Refas
11 anos
Revista Fatec Zona Sul
1ª. Edição Outubro 2014
ISSN 2359-182X

DETECÇÃO DE *DEEPFAKES* EM REDES SOCIAIS COM FERRAMENTAS DE INTELIGÊNCIA ARTIFICIAL

DETECTION OF *DEEPFAKES* ON SOCIAL MEDIA USING ARTIFICIAL INTELLIGENCE TOOLS

DETECCIÓN DE *DEEPFAKES* EN REDES SOCIALES CON HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL

João Emmanuel D Alkmin Neves[1]
Luiz Felipe Medeiros Pio[2]
Lucas Rodrigues de Oliveira Souza[3]
Kayque Wendell de Lima[4]

DOI: 10.26853/Refas_ISSN-2359-182X_v12n05_02

Artigo recebido em janeiro de 2026
Artigo aceito em maio de 2026

RESUMO

Este artigo investiga a eficácia e a aplicabilidade de ferramentas acessíveis para a detecção de *deepfakes* em vídeos compartilhados em redes sociais, considerando cenários de uso por usuários sem conhecimento técnico especializado. O avanço das tecnologias de inteligência artificial tem ampliado a circulação de conteúdos audiovisuais manipulados, representando desafios relevantes para a segurança da informação e a confiabilidade das comunicações digitais. A pesquisa adotou abordagem aplicada de natureza quantitativa, combinando revisão bibliográfica e experimentação prática. Para a avaliação empírica, foi construído um conjunto composto por cem vídeos, sendo cinquenta conteúdos sintéticos gerados por inteligência artificial e cinquenta vídeos autênticos. Os materiais foram analisados por meio da ferramenta *Deepfake* Video Detection, e o desempenho foi avaliado com base em métricas clássicas de classificação. Os resultados indicaram acurácia geral de 66%, evidenciando melhor desempenho na identificação de manipulações perceptíveis e limitações na diferenciação de conteúdos autênticos e *deepfakes* de alta qualidade. Conclui-se que a ferramenta apresenta potencial como mecanismo inicial de triagem de conteúdos suspeitos, embora não substitua análises especializadas em contextos que demandam maior precisão. O estudo reforça a necessidade de aprimoramento das tecnologias de detecção e de iniciativas de educação midiática para mitigar os impactos da desinformação audiovisual.

Palavras-chave: Inteligência Artificial; Segurança Digital; Redes Sociais, Detecção; *Deepfakes*.

[1] Doutor em Tecnologia pela Universidade Estadual de Campinas. Mestre em Tecnologia pela Universidade Estadual de Campinas. Tecnólogo em Análise e Desenvolvimento de Sistemas pela Fatec/Americana, com Graduação Sanduíche em Computer Science pela SUNY - State University of New York. Docente do Ensino Superior na Fatec Americana e Editor da Revista Tecnológica da Fatec Americana. E-mail: joao.neves11@fatec.sp.gov.br. Lattes: http://lattes.cnpq.br/4815333898210629. OrcId: 0000-0002-9472-9753.
[2] Graduando em Segurança da Informação. E-mail: luiz.pio@fatec.sp.gov.br. OrcId: 0009-0001-3124-6180.
[3] Graduando em Segurança da Informação. E-mail: lucas.souza250@fatec.sp.gov.br. OrcId: 0009-0003-8707-9275.
[4] Graduando em Segurança da Informação. E-mail: kayque.lima@fatec.sp.gov.br. OrcId: 0009-0007-5992-4979.

ABSTRACT

This article investigates the effectiveness and practical applicability of accessible tools for detecting *deepfake* videos shared on social media, considering scenarios involving non-technical users. The rapid advancement of artificial intelligence technologies has increased the circulation of manipulated audiovisual content, posing challenges to information security and digital trust. The study adopted an applied quantitative approach, combining a literature review with an experimental evaluation. A dataset of one hundred videos was assembled, including fifty synthetic videos generated using artificial intelligence and fifty authentic videos. The materials were analyzed using the Deepfake Video Detection tool, and performance was assessed through standard classification metrics. The results showed an overall accuracy of 66%, with better performance in detecting perceptible manipulations and limitations in distinguishing authentic content from high-quality deepfakes. The findings suggest that such tools can support the initial screening of suspicious content, although they do not replace more robust analyses in high-precision contexts. The study highlights the need for continued advances in detection technologies and media literacy initiatives to mitigate the impact of audiovisual disinformation.

Keywords: Artificial Intelligence; Digital Security; Social Networks, Detection; *Deepfakes*.

RESUMEN

Este artículo investiga la eficacia y la aplicabilidad de herramientas accesibles para la detección de *deepfakes* en videos compartidos en redes sociales, considerando escenarios de uso por parte de usuarios sin conocimientos técnicos especializados. El rápido avance de las tecnologías de inteligencia artificial ha incrementado la circulación de contenidos audiovisuales manipulados, generando desafíos relevantes para la seguridad de la información y la confiabilidad de las comunicaciones digitales. El estudio adoptó un enfoque aplicado de naturaleza cuantitativa, combinando revisión bibliográfica y evaluación experimental. Para el análisis empírico, se construyó un conjunto de cien videos, incluyendo cincuenta contenidos sintéticos generados mediante inteligencia artificial y cincuenta videos auténticos. Los materiales fueron analizados utilizando la herramienta *Deepfake* Video Detection, y el desempeño fue evaluado a partir de métricas clásicas de clasificación. Los resultados mostraron una precisión global del 66%, evidenciando mejor desempeño en la identificación de manipulaciones perceptibles y limitaciones en la diferenciación entre contenidos auténticos y *deepfakes* de alta calidad. Se concluye que la herramienta presenta potencial como mecanismo inicial de triage de contenidos sospechosos, aunque no sustituye análisis más robustos en contextos que requieren mayor precisión. El estudio destaca la necesidad de avances continuos en las tecnologías de detección y de iniciativas de alfabetización mediática para mitigar los impactos de la desinformación audiovisual.

Palabras clave: Inteligencia Artificial; Seguridad Digital; Redes Sociales; Detección, *Deepfakes*.

1 INTRODUÇÃO

O avanço acelerado das tecnologias digitais e a popularização das redes sociais transformaram profundamente a maneira como as pessoas consomem, interagem e compartilham informações. Esse novo cenário comunicacional, embora tenha ampliado o acesso e a conectividade, também trouxe desafios significativos relacionados à veracidade, rastreabilidade e confiabilidade do conteúdo digital, aspectos diretamente associados à segurança da informação em ambientes sociotécnicos contemporâneos. Nesse contexto, a manipulação digital de mídias, especialmente os chamados *deepfakes*, tem se destacado como uma das formas mais sofisticadas e preocupantes de desinformação potencializada pelo avanço recente de técnicas de aprendizado profundo (Verdoliva, 2020; Mirsky; Lee, 2021).

