Deepfakes: uma nova ameaça à segurança e à confiança na informação

Filipa Lopes | Inês Aparício | Sara Esteves

up200506689@edu.fe.up.pt | up200400711@edu.letras.up.pt | up202103882@edu.letras.up.pt

Relatório do mini-projeto, realizado no âmbito da unidade curricular de Segurança da Informação, do Mestrado em Ciência da Informação, lecionada pelo Prof. Doutor José Manuel de Magalhães Cruz

Faculdade de Engenharia da Universidade do Porto

Dezembro de 2024

Resumo

Deepfakes são ficheiros de vídeo, imagem ou voz manipulados usando Inteligência Artificial (IA), com o intuito de fazer com que o conteúdo falso pareça autêntico. Estes tornaram-se numa nova forma de disseminar desinformação, especialmente desafiadora devido ao seu realismo e à rapidez com que se propagam em notícias *online* e redes sociais. Este relatório analisa a tecnologia *deepfake* em vídeo, examinando a sua natureza, os seus usos e as ameaças que podem representar para a segurança e confiança na informação. Começa por abordar o contexto e a dualidade dos *deepfakes*, explorando as suas aplicações positivas (no entretenimento, educação, investigação) e negativas (desinformação, crimes e fraudes), sublinhando a importância de uma reflexão ética sobre o uso desta tecnologia. Seguidamente, são descritos os dois principais modelos de IA utilizados na criação de *deepfakes*: os *autoencoders* e as redes adversárias generativas. São, igualmente, abordadas as técnicas de deteção de *deepfakes* com recurso à IA, testando-se três ferramentas *online*: o *Deepware*, o *Deepfake-O-Meter* e o *TrueMedia.org*. De seguida, são descritos outros métodos de verificação de autenticidade, complementares à IA, que incluem a análise visual, fisiológica, de metadados, de assinaturas digitais, marcas de água e *blockchain*. A análise realizada destaca a necessidade de estratégias mais abrangentes para lidar com *deepfakes*, reconhecendo as limitações dos métodos de deteção atuais, com ou sem IA. Nesta sequência, é enfatizada a importância das literacias digital, mediática e informacional como armas poderosas contra a desinformação e a fraude, pois capacitam os cidadãos a navegar no mundo digital de forma crítica e segura e sensibilizam-nos para a necessidade de adoção de mecanismos para proteger a autenticidade e integridade da informação que ajudem a combater o uso malicioso desta tecnologia. São sugeridos alguns recursos para formação nesse sentido, destacando-se o papel das bibliotecas universitárias e escolares na disseminação destas literacias. Conclui-se que os *deepfakes* são um fenómeno complexo que requer uma resposta coordenada de toda a sociedade, envolvendo cidadãos, plataformas de comunicação social, governos e instituições responsáveis pela regulamentação e criação de políticas éticas.

Palavras-chave:
Deepfakes; Inteligência Artificial; Desinformação; Literacia Digital; Autenticidade da informação

Abstract

Deepfakes are modified video, image, or audio files created using artificial intelligence (AI) to make the fabricated information appear authentic. They have emerged as a new method of disseminating disinformation and are particularly problematic due to their realism and the speed with which they spread in online news and social media. This report examines video deepfake technology, its characteristics, applications, and the potential threats it poses to information security and trust. The first part looks at the background and duality of deepfakes, exploring their beneficial uses (in entertainment, education, and research) alongside their harmful uses in disinformation, criminal activity, and fraud, highlighting the need for ethical reflection on the use of this technology. The following section describes the two primary AI models used to generate deepfakes: autoencoders and generative adversarial networks. The methods of identifying deepfakes using AI are also examined by evaluating three online tools: Deepware, Deepfake-O-Meter, and TrueMedia.org. It is then outlined other approaches to authenticity verification that complement AI, such as visual and physiological analysis, metadata examination, digital signatures, watermarks, and blockchain technology. The research points to the need for more robust methods to combat deepfakes, acknowledging the shortcomings of current detection methods, regardless of the involvement of AI. In this context, it highlights the importance of digital, media, and information literacies as essential tools in the battle against disinformation and fraud, empowering citizens to navigate the digital landscape critically and safely, while at the same time raising awareness of the need to implement measures that protect the authenticity and integrity of information, thereby preventing the nefarious exploitation of technology. The final section discusses the importance of university and school libraries in promoting these skills. Some resources for training in this area are also recommended. We conclude that deepfakes are a multifaceted issue that requires a unified response from society, including citizens, media platforms, governments, and other bodies tasked with regulating and formulating ethical policies.

Keywords:
Deepfakes; Artificial Intelligence; Disinformation; Digital Literacy; Information Authenticity

Índice

Introdução 4
1.Usos positivos e negativos de *deepfakes* 5
2.Inteligência Artificial e *deepfakes* 9
2.1. Criação de *deepfakes* 9
2.1.1. *Autoencoders* 10
2.1.2. Redes Adversárias Generativas 11
2.2. Deteção de *deepfakes* 11
2.2.1. Teste de ferramentas de deteção de *deepfakes* 14
3.Outras técnicas de deteção de *deepfakes* 17
3.1. Análise de inconsistências visuais 17
3.2. Análise de características fisiológicas e biológicas 18
3.3. Análise de metadados e dados auxiliares 18
3.3.1. Assinatura digital 18
3.3.2. Marca de água 19
3.3.3. Utilização da tecnologia *Blockchain* 21
4.Literacias digital, mediática e informacional 23
Conclusões 25
Referências Bibliográficas 26
Apêndices 30
Apêndice 1 – Recursos informacionais 31
A. O que são deepfakes – conteúdos gerais 31
B. Aplicações positivas, aplicações potencialmente positivas (mas que podem gerar polémica) e aplicações negativas 31
C. Criação de *deepfakes* 33
D. Deteção de *deepfakes* com IA 33
E. Outras técnicas de deteção de *deepfakes* 34
Apêndice 2 – Testes no Deepware, Deepfake-O-Meter e TrueMedia.org 36

Introdução

Deepfakes são ficheiros de vídeo, imagem ou voz manipulados usando Inteligência Artificial (IA), que pretendem fazer o seu conteúdo falso passar por autêntico. O termo resulta da junção das palavras *fake* (falsificação) e *deep learning* (aprendizagem profunda), uma técnica de IA que usa redes neuronais, ou seja, baseia-se em modelos matemáticos e computacionais inspirados na estrutura do cérebro humano. Estas redes aprendem com grandes quantidades de dados e geram conteúdo sintético manipulado a partir desses mesmos dados (Kaswan et al., 2023). O nosso mini-projeto teve como principal objetivo compreender o que são *deepfakes*, nomeadamente em vídeo, e como podem ser uma ameaça à segurança e à confiança na informação. A atualidade deste fenómeno e as implicações negativas que o seu uso malicioso pode ter nos indivíduos e nas sociedades levou-nos igualmente a preparar uma lista de recursos (Apêndice 1) que possam ser utilizados em formações de literacia digital, mediática e informacional. Na nossa perspetiva, o conhecimento dos mecanismos de geração e de deteção de *deepfakes* e a consciencialização sobre os contextos dos seus usos positivos e negativos contribuem para o pensamento crítico e o uso ético, cauteloso e responsável dos meios digitais e das ferramentas de IA. Estas literacias são armas poderosas contra a desinformação e a fraude, e devem estar ao alcance do cidadão comum.

Estruturamos o relatório do nosso mini-projeto em quatro partes. Na primeira parte, contextualizamos o tema, mostrando a dualidade dos *deepfakes* através da identificação das aplicações positivas e negativas. Na segunda parte, focamos a nossa atenção no uso da IA para a criação e a deteção de *deepfakes* em vídeo, testando algumas ferramentas disponíveis *online*. Na terceira parte, exploramos outras formas de deteção que complementam a análise realizada pelas ferramentas que usam IA. Na última parte, destacamos a necessidade de se investir mais em literacia digital, mediática e informacional, remetendo para o Apêndice 1, que compila o conjunto de recursos já mencionado. No Apêndice 2, estão os resultados de alguns testes realizados nas ferramentas de deteção de *deepfakes* disponíveis *online*: Deepware, Deepfake-O-Meter e TrueMedia.org.

1. Usos positivos e negativos de *deepfakes*

O termo *deepfake* surgiu em 2017, associado à publicação de vídeos falsos, de cariz pornográfico, de atrizes de Hollywood no *website* Reddit (Rancourt-Raymond & Smaili, 2023). Desde aí, o tipo de tecnologia que permite criar *deepfakes* converteu-se numa nova forma de fazer proliferar desinformação, particularmente desafiadora pelo seu realismo e pela rapidez com que se propaga por notícias *online* e redes sociais (Vizoso et al., 2021).

No entanto, a tecnologia que gera *deepfakes* com conteúdos desinformativos, muitas vezes usurpando a identidade de pessoas reais, quando usada de forma ética e em prol do bem comum, pode gerar rostos ou corpos completos sintéticos com aplicações positivas em diferentes áreas, desde o entretenimento à investigação científica.

Os *deepfakes* começaram por ser usados no entretenimento, na criação de efeitos realistas em videojogos e em filmes (Stanciu & Ciuperca, 2024). Possibilitam não só a criação de personagens fictícias, como também transformações nos corpos e nos rostos de atores humanos no momento da edição audiovisual ou a recriação da aparência de atores já falecidos (Murphy et al., 2023; Renier et al., 2024). Por exemplo, em 2019, no filme de Martin Scorsese *O Irlandês* foi possível usar esta tecnologia para rejuvenescer digitalmente os atores Robert De Niro e Al Pacino (Kaswan et al., 2023, p. 293). Ela também pode ser usada para proteger a identidade de determinados indivíduos, como aconteceu com as testemunhas que participaram no documentário *Welcome to Chechnya*, em 2020, realizado por David France, centrado na perseguição perpetrada contra homossexuais na Chechénia, sob o domínio russo (Danry et al., 2022). Outro uso bastante promissor é nas dobragens, permitindo sincronizar os movimentos labiais dos atores com o áudio dobrado em diferentes idiomas, o que torna a experiência mais realista para o público estrangeiro (Kaswan et al., 2023, p. 293).

