consideravelmente mais complexa.

O risco não é apenas teórico. Um relatório da empresa de cibersegurança Recorded Future identificou 82 casos documentados de *deepfakes* envolvendo figuras públicas em 38 países entre julho de 2023 e julho de 2024, com proporção expressiva associada a períodos eleitorais. A ameaça deixou de ser pontual e passou a operar em escala global, exigindo respostas técnicas e regulatórias igualmente sistemáticas (Recorded Future, 2024).

A dimensão do problema em plataformas de vídeo curto é expressiva: o TikTok rotulou mais de 1,3 bilhão de vídeos gerados por IA somente em 2024, sinalizando a escala com que conteúdos sintéticos circulam na plataforma (TikTok, 2024). Dados de 2025 apontam que Meta e TikTok figuram entre as plataformas com maior incidência de deepfakes, responsáveis por 49% e 47% dos casos documentados, respectivamente (Programs.com, 2026). Os processos de compressão de vídeo aplicados automaticamente pelas plataformas, como os algoritmos H.264 e H.265 utilizados por YouTube, TikTok e Instagram, reduzem a eficácia das técnicas de verificação, pois eliminam parte dos artefatos introduzidos durante a síntese, tornando o conteúdo manipulado ainda mais difícil de identificar (Verdoliva, 2020). O problema se agrava quando vídeos são repostados repetidamente, acumulando múltiplos ciclos de recompressão. Sun *et al.* (2023) apontam que esse fenômeno compromete seriamente a capacidade de generalização dos modelos de detecção, já que a maioria dos *datasets* de treinamento disponíveis é composta por vídeos de alta qualidade, sem as degradações características do ambiente real das redes sociais.

O ambiente das redes sociais impõe, assim, aos detectores de *deepfakes* restrições inexistentes em contextos laboratoriais: resolução reduzida, compressão agressiva, diversidade de dispositivos de captura e ausência de metadados confiáveis. O desempenho reportado em condições controladas tende a ser superestimado em relação ao desempenho real em produção, o que reforça a necessidade de *benchmarks* específicos para o contexto de plataformas sociais, como o *dataset* Celeb-DF (Li *et al.*, 2020) e o Desafio de Detecção de *Deepfakes* (DFDC, do inglês *DeepFake Detection Challenge*), organizado por Dolhansky *et al.* (2020), ambos detalhados na seção 2.4.

2.3 Técnicas de detecção de deepfakes baseadas em Inteligência Artificial

A detecção automática de *deepfakes* por meio de IA constitui um campo em rápida evolução. Verdoliva (2020) organiza as abordagens em dois grandes grupos: métodos convencionais, baseados em características extraídas manualmente, e métodos baseados em aprendizado profundo (*deep learning*), que aprendem representações diretamente dos dados. As subseções a seguir descrevem as quatro técnicas centrais discutidas na literatura recente, com ênfase em suas implicações para o contexto das redes sociais.

2.3.1 Redes Neurais Convolucionais (CNNs)

As Redes Neurais Convolucionais (CNNs, do inglês *Convolutional Neural Networks*) representam a abordagem mais consolidada para detecção de *deepfakes*. Sua arquitetura é composta por camadas de convolução que aplicam filtros sobre a imagem de entrada, extraindo progressivamente características de baixo nível, como bordas e texturas, até representações mais abstratas, como padrões faciais e inconsistências de iluminação. Camadas densas ao final do processo realizam a classificação binária entre conteúdo real e manipulado (Verdoliva, 2020).

A arquitetura XceptionNet tornou-se referência a partir do trabalho de Rössler *et al.* (2019), que a utilizou para treinar um detector sobre o *dataset* FaceForensics++, obtendo resultados expressivos em vídeos de alta qualidade. A queda de desempenho sob compressão, porém, é significativa: o modelo pode superar 95% de acurácia em vídeos sem compressão, mas esse valor recua para cerca de 80% quando os vídeos passam pelos algoritmos de codificação típicos das plataformas sociais, o que configura uma limitação estrutural para o ambiente real de uso. Outros modelos, como EfficientNet, DenseNet e ResNet, também foram amplamente explorados, cada um apresentando compromissos distintos entre custo computacional, acurácia e capacidade de generalização.

2.3.2 Transformadores de Visão (ViT)

Os Transformadores de Visão (ViT, do inglês *Vision Transformers*) representam uma evolução mais recente, originalmente desenvolvida para tarefas gerais de visão computacional e posteriormente adaptada para detecção de *deepfakes*. Diferentemente das CNNs, que processam a imagem por filtros convolucionais locais, o ViT divide a imagem em pequenos blocos (*patches*) e aplica mecanismos de autoatenção (*self-attention*) para modelar relações de longo alcance entre diferentes

regiões da face. Essa característica permite identificar inconsistências espaciais sutis que as CNNs, por sua natureza local, tendem a não capturar com a mesma eficácia.