Os *deepfakes* consistem em vídeos, imagens ou áudios gerados por algoritmos de inteligência artificial capazes de reproduzir de forma extremamente realista a aparência, a voz e os gestos de pessoas reais (Bae; Kim; Ryu, 2024). Tais conteúdos sintéticos são frequentemente produzidos por modelos generativos, como Redes Adversariais Generativas e arquiteturas multimodais, o que amplia sua capacidade de enganar usuários e sistemas automatizados de verificação (Tolosana et al., 2020). Essa tecnologia, cada vez mais acessível, tem o potencial de comprometer a integridade das informações e abalar a confiança pública, influenciando percepções sociais e políticas. Além disso, *deepfakes* têm sido associados a riscos emergentes para a privacidade, a segurança nacional e a estabilidade democrática, conforme discutido por Chesney e Citron (2019). Estudos recentes alertam que o avanço rápido das mídias sintéticas e da inteligência artificial, aliado à disseminação em larga escala proporcionada por redes de alta velocidade e plataformas digitais, intensifica o risco de manipulação informacional e demanda novas estratégias de monitoramento e mitigação (Hendrix; Morozoff, 2022).

No campo técnico, pesquisas em forense digital e visão computacional têm investigado métodos automatizados capazes de detectar inconsistências espaciais, temporais e multimodais em conteúdos manipulados, contribuindo para o desenvolvimento de ferramentas de apoio à verificação de autenticidade (Rossler et al., 2019; Li et al., 2020; Zhou; Lim, 2021). Diante desse cenário, parte-se da hipótese de que é possível detectar *deepfakes* a partir de métodos automatizados de análise de vídeo, capazes de identificar padrões sutis de manipulação digital, ainda que com limitações relacionadas à qualidade do material analisado e à evolução contínua das técnicas de geração sintética (Verdoliva, 2020).

O objetivo deste estudo é investigar ferramentas para a identificação de *deepfakes* em vídeos compartilhados em redes sociais, considerando sua eficácia técnica e aplicabilidade prática em contextos reais de uso. Busca-se, portanto, selecionar uma ferramenta com base em critérios de acessibilidade e desempenho, avaliando empiricamente sua capacidade de contribuir para a proteção informacional em ambientes digitais.

A relevância desta investigação reside na necessidade urgente de enfrentar a crescente circulação de informações manipuladas e na importância de preservar a credibilidade dos meios digitais. Relatórios de empresas especializadas em segurança digital indicam que o conhecimento público sobre *deepfakes* ainda é limitado, o que pode ampliar a vulnerabilidade de usuários a campanhas de desinformação e ataques de engenharia social (Kaspersky, 2022). Ao explorar mecanismos tecnológicos capazes de identificar manipulações audiovisuais, este trabalho busca colaborar para o fortalecimento da segurança informacional e da confiança nas interações mediadas por tecnologia, aspectos essenciais para a manutenção da integridade das instituições e da democracia contemporânea. Além disso, a investigação contribui para a compreensão das limitações atuais das ferramentas automatizadas de detecção, oferecendo subsídios para o desenvolvimento de estratégias mais robustas de combate à desinformação digital.

Este estudo contribui para a literatura ao realizar uma avaliação empírica de uma ferramenta acessível de detecção de *deepfakes* em um cenário controlado que simula o uso real por usuários não especializados, evidenciando suas capacidades, limitações e implicações práticas para a segurança da informação em ambientes digitais.

2 REFERENCIAL TEÓRICO

O avanço das tecnologias digitais e da inteligência artificial tem transformado significativamente a forma como as informações são produzidas, disseminadas e consumidas em ambientes online. Nesse contexto, esta seção apresenta os principais conceitos relacionados ao fenômeno dos *deepfakes*, suas implicações para a segurança digital e o papel das ferramentas tecnológicas no apoio à verificação da autenticidade de conteúdos audiovisuais. Busca-se, assim, estabelecer a base teórica necessária para compreender os desafios associados à manipulação informacional nas redes sociais e as possibilidades de mitigação por meio de soluções computacionais.

2.1 *Deepfakes* e a segurança digital

A evolução da inteligência artificial tem possibilitado a criação de conteúdos artificiais altamente realistas, conhecidos como *deepfakes*. Essa técnica utiliza redes neurais profundas, especialmente arquiteturas baseadas em aprendizado generativo, como Redes Adversariais Generativas e modelos de síntese multimodal, para criar vídeos, imagens ou áudios manipulados com aparência realista (Tolosana et al., 2020; Mirsky; Lee, 2021). A acessibilidade e o baixo custo dessas ferramentas ampliaram sua disseminação em redes sociais, ambiente muito favorável para a circulação de desinformação audiovisual em larga escala e com elevado potencial de impacto social.

O uso de *deepfakes* atrai sérias preocupações para a segurança digital e a confiança pública. Sua aplicação de forma maliciosa pode afetar eleições, arruinar reputações e estimular a polarização política. Conforme apontam Bae, Kim e Ryu (2024), o impacto desses conteúdos vai além do meio técnico, ameaçando a credibilidade de instituições e a integridade da informação compartilhada em ambientes democráticos. Ademais, tais conteúdos representam riscos relevantes para a privacidade e a segurança digital, ampliando o potencial de uso malicioso em diferentes contextos, conforme discutido por Chesney e Citron (2019).

Além disso, o uso de mídias falsas e manipuladas interfere diretamente no ambiente da informação digital. Segundo Hendrix e Morozoff (2022), o avanço de tecnologias como redes 5G, realidade aumentada e mídias artificiais deve causar transformações significativas na forma como as informações são produzidas, distribuídas e interpretadas, exigindo mecanismos e estratégias eficazes de monitoramento, regulação e compreensão para preservar a integridade digital. Nesse contexto, a área de forense digital tem se consolidado como campo estratégico para o desenvolvimento de métodos computacionais capazes de identificar manipulações em conteúdo multimídia (Verdoliva, 2020).

Diante disso, diversos autores apontam para a necessidade de políticas públicas, regulamentações sobre a tecnologia, iniciativas educacionais e o desenvolvimento de ferramentas de detecção que possam mitigar os impactos da desinformação gerada por *deepfakes*. Portanto, a defesa da democracia não depende apenas de mecanismos institucionais, mas também da capacidade de reconhecer, prevenir e eliminar ameaças informacionais estimuladas pela inteligência artificial. Essa problemática reforça a importância de abordagens interdisciplinares que integrem segurança da informação, ciência de dados e políticas públicas digitais.

A circulação massiva de conteúdos manipulados, incluindo *deepfakes*, tem degradado a qualidade do debate público e intensificado a polarização social (Tonezer et al., 2024). Nesse cenário, a desinformação compromete a integridade dos processos informacionais, exigindo

respostas que integrem tecnologias de detecção, educação midiática e mecanismos regulatórios mais robustos (Alves; Salandin; Neves, 2025).