Na área da publicidade e do *marketing*, podem realizar-se aplicações semelhantes, criando experiências personalizadas e interativas para os consumidores (Stanciu & Ciuperca, 2024). Embora a literatura que selecionámos e consultámos não o mencione, o uso de *deepfakes* pode ser particularmente importante para proteger as identidades de crianças que participem em campanhas publicitárias ou no cinema. No entanto, é importante ressalvar que a utilização de *deepfakes* no entretenimento e na publicidade pode levantar questões éticas (Campbell et al., 2022; Murphy et al., 2023), especialmente quando se trata da recriação de pessoas falecidas sem o seu consentimento, ou quando se pretende enganar os consumidores ou perpetuar padrões de beleza tóxicos que afetam a autoestima dos consumidores, especialmente dos mais jovens. A linha entre o positivo e o negativo pode ser muito ténue nestes casos. Como se constata do relatório e da campanha levada a cabo pela Dove, através do *Dove Self-*

Esteem Project[1], a tecnologia de *deepfakes* pode ser utilizada para a proliferação de imagens "perfeitas", que criam pressão social sobre os jovens, levando-os à insatisfação com a sua aparência e gerando problemas de autoestima. Já no contexto da recriação de pessoas falecidas, o anúncio publicitário brasileiro da *Volkswagen 70 anos*, que contou com a participação da cantora Maria Rita e de um *deepfake* da sua falecida mãe, a cantora Elis Regina, é um bom exemplo. O vídeo gerou polémica não só pela falta de menção explícita ao uso de IA, mas também pela associação da imagem de Elis Regina a uma empresa que, no passado, manteve vínculos e apoiou a ditadura militar brasileira (1964-1985) que a cantora tão veemente combateu (Haddad, 2023).

No domínio da educação e da formação, pode-se tirar partido da vertente lúdica e interativa dos *deepfakes*. É possível a criação de materiais educativos multilíngues, *on demand*, a baixo custo, democratizando o acesso à educação (Roe et al., 2024, p. 7). Também se podem realizar simulações realistas de eventos e personagens históricos (Stanciu & Ciuperca, 2024, p. 64). A oportunidade de interagir com essas representações virtuais ou de visitar recriações digitais de cidades antigas, feitas com rigor científico, pode proporcionar aos alunos e aos formandos uma compreensão mais abrangente do passado, tornando a aprendizagem numa experiência envolvente e tornando mais acessível o conhecimento da história ou de outras matérias.

O mesmo pode acontecer no domínio da arte e dos museus, *deepfakes* podem desempenhar funções didáticas e de entretenimento. É o caso do *deepfake* de Salvador Dalí que foi criado no museu que recebeu o seu nome na Flórida (EUA), proporcionando uma experiência educativa e interativa aos visitantes (Kwow & Koh, 2021)[2]. A tecnologia *deepfake* permite igualmente aos artistas experimentar novas formas de expressão e criar obras inovadoras, desafiando as fronteiras entre a realidade e a ficção ou explorando temas complexos de uma forma envolvente e acessível ao público, como é o caso dos perigos das redes sociais (Cheres & Groza, 2023)[3].

A investigação científica é outro setor que pode beneficiar muito com este género de tecnologia. A pesquisa na área da saúde pode ser uma delas, mas também a área dos estudos comportamentais não verbais. Um estudo, publicado em 2024, investigou o impacto da expressividade facial em entrevistas de emprego, utilizando tecnologia *deepfake* para criar vídeos experimentais (Renier et al., 2024). Os autores geraram vídeos de indivíduos exibindo diferentes níveis de expressividade (olhar, acenar com a cabeça, sorrir) e analisaram como os observadores percebiam a competência, o carisma e qual a sua impressão geral dos candidatos. Os resultados confirmaram que candidatos mais expressivos foram avaliados mais favoravelmente, demonstrando a utilidade da

[1] Projeto dedicado à promoção da autoestima e da imagem corporal positiva, especialmente entre jovens. Visa combater a influência das redes sociais e da publicidade na autoperceção, principalmente no que toca a padrões de beleza irrealistas (Apêndice 1, B.2.2).
[2] Apêndice 1, B.3.1.
[3] Apêndice 1, B.3.2.

tecnologia *deepfake* na padronização de expressões faciais, o que se torna particularmente útil em cenários de entrevistas. Em suma, os entrevistados pareciam pessoas reais, mas não o eram e tal tornava possível o controlo de variáveis ligadas à expressividade facial, o que seria extremamente mais difícil com pessoas reais.

A tecnologia que gera *deepfakes* também pode ter um impacto positivo noutras áreas. A humanização da IA é uma delas. A criação de avatares virtuais mais realistas e expressivos com recurso a *deepfakes* pode tornar a interação com assistentes virtuais e sistemas de IA mais natural e envolvente, enquanto a sua utilização em treinos de realidade virtual pode promover a empatia e o desenvolvimento de competências sociais (Danry et al., 2022; Stanciu & Ciuperca, 2024).

Todas as utilizações já mencionadas apenas podem contribuir para o bem individual e coletivo se forem acompanhadas de uma reflexão ética cuidadosa, garantindo a transparência, o consentimento informado e a proteção da privacidade dos indivíduos envolvidos (Kaswan et al., 2023). Esta chamada de atenção é muito importante porque, não obstante todas as potencialidades didáticas, sociais e de entretenimento mencionadas, a maioria das aplicações de *deepfakes* em vídeo tem sido negativa, estando especialmente ligada a casos de desinformação, manipulação, crime e fraude (Peters, 2024).

A propagação de notícias falsas ou de conteúdo enganoso é aquela que tem sido mais noticiada, tendo como objetivo influenciar eleições e processos políticos, difamar pessoas ou grupos e manipular a opinião pública. Os exemplos têm-se somado, criando uma atmosfera de suspeita resultante da erosão da confiança nas instituições e nas figuras públicas. Por exemplo, desde 2022, após a invasão russa à Ucrânia, a figura do presidente Volodymyr Zelensky tem sido objeto de diversos vídeos manipulados. A sua utilização tem sido uma tática de desinformação para prejudicar a sua imagem e manipular a perceção pública sobre a guerra. Num desses vídeos manipulados, criado em 2022, o *deepfake* do presidente ucraniano fazia um apelo público de capitulação à Rússia, instruindo as forças ucranianas a renderem-se (Apêndice 1, B.4.1).

Outra aplicação negativa dos *deepfakes* é no contexto de exploração e abuso sexual, um problema crescente que tem afetado tanto pessoas famosas quanto pessoas comuns, especialmente crianças e adolescentes. Os *deepfakes* podem ser usados para criar vídeos pornográficos falsos sem o consentimento das vítimas, colocando o rosto de alguém em situações sexualmente explícitas. Este tipo de conteúdo é usado para humilhar, envergonhar e extorquir as vítimas. Existem diferentes formas de exploração sexual com *deepfakes*, como a pornografia não consensual, a "sextorsão" e a criação de material de abuso sexual infantil falso. A pornografia não consensual utiliza *deepfakes* para criar vídeos pornográficos falsos com rostos de pessoas sem o seu consentimento. Na "sextorsão", jovens, em 90% dos casos rapazes adolescentes, são enganados a partilhar imagens íntimas *online* e depois chantageados. O FBI alertou para o facto de os jovens, especialmente menores, serem as principais vítimas de "sextorsão" com o uso

de *deepfakes* sexualmente explícitos, e registou, desde 2021, mais de 12600 vítimas menores de idade (Peters, 2024, p. 14). Por sua vez, a criação e distribuição de simulações de abuso sexual infantil, apesar de não envolver diretamente crianças reais, é ilegal e extremamente prejudicial, pois alimenta a procura deste tipo de conteúdo abusivo e contribui para a normalização da exploração sexual de menores. A exploração sexual com recurso a *deepfakes* representa uma grave violação de privacidade, com a criação de conteúdo falso. Estes causam danos psicológicos e à reputação das vítimas, por vezes irreparáveis, podendo levar, em alguns casos, ao suicídio (Peters, 2024, p. 14).

O uso de *deepfakes* no contexto de crimes e fraudes financeiras tem-se tornado uma preocupação crescente. Um inquérito de 2024 da *Business.com* dirigido a executivos concluiu que 10% já tinha sido alvo de *deepfakes* e golpes gerados por IA; outro estudo do mesmo ano, focado na indústria financeira nos EUA e no Reino Unido, descobriu que 53% das empresas tinham sofrido alguma forma de fraude com *deepfakes* (Peters, 2024, p. 13). Esta tecnologia tem facilitado a prática de diversos crimes, como o roubo de identidade (para criar identidades falsas e abrir contas bancárias fraudulentas), a manipulação de mercados financeiros (com a divulgação de notícias falsas usando *deepfakes* de figuras influentes) e as burlas e extorsões (com os golpistas a usarem *deepfakes* para se fazerem passar por pessoas conhecidas das vítimas) (Kaswan et al., 2023, pp. 293–294; Peters, 2024).

Durante séculos, os humanos procuraram forjar conteúdo, falsificando documentos. Com o desenvolvimento de aplicações informáticas como o Adobe Photoshop, tornou-se mais fácil manipular imagens. A tecnologia *deepfake* permite agora manipular vídeos e áudios, criando conteúdos sintéticos extremamente realistas (Zhang et al., 2021). Para nos protegermos das aplicações maliciosas dos *deepfakes*, é fundamental compreender como estes são gerados e como podem ser detetados.