Para a detecção de *deepfakes* em vídeo, modelos como o ISTVT (*Interpretable Spatial-Temporal Video Transformer*) combinam atenção espacial e temporal em um único *framework*, identificando tanto artefatos visuais por quadro quanto inconsistências de movimento entre quadros consecutivos. Arquiteturas baseadas em ViT tendem a generalizar melhor entre diferentes *datasets* do que CNNs, pois capturam padrões globais em vez de artefatos específicos de um conjunto de dados particular (Haliassos *et al.*, 2022). Esse ganho, contudo, vem acompanhado de maior custo computacional, o que ainda limita aplicações em tempo real em larga escala.

2.3.3 Análise de Sinais Fisiológicos Remotos (rPPG)

A Fotopletismografia Remota (*rPPG*, do inglês *remote Photoplethysmography*) constitui uma abordagem biologicamente fundamentada para a detecção de *deepfakes*. Em um vídeo de um rosto humano real, variações sutis de cor da pele causadas pela circulação sanguínea produzem um sinal periódico correspondente ao batimento cardíaco. Como os algoritmos de síntese de *deepfakes* manipulam os pixels faciais sem preservar esses padrões fisiológicos, o sinal de *rPPG* extraído de vídeos falsos tende a apresentar irregularidades ou ausência das características rítmicas esperadas (Ciftci *et al.*, 2020).

A principal vantagem dessa abordagem reside em sua fundamentação em propriedades biológicas que os geradores de *deepfakes* não replicam intencionalmente, tornando-a mais robusta a ataques adversariais direcionados a detectores baseados em artefatos visuais. A sensibilidade à compressão de vídeo, porém, é um desafio relevante: processos de codificação agressivos, como os aplicados por TikTok e Instagram, degradam as variações sutis de cor necessárias para a extração confiável do sinal fisiológico, reduzindo a precisão do método em condições reais (Verdoliva, 2020). A aplicação do *rPPG* em redes sociais, portanto, ainda exige adaptações metodológicas que permanecem em aberto na literatura.

2.3.4 Análise Espectral e no Domínio da Frequência

A análise espectral parte do princípio de que algoritmos geradores de *deepfakes*, ao sintetizarem imagens, introduzem padrões periódicos no domínio da

frequência que não estão presentes em fotografias ou vídeos naturais. Esses artefatos, invisíveis no domínio espacial, tornam-se detectáveis pela aplicação de transformações como a Transformada de Fourier ou a Transformada Wavelet sobre os quadros do vídeo. Sun *et al.* (2023) exploram essa abordagem especificamente para lidar com o problema da generalização entre domínios: como os artefatos espectrais são menos sensíveis a variações de estilo visual entre *datasets* distintos, modelos treinados nessa perspectiva tendem a manter desempenho mais estável em cenários de distribuição desconhecida (*out-of-distribution*).

No ambiente das redes sociais, a análise espectral enfrenta desafio análogo ao do *rPPG*: os algoritmos H.264 e H.265 eliminam parte das altas frequências do sinal, onde os artefatos de síntese tipicamente se concentram. Modelos treinados exclusivamente em vídeos de alta qualidade perdem capacidade discriminativa quando o conteúdo foi recomprimido pelas plataformas. Essa limitação reforça a necessidade de estratégias híbridas que combinem análise espectral com outras técnicas, além de *datasets* de treinamento que contemplem vídeos em múltiplos níveis de compressão, aspecto retomado na análise comparativa da seção de Resultados.

2.4 Conjuntos de dados de referência

A avaliação de técnicas de detecção de *deepfakes* depende diretamente da qualidade e da diversidade dos conjuntos de dados utilizados. Três *datasets* consolidaram-se como referências na literatura: FaceForensics++, o Desafio de Detecção de *Deepfakes* (DFDC) e o Celeb-DF. Cada um apresenta características distintas quanto ao volume, à qualidade dos vídeos, às técnicas de manipulação empregadas e ao grau de dificuldade que impõem aos detectores.

O FaceForensics++ (FF++), proposto por Rössler *et al.* (2019), é o *benchmark* mais amplamente utilizado na área. Composto por 1.000 vídeos originais e 4.000 manipulados por quatro métodos distintos (DeepFakes, Face2Face, FaceSwap e NeuralTextures), o FF++ totaliza mais de 1,8 milhão de imagens distribuídas em três níveis de qualidade: RAW, alta qualidade (c23) e baixa qualidade (c40). Essa variação permite avaliar o comportamento dos detectores sob diferentes graus de degradação, funcionando também como protocolo padronizado de comparação entre abordagens.

O DFDC (sigla para *DeepFake Detection Challenge*), organizado por Dolhansky *et al.* (2020) em parceria entre Facebook, Microsoft, Amazon e outros parceiros, foi

desenvolvido para estimular detectores robustos. Com mais de 100.000 vídeos, sendo aproximadamente 23.000 autênticos e mais de 100.000 manipulados, o *dataset* cobre ampla variedade de técnicas de síntese facial, condições de iluminação, etnias e ângulos de câmera. Estruturado como competição aberta com parcela dos dados reservada para avaliação cega, o DFDC garante maior confiabilidade nos resultados reportados. Outro aspecto que o diferencia do FF++ é que seus vídeos já foram submetidos a diferentes processos de compressão, tornando os resultados obtidos nele mais representativos do cenário real de redes sociais do que os obtidos exclusivamente com o FF++ em alta qualidade.