Os documentos audiovisuais gerados artificialmente vêm sendo utilizados como instrumentos para a propagação de desinformação. Segundo Soella e Maimone (2022), o crescente interesse acadêmico pela detecção de *deepfakes* reflete a necessidade de aprimorar métodos computacionais capazes de identificar manipulações com maior precisão. Nesse sentido, benchmarks experimentais amplamente utilizados na literatura, como o FaceForensics++ e o Celeb-DF, têm contribuído para o avanço de pesquisas em detecção automática de conteúdos sintéticos (Rossler et al., 2019; Li et al., 2020).

Embora a Inteligência Artificial Generativa possibilite resultados altamente realistas, o principal problema reside na desinformação que pode ser gerada por seu uso indevido. A acessibilidade dessas tecnologias permite a rápida criação de conteúdos falsos, aumentando o potencial de exploração em ataques de engenharia social e campanhas de desinformação digital. Tais riscos evidenciam a necessidade de mecanismos tecnológicos capazes de apoiar usuários na verificação da autenticidade de conteúdo multimídia.

2.2 Detecção e verificação de *deepfakes* nas redes Sociais

Devido ao seu foco em conteúdos audiovisuais, as redes sociais apresentam desafios específicos para a detecção de *deepfakes*. Esses ambientes favorecem a disseminação rápida de vídeos curtos, frequentemente submetidos a compressão, edição e aplicação de filtros, o que pode comprometer a eficácia dos métodos automatizados de detecção. Nesse contexto, a robustez e a capacidade de generalização dos modelos de detecção tornam-se fatores críticos para o desempenho em cenários reais (Verdoliva, 2020).

Os métodos de detecção de *deepfakes* evoluíram significativamente nos últimos anos. De forma geral, podem ser classificados em abordagens baseadas em aprendizado profundo, análise de artefatos visuais e detecção de inconsistências temporais ou multimodais (Tolosana et al., 2020).

Nos métodos baseados em inteligência artificial, algoritmos como redes neurais convolucionais e modelos multimodais são treinados para reconhecer padrões sutis de manipulação digital. Esses modelos analisam características como texturas faciais, reflexos oculares, inconsistências de iluminação e padrões de compressão. Apesar de apresentarem desempenho elevado em ambientes controlados, tais métodos dependem fortemente da qualidade dos dados de treinamento e podem sofrer degradação em cenários reais (Mirsky; Lee, 2021).

A análise de artefatos visuais constitui outra estratégia relevante. Essa abordagem busca identificar irregularidades em regiões específicas do vídeo, como distorções nas bordas faciais, piscadas não naturais ou desalinhamentos de pixels. Esses indícios são frequentemente explorados por sistemas de forense digital para auxiliar na identificação de conteúdo manipulados (Verdoliva, 2020).

Já os métodos baseados em inconsistências temporais e multimodais investigam discrepâncias entre áudio e vídeo, como falta de sincronização labial ou movimentos corporais incoerentes. Abordagens multimodais têm ganhado destaque por permitir a análise conjunta de diferentes fontes de informação, aumentando a confiabilidade dos sistemas de detecção (Zhou; Lim, 2021).

Além das técnicas automatizadas, iniciativas voltadas à autenticação da origem do

conteúdo digital também têm sido propostas. Métodos baseados em assinaturas digitais, metadados criptografados e tecnologias de registro distribuído, como blockchain, buscam garantir a rastreabilidade e a integridade de conteúdos desde sua criação. Essas estratégias complementam os sistemas de detecção ao atuar na prevenção da manipulação digital e no fortalecimento da confiança em plataformas online (Hendrix; Morozoff, 2022). Relatórios de empresas especializadas em segurança digital indicam que o nível de conhecimento público sobre tecnologias de manipulação audiovisual ainda é limitado. Dados divulgados pela Kaspersky (2022) apontam que uma parcela significativa da população brasileira desconhece o conceito de *deepfake*, evidenciando vulnerabilidades informacionais que podem ser exploradas em campanhas de desinformação e ataques de engenharia social. Esse cenário reforça a necessidade de iniciativas voltadas à educação midiática e ao desenvolvimento de mecanismos tecnológicos capazes de auxiliar usuários na verificação da autenticidade de conteúdos digitais.

Portanto, a combinação entre técnicas automatizadas de detecção e mecanismos de autenticação digital representa um caminho promissor para reduzir o impacto dos *deepfakes* nas redes sociais. A integração dessas abordagens pode contribuir para o desenvolvimento de soluções mais robustas de segurança informacional, capazes de mitigar riscos associados à manipulação audiovisual em ambientes digitais contemporâneos.

3 MÉTODO

Este estudo combina uma abordagem bibliográfica e experimental. Do ponto de vista metodológico, a pesquisa caracteriza-se como aplicada, de abordagem quantitativa e estratégia experimental controlada, voltada à avaliação de desempenho de ferramentas computacionais em cenários simulados, conforme princípios adotados em estudos empíricos em Segurança da Informação e Engenharia de Software Experimental. Esse delineamento metodológico está alinhado a estudos empíricos em Engenharia de Software Experimental e avaliação de sistemas computacionais, nos quais abordagens quantitativas controladas são utilizadas para mensurar desempenho e comportamento de soluções tecnológicas em cenários simulados (Wohlin et al., 2012).

Inicialmente, foi realizada uma revisão bibliográfica para fundamentar o estudo sobre *deepfakes* e desinformação audiovisual, possibilitando a construção do referencial teórico e a definição dos critérios experimentais. Na etapa experimental, analisou-se a usabilidade e a eficácia da ferramenta de detecção de *deepfakes* "*Deepfake* Video Detection", desenvolvida pela empresa Attestiv. O foco principal foi verificar se essa ferramenta apresenta desempenho satisfatório em um cenário de uso próximo ao contexto real de usuários finais, considerando indivíduos que não possuem conhecimento técnico avançado, mas que necessitam verificar a autenticidade de vídeos compartilhados em redes sociais.

3.1 Ferramentas de detecção

Nos últimos anos, diversas ferramentas baseadas em inteligência artificial têm sido propostas como resposta ao avanço dos *deepfakes*. Essas soluções adotam metodologias distintas para verificar a autenticidade de conteúdos digitais, incluindo técnicas de aprendizado profundo, análise forense de artefatos visuais e verificação de inconsistências multimodais (Tolosana et al., 2020; Verdoliva, 2020).

Entre as plataformas identificadas estão *Sentinel, Sensity, WeVerify, Hive Moderation e DuckDuckGo Deepfake Detector*. Embora apresentem abordagens variadas e diferentes níveis de maturidade tecnológica, tais ferramentas compartilham o objetivo comum de apoiar a detecção automatizada de conteúdos manipulados em ambientes digitais.

Entre as soluções analisadas, destaca-se a ferramenta *Deepfake Video Detection*, desenvolvida pela empresa *Attestiv*, que combina algoritmos de aprendizado de máquina e técnicas de análise forense digital para autenticar conteúdos audiovisuais. A escolha dessa ferramenta considerou critérios como acessibilidade, disponibilidade pública, aplicabilidade em vídeos curtos e potencial de uso por usuários não especializados.