2. Inteligência Artificial e *deepfakes*

2.1. Criação de *deepfakes*

Nos *deepfakes* de vídeo, as imagens são manipuladas fotograma a fotograma. A modificação pode alterar o conteúdo ou o contexto do vídeo, envolvendo, por exemplo, a eliminação ou a inserção de objetos ou o ajuste de efeitos visuais e sonoros (Kaur et al., 2024). No entanto, para a criação de *deepfakes* de imagem e vídeo, a manipulação do rosto é o método mais utilizado, existindo vários tipos de manipulação, nomeadamente:

- Criação de um rosto totalmente novo;
- Substituição de um rosto por um outro;
- Modificação de um atributo específico, como a cor da pele ou a cor do cabelo;
- Alteração de expressões faciais, nomeadamente para permitir uma sincronização entre o movimento dos lábios e a fala (Patel et al., 2023).

Além do rosto, pode existir uma manipulação em que os movimentos do corpo de uma pessoa são transferidos para o corpo de outra, sendo depois usados para criar vídeos manipulados (Patel et al., 2023), como no caso de vídeos em que uma pessoa aparenta estar a realizar uma dança quando, na realidade, os movimentos foram capturados de um vídeo de outra pessoa[4].

Existem várias ferramentas disponíveis na Web para a criação de *deepfakes*, tanto pagas como gratuitas. Uma das principais e mais populares ferramentas para a criação de *deepfakes* é o DeepFaceLab[5], uma ferramenta de código aberto (Rajput & Arora, 2024). O *website* DeepFake Lab[6] é uma excelente fonte para se perceber como é que se pode fazer um *deepfake* com este *software*. Outras ferramentas de produção de *deepfakes* de código aberto mencionadas nas fontes consultadas são, por exemplo, o Faceswap[7] e o Style-GAN[8] (Rajput & Arora, 2024).

No que diz respeito ao processo de criação de *deepfakes* de forma mais específica, são usadas, na maior parte das vezes, redes neuronais profundas (*deep neural networks*) para manipular vídeos e imagens, sendo necessárias grandes quantidades de imagens e vídeos verdadeiros e falsos para treinar tais modelos.

---

[4] Veja-se: https://www.youtube.com/watch?v=PCBTZh41Ris&t=37s

[5] Este *software* estava disponível no Github até há pouco tempo em: https://github.com/iperov/DeepFaceLab. Ao visitá-lo atualmente surge a mensagem que o proprietário o arquivou em novembro de 2024, permitindo agora apenas a leitura.

[6] Disponível em: https://deepfakelab.theglassroom.org/index.html

[7] Disponível em: https://faceswap.dev/

[8] Disponível em: https://github.com/NVlabs/stylegan

A literatura destaca dois modelos usados na criação de *deepfakes*: os ***autoencoders*** e as **redes adversárias generativas** (Patel et al., 2023).

2.1.1. *Autoencoders*

O *autoencoder* é um tipo de rede neuronal artificial constituída por um *encoder* (codificador), que comprime uma imagem para uma versão menor, mantendo as informações principais da mesma; e por um *decoder* (descodificador), que tenta recriar a imagem original com base na versão comprimida. No contexto dos *deepfakes*, este modelo é usado para substituir o rosto de uma pessoa por outro em vídeos ou em fotografias. Como se pode ver na Figura 1, o processo é feito da seguinte forma: dois *autoencoders* são treinados, um para cada rosto (da Pessoa A e da Pessoa B). Neste processo de treino, ambos os modelos usam o mesmo codificador para criarem representações latentes das imagens originais, ou seja, formas compactas e abstratas das características principais da imagem. A seguir, descodificadores diferentes são usados para reconstruir as imagens originais. Para a geração de *deepfakes*, ou seja, para se trocar o rosto da Pessoa A pelo rosto da Pessoa B, o que acontece é que o rosto da Pessoa A é descodificado não pelo seu descodificador, mas pelo descodificador da Pessoa B. Desta forma, é criada uma imagem em que o rosto da Pessoa A é transformado no rosto da Pessoa B, mantendo, no entanto, as expressões e movimentos originais. Para criar um *deepfake* de vídeo, esse processo é repetido para cada imagem do vídeo, substituindo o rosto da Pessoa A pelo rosto da Pessoa B em todas as imagens (Patel et al., 2023).

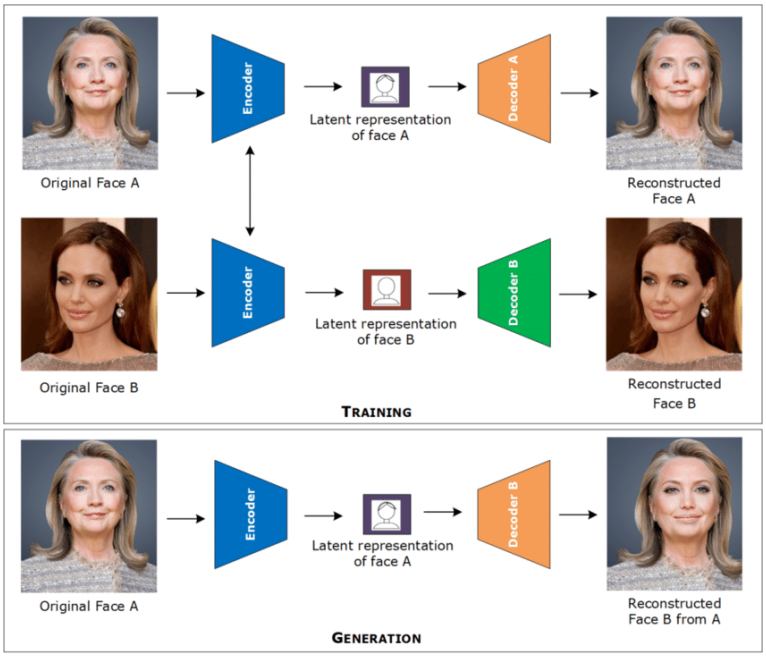

Figura 1- *Funcionamento do Autoencoder*
Fonte: Masood et al., 2021.

2.1.2. Redes Adversárias Generativas

A rede adversária generativa é um modelo de IA que tanto é usado na criação como na deteção de *deepfakes*. O modelo é constituído por um gerador (*generator*) e um discriminador (*discriminator*). O papel do gerador é criar conteúdo sintético que pareça real. Depois, este conteúdo é misturado com conteúdo original e transmitido ao discriminador. O discriminador analisa e diferencia entre conteúdo sintético e original (Figura 2). O gerador recebe *feedback* do discriminador e com isso vai aprendendo a corrigir os erros e a fazer melhores criações. Após muitos ciclos de aprendizagem, o gerador torna-se capaz de criar conteúdo falso que é praticamente indistinguível do conteúdo original. Quando o gerador consegue enganar o discriminador em cerca de 50% dos *inputs*, a aprendizagem considera-se completa (Patel et al., 2023).

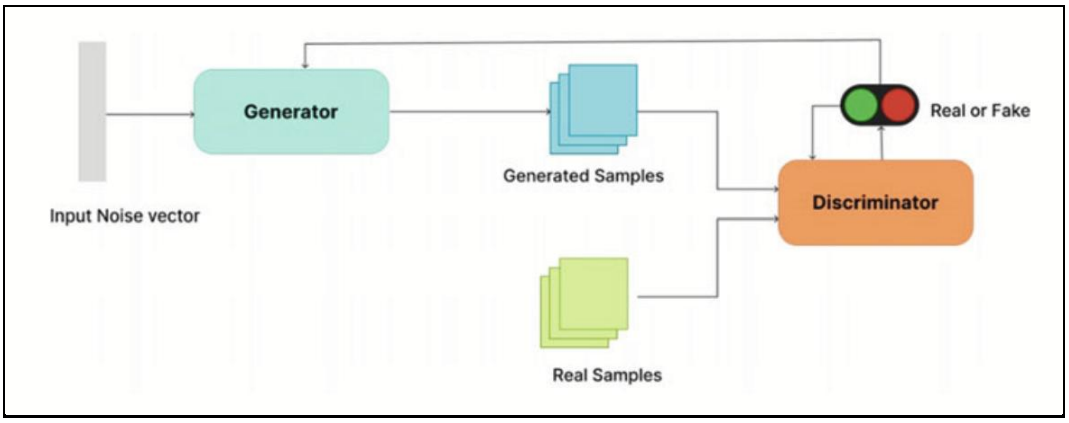

Figura 2- Funcionamento das redes adversárias generativas
Fonte: Chawki, 2024.

2.2. Deteção de *deepfakes*

A deteção de *deepfakes* exige uma metodologia abrangente que integre diversas abordagens, incluindo técnicas que tenham em conta aspetos espaciais, temporais e de frequência presentes nos vídeos (Sandotra & Arora, 2024). Entre as principais estratégias utilizadas, destaca-se o uso de modelos de aprendizagem automática e de aprendizagem profunda treinados para distinguir conteúdo real de falso e que têm em consideração estes diferentes aspetos (Chanda et al., 2024). Estes modelos necessitam de um grande volume de dados para serem desenvolvidos. O treino e avaliação de modelos de deteção criados neste âmbito depende, assim, da existência de grandes bases de dados constituídas por vídeos falsos e vídeos verdadeiros, sendo que existe uma disparidade entre a quantidade de vídeos rotulados como *deepfakes* e a quantidade de vídeos autênticos disponíveis, existindo menos vídeos de *deepfakes* disponíveis para treino do que vídeos autênticos (Abbas & Taeihagh, 2024). Ainda assim,

existem algumas bases de dados públicas usadas para treinar estes modelos, como a FaceForensics++[9], a Celeb-DF[10] e o DeeperForensics-1.0.[11].