O Celeb-DF, apresentado por Li *et al.* (2020), surgiu como resposta a uma limitação recorrente nos *datasets* anteriores: a baixa qualidade visual dos vídeos falsos, que facilita a detecção e não representa adequadamente o material circulante na Internet. O *dataset* contém 5.639 vídeos *deepfake* de alta qualidade gerados a partir de 590 vídeos reais de 59 celebridades, com processo aprimorado de síntese facial que reduz artefatos visíveis como bordas irregulares e inconsistências de cor. Ao avaliar nove detectores do estado da arte sobre o Celeb-DF, os autores demonstraram que as taxas de detecção caíam expressivamente em comparação com o FF++, evidenciando que modelos treinados em conjuntos de dados mais antigos não generalizam para *deepfakes* de geração mais recente. A escolha do *dataset* de treinamento é, portanto, tão determinante quanto a arquitetura do modelo.

2.5 Métricas de avaliação

A avaliação quantitativa de modelos de detecção de *deepfakes* exige métricas capazes de capturar diferentes aspectos do desempenho, uma vez que alta acurácia geral não garante eficácia na identificação de vídeos falsos. As cinco métricas mais utilizadas na literatura são: Acurácia, Precisão, *Recall*, F1-*Score* e AUC-ROC.

A Acurácia (*Accuracy*) mede a proporção de classificações corretas sobre o total de amostras. Em conjuntos desbalanceados, pode ser enganosa: um detector que classifique tudo como real atingiria alta acurácia sem detectar sequer um *deepfake*.

A Precisão (*Precision*) indica, entre os vídeos classificados como falsos, qual proporção realmente o é. O *Recall* (também denominado sensibilidade ou taxa de verdadeiros positivos) indica, entre todos os vídeos genuinamente falsos, qual

proporção foi corretamente identificada. Essas duas métricas apresentam relação de compromisso: modelos ajustados para alta precisão tendem a ser conservadores e, por consequência, deixam escapar mais *deepfakes* reais, reduzindo o *recall*. O ponto de equilíbrio adequado depende do contexto de aplicação.

O F1-*Score* resolve essa tensão ao calcular a média harmônica entre precisão e *recall*, fornecendo um valor único que penaliza desequilíbrios entre as duas métricas. É especialmente útil no contexto de moderação automatizada de conteúdo em redes sociais, onde tanto falsos negativos (*deepfakes* não detectados) quanto falsos positivos (conteúdo legítimo removido indevidamente) geram consequências práticas relevantes.

A AUC-ROC (Área sob a Curva ROC, do inglês *Area Under the Receiver Operating Characteristic Curve*) é a métrica mais robusta para comparação entre modelos, pois avalia o desempenho do classificador em todos os limiares de decisão possíveis, e não apenas em um ponto específico. Um valor de AUC-ROC igual a 1,0 indica separação perfeita entre classes reais e falsas; um valor de 0,5 equivale à classificação aleatória. Na literatura de detecção de *deepfakes*, a AUC-ROC é amplamente adotada como principal métrica de comparação entre abordagens, por permitir análises *cross-dataset* mais justas, independentemente do limiar de classificação adotado por cada modelo.

3 MATERIAIS E MÉTODOS

A natureza do problema investigado neste artigo orienta diretamente as escolhas metodológicas. Trata-se de revisão bibliográfica comparativa de natureza qualitativa, com análise quantitativa indireta baseada em resultados reportados na literatura científica, nos termos de Gil (2002). O estudo é voltado à análise crítica e interpretativa de produções científicas existentes sobre detecção de *deepfakes* por meio de IA, sem implementação de detectores ou treinamento de modelos. Como complemento, foi realizado um estudo empírico de caráter ilustrativo descrito na seção 3.1.

O corpus bibliográfico foi construído a partir de buscas nas bases IEEE Xplore, ACM Digital Library e Google Scholar, com publicações predominantemente entre 2018 e 2024, priorizando trabalhos com avaliação quantitativa em condições de compressão de vídeo. As buscas foram realizadas com os seguintes termos e

operadores booleanos: ("deepfake detection" OR "synthetic face detection") AND ("neural network" OR "deep learning"); ("facial manipulation detection") AND ("compression" OR "H.264" OR "video codec"); ("AI-generated media forensics") AND ("social media"). Foram identificados inicialmente 147 artigos por meio das combinações de termos de busca *deepfake detection*, *facial manipulation detection*, *synthetic media forensics* e *deep learning forgery detection*. Após leitura de título e resumo (primeira triagem), foram mantidos 68 artigos; após leitura integral e aplicação dos critérios de inclusão e exclusão descritos a seguir, 31 trabalhos foram selecionados para análise aprofundada.