Com o objetivo de contextualizar a ferramenta analisada em relação a outras soluções disponíveis, apresenta-se Quadro 1, que realiza uma comparação qualitativa entre diferentes ferramentas de detecção de *deepfakes*, considerando aspectos como tipo de abordagem, público-alvo e forma de acesso. Ressalta-se que essa comparação possui caráter descritivo e não experimental.

Quadro 1 - Comparação qualitativa entre ferramentas de detecção de *deepfakes*

Ferramenta	**Tipo de abordagem**	**Público-alvo**	**Forma de acesso**	**Foco principal**
Attestiv	IA + análise forense	Usuário geral	Web (upload / manual)	Verificação de autenticidade
Sentinel	IA avançada (enterprise)	Corporativo / governo	API / Plataforma	Monitoramento em larga escala
Hive Moderation	IA multimodal	Empresas / plataformas	API	Moderação automatizada de conteúdo
Sensity	Inteligência de ameaças	Empresas / segurança	Plataforma própria	Análise de campanhas de *deepfake*
WeVerify	Ferramentas jornalísticas	Jornalistas / OSINT	Web / plugins	Verificação de mídia e fact-checking

Fonte: Elaborado pelos autores (2026)

Observa-se que diferentes ferramentas apresentam propostas distintas, variando entre soluções voltadas ao público geral e plataformas corporativas com foco em monitoramento em larga escala. Nesse contexto, a ferramenta *Deepfake* Video Detection destaca-se por sua acessibilidade e simplicidade de uso, características que justificam sua seleção neste estudo para avaliação em cenários de uso não especializado, alinhando-se ao objetivo de analisar soluções aplicáveis a usuários não técnicos.

Diante desses aspectos, a ferramenta foi selecionada para a aplicação prática deste estudo, com o objetivo de avaliar empiricamente sua eficácia como mecanismo de triagem inicial na identificação de *deepfakes* em redes sociais.

3.2 Justificativa do método

A escolha por uma metodologia experimental prática justifica-se pelo objetivo de avaliar não apenas o desempenho técnico da ferramenta, mas também sua aplicabilidade em contextos reais de uso. Em ambientes digitais caracterizados pela rápida disseminação de conteúdos manipulados, é fundamental que soluções de verificação apresentem equilíbrio entre precisão, acessibilidade e facilidade de uso.

Ressalta-se que a ausência de utilização de bases públicas consolidadas decorreu de restrições relacionadas a direitos de imagem e à necessidade de controle experimental sobre as variáveis analisadas. Para garantir conformidade ética e legal, utilizaram-se imagens e vídeos licenciados da plataforma Envato Elements, combinados com a geração controlada de vídeos sintéticos por meio da ferramenta HeyGen. Essa estratégia permitiu a construção de um conjunto experimental com variação de cenários e perfis, ainda que reconhecidamente limitado em termos de representatividade em relação a bases amplamente utilizadas na literatura científica.

3.3 Materiais utilizados

Envato Elements: Fonte das imagens e dos vídeos de base. Foram selecionados imagens e vídeos sob licença de uso comercial/derivações (*royalty-free*) disponíveis na plataforma, garantindo que os direitos de imagem e de uso fossem respeitados durante o processo de criação de conteúdo sintético. A seleção no Envato priorizou variedade de cenários e perfis (homens, mulheres, grupos, imagens que simulam debates, diferentes faixas etárias e etnias) para cobrir distintos casos de uso.

HeyGen: Ferramenta de criação de avatares e geração de vídeos sintéticos a partir de imagens e *prompts* textuais. Utilizada para transformar as imagens licenciadas em vídeos nos quais o avatar realiza expressões e falas descritas por *prompts* (controle das expressões, movimentos labiais, tom de voz e fala).

Attestiv - *Deepfake Video Detection*: Ferramenta de detecção utilizada para a análise dos vídeos gerados. Os vídeos produzidos foram submetidos ao sistema da Attestiv para avaliação de autenticidade e geração de relatórios técnicos.

Ambiente: *Notebook* com Windows 10, processador Intel Core i5, 8 GB de RAM, navegador Google Chrome. Os experimentos foram conduzidos em ambiente computacional representativo de um usuário comum, utilizando notebook com sistema operacional Windows, navegador web e conexão padrão à internet, buscando simular condições realistas de uso.

3.4 Procedimentos Metodológicos

A criação e seleção dos vídeos analisados foi orientada por critérios descritos na literatura científica como indicativos frequentes de manipulação por *deepfakes*, incluindo inconsistências faciais, falhas de sincronização labial e movimentos artificiais (Rossler et al., 2019; Li et al., 2020).

Os vídeos produzidos foram submetidos à ferramenta de detecção, que gerou relatórios técnicos indicando probabilidade de manipulação. Os resultados foram posteriormente organizados em métricas de desempenho, permitindo a análise da capacidade da ferramenta em diferenciar conteúdos autênticos e sintéticos no conjunto experimental proposto.

O procedimento passo a passo adotado foi o seguinte:

1. Seleção das imagens/vídeos base (Envato Elements):

a) Foram identificadas e baixadas do Envato Elements imagens e clipes licenciados que representassem perfis variados: homens, mulheres, grupos, cenas tipo debate, diferentes idades e etnias. A seleção priorizou material que permitisse a geração de avatares fidedignos e que representasse uma gama de contextos de publicação em redes sociais;

b) Cada arquivo recebeu um identificador único (ID) e metadados registrados na planilha de amostras (categoria, descrição breve, nome do arquivo, licença);

2. Geração de vídeos sintéticos (HeyGen):

a) Para cada imagem selecionada, realizou-se o *upload* na ferramenta de criação de avatares HeyGen. Em seguida, foram elaborados *prompts* textuais padronizados descrevendo o que o avatar deveria executar (texto falado), bem como especificações de expressões faciais, movimentos labiais e tom de voz. Exemplo de *prompt*: "Falar por 10 segundos: 'Esta é uma demonstração de teste', com sorriso leve, olhar direto para a câmera, piscada natural a cada 4-6 segundos.";

b) Os *prompts* foram projetados para gerar vídeos com variações de expressão e fala de modo a simular diferentes tipos de conteúdo (mensagem curta, discurso, diálogo, debate fictício);

c) Os vídeos gerados foram baixados e nomeados conforme o ID da amostra;

3. Composição do conjunto de teste:

a) O conjunto experimental foi composto por duas classes: (a) vídeos sintéticos criados pelo HeyGen (classe *fake*) e (b) vídeos de controle autênticos, formados por clipes licenciados do Envato que não foram submetidos a manipulação adicional (classe autêntico). Essa separação permite comparar a capacidade da ferramenta Attestiv em distinguir vídeos gerados (*deepfakes* sintéticos) de vídeos genuínos (*stock videos* não manipulados);

b) Para assegurar diversidade, as amostras cobriram variações de gênero, composição (individual/grupo), cenário (debate e fala isolada) e características demográficas; o Quadro 2 apresenta o agrupamento das amostras de vídeo de acordo com o tipo de conteúdo e finalidade do teste.