Além da questão da quantidade de dados disponíveis para treino, a deteção de *deepfakes* em vídeo apresenta maiores desafios que a deteção em imagens, visto implicar:

- a análise de uma sequência de imagens, ou seja, maior processamento dados;
- a análise da informação temporal e da informação de áudio;
- a existência de compressão dos vídeos, que altera os dados e pode dificultar o processo de deteção;
- a necessidade de modelos mais complexos e com elevados custos computacionais (Kaur et al., 2024).

O maior desafio será, por fim, garantir que os modelos de deteção de *deepfakes* sejam capazes de generalizar para diferentes tipos de manipulação e conjuntos de dados (Xia et al., 2024).

No contexto da deteção de *deepfakes*, são combinados diferentes modelos de IA para analisar vários aspetos do vídeo e identificar manipulações. Além das **redes adversárias generativas** já referidas anteriormente, podem ser utilizadas, por exemplo, **redes neuronais convolucionais** para extrair características visuais detalhadas de cada fotograma, ajudando a identificar padrões e anomalias nas imagens (Diwan et al., 2024). Da mesma forma, podem ser utilizadas **redes neuronais recorrentes** para analisar padrões temporais, o que permite detetar anomalias como movimentos não naturais ou alterações súbitas entre fotogramas (Sandotra & Arora, 2024).

Este é um campo em constante evolução, onde novas abordagens e técnicas são continuamente desenvolvidas para acompanhar a crescente sofisticação dos *deepfakes*. A "batalha" existente entre a criação e a deteção de *deepfakes* que recorrem a modelos de IA está bem representada na Figura 3, retirada do artigo de Juefei-Xu et al. (2022).

A figura apresenta um diagrama que mapeia a relação entre métodos de geração e detecção de *deepfakes*. Na coluna da esquerda estão os métodos de geração (incluindo conjuntos de dados como o FaceForensics++), enquanto na coluna da direita estão os métodos de deteção. As curvas conectam os métodos de geração aos métodos de deteção que os avaliaram, representando a interação entre eles, e são coloridas para representar diferentes tipos de métodos de deteção.

[9] Disponível em: https://github.com/ondyari/FaceForensics
[10] Disponível em: https://github.com/yuezunli/celeb-deepfakeforensics/tree/master/Celeb-DF-v1
[11]Disponível em: https://github.com/EndlessSora/DeeperForensics-1.0

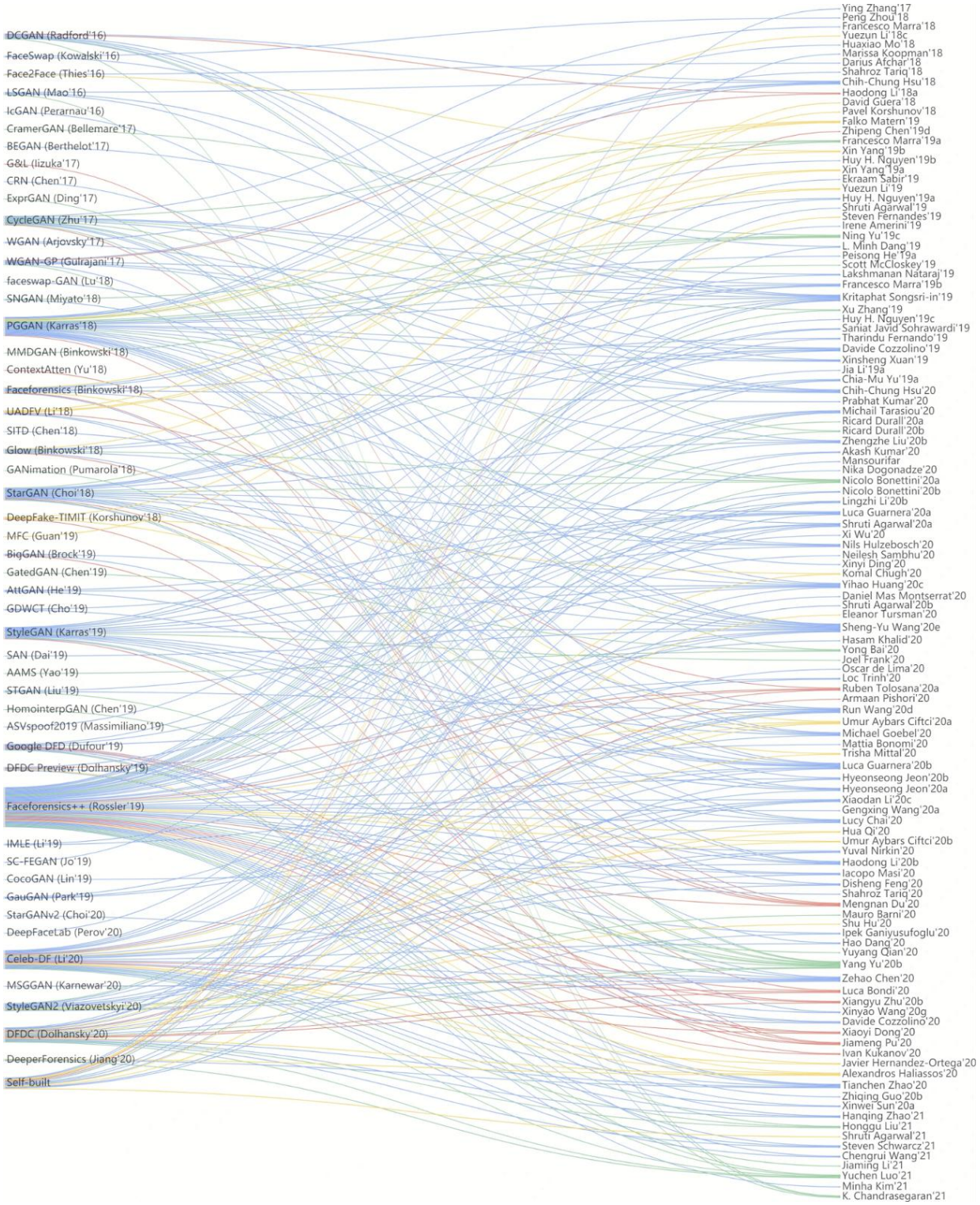

Figura 3- A "batalha" entre a geração e deteção de *deepfakes*
Fonte: Juefei-Xu et al., 2022.

2.2.1. Teste de ferramentas de deteção de *deepfakes*

Existem várias ferramentas disponíveis *online* que utilizam modelos de IA para detetar *deepfakes*, sendo que a própria deteção se tornou também um negócio. No âmbito do nosso mini-projeto, testámos dois vídeos (um autêntico e um *deepfake*) em três plataformas gratuitas: o Deepware[12], o Deepfake-O-Meter[13] e o TrueMedia.org[14]. Os dois vídeos escolhidos foram de personalidades políticas, dado a última das ferramentas ser dedicada exclusivamente à deteção de *deepfakes* políticos. O vídeo autêntico diz respeito a parte de um debate entre Biden e Trump[15], largamente difundido em canais noticiosos e retirado da conta de Youtube oficial do *Wall Street Journal*. O vídeo que se trata de um *deepfake* foi retirado de uma conta do Youtube e apresenta o presidente Biden a narrar uma história sobre um pistácio mágico[16]. No Apêndice 2, é possível consultar as imagens capturadas durante o processo de teste das ferramentas.

As três ferramentas utilizam múltiplos algoritmos para avaliar vídeos. O Deepware fornece informações gerais sobre eles, enquanto o Deepfake-O-Meter apresenta detalhes mais específicos. O TrueMedia.org apenas indica a quantidade de detetores utilizados na análise, dividindo-os em imagem e áudio. No Deepware e no Deepfake-O-Meter, os resultados são apresentados sob a forma de percentagem, sendo que valores mais altos significam uma maior probabilidade de o vídeo ter sido manipulado. O Deepware combina esses valores e apresenta uma avaliação final, o que não é feito no Deepfake-O-Meter. Em relação ao TrueMedia.org, não são apresentadas percentagens, existe apenas informação escrita relativa à avaliação da imagem e do áudio, sendo dada uma avaliação qualitativa final.

Quanto ao Deepware, os resultados obtidos podem ser consultados na Tabela 1. Em relação ao vídeo autêntico, os valores indicam que o vídeo foi corretamente identificado como não sendo um *deepfake.* No que diz respeito à avaliação do vídeo *deepfake*, o desempenho ficou aquém do esperado. O vídeo apresentado tem um conteúdo facilmente reconhecível como falso dada a história sem sentido que é apresentada, sendo que qualquer pesquisa simples no Google o confirmaria de imediato. No entanto, a plataforma apenas o identifica como sendo "suspeito".

[12] Disponível em: https://deepware.ai/
[13] Disponível em: https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/home_login
[14] Disponível em: https://detect.truemedia.org
[15] Vídeo retirado de: https://www.youtube.com/watch?v=KdOfiPyY87I&t=1s
[16] Vídeo retirado de: https://www.youtube.com/watch?v=yVEhrIMc-ps

Tabela 1 – Resultados obtidos na plataforma Deepware

	Vídeo autêntico	***Deepfake***
Modelo Avatarify	6%	12%
Modelo Deepware	6%	48%
Modelo Seferbekov	1%	55%
Modelo Ensemble	5%	53%
Avaliação final	No deepfake detected	Suspicious

A análise pelo Deepfake-O-Meter (ver Tabela 2) também foi feita por múltiplos algoritmos, que usam modelos que procuram aspetos diferentes no vídeo e serão mais ou menos suscetíveis a alterações subtis. Em relação à análise do vídeo autêntico, a análise feita pela plataforma levanta uma questão séria: a identificação de vídeos autênticos como tendo uma probabilidade de 100% de terem sido criados por inteligência artificial. Um conhecimento mais aprofundado dos modelos e do seu funcionamento permitiria certamente uma interpretação diferente desse valor. No entanto, estimar esta probabilidade na análise de um vídeo que é verdadeiro, leva-nos a questionar até que ponto certos modelos de deteção estão a levar os utilizadores comuns a duvidar do que é real. O problema já não é acreditar no que é falso, mas também começar a duvidar do que é verdadeiro. Quanto à análise do vídeo *deepfake*, verificamos que as percentagens também variam muito, sendo que apenas três dos sete detetores apresentam percentagens que não deixam dúvidas de que se trata de um vídeo falso.