Os critérios de inclusão adotados foram: (i) trabalhos que descrevam ou avaliem técnicas de detecção de *deepfakes* em imagens ou vídeos faciais; (ii) estudos que reportem métricas de desempenho padronizadas, como Acurácia, Precisão, *Recall*, F1-*Score* ou AUC-ROC; e (iii) publicações que utilizem ao menos um dos *datasets* de referência FaceForensics++, DFDC ou Celeb-DF. Como critérios de exclusão, foram desconsiderados trabalhos sem avaliação quantitativa, artigos não revisados por pares e estudos restritos a modalidades não visuais de *deepfake*, como clonagem de voz exclusivamente.

A análise seguiu abordagem comparativa e interpretativa. Cada técnica identificada foi avaliada segundo quatro dimensões: (a) arquitetura do modelo e princípio de funcionamento; (b) *dataset* utilizado para treinamento e avaliação; (c) métricas reportadas e desempenho em cenários com e sem compressão; e (d) limitações identificadas pelos próprios autores quanto à generalização entre domínios distintos. Os resultados foram consolidados em tabela comparativa, de modo a facilitar a visualização das diferenças e complementaridades entre as abordagens estudadas.

A delimitação do estudo concentra-se no contexto de plataformas de redes sociais de grande alcance, especificamente Instagram, YouTube, TikTok, Facebook e X (antigo Twitter), com ênfase em conteúdos que envolvam figuras públicas, autoridades governamentais e processos democráticos. Essa delimitação justifica-se pelo fato de que os algoritmos de compressão empregados por essas plataformas introduzem degradações específicas nos vídeos que afetam de forma diferenciada o desempenho dos detectores, tornando este contexto particularmente relevante para a Segurança da Informação.

3.1 Estudo empírico complementar

Com o objetivo de verificar empiricamente o fenômeno de degradação de sinais visuais descrito na literatura, foi conduzido um experimento exploratório de compressão progressiva. Cabe destacar que a implementação de um ambiente de teste com vídeos reais submetidos à codificação H.264 e H.265 por meio de ferramentas como FFmpeg, seguida de análise por detectores baseados em modelos de aprendizado profundo, demandaria infraestrutura computacional significativa: unidades de processamento gráfico (GPUs) de alto desempenho — como NVIDIA A100, cujo custo de aluguel em serviços de nuvem situa-se entre US$ 2,00 e US$ 4,00 por hora — e acesso a APIs de modelos de visão computacional, cujo custo por volume de tokens processados inviabiliza testes em escala sem financiamento de pesquisa. Adicionalmente, a coleta de vídeos diretamente de plataformas como TikTok, que rotulou mais de 1,3 bilhão de vídeos gerados por IA somente em 2024 (TikTok, 2024) e figura entre as plataformas com maior incidência de deepfakes circulantes (Programs.com, 2026), envolveria restrições de termos de uso e ausência de ambientes de sandbox que permitissem análise controlada. Diante dessas limitações de escopo e recursos, o experimento foi delimitado ao processamento de imagens estáticas com compressão JPEG progressiva, como aproximação controlada do fenômeno real. O estudo não envolveu implementação de detectores baseados em IA nem treinamento de modelos, sendo circunscrito à mensuração do impacto da compressão sobre propriedades de imagem relevantes para a detecção de *deepfakes*.

Foram selecionadas quatro imagens para o experimento: duas imagens sintéticas geradas pelo modelo StyleGAN3, obtidas na plataforma *thispersondoesnotexist.com*, e duas fotografias reais, sendo uma imagem de rosto humano e uma imagem de paisagem. Cada imagem foi processada em quatro cenários de compressão: original sem compressão (RAW), compressão simulando YouTube (qualidade JPEG 65%), compressão simulando Instagram (qualidade JPEG 40%) e recompressão agressiva simulando repostagem no TikTok (qualidade JPEG 15%). Para cada cenário, foram calculadas as seguintes métricas: Relação Sinal-Ruído de Pico (PSNR, do inglês *Peak Signal-to-Noise Ratio*), Erro Quadrático Médio (MSE, do inglês *Mean Squared Error*), índice de artefatos de bloco 8×8, perda de densidade de bordas pelo operador Sobel e perda de ruído fino de alta frequência. Os resultados obtidos são apresentados na Tabela 2, na seção 4.3. A ferramenta utilizada

para o experimento, denominada CompressTest, foi desenvolvida pelos autores em HTML com processamento local via JavaScript e está disponível em: https://drive.google.com/drive/folders/1_Pc2oUHtCJuAwmpF0_mib9lZWsIMPOcF?usp=drive_link. A Figura 1 apresenta a interface da ferramenta com exemplo de resultado gerado durante o experimento.

4 RESULTADOS E DISCUSSÕES

A análise da literatura recente sobre detecção de *deepfakes* por meio de IA revela um panorama técnico diversificado, no qual diferentes arquiteturas apresentam vantagens e limitações complementares. Os resultados são discutidos a seguir, organizados em torno das técnicas analisadas, dos *datasets* utilizados e dos desafios que persistem para a aplicação efetiva dessas soluções no ambiente real das redes sociais.