Quadro 2 - Agrupamento das amostras de vídeos utilizadas no experimento

Categoria	Descrição Geral	Quantidade	Origem (Envato)	Perfis Representados	Tipo de Cena / Contexto	Duração Média (s)	Uso / Finalidade
Fala Individual Masculina	Vídeos de homens falando diretamente para a câmera (estilo vlog ou discurso curto).	39	Clipes de vídeo + imagens	Masculino / 25-45 anos / Diversas etnias	Individual / Mensagem curta	15-30	Testes de detecção facial e movimento labial
Fala Individual Feminina	Mulheres apresentando mensagens ou produtos (expressão leve e sorriso).	33	Clipes de vídeo + imagens	Feminino / 20-40 anos / Diversas etnias	Individual / Apresentação	15-30	Testes de sincronização labial
Debates / Interações em Grupo	Dois ou mais indivíduos em simulação de diálogo ou debate.	28	Clipes de vídeo + imagens	Masculino e feminino / 30-55 anos	Grupo / Interativo	10-25	Testes de coerência gestual e áudio

Fonte: Elaborado pelos autores (2025)

4. *Upload* para análise (Attestiv):
a) Cada vídeo do conjunto de teste foi submetido à ferramenta *Deepfake Video Detection* da Attestiv via *upload* de arquivo. Em cada submissão registrou-se: ID do vídeo, nome do arquivo, *prompt* / descrição da geração, data/hora do *upload*, e o relatório retornado pela ferramenta (pontuação final de suspeita, gráficos temporais e classificações indicadas);
b) Quando aplicável, realizou-se o *upload* via *link* e, em casos de instabilidade, via envio direto do arquivo.
5. Registro e tratamento dos resultados:
a) Os *outputs* da Attestiv foram consolidados em planilha, onde se registraram as previsões (*fake* / autêntico), probabilidades/pontuações e observações (por exemplo, trechos do vídeo onde o índice de suspeita foi mais alto);
b) A partir desses registros construiu-se a matriz de confusão (Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos) e calculou-se métricas de desempenho: acurácia, sensibilidade (*recall*), especificidade, taxa de falsos positivos e *F1-score*.
6. Análise de usabilidade e observações qualitativas:
a) Além das métricas automatizadas, avaliou-se a clareza dos relatórios da Attestiv, a facilidade de interpretação por usuários leigos, o tempo médio de análise por vídeo e quaisquer limitações operacionais (por exemplo, instabilidade no *upload* via *link*);
b) Foram anotadas situações em que a ferramenta indicou alto índice de suspeita em vídeos aparentemente simples e casos em que classificou erroneamente vídeos autênticos como suspeitos. Essas situações foram descritas qualitativamente na seção de Resultados e Discussão.

Ressalta-se que a análise de usabilidade possui caráter exploratório e qualitativo, não sendo baseada em instrumentos formais de avaliação de experiência do usuário, como questionários padronizados. Essa limitação deve ser considerada na interpretação dos resultados.

7. Garantia de diversidade no teste:
a) Foi construída uma tabela com amostras representativas que demonstram que a ferramenta foi avaliada em diversos cenários (homens, mulheres, grupos, cenas que simulam debates etc.). Essa tabela contém, para cada amostra, resumo do *prompt* utilizado e a categoria do cenário, garantindo que a avaliação não está restrita a um único tipo de rosto ou contexto.

Ao final desse processo, obteve-se um conjunto estruturado de dados contendo as classificações realizadas pela ferramenta, permitindo a construção de métricas quantitativas de desempenho, como acurácia, sensibilidade, especificidade e F1-score. Esses indicadores foram utilizados para avaliar a capacidade da ferramenta em distinguir conteúdos autênticos e sintéticos em um cenário controlado. Os resultados obtidos a partir dessa análise são apresentados e discutidos na seção seguinte.

4 RESULTADOS E DISCUSSÃO

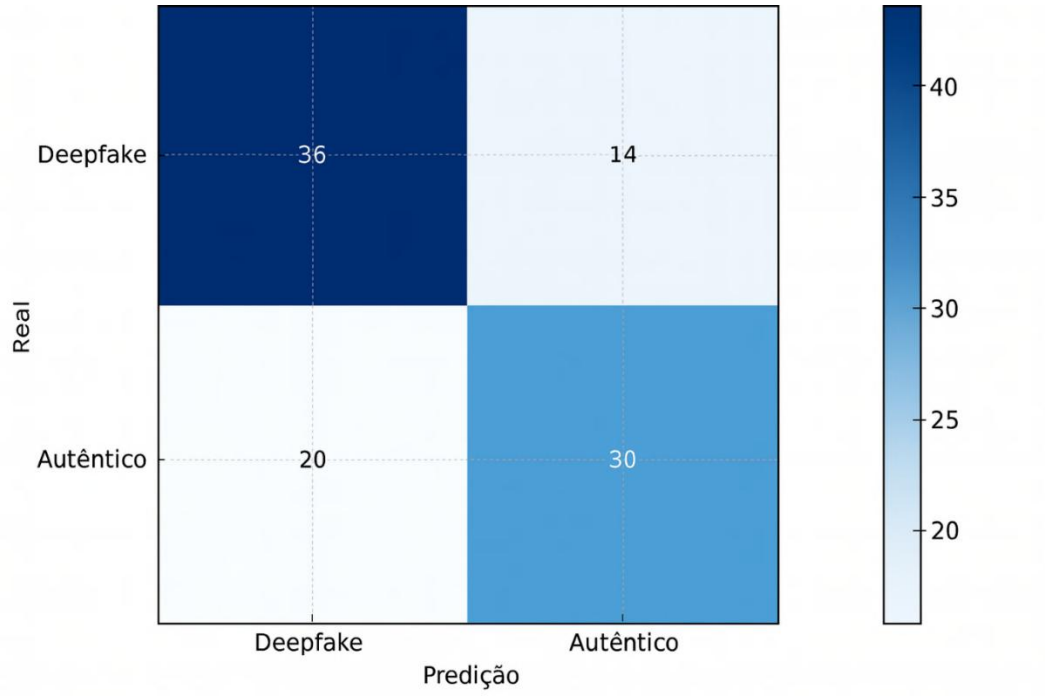
Os testes realizados com a ferramenta *Deepfake* Video Detection, desenvolvida pela Attestiv, possibilitaram uma análise prática sobre sua eficácia e aplicabilidade na detecção de conteúdos manipulados. Ao todo, foram testados cem vídeos, cinquenta classificados previamente como *deepfakes* e cinquenta autênticos, todos em resolução média HD (720p) e com foco em falsificações faciais, de voz e expressões. A definição desse conjunto experimental buscou simular o formato predominante de conteúdos audiovisuais compartilhados em redes sociais, caracterizados por curta duração e qualidade variável.