Tabela 2 – Resultados obtidos na plataforma Deepfake-O-Meter

	Vídeo autêntico	***Deepfake***
Detetor AltFreezing	12,2%	18,3%
Detetor DSP-FWA	27,5%	97,5%
Detetor FTCN	1,4%	1,1%
Detetor LIPINC	100%	100%
Detetor LSDA	13,4%	48,8%
Detetor SBI	4,5%	17,6%
Detetor WAV2LIP-STA	7,8%	89,1%

Quanto ao TrueMedia.org (ver Tabela 3), a análise do vídeo autêntico levanta o mesmo tipo de questão discutida anteriormente. Um vídeo verdadeiro pode ser avaliado como autêntico ou manipulado. No que diz respeito à deteção do vídeo *deepfake*, a plataforma identificou evidências substanciais que permitiram avaliá-lo como tal. Também solicita ao utilizador que partilhe a sua opinião sobre a veracidade do vídeo, pedindo que deixe um comentário em que junte, por exemplo, artigos de notícias e a fonte original do vídeo.

Tabela 3 – Resultados obtidos na plataforma TrueMedia.org

	Vídeo autêntico	***Deepfake***
3 detetores de rostos	Uncertain	substantial evidence
3 detetores de vozes	little evidence	substantial evidence
Avaliação final	Uncertain: could be authentic or manipulated	Substantial evidence of manipulation

Tendo em conta que testámos apenas três plataformas e dois vídeos (uma amostra demasiado pequena), não podemos tirar conclusões definitivas sobre a deteção que é realizada por este tipo de plataformas. No entanto, ficámos com a impressão de que não são um recurso suficientemente satisfatório para se avaliar se um vídeo foi manipulado ou não. Um utilizador que não tenha um conhecimento aprofundado destes modelos, com base nos resultados que lhe são apresentados, fica sem saber se o vídeo é realmente verdadeiro ou falso. Sendo assim, outras técnicas devem ser consideradas para a deteção de *deepfakes*, não se devendo depositar toda a confiança num único modelo de IA ou numa única plataforma mesmo que trabalhe como um conjunto de algoritmos diferentes.

3. Outras técnicas de deteção de *deepfakes*

A literatura consultada foca-se sobretudo nas técnicas de deteção de *deepfakes* que recorrem à IA. No entanto, é importante conhecer outras técnicas que permitam ao utilizador comum avaliar um vídeo e complementar a avaliação dos resultados obtidos através de ferramentas de deteção com IA, uma vez que nenhum método é 100% preciso (Chanda et al., 2024). Estas técnicas passam, sobretudo, por analisar características específicas de uma imagem ou vídeo para determinar se o seu conteúdo é autêntico ou manipulado (Chanda et al., 2024), podendo ser realizadas com observação cuidadosa ou usando ferramentas básicas de edição e análise de imagem (Sandotra & Arora, 2024). As características elencadas de seguida serão em grande medida as mesmas que os modelos de IA procuram.

3.1. Análise de inconsistências visuais

A análise de um vídeo deve passar pela análise do contexto e da consistência da narrativa apresentada. A verificação desses factos é fundamental. Após isso, devem-se procurar inconsistências visuais que possam indicar a existência de algum tipo de manipulação. A manipulação de vídeos pode ser identificada, em alguns casos, com a observação direta e cuidadosa de certos elementos visuais. Contudo, é importante salientar que tal tarefa se torna cada vez mais desafiadora devido à crescente sofisticação dos *deepfakes*.

Algumas características servem de indicadores de manipulação, destacando-se (Edwards et al., 2024; Sandotra & Arora, 2024):

- **Anomalias na iluminação**, nomeadamente inconsistências na direção e intensidade da luz, existência de sombras inconsistentes e falta de reflexos em superfícies espelhadas ou objetos brilhantes.
- **Distorção da imagem**, em que se incluem alterações na geometria e perspetiva (objetos esticados, comprimidos ou distorcidos de forma não natural), deformações e alterações no aspeto dos objetos.
- **Áreas excessivamente suaves ou desfocadas**, devendo-se prestar atenção à existência de bordas sem definição entre diferentes partes da imagem e à eventual falta de detalhe. É muito importante verificar se há **bordas de mistura** (*blending*), ou seja, áreas onde um rosto manipulado foi sobreposto a outro.
- **Descontinuidade na cor da pele e nas texturas**, ou seja, diferenças abruptas na tonalidade da pele ou nas características da textura entre áreas próximas, particularmente no rosto ou corpo.

- **Alterações na sequência de fotogramas de um vídeo**, nomeadamente a existência de movimentos bruscos ou mudanças repentinas na continuação das ações que interrompam a fluidez natural.

3.2. Análise de características fisiológicas e biológicas

Além das inconsistências visuais, muitas vezes detetadas por observação direta, existem outras características a ter em conta, tais como:

- **Inconsistências nos movimentos dos lábios e do queixo,** sendo que movimentos não sincronizados e artificiais podem indicar a presença de manipulações. Neste caso, além da análise visual, pode ser usada a análise do fluxo ótico, que passa essencialmente pela estimativa de deslocamento de cada pixel entre duas imagens, sendo este processo fundamental para perceber o movimento (Nassif et al., 2022).
- **Anomalias na frequência e nos padrões de piscar dos olhos**, sendo que uma frequência muito alta ou muito baixa pode ser sinal de manipulação, especialmente se contrastar com o comportamento que é normal na pessoa. É igualmente importante ver se o piscar é sincronizado com outros movimentos faciais e com a fala. Esta característica é particularmente importante por ser de difícil replicação (Sandotra & Arora, 2024).
- **Inconsistências na frequência cardíaca**, observadas através de variações no tom de pele e fluxo sanguíneo. Este tipo de análise será muito difícil de ser realizada a olho nu, sendo mais fácil com o uso da IA. No entanto, é possível recorrer a outras técnicas de processamento como a EVM (Magnificação de Vídeo Euleriana), que permite amplificar pequenas variações de movimento e de cor em vídeos (Rehaan et al., 2024).

3.3. Análise de metadados e dados auxiliares

Além do conteúdo visível no vídeo, é fundamental analisar os metadados dos ficheiros. Devem-se verificar inconsistências ou alterações, assim como a eventual presença de assinaturas digitais e marcas de água. De seguida, aprofundamos estas duas técnicas, assim como a verificação através da utilização da tecnologia *blockchain*.

3.3.1. Assinatura digital

A assinatura digital é uma técnica criptográfica utilizada para garantir a autenticidade e integridade do conteúdo de um documento digital (Kaur & Kaur, 2012). Esta assinatura é muito utilizada em documentos textuais; no caso dos vídeos, é mais comum que se assine um ficheiro de *hash* do vídeo em vez do vídeo em si, por ser mais eficiente este

processo em ficheiros de grandes dimensões. Através de um algoritmo de *hash* cria-se um resumo criptográfico de tamanho fixo ("impressão digital", *hash*) do documento vídeo. Esse identificador único é cifrado com a chave privada do criador. Esse *hash* cifrado é uma assinatura digital, que é guardada ou enviada em anexo ao ficheiro original como prova de autenticidade (Kaur & Kaur, 2012).

É importante notar que um *hash* é único para cada entrada específica; assim, qualquer alteração mínima no vídeo resultará num *hash* completamente diferente. Deste modo, esta "impressão digital" permite verificar, posteriormente, se a integridade do ficheiro foi comprometida, ou seja, se o ficheiro sofreu qualquer alteração. Primeiro, no local de armazenamento ou no destino (no caso de o documento ter sido transmitido), calcula-se um novo resumo criptográfico do vídeo. De seguida, utilizando a chave pública do criador, decifra-se o resumo criptográfico original, que se encontrava cifrado. Por fim, comparam-se os produtos das duas operações indicadas. Se forem iguais, o documento (vídeo), que estava guardado ou o que foi recebido, corresponde ao original. Caso não sejam iguais, foi realizada alguma alteração indevida no documento original durante o armazenamento ou durante a transmissão (Kaur & Kaur, 2012).

A autenticação por assinatura digital é considerada segura devido à sua natureza criptográfica. No entanto, também tem limitações: depende da proteção da chave privada do criador; requer equipamento e *software* especializado para gerar o *hash* e inserir a assinatura no vídeo durante a fase de criação, o que pode aumentar custos e limitar a disponibilidade dos sistemas; e, embora possa proteger o vídeo de acessos não autorizados por terceiros, nada impede que o criador do vídeo o altere (Diwan et al., 2024).

3.3.2. Marca de água

A marca de água é uma técnica que permite incorporar informação nos ficheiros multimédia de forma discreta, com o objetivo de proteger a sua autenticidade e integridade. Aberna e Agilandeeswari (2024) explicam que a marca de água pode ser incorporada de duas formas principais: no **domínio espacial** (SDW), modificando diretamente os valores dos pixéis, ou no **domínio da frequência** (TDW), transformando a imagem em representações de frequência. A marca de água no domínio da frequência é geralmente mais robusta contra ataques, pois é menos percetível e pode resistir a manipulações como compressão e recorte. Além disso, podem ser utilizadas técnicas híbridas que combinam ambos os domínios, aumentando a eficácia da proteção ao tirar proveito das vantagens de cada abordagem.