4.1 Análise comparativa das técnicas

A Tabela 1 consolida os principais resultados reportados pelos estudos analisados, organizando cada técnica segundo a arquitetura empregada, o *dataset* de avaliação, a acurácia ou AUC-ROC aproximada e a referência correspondente. Os valores referem-se aos resultados mais representativos encontrados na literatura, considerando os cenários de maior relevância para o contexto de redes sociais.

Tabela 1: Comparativo de técnicas de detecção de *deepfakes* baseadas em IA

Técnica	**Arquitetura principal**	**Dataset de avaliação**	**Acurácia aproximada**	**Referência**
CNN (XceptionNet)	*XceptionNet*	*FaceForensics*++	95% (sem compressão) / ~80% (com compressão H.264)	Rössler *et al.* (2019)
CNN (EfficientNet-B7)	*EfficientNet-B7*	*DFDC*	~88% AUC	Dolhansky *et al.* (2020)
ViT (ISTVT)	*Vision Transformer* espaço-temporal	*FaceForensics*++ / Celeb-DF	~99% AUC (FF++)	Haliassos *et al.* (2022)
rPPG (FakeCatcher)	CNN + mapas PPG multirregião	*FaceForensics*++	~82-88% (vídeos sem compressão)	Ciftci *et al.* (2020)

Análise espectral	CNN + gargalo de estilo	*FaceForensics*++ / Celeb-DF	~91% AUC (generalização *cross-dataset*)	Sun *et al.* (2023)

Fonte: elaborado pelos autores com base em Rössler *et al.* (2019), Dolhansky *et al.* (2020), Haliassos *et al.* (2022), Ciftci *et al.* (2020) e Sun *et al.* (2023).

Os dados da Tabela 1 mostram que o desempenho dos detectores varia expressivamente conforme o *dataset* e o nível de compressão. O XceptionNet, por exemplo, ultrapassa 95% de acurácia no FaceForensics++ sem compressão, mas recua para cerca de 80% com o nível c40, que simula condições típicas de plataformas sociais (Rössler *et al.*, 2019). Essa distância entre os dois cenários coloca em xeque qualquer afirmação de que um detector está pronto para operação em produção apenas com base em resultados laboratoriais.

As arquiteturas baseadas em ViT apresentam resultados promissores em termos de generalização. O modelo ISTVT alcança AUC próxima a 99% no FaceForensics++. Manter esse nível ao mudar de *datasets* distintos é mais variável, dependendo das condições de treinamento (Haliassos *et al.*, 2022). A natureza global da autoatenção nos ViT os torna, em princípio, menos suscetíveis a artefatos locais específicos de um único *dataset*, o que confere uma vantagem metodológica em relação às CNNs para cenários de generalização.

A abordagem por *rPPG* distingue-se das demais por não depender da detecção de artefatos visuais introduzidos durante a síntese. Ao explorar a ausência de sinais fisiológicos periódicos em vídeos falsos, oferece uma perspectiva complementar que, em princípio, é mais difícil de ser contornada por ataques adversariais. Ciftci *et al.* (2020) reportam resultados satisfatórios em vídeos não comprimidos do FaceForensics++, mas a eficácia do método cai consideravelmente sob compressão H.264, dado que as variações sutis de cor da pele necessárias para a extração do sinal são atenuadas. O *rPPG* ainda depende de adaptações metodológicas que não foram resolvidas na literatura para funcionar de forma confiável em redes sociais.

A análise espectral mostra-se especialmente relevante para cenários de generalização entre domínios. Sun *et al.* (2023) demonstram que modelos treinados com estratégias baseadas em gargalo de estilo mantêm desempenho mais estável quando avaliados em *datasets* não vistos durante o treinamento, atingindo AUC próxima a 91% em avaliações *cross-dataset*. Esse resultado tem relevância particular considerando que novos métodos de geração de *deepfakes* surgem continuamente,

e detectores sem capacidade de generalização ficam rapidamente desatualizados. Essa dinâmica impõe um limite prático à validade temporal dos resultados reportados na literatura: um modelo avaliado sobre o FaceForensics++ em 2019 pode apresentar desempenho significativamente inferior diante de deepfakes gerados por modelos de difusão disponíveis em 2024 e 2025, uma vez que esses geradores mais recentes produzem artefatos distintos dos que o detector aprendeu a identificar. A vida útil efetiva de um detector, portanto, está diretamente condicionada à velocidade de evolução dos modelos generativos adversariais — o que reforça a necessidade de atualização contínua dos conjuntos de dados de treinamento e de reavaliação periódica dos detectores em produção.

4.2 Desafios técnicos persistentes

Além dos resultados quantitativos, a análise da literatura permite identificar três desafios técnicos que ainda limitam a aplicação efetiva de detectores de *deepfakes* em redes sociais: a compressão de vídeo, os ataques adversariais e a generalização entre *datasets*.

A compressão de conteúdo aplicada pelas plataformas de redes sociais é o desafio mais imediato para os detectores de deepfakes. Ao processar automaticamente o material enviado pelos usuários, essas plataformas eliminam parte dos artefatos visuais de síntese que os detectores utilizam como sinal discriminativo. Detectores treinados exclusivamente em conteúdo de alta qualidade perdem capacidade discriminativa em ambientes reais. Verdoliva (2020) aponta que esse problema exige estratégias específicas de *data augmentation* durante o treinamento, como a simulação de diferentes níveis de compressão, para que o modelo aprenda a identificar manipulações mesmo em condições degradadas.