A maioria dos vídeos possuía curta duração, refletindo o formato predominante na plataforma. A ferramenta de detecção retorna para o usuário a probabilidade de o vídeo ser ou não uma *deepfake*, que, quando analisada em conjunto com a origem do vídeo, permite a definição de quatro categorias de classificação:

a) Verdadeiro Positivo (36): A ferramenta classificou corretamente o vídeo como *deepfake.*
b) Verdadeiro Negativo (30): A ferramenta classificou corretamente o vídeo como autêntico.
c) Falso Positivo (20): A ferramenta classificou o vídeo como *deepfake*, entretanto, o vídeo é autêntico.
d) Falso Negativo (14): A ferramenta classificou o vídeo como autêntico, entretanto o vídeo é *deepfake*.

Essa categorização segue métricas amplamente utilizadas na avaliação de sistemas de detecção automatizada, permitindo mensurar o desempenho do modelo em termos de acertos e erros de classificação (Verdoliva, 2020). O resultado pode ser ilustrado a partir de uma matriz de confusão, conforme exibido na Figura 1.

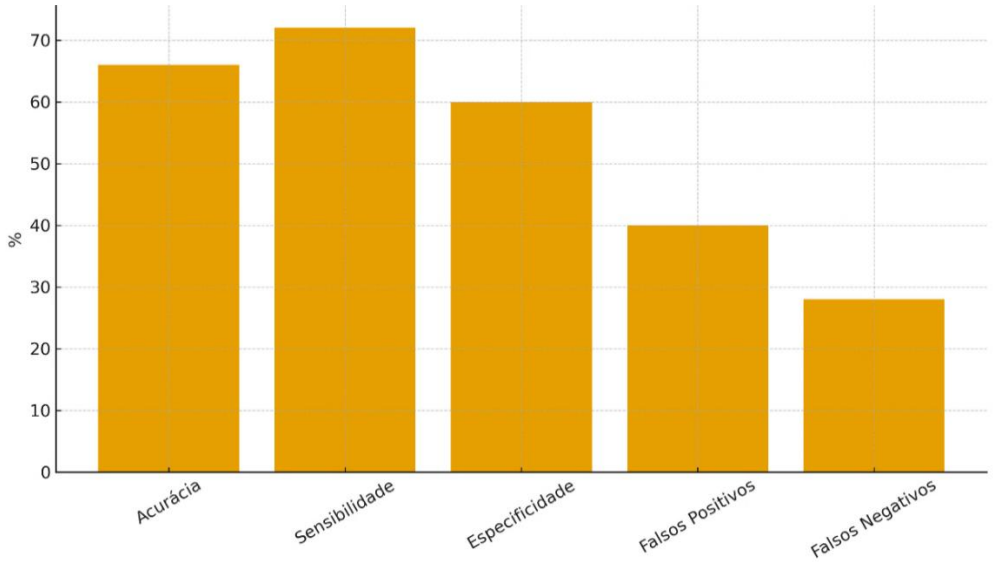
Figura 1 - Matriz de confusão obtida na análise dos vídeos pela ferramenta *Deepfake* Video Detection.



Fonte: Elaborado pelos autores (2025)

Os resultados obtidos revelaram que a ferramenta identificou corretamente 36 dos 50 vídeos falsos e 30 dos 50 vídeos autênticos, atingindo uma acurácia geral de 66%. A taxa de sensibilidade foi de 72%, demonstrando um desempenho moderado na detecção de *deepfakes*, enquanto a especificidade foi de 60%, indicando limitação na diferenciação de vídeos genuínos. Resultados dessa natureza são compatíveis com estudos que apontam redução de desempenho de detectores automatizados em cenários realistas, especialmente quando submetidos a compressão, variações de iluminação e limitações de resolução (Mirsky; Lee, 2021). Essas métricas mostram que a ferramenta é eficiente em reconhecer manipulações evidentes, mas apresenta margens de erro, especialmente em vídeos autênticos, como mostrado na Figura 2.

Figura 2 - Métricas de desempenho calculadas para a ferramenta *Deepfake Video Detection*

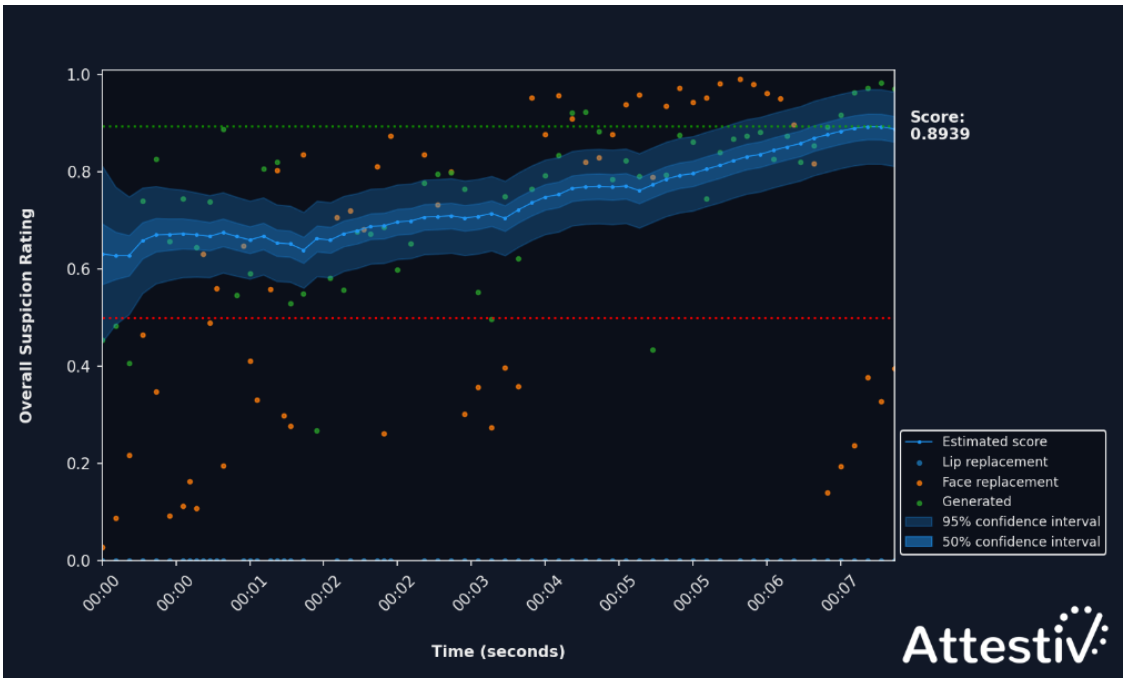


Fonte: Elaborado pelos autores (2025)

Durante a análise prática, verificou-se que a ferramenta é mais assertiva em identificar *deepfakes* com alterações perceptíveis, como falhas de sincronização labial, substituição facial e distorções em reflexos ou iluminação. Esse comportamento está em consonância com resultados obtidos em benchmarks experimentais amplamente utilizados na literatura, nos quais detectores apresentam maior eficácia diante de manipulações com artefatos visuais evidentes (Rossler et al., 2019; Li et al., 2020). O gráfico apresentado no relatório ilustra a variação do índice de suspeita ao longo do tempo do vídeo, permitindo acompanhar como a ferramenta avalia possíveis sinais de manipulação quadro a quadro.