A Figura 4 ilustra o processo de aplicação de uma marca de água digital para proteger a integridade e autenticidade de uma imagem. No lado do emissor, a marca de água é incorporada na imagem original (*host image*) através de um processo de inserção, utilizando uma chave secreta, o que eleva a segurança do processo. Durante a

transmissão, a imagem com a marca de água inserida pode estar sujeita a ataques ou modificações. No lado do recetor, utilizando a mesma chave secreta, a marca de água é extraída da imagem recebida e comparada à marca de original para se verificar a integridade do conteúdo.

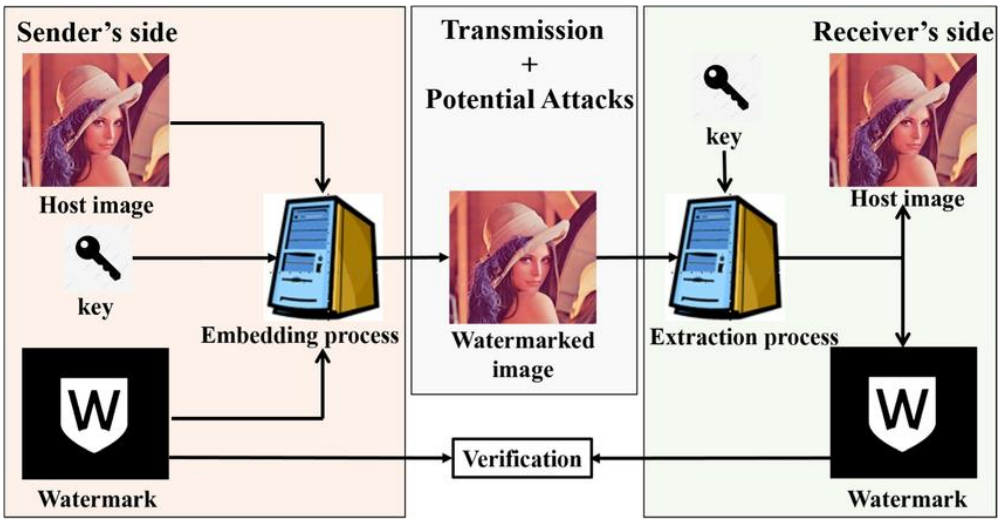

Figura 4- *Processo geral da marca de água*
Fonte: Sharma et al., 2024.

No contexto do vídeo, a marca de água é especialmente relevante para autenticar e garantir a integridade do conteúdo. Tal como no processo acima descrito relativo às imagens, a incorporação da marca de água em vídeos envolve igualmente dois passos principais: a inserção da marca no vídeo original e a sua extração.

As marcas de água podem ser projetadas para resistir a modificações, como compressão, escalonamento e ruído. Embora as marcas de água digitais ofereçam vantagens em relação a outras técnicas de autenticação, como as assinaturas digitais, devido à sua capacidade de sobreviver a diversos tipos de processamento de vídeo, elas apresentam igualmente limitações. Entre essas limitações, destaca-se o compromisso entre invisibilidade e robustez, o que significa que uma marca de água mais visível pode ser mais fácil de detetar, mas também mais suscetível a remoções não autorizadas. Além disso, a marca de água pode ser vulnerável a alterações não intencionais, como em casos de processamento excessivo do vídeo, e pode resultar na degradação da qualidade do conteúdo (Diwan et al., 2024).

3.3.3. Utilização da tecnologia *Blockchain*

A tecnologia *blockchain*, embora não detete *deepfakes* diretamente, oferece uma infraestrutura segura e descentralizada para garantir a integridade e autenticidade do conteúdo digital. Funciona como um registo imutável e transparente, permitindo rastrear as alterações e a origem dos conteúdos de forma fiável (Hasan & Salah, 2019).

Uma *blockchain* é, essencialmente, uma base de dados distribuída onde as transações são agrupadas em blocos sequenciais, encadeados criptograficamente para garantir a sua imutabilidade. Embora seja frequentemente associada a criptomoedas, a sua funcionalidade pode ser aplicada para autenticar diversos tipos de conteúdo digital, incluindo vídeos. No contexto de vídeos, a *blockchain* pode ser utilizada para registar alterações autorizadas, armazenando resumos criptográficos (*hashes*) de cada versão do conteúdo. O histórico de edições torna-se rastreável e verificável, garantindo que qualquer tentativa de manipulação indevida seja detetada. A proteção da integridade na *blockchain* é garantida pela utilização de assinaturas digitais e pelo encadeamento criptográfico dos blocos. Cada bloco contém uma assinatura digital gerada com base na nova informação desse bloco e no *hash* do bloco anterior. Essa estrutura sequencial assegura que qualquer alteração não autorizada num dos blocos invalida automaticamente as assinaturas digitais dos blocos subsequentes, tornando evidente qualquer adulteração (Hasan & Salah, 2019).

A tecnologia *blockchain* também incorpora o sistema de registo de datas, que arrola de forma cronológica cada transação associada aos vídeos. Esse sistema é também fundamental para a autenticação, pois permite verificar com precisão quando o conteúdo foi criado ou modificado, ajudando a confirmar sua integridade, origem e histórico de alterações.

Ao criar um registo imutável e transparente da proveniência dos conteúdos digitais, a tecnologia *blockchain* oferece uma ferramenta poderosa para combater a disseminação de *deepfakes*. No entanto, a sua implementação nesta deteção enfrenta desafios. Um dos principais é a escalabilidade, devido à complexidade e ao custo elevado com o armazenamento de grandes arquivos de vídeo na *blockchain*, o que limita a sua viabilidade em larga escala (Diwan et al., 2024). Para solucionar esse problema, é comum armazenar apenas os *hashes* dos vídeos na *blockchain*. As questões de privacidade também são motivo de preocupação, especialmente quando o conteúdo envolve informações sensíveis ou dados pessoais. Para mitigar isso, pode ser usada, por exemplo, criptografia para proteger dados confidenciais, possibilitando o acesso apenas a agentes autorizados. Para que a *blockchain* se torne uma solução viável e amplamente utilizada na deteção de *deepfakes*, é essencial superar estas limitações relacionadas tanto com o desempenho como com a privacidade (Diwan et al., 2024; Heidari et al., 2024).

Pode ser igualmente importante complementar a *blockchain* com outras técnicas de deteção de manipulação de vídeo. No artigo *Combating Deepfake Videos Using Blockchain and Smart Contracts*, Hasan e Salah (2019) apresentam algumas soluções concretas para combater a disseminação de *deepfakes*. Propõem o sistema *Proof of Authenticity* (PoA) que, implementado na plataforma *Ethereum*, utiliza a tecnologia *blockchain* para rastrear a origem e o histórico do conteúdo digital. Este sistema baseia-se em contratos inteligentes e no IPFS (*InterPlanetary File System*) para armazenar vídeos e os seus metadados, garantindo rastreabilidade e autenticidade, mesmo após múltiplas edições. Cada vídeo é associado a um contrato inteligente que regista de forma imutável todas as transações e alterações. O IPFS é um sistema de armazenamento descentralizado, em que cada vídeo e os seus metadados são armazenados e endereçados pelo seu *hash*. O *hash* ou "impressão digital" do conteúdo digital, como já foi mencionado, é único, garantindo-se, assim, que qualquer modificação resulta num *hash* diferente. Os metadados referidos, em formato EXIF, incluem informações sobre o dispositivo, configurações, data e hora, e *logs* adicionais, o contrato inteligente e o endereço *Ethereum* do artista. O vídeo e os seus metadados são armazenados no IPFS, e o *hash* gerado pelo IPFS é usado no contrato inteligente. O IPFS permite também armazenar um contrato com termos e condições para cópia e edição do vídeo. Após a criação do *hash* no IPFS, é criado um contrato inteligente na *blockchain Ethereum*, onde são armazenadas informações sobre o vídeo e o seu proprietário. Os contratos de vídeos editados também são ligados ao contrato do vídeo original. O *Ethereum Name Service* (ENS) associa endereços *Ethereum* a identidades legíveis por humanos, capturando a identidade do artista. Uma base de dados externa com um sistema de reputação descentralizado permite avaliar a credibilidade dos artistas.

Embora o estudo tenha o foco nos vídeos, o sistema PoA é genérico e pode ser aplicado a outros formatos de conteúdo digital, como áudios, fotos, imagens e manuscritos. Segundo os autores, a solução PoA atende aos requisitos de segurança e, ao mesmo tempo, é resiliente a ataques cibernéticos conhecidos, mantendo um custo de operação baixo. A segurança é garantida através da natureza imutável da *blockchain*, da utilização de *hashes* criptográficos e da assinatura das transações. A combinação da *blockchain* com contratos inteligentes, IPFS, ENS e sistemas de avaliação da reputação do criador oferece, assim, uma solução promissora para combater a manipulação indevida de vídeos e outros tipos de conteúdo digital, embora seja necessário continuar a melhorar a escalabilidade e a privacidade (Hasan & Salah, 2019).

4. Literacias digital, mediática e informacional

Conhecer todas estas tecnologias e estratégias de deteção representa um verdadeiro desafio para o cidadão comum. As literacias digital, mediática e informacional são fundamentais para o capacitar a navegar no mundo digital de forma crítica e segura, especialmente diante dos desafios colocados pelos *deepfakes* criados ou utilizados com fins maliciosos (Alencar et al., 2022; Garriga et al., 2024; Martínez et al., 2020; Paisana et al., 2024). Em complemento, a literacia em IA e dados emerge como uma área fundamental ao fornecer ferramentas para compreender e questionar o papel dos algoritmos e da análise de dados na produção e disseminação de *deepfakes* (McCosker, 2024).