Os ataques adversariais representam ameaça crescente: perturbações invisíveis ao olho humano podem ser adicionadas a vídeos falsos para enganar classificadores de aprendizado profundo, reduzindo drasticamente a taxa de detecção. A pesquisa sobre robustez adversarial no contexto específico de *deepfakes* ainda é incipiente, configurando uma lacuna relevante na literatura.

A generalização entre *datasets* permanece como o problema mais estrutural da área. Modelos treinados no FF++ e avaliados no Celeb-DF frequentemente apresentam queda expressiva de desempenho, evidenciando que os detectores

aprendem características específicas do processo de geração de um conjunto de dados, e não propriedades universais das manipulações. Sun *et al.* (2023) e Haliassos *et al.* (2022) avançam nessa direção, mas os resultados indicam que nenhuma abordagem isolada resolve o problema de forma satisfatória. Estratégias híbridas que combinem múltiplas técnicas, como a fusão de análise espectral com ViT, tendem a ser o caminho mais promissor para detectores robustos e generalizáveis.

4.3 Verificação empírica do impacto da compressão

O estudo exploratório descrito na seção 3.1 permitiu verificar, de forma ilustrativa e quantitativa, o fenômeno de degradação dos sinais visuais relevantes para a detecção de *deepfakes* sob compressão progressiva. A Tabela 2 consolida os resultados obtidos para as quatro imagens analisadas em cada cenário de compressão simulado.

Tabela 2 — Resultados do experimento de compressão progressiva sobre sinais visuais de detecção

Imagem	Cenário	Qualidade (%)	PSNR (dB)	MSE	Art. Bloco	Perda Borda (%)	Perda Ruído (%)	Impacto
Sintética 1 (rosto)	Original (RAW)	100	∞ (ref.)	0,00	9,41	0,00	0,00	Nenhum
	YouTube (H.264)	65	37,63	11,22	11,12	19,4	0,0	Moderado
	Instagram (H.264)	40	35,96	16,49	11,97	24,2	0,0	Moderado
	TikTok (repost)	15	32,21	39,13	14,83	29,2	4,7	Alto
Real 1 (rosto)	Original (RAW)	100	∞ (ref.)	0,00	24,48	0,00	0,00	Nenhum
	TikTok (repost)	15	27,71	110,11	29,75	41,9	8,1	Alto
Sintética 2 (rosto)	TikTok (repost)	15	30,94	52,41	18,98	31,3	3,8	Alto
Real 2 (paisagem)	Original (RAW)	100	∞ (ref.)	0,00	36,87	0,00	0,00	Nenhum
	TikTok (repost)	15	22,64	354,13	39,54	34,7	11,8	Alto

Fonte: elaborado pelos autores. Nota: para as imagens Real 1, Sintética 2 e Real 2, registraram-se apenas os cenários extremos (RAW e TikTok) por economia de espaço; os cenários intermediários apresentaram progressão proporcional aos valores observados na Sintética 1.

Os dados da Tabela 2 confirmam, de forma empírica, o fenômeno descrito na literatura: a compressão progressiva degrada de maneira consistente os sinais visuais que os detectores de *deepfakes* baseados em CNN e ViT utilizam para identificar manipulações. No cenário de repostagem no TikTok (qualidade 15%), a perda de

densidade de bordas pelo operador Sobel atingiu 29,2% nas imagens sintéticas e 41,9% na fotografia real, enquanto o PSNR recuou para 32,21 dB e 27,71 dB, respectivamente. Esses valores corroboram os resultados de Rössler *et al.* (2019), que documentaram queda de acurácia do XceptionNet de 95% para cerca de 80% sob compressão H.264, e replicam em escala controlada o fenômeno de destruição de sinais espectrais apontado por Sun *et al.* (2023) como principal limitador da generalização de detectores em ambientes de redes sociais.

Os dados revelam ainda uma assimetria relevante entre os dois tipos de imagem analisados. As imagens sintéticas geradas pelo modelo StyleGAN3 sofreram degradação consistentemente inferior à das fotografias reais sob os mesmos níveis de compressão. Enquanto a fotografia real (rosto) registrou PSNR de 27,71 dB e perda de borda de 41,9% no cenário TikTok, as imagens sintéticas atingiram PSNR entre 30,94 dB e 32,21 dB com perdas de borda entre 29,2% e 31,3%. Esse comportamento decorre da natureza dos modelos generativos: imagens produzidas por GAN já apresentam, em sua versão original, níveis mais baixos de ruído de alta frequência e bordas mais suavizadas do que fotografias reais, reduzindo o impacto relativo da compressão sobre essas propriedades. O efeito paradoxal é que, em condições de compressão agressiva, imagens sintéticas e fotografias reais tendem a convergir para perfis de degradação semelhantes, tornando a discriminação entre elas ainda mais difícil para detectores baseados em análise de artefatos visuais.