O gráfico apresentado no relatório ilustra a variação do índice de suspeita ao longo do tempo do vídeo, permitindo identificar padrões de possível manipulação em diferentes momentos da análise. A representação inclui a estimativa central da probabilidade de *deepfake*, acompanhada por intervalos de confiança que indicam a estabilidade da detecção, além de marcações associadas a eventos específicos, como possíveis substituições faciais e inconsistências labiais. Esse conjunto de elementos permite uma análise temporal detalhada do comportamento do modelo ao longo do vídeo, conforme exemplificado na Figura 3.

Figura 3 - Exemplo de gráfico de análise de autenticidade gerado pela ferramenta *Deepfake* Video Detection, indicando o índice de suspeita ao longo do tempo.



Fonte: Attestiv (2025)

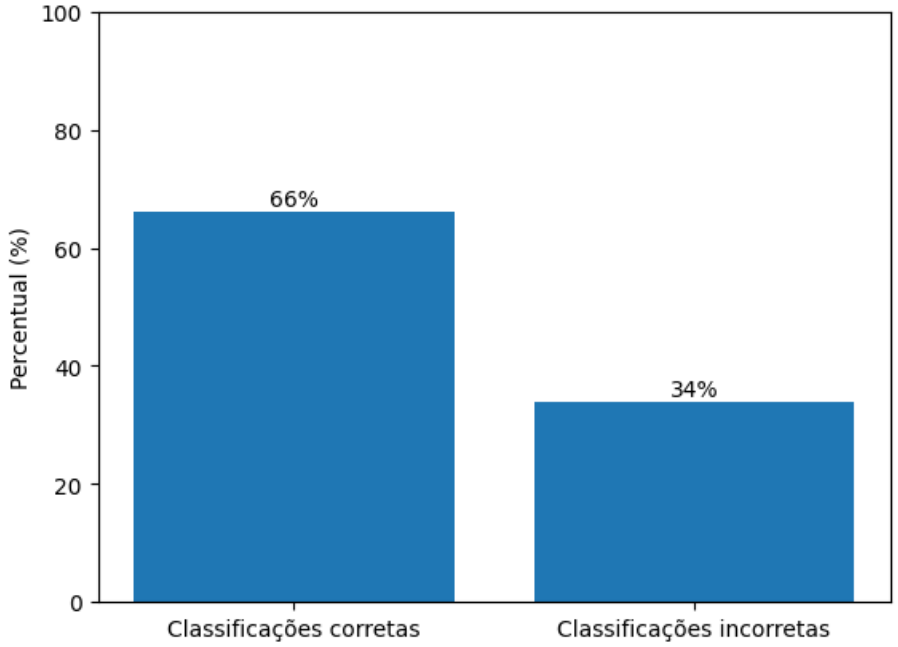
Embora esse tipo de relatório forneça informações detalhadas, observou-se que a interpretação dos resultados requer atenção e um mínimo de familiaridade com gráficos e indicadores técnicos. Esse aspecto evidencia uma limitação relacionada à usabilidade plena da ferramenta, especialmente em contextos de uso por usuários não especializados. O tempo médio de análise foi de aproximadamente cinco minutos por vídeo, um desempenho considerado adequado para uma ferramenta baseada em aprendizado de máquina operando em ambiente online. No entanto, foi constatado que a utilização de links diretos da web pode ocasionar falhas ocasionais de carregamento, exigindo, em alguns casos, o download prévio do vídeo e o upload manual do arquivo para que a verificação seja concluída. Ainda assim, o processo de uso é simples: basta arrastar o arquivo ou inserir o link, o que torna a interface acessível a usuários não técnicos.

Apesar dessa facilidade operacional, o relatório final ainda apresenta traços técnicos como terminologias e indicadores estatísticos que podem gerar dúvidas entre usuários leigos. Esse resultado reforça a necessidade de aprimoramento na comunicação dos resultados, de modo a ampliar a efetividade dessas ferramentas como apoio à verificação de autenticidade em ambientes digitais. A análise comparativa dos resultados confirma que a ferramenta adota métodos já reconhecidos na literatura científica, como a detecção de inconsistências temporais entre áudio e movimento labial, além da análise de artefatos de compressão e substituição facial. Todavia, o desempenho observado indica limitações na identificação de *deepfakes* de alta qualidade, corroborando estudos que destacam a evolução contínua das técnicas de geração sintética e os desafios associados à generalização dos detectores automatizados (Tolosana et al., 2020; Zhou; Lim, 2021).

O desempenho obtido, embora satisfatório em contexto experimental, permanece inferior às taxas superiores a 85% reportadas em modelos mais avançados, indicando limitações na detecção automatizada de *deepfakes*. Ainda assim, a ferramenta apresenta bom equilíbrio

entre usabilidade e capacidade técnica, sendo adequada para triagem inicial de conteúdos suspeitos. Em síntese, demonstra eficácia na identificação de manipulações evidentes, mas a presença de falsos positivos e falsos negativos exige cautela em sua aplicação, especialmente em contextos sensíveis. A Figura 4 apresenta a distribuição percentual das classificações corretas e incorretas, sintetizando o desempenho global da ferramenta no conjunto experimental.

Figura 4 - Distribuição percentual das classificações corretas e incorretas nos testes realizados.



Fonte: Elaborado pelos autores (2025)

A análise global dos resultados evidencia que, embora a ferramenta apresente desempenho funcional na identificação de *deepfakes* com manipulações perceptíveis, sua eficácia permanece limitada diante de conteúdos sintéticos mais sofisticados. Esse comportamento reforça achados da literatura que apontam a dificuldade de generalização de detectores automatizados em cenários reais, nos quais fatores como compressão de vídeo, variações de iluminação e qualidade reduzida influenciam negativamente a precisão dos sistemas (Verdoliva, 2020; Mirsky; Lee, 2021).

Do ponto de vista prático, a ocorrência de falsos positivos pode comprometer a credibilidade de sistemas automatizados ao classificar conteúdos autênticos como manipulados. Por outro lado, os falsos negativos representam um risco ainda mais crítico, uma vez que conteúdos falsificados podem não ser detectados, especialmente em cenários de segurança digital e combate à desinformação.

Como limitação, destaca-se que a avaliação foi realizada em um ambiente controlado, com base em um conjunto de dados sintético e na análise de uma única ferramenta de detecção. Essas condições restringem a generalização dos resultados para cenários reais e dificultam comparações mais amplas com outras soluções tecnológicas.

Apesar dessas restrições, os resultados indicam que ferramentas acessíveis podem desempenhar papel relevante como mecanismos iniciais de triagem de conteúdos suspeitos.

Nesse sentido, recomenda-se que investigações futuras incluam análises comparativas entre diferentes ferramentas e a utilização de bases de dados padronizadas, visando aumentar a robustez experimental e a validade externa dos achados.