Relatórios como o da ACCO (*Alliance to Counter Crime Online*) reforçam a importância da consciencialização e da educação para ajudar as pessoas a identificar e evitar golpes e manipulações que usam tecnologias como os *deepfakes* (Peters, 2024, p. 19). A finalidade é incentivar os cidadãos a analisar criticamente as informações que consomem e a refletir sobre o conteúdo que escolhem partilhar *online*, além de capacitá-los para identificar fontes de informação fidedignas e a saber onde procurar apoio em caso de dúvida. Um vídeo disponibilizado pela *Deutsche Telekom*, uma companhia de telecomunicações alemã, intitulado "A message from Ella without consent"[17], serve como um alerta claro sobre a necessidade de cautela ao partilhar conteúdos *online*. Ele ilustra como os *deepfakes* podem impactar profundamente a nossa segurança, bem-estar e confiança na informação.

Em Portugal, a promoção destas literacias conta com a participação de diversos atores, desde órgãos do Estado a universidades, centros de investigação, centros de formação e órgãos de comunicação social (Paisana et al., 2024). É essencial reconhecer que a rápida evolução tecnológica impõe desafios constantes a estes atores, exigindo uma adaptação contínua das suas estratégias e práticas. Reconhecendo a importância do investimento em literacia sobre IA e competências digitais, o relatório da SAPEA (*Science Advice for Policy by European Academies*) de 2024 destaca o papel fundamental das universidades e da comunidade científica na educação dos indivíduos. Este investimento deve abranger não só os aspetos tecnológicos, mas também a dimensão ética da utilização da IA, preparando os cidadãos para os desafios colocados pela desinformação e manipulação (SAPEA, 2024, pp. 14–15). Neste contexto, torna-se ainda mais relevante a promoção destas literacias junto de estudantes de instituições como a FEUP ou a FLUP, duas unidades orgânicas da Universidade do Porto, à qual o Mestrado em Ciência da Informação está associado.

[17] Apêndice 1, B.5.2.

Com o objetivo de contribuir para essa promoção, foi compilada uma seleção de recursos, disponíveis no Apêndice 1 e organizados de acordo com os capítulos deste relatório, que podem ser utilizados em ações de formação e sensibilização sobre essas questões.

Estas ações e formações podem eventualmente ser desenvolvidas pelas bibliotecas das respetivas faculdades em colaboração com investigadores de Ciência da Informação e de Engenharia Informática. As bibliotecas universitárias têm assumido um papel cada vez mais ativo no combate à desinformação e na promoção do pensamento crítico entre os estudantes universitários[18]. Esta preocupação deve ser partilhada pelas bibliotecas escolares, que, através dos profissionais da informação, podem oferecer um apoio fundamental a professores e alunos de todos os ciclos de ensino. A colaboração entre bibliotecas universitárias e escolares poderá potenciar ainda mais a disseminação destas literacias, criando uma rede de apoio e aprendizagem que abranja toda a comunidade educativa.

[18] Como, por exemplo, no projeto *Literacia da Informação e pensamento crítico no Ensino Superior: combater a desinformação* e no *Referencial da Literacia da Informação para o Ensino Superior* (Sanches et al., 2022).

Conclusões

Os *deepfakes* representam um fenómeno complexo e desafiador no panorama digital atual. Embora possuam um potencial positivo, a sua capacidade de manipular as perceções exige uma resposta coordenada e abrangente da sociedade.

As literacias digital, mediática e informacional são fundamentais para capacitar os cidadãos a enfrentar esta nova ameaça. É essencial que os indivíduos compreendam os mecanismos de criação dos *deepfakes*, aprendendo a identificar, analisar e avaliar criticamente a informação que esses vídeos contêm. Além disso, a verificação rigorosa dos factos e a análise crítica do contexto dos vídeos são fundamentais para mitigar os riscos de engano.

É importante reconhecer as limitações dos métodos de deteção atuais — com ou sem recurso a IA — e destacar a crescente dificuldade de diferenciar a realidade do conteúdo manipulado, devido aos avanços da IA generativa. A adoção generalizada de mecanismos para proteger a autenticidade e a integridade da informação, bem como o desenvolvimento de novos mecanismos para superar as limitações existentes e atrás enunciadas, pode ser fundamental para reduzir a proliferação de *deepfakes*. Nesse contexto, a literacia digital desempenha um papel essencial, capacitando a população para utilizar essas ferramentas de forma crítica. A consciencialização sobre a importância da literacia digital deve estar acompanhada de uma mudança de atitude por parte dos cidadãos, promovendo o cultivo do espírito crítico, o questionamento das fontes de informação e a cautela ao ver e partilhar conteúdos sem verificação.

A construção de um ambiente digital mais seguro e confiável depende da participação ativa de todos os atores da sociedade — cidadãos, plataformas de comunicação social, governos e instituições. Para resolver um problema tão complexo, é necessário que várias estratégias se articulem para proteger as possíveis vítimas do uso malicioso de *deepfakes*. É fundamental reconhecer que, embora as literacias digitais, mediáticas e informacionais sejam ferramentas poderosas no combate à desinformação e à fraude, elas por si só não protegem todas as vítimas.

Embora não tenha sido esse o foco deste mini-projeto, é importante referir a regulamentação e a criação de políticas éticas para a utilização de *deepfakes* (Peters, 2024, p. 19). A este nível, os governos e as instituições têm a responsabilidade de legislar e de fiscalizar o cumprimento da lei. Também as plataformas de comunicação social têm a responsabilidade de implementar mecanismos de deteção e verificação de factos, além de adotar diretrizes éticas que promovam o uso responsável destas tecnologias para proteger os cidadãos dos potenciais danos causados pelos *deepfakes*.

Referências Bibliográficas

Abbas, F., & Taeihagh, A. (2024). Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. *Expert Systems with Applications*, 252, Part B. https://doi.org/10.1016/j.eswa.2024.124260

Aberna, P., & Agilandeeswari, L. (2024). Digital image and video watermarking: Methodologies, attacks, applications, and future directions. *Multimedia Tools and Applications*, 83(2), 5531–5591. https://doi.org/10.1007/s11042-023-15806-y

Alencar, A. P., Marques, J. F., Schneider, M., & Alves, E. C. (2022). Competência Crítica em Informação e Educomunicação: Proposta interdominial no combate à desinformação. *Palabra Clave (La Plata)*, 11(2), e153. https://doi.org/10.24215/18539912e153

Campbell, C., Plangger, K., Sands, S., & Kietzmann, J. (2022). Preparing for an Era of Deepfakes and AI-Generated Ads: A Framework for Understanding Responses to Manipulated Advertising. *Journal of Advertising*, 51(1), 22–38. https://doi.org/10.1080/00913367.2021.1909515

Chanda, K., Ahmed, W., & Banik, S. (2024). Deepfake Image Forgery Detection for Suspicious Images. In S. Tanwar, P. K. Singh, M. Ganzha, & G. Epiphaniou (Eds.), *Proceedings of Fifth International Conference on Computing, Communications, and Cyber-Security. IC4S 2023. Lecture Notes in Networks and Systems*, vol. 991 (pp. 891–904). Springer. https://doi.org/10.1007/978-981-97-2550-2_63

Chawki, M. (2024). Navigating legal challenges of deepfakes in the American context: A call to action. *Cogent Engineering*, 11(1), 1–13. https://doi.org/10.1080/23311916.2024.2320971

Cheres, I., & Groza, A. (2023). The profile: Unleashing your deepfake self. *Multimedia Tools Applications*, 82, 31839–31854. https://doi.org/10.1007/s11042-023-14568-x

Danry, V., Leong, J., Pataranutaporn, P., Tandon, P., Liu, Y., Shilkrot, R., Punpongsanon, P., Weissman, T., Maes, P., & Sra, M. (2022). AI-Generated Characters: Putting Deepfakes to Good Use. In Barbara, S., Lamp, C., & Appert, C. (Eds.), *Extended Abstracts of CHI '22: CHI Conference on Human Factors in Computing Systems* (pp. 1–5). Association for Computing Machinery. https://doi.org/10.1145/3491101.3503736

Diwan, A., Dixit, S., Subbiah, R., & Mahadeva, R. (2024). Systematic analysis of video tampering and detection techniques. *Cogent Engineering*, 11(1). https://doi.org/10.1080/23311916.2024.2424466

Edwards, P., Nebel, J.-C., Greenhill, D., & Liang, X. (2024). A Review of Deepfake Techniques: Architecture, Detection, and Datasets. *IEEE Access*, 12, 154718–154742. https://doi.org/10.1109/ACCESS.2024.3477257

Garriga, M., Ruiz-Incertis, R., & Magallón-Rosa, R. (2024). Artificial intelligence, disinformation and media literacy proposals around deepfakes. *Observatorio (OBS*)*, 18(5), 175–194. https://doi.org/10.15847/obsOBS18520242445

Haddad, Ana Carolina. (2023, julho 5). Elis recriada para vender carro de empresa que apoiou a ditadura, o que pode ter de errado? *Brasil de Fato*. https://www.brasildefato.com.br/2023/07/05/elis-recriada-para-vender-carro-de-empresa-que-apoiou-a-ditadura-o-que-pode-ter-de-errado

Hasan, H. R., & Salah, K. (2019). Combating Deepfake Videos Using Blockchain and Smart Contracts. *IEEE Access*, 7, 41596–41606. https://doi.org/10.1109/ACCESS.2019.2905689

Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs: Data Mining & Knowledge Discovery*, 14(2), 1–45. https://doi.org/10.1002/widm.1520

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *International Journal of Computer Vision*, 130(7), 1678–1734. https://doi.org/10.1007/s11263-022-01606-8