5 CONSIDERAÇÕES FINAIS

O exame do conjunto de técnicas analisadas neste trabalho conduz a uma constatação central: o problema da detecção de *deepfakes* em redes sociais é substancialmente mais complexo do que os índices de laboratório sugerem. CNNs, ViT, *rPPG* e análise espectral oferecem, cada uma a seu modo, contribuições relevantes. Nenhuma delas, porém, resolve o problema de forma completa quando confrontada com as condições reais das plataformas digitais, especialmente a compressão agressiva de vídeo e a constante evolução das técnicas de geração.

A principal contribuição deste trabalho é consolidar comparativamente técnicas recentes de detecção considerando especificamente o impacto da compressão e do ambiente de redes sociais sobre o desempenho dos modelos, dimensão frequentemente secundarizada em estudos que avaliam detectores apenas em

condições controladas. O experimento empírico conduzido na seção 4.3 reforça essa contribuição ao verificar, com dados produzidos pelos próprios autores, que a degradação de sinais visuais sob compressão progressiva é mensurável e coerente com o que a literatura prevê. Essa perspectiva aplicada permite identificar com maior clareza as lacunas entre o desempenho reportado na literatura e o desempenho esperado em produção.

Em relação ao primeiro objetivo específico, a análise revelou que a Segurança da Informação e a IA estão profundamente interligadas no contexto das redes sociais: os algoritmos que permitem curar e recomendar conteúdo automaticamente também amplificam a visibilidade de mídias sintéticas, criando uma vulnerabilidade estrutural que transcende os aspectos puramente técnicos. A lógica do *capitalismo de vigilância*, discutida por Zuboff (2019), contribui para que o engajamento seja priorizado em detrimento da integridade informacional, favorecendo a propagação de *deepfakes* de cunho manipulativo.

Quanto ao segundo objetivo, foram identificadas quatro modalidades principais de *deepfakes* visuais: troca de rosto, sincronização labial, reencenação facial e síntese completa. Cada uma produz artefatos distintos, o que implica a inexistência de uma abordagem universal de detecção capaz de cobrir todos os tipos de manipulação com eficácia uniforme. Essa heterogeneidade constitui, em si, uma vulnerabilidade relevante para o contexto de plataformas públicas, onde conteúdos de múltiplas origens e técnicas circulam simultaneamente.

No que diz respeito ao terceiro objetivo, a comparação entre técnicas mostrou que CNNs oferecem alta acurácia em condições controladas, mas perdem desempenho expressivo sob compressão. Os ViT apresentam melhor capacidade de generalização, porém com maior custo computacional. O *rPPG* é conceitualmente robusto a ataques adversariais, mas vulnerável à compressão. A análise espectral demonstra potencial para generalização entre domínios, mas também sofre os efeitos da codificação de vídeo. Nenhuma técnica isolada supera as demais em todos os critérios avaliados, o que reforça a necessidade de abordagens híbridas e multimodais.

Em relação ao quarto objetivo, os três *datasets* analisados, FaceForensics++, DFDC e Celeb-DF, demonstraram que a escolha do conjunto de treinamento é tão

determinante quanto a arquitetura do modelo. O Celeb-DF, especificamente, evidenciou que detectores treinados em *datasets* de geração mais antiga não generalizam adequadamente para *deepfakes* de qualidade mais elevada, sinalizando a necessidade de atualização contínua dos conjuntos de dados utilizados no treinamento de novos modelos.

Como limitação do presente trabalho, registra-se que a revisão bibliográfica, por sua natureza qualitativa, não permite generalizações estatísticas sobre o desempenho médio das técnicas no ambiente real de redes sociais. Os valores de acurácia e AUC-ROC reportados referem-se a condições de laboratório que nem sempre reproduzem fielmente os cenários de produção. O estudo empírico exploratório realizado na seção 4.3 operou sobre um conjunto reduzido de imagens estáticas com compressão JPEG, o que constitui uma aproximação do fenômeno real, mas não substitui testes com vídeos codificados por H.264 e H.265 em condições de upload real nas plataformas. A ampliação desse experimento encontra barreiras concretas de infraestrutura e custo: o processamento de vídeos com codificadores reais, a coleta de amostras diretamente de plataformas como TikTok — que não disponibiliza ambientes de sandbox para pesquisa acadêmica — e o uso de detectores baseados em modelos de visão computacional demandam GPUs de alto desempenho e acesso a APIs de IA com custo por volume de processamento, investimentos que extrapolam o escopo e os recursos disponíveis em um trabalho de conclusão de curso de graduação sem financiamento externo. Trabalhos futuros poderiam avançar nessa direção com o apoio de laboratórios de pesquisa ou parcerias institucionais, além de explorar arquiteturas híbridas que combinem análise espectral, ViT e sinais fisiológicos em um único *pipeline* de detecção.