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos neste estudo evidenciam que a ferramenta *Deepfake* Video Detection representa uma solução tecnicamente funcional para a detecção de conteúdos audiovisuais manipulados, especialmente em situações nas quais as alterações audiovisuais apresentam indícios perceptíveis. A acurácia geral de 66% indica um desempenho moderado, sugerindo potencial de aplicação como mecanismo inicial de triagem, embora ainda insuficiente para contextos que exigem elevado nível de confiabilidade, como investigações periciais, processos judiciais ou verificação jornalística especializada.

O estudo demonstrou que a ferramenta é capaz de identificar padrões recorrentes associados à manipulação digital, incluindo inconsistências faciais, substituição de voz e falhas de sincronização entre áudio e imagem. Esse resultado reforça o potencial das abordagens automatizadas baseadas em inteligência artificial como apoio à verificação de autenticidade de conteúdos audiovisuais, sobretudo em ambientes digitais caracterizados por grande volume e velocidade de disseminação de informações. Entretanto, observou-se que a interpretação dos relatórios gerados pela ferramenta requer atenção e familiaridade mínima com indicadores técnicos, o que pode representar uma barreira para usuários sem conhecimento especializado.

Entre as limitações identificadas, destaca-se a ocorrência de falsos positivos e falsos negativos, evidenciando que o uso exclusivo de sistemas automatizados pode resultar em classificações equivocadas. Além disso, a análise foi conduzida com base em um conjunto experimental controlado e em uma única ferramenta de detecção, o que restringe a generalização dos resultados para outros cenários e soluções tecnológicas. Também foram observadas instabilidades operacionais pontuais relacionadas ao envio de vídeos por meio de links diretos, o que pode impactar a experiência de uso em situações reais.

Em um cenário marcado pela crescente circulação de conteúdos manipulados e pelo avanço contínuo das técnicas de geração sintética, estudos dessa natureza contribuem para ampliar a compreensão sobre as capacidades e limitações das ferramentas de detecção de *deepfakes*. Os achados indicam que soluções acessíveis ao público podem desempenhar papel relevante como instrumentos de apoio à verificação inicial de conteúdos suspeitos, embora devam ser utilizadas de forma complementar a análises humanas e a estratégias mais robustas de verificação digital.

Como perspectivas futuras, recomenda-se a realização de investigações comparativas envolvendo múltiplas ferramentas de detecção, a utilização de bases de dados mais amplas e representativas e a análise do impacto de fatores como compressão de vídeo, qualidade de imagem e diversidade de cenários sobre o desempenho dos sistemas. Tais avanços são fundamentais para o desenvolvimento de soluções mais confiáveis e para o fortalecimento da segurança informacional em ambientes digitais contemporâneos.

6 REFERÊNCIAS

ALVES, J.; SALANDIN, A. T.; NEVES, J. E. D. Clonagem de Voz por IA: Avaliando a Eficácia em Ataques de Phishing por Voz. Revista Tecnológica da Fatec de Americana, v. 13, n. 01, 2025. Disponível em: https://fatec.edu.br/revista/index.php/RTecFatecAM/article/view/412. Acesso em: 01 nov. 2025.

BAE, J.; KIM, S.; RYU, Y. Understanding the impact of *deepfake* technology on social media platforms. Journal of Digital Ethics, 2024. Disponível em: https://ieeexplore.ieee.org/document/10552098. Acesso em: 7 nov. 2025.

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. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3213954. Acesso em: 23 mar. 2026.

HENDRIX, J.; MOROZOFF, D. Media forensics in the age of disinformation. In: CHANDRASEKARAN, S. (Org.). Multimedia forensics. Singapore: Springer Singapore, 2022.

KASPERSKY. Brasileiros desconhecem *deepfakes*. Kaspersky Blog, 2022. Disponível em: https://www.kaspersky.com.br/blog/brasileiros-desconhecem-*deepfake*/18834. Acesso em: 7 nov. 2025.

LI, Y.; YANG, X.; SUN, P.; QI, H.; LIAO, S. Celeb-DF: a large-scale challenging dataset for *deepfake* forensics. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. Disponível em: https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Celeb-DF_A_Large-Scale_Challenging_Dataset_for_*Deepfake*_Forensics_CVPR_2020_paper.html. Acesso em: 23 mar. 2026.

MEDON, F. O direito à imagem na era das *deepfakes*. Revista Brasileira de Direito Civil, v. 27, p. 251-277, 2021. Disponível em: https://rbdcivil.ibdcivil.org.br/rbdc/article/viewFile/438/447. Acesso em: 7 nov. 2025.

MIRSKY, Y.; LEE, W. The creation and detection of *deepfakes*: a survey. ACM Computing Surveys, v. 54, n. 1, p. 1-41, 2021. Disponível em: https://dl.acm.org/doi/10.1145/3425780. Acesso em: 23 mar. 2026.

ROSSLER, A.; COZZOLINO, D.; VERDOLIVA, L.; RIESS, C.; THIES, J.; NIESSNER, M. FaceForensics++: learning to detect manipulated facial images. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019. Disponível em: https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html. Acesso em: 23 mar. 2026.

SOELLA, G. M.; MAIMONE, G. D. Mapeamento da detecção de *deepfakes*: um trabalho terminológico. Brazilian Journal of Information Studies: Research Trends, v. 16, 2022. Disponível em: https://dialnet.unirioja.es/descarga/articulo/8506171.pdf. Acesso em: 7 nov. 2025.

TOLOSANA, R.; VERDOLIVA, L.; MORALES, A.; FIERREZ, J.; KOMPELLA, R. *Deepfakes* and beyond: a survey of face manipulation and fake detection. Information Fusion, v. 64, p. 131-148, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S1566253520303110. Acesso em: 23 mar. 2026.

TONEZER, L. N.; SILVA, A. C. M.; ALMEIDA, A. H.; NEVES, J. E. D. Simulações Multiagentes e Phishing: Explorando a Segurança em Ambientes de Nuvem. Revista Tecnológica da Fatec de Americana, v. 11, n. 02, 2024. Disponível em: https://fatec.edu.br/revista/index.php/RTecFatecAM/article/view/393. Acesso em: 01 nov. 2025.

VERDOLIVA, L. Media forensics and *deepfakes*: an overview. IEEE Journal of Selected Topics in Signal Processing, v. 14, n. 5, p. 910-932, 2020. Disponível em: https://ieeexplore.ieee.org/document/9085890. Acesso em: 23 mar. 2026.

WOHLIN, C.; RUNESON, P.; HÖST, M.; OHLSSON, M.; REGNELL, B.; WESSLÉN, A. Experimentation in Software Engineering. Berlin: Springer, 2012.

ZHOU, P.; LIM, S. Joint audio-visual *deepfake* detection. IEEE International Conference on Computer Vision (ICCV), p. 14800-14809, 2021. Disponível em: https://openaccess.thecvf.com/content/ICCV2021/html/Zhou_Joint_Audio-Visual_*Deepfake*_Detection_ICCV_2021_paper.html. Acesso em: 23 mar. 2026.