A detecção de *deepfakes* obedece a uma lógica de corrida: cada avanço nos detectores motiva uma resposta dos sistemas generativos, que produzem conteúdo cada vez mais difícil de identificar. Esse ciclo impõe também um limite à validade temporal dos resultados reportados na literatura: detectores treinados sobre conjuntos de dados de 2019 podem apresentar desempenho significativamente inferior diante de deepfakes gerados por modelos de difusão mais recentes, cujos artefatos diferem dos padrões que esses detectores aprenderam a identificar. Isso significa que qualquer avaliação comparativa de técnicas — incluindo a realizada neste trabalho — deve ser interpretada como retrato de um estado da arte em determinado momento,

e não como referência permanente. Esse ciclo não tem resolução puramente técnica. A construção de um ambiente digital mais seguro depende também de respostas regulatórias, de mecanismos de autenticidade integrados às plataformas e de uma população de usuários capaz de questionar a procedência do que consome. A Segurança da Informação tem papel central nessa construção, tanto no desenvolvimento de ferramentas de detecção quanto na formação de profissionais preparados para atuar nesse ambiente.

REFERÊNCIAS

ANDERSON, R.; MOORE, T. The economics of information security. **Science**, v. 314, n. 5799, p. 610–613, 2006. DOI: 10.1126/science.1130992.

BRUNDAGE, M. *et al.* The malicious use of artificial intelligence: forecasting, prevention, and mitigation. **arXiv preprint arXiv:1802.07228**, 2018. Disponível em: https://arxiv.org/abs/1802.07228.

CHESNEY, R.; CITRON, D. Deep fakes: a looming challenge for privacy, democracy, and national security. **California Law Review**, v. 107, n. 6, p. 1753–1820, 2019. DOI: 10.15779/Z38RV0D15J.

CIFTCI, U. A.; DEMIR, I.; YIN, L. FakeCatcher: detection of synthetic portrait videos using biological signals. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 43, n. 3, p. 1009–1018, 2020. DOI: 10.1109/TPAMI.2020.3009287.

DELOITTE CENTER FOR FINANCIAL SERVICES. Generative AI is expected to magnify the risk of deepfakes and other fraud in banking. **Deloitte Insights**, 2024. Disponível em: https://www2.deloitte.com/us/en/insights/industry/financial-services/financial-services-industry-predictions/2024/deepfake-banking-fraud-risk-on-the-rise.html. Acesso em: 28 maio 2026.

DOLHANSKY, B. *et al.* The DeepFake Detection Challenge (DFDC) dataset. **arXiv preprint arXiv:2006.07397**, 2020. DOI: 10.48550/arXiv.2006.07397.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

PROGRAMS.COM. The latest deepfake facts & statistics. 2026. Disponível em: https://programs.com/resources/deepfake-stats/. Acesso em: 10 jun. 2026.

GOODFELLOW, I. *et al.* Generative adversarial nets. In: **ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS**, v. 27, 2014. p. 2672–2680. Disponível em: https://arxiv.org/abs/1406.2661.

HALIASSOS, A. *et al.* Leveraging real talking faces via self-supervision for robust forgery detection. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2022, Nova Orleans. **Anais**[...]. Nova Orleans: IEEE, 2022. DOI: 10.1109/CVPR52688.2022.00490.

LI, Y. *et al.* Celeb-DF: a large-scale challenging dataset for deepfake forensics. In: IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), 2020, Seattle. **Anais**[...]. Seattle: IEEE, 2020. p. 3207–3216. DOI: 10.1109/CVPR42600.2020.00275.

RECORDED FUTURE. **Targets, objectives, and emerging tactics of political deepfakes**. Insikt Group, 2024. Disponível em: https://www.recordedfuture.com/research/targets-objectives-emerging-tactics-political-deepfakes. Acesso em: 28 maio 2026.

RÖSSLER, A. *et al.* FaceForensics++: learning to detect manipulated facial images. In: IEEE/CVF INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2019, Seul. **Anais**[...]. Seul: IEEE, 2019. DOI: 10.1109/ICCV.2019.00084.

RUSSELL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. 4. ed. Hoboken: Pearson, 2021.

SCHNEIER, B. **Data and Goliath: the hidden battles to collect your data and control your world**. New York: W. W. Norton & Company, 2015.

SUN, Y. *et al.* Domain generalization for deepfake detection via style-based information bottleneck. **IEEE Transactions on Multimedia**, v. 25, p. 8760–8772, 2023. DOI: 10.1109/TMM.2023.3238703.

TIKTOK. TikTok transparency report 2024: AI-generated content labeling. 2024. Disponível em: https://www.tiktok.com/transparency/. Acesso em: 10 jun. 2026.

VERDOLIVA, L. Media forensics and deepfakes: an overview. **IEEE Journal of Selected Topics in Signal Processing**, v. 14, n. 5, p. 910–932, 2020. DOI: 10.1109/JSTSP.2020.3009287.

WHITMAN, M. E.; MATTORD, H. J. **Principles of information security**. 6. ed. Boston: Cengage Learning, 2018.

ZUBOFF, S. **The age of surveillance capitalism: the fight for a human future at the new frontier of power**. New York: PublicAffairs, 2019.