Kaswan, K. S., Malik, K., Dhatterwal, J. S., Naruka, M. S., & Govardhan, D. (2023). Deepfakes: A Review on Technologies, Applications and Strategies. *2023 International Conference on Power Energy, Environment & Intelligent Control (PEEIC)*, 292–297. https://doi.org/10.1109/PEEIC59336.2023.10450604

Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., & Xia, F. (2024). Deepfake video detection: Challenges and opportunities. *Artificial Intelligence Review*, 57(6), 159. https://doi.org/10.1007/s10462-024-10810-6

Kaur, R. & Kaur, A. (2012). Digital Signature. *2012 International Conference on Computing Sciences*, 295–301. https://doi.org/10.1109/ICCS.2012.25

Kwow, A. O. J., & Koh, S. G. M. (2021). Deepfake: A social construction of technology perspective. *Current Issues in Tourism*, 24(13), 1798–1802. https://doi.org/10.1080/13683500.2020.1738357

Martínez, V. C., Guardia, M. L. G., & Castillo, G. P. (2020). Alfabetización moral digital para la detección de deepfakes y fakes audiovisuales. *CIC. Cuadernos de Información y Comunicación*, 25, 165–181. https://doi.org/10.5209/ciyc.68762

Masood, M., Nawaz, M., Malik, K., Javed, A., & Irtaza, A. (2021). Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. *arXiv*. https://doi.org/10.48550/arXiv.2103.00484

Murphy, G., Ching, D., Twomey, J., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *PLOS ONE*, 18(7), e0287503. https://doi.org/10.1371/journal.pone.0287503

Nassif, A. B., Nasir, Q., Talib, M. A., & Gouda, O. M. (2022). Improved Optical Flow Estimation Method for Deepfake Videos. *Sensors*, 22(7), 2500. https://doi.org/10.3390/s22072500

Paisana, M., Foa, C., Vasconcelos, A., Couraceiro, P., Santos, S. F., Margato, A., & Crespo, M. (2024). Uma taxonomia para a literacia para os media em Portugal—Caracterização de atores, iniciativas e linhas de intervenção. *Observatorio (OBS*)*, 18(5), 58–82. https://doi.org/10.15847/obsOBS18520242439

Patel, Y., Tanwar, S., Gupta, R., Bhattacharya, P., Davidson, I. E., Nyameko, R., Aluvala, S., & Vimal, V. (2023). Deepfake Generation and Detection: Case Study and Challenges. *IEEE Access*, 11, 143296–143323. https://doi.org/10.1109/ACCESS.2023.3342107

Peters, G. (2024). *Deep Fake Frauds. When You Lose Trust in Your Own Eyes & Ears*. Alliance to Counter Crime Online.

Rajput, T., & Arora, B. (2024). A Systematic Review of Deepfake Detection Using Learning Techniques and Vision Transformer. In Tanwar, S., Singh, P.K., Ganzha, M., Epiphaniou, G. (Eds.), *Proceedings of Fifth International Conference on Computing, Communications, and Cyber-Security. IC4S 2023. Lecture Notes in Networks and Systems*, vol. 991 (pp. 217–235). Springer. https://doi.org/10.1007/978-981-97-2550-2_17

Rancourt-Raymond, A. de, & Smaili, N. (2023). The unethical use of deepfakes. *Journal of Financial Crime*, *30*(4), 1066–1077. https://doi.org/10.1108/JFC-04-2022-0090

Rehaan, M., Kaur, N., & Kingra, S. (2024). Face manipulated deepfake generation and recognition approaches: A survey. *Smart Science*, *12*(1), 53–73. https://doi.org/10.1080/23080477.2023.2268380

Renier, L. A., Shubham, K., Vijay, R. S., Mishra, S. S., Kleinlogel, E. P., Jayagopi, D. B., & Schmid Mast, M. (2024). A deepfake-based study on facial expressiveness and social outcomes. *Scientific Reports*, *14*(1). https://doi.org/10.1038/s41598-024-53475-5

Roe, J., Perkins, M., & Furze, L. (2024). Deepfakes and higher education: A research agenda and scoping review of synthetic media. *Journal of University Teaching and Learning Practice*, *21*(10), 1–22. https://doi.org/10.53761/2y2np178

Sanches, T., Antunes, M. da L., & Lopes, C. (2022). *Referencial da literacia da informação para o ensino superior: Versão portuguesa*. BAD. http://hdl.handle.net/10451/57509

Sandotra, N., & Arora, B. (2024). A comprehensive evaluation of feature-based AI techniques for deepfake detection. *Neural Computing and Applications*, *36*(8), 3859–3887. https://doi.org/10.1007/s00521-023-09288-0

SAPEA. (2024). *Successful and timely uptake of artificial intelligence in science in the EU: Evidence review report*. SAPEA. DOI:10.5281/zenodo.10849580. https://scientificadvice.eu/advice/artificial-intelligence-in-science/

Sharma, S., Zou, J. J., Fang, G., Shukla, P., & Cai, W. (2024). A review of image watermarking for identity protection and verification. *Multimedia Tools and Applications*, 83(11), 31829–31891. https://doi.org/10.1007/s11042-023-16843-3

Stanciu, A., & Ciuperca, E.-M. (2024). Can Deepfakes Benefit the Metaverse in an Era of Disinformation? Insights from a Systematic Review. *IFAC-PapersOnLine*, 58(3), 61–65. https://doi.org/10.1016/j.ifacol.2024.07.125

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech

misinformation. *Media and Communication*, 9(1), 291–300. https://doi.org/10.17645/MAC.V9I1.3494

Xia, R., Zhou, D., Liu, D., Yuan, L., Wang, S., Li, J., Wang, N., & Gao, X. (2024). Advancing Generalized Deepfake Detector with Forgery Perception Guidance. In Cai, J., & Kankanhalli, M. (Eds.), *MM'24: Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 6676–6685). Association for Computing Machinery. https://doi.org/10.1145/3664647.3680713

Zhang, B., Zhou, J. P., Shumailov, I., & Papernot, N. (2021). On Attribution of Deepfakes. *arXiv*. https://doi.org/10.48550/arXiv.2008.09194

Apêndices

Apêndice 1 – Recursos informacionais

A. O que são deepfakes – conteúdos gerais

A.1. Notícia da NBC News: *New warnings about A.I.-generated video deep fakes*, https://www.youtube.com/watch?v=yjRZmWs8Ji8

A.2. *Unmask The DeepFake: Defending Against Generative AI Deception*,

https://www.youtube.com/watch?v=cVvJgdm19Ak

B. Aplicações positivas, aplicações potencialmente positivas (mas que podem gerar polémica) e aplicações negativas

B.1. Usos no cinema

B.1.1. No filme *O Irlandês*, para rejuvenescimento de atores:

- *De-aging Robert Deniro in The Irishman [DeepFake]*: https://www.youtube.com/watch?v=dHSTWepkp_M
- *De-Aging Al Pacino in The Irishman [DeepFake]*: https://www.youtube.com/watch?v=MiTxoJ4sBfY
- Crítica: *The Problem with De-Aging and the Irishman*, https://www.youtube.com/watch?v=4B5VOd533tk

B.1.2. Documentário *Welcome to Chechnya*, para a proteção da identidade de ativistas, testemunhas e vítimas da perseguição a homossexuais na Chechénia:

- Notícia publicada no *The New York Times*: https://www.nytimes.com/2020/07/01/movies/deepfakes-documentary-welcome-to-chechnya.html
- *Trailer* oficial do documentário: https://www.youtube.com/watch?v=_2KMm49B6pE
- Conversa com o realizador do documentário que foi convidado para o episódio *Deepfakery* do *MIT Open Documentary Lab Fall 2020 lecture series*: *David France | Identity protection with deepfakes,* https://opendoclab.mit.edu/presents/david-france-identity-protection-deepfakes/

B.2. Usos na publicidade e marketing

B.2.1. Campanha publicitária *Volkswagen 70 anos*, com recriação da falecida cantora Elis Regina:

- Vídeo: https://www.youtube.com/watch?v=aMl54-kqphE&t=120s
- Notícia: https://www.nit.pt/cultura/musica/ia-ressuscita-elis-regina-para-cantar-com-a-filha-tinha-5-anos-quando-a-mae-morreu
- Polémica: https://www.cnnbrasil.com.br/nacional/conar-abre-representacao-etica-contra-propaganda-da-volkswagen-com-elis-regina/ ; https://www.brasildefato.com.br/2023/07/05/elis-recriada-para-vender-carro-de-empresa-que-apoiou-a-ditadura-o-que-pode-ter-de-errado

B.2.2. *The Dove Self-Esteem Project - Dove's Toxic Advice with deepfakes*, para consciencializar as mães das mensagens de beleza tóxica a que as filhas são expostas nas redes sociais:

- *Dove Toxic Influence: Mothers & Daughters Confront Toxic Social Media*, https://www.youtube.com/watch?v=sF3iRZtkyAQ&t=168s
- *The Dove Self-Esteem Project*: https://www.dove.com/uk/dove-self-esteem-project.html

B.3. Usos na Arte e nos Museus

B.3.1. *Deepfake* do falecido pintor Salvador Dalí, no Museu Dalí em São Petersburgo, Flórida (EUA):

- *Deepfake* em vídeo: *Dalí Lives – Art Meets Artificial Intelligence*: https://www.youtube.com/watch?v=mPtcU9VmIIE
- Notícia sobre a criação deste *deepfake*, em 2019, disponível em https://www.dezeen.com/2019/05/24/salvador-dali-deepfake-dali-musuem-florida/

B.3.2. Instalação artística *The Profile*, baseada em tecnologias que geram *deepfakes*:

- Vídeo com explicação da artista: https://www.youtube.com/watch?v=RgU3i4qAVaI
- Artigo onde é apresentado e explicado este trabalho artístico: Cheres & Groza, 2023.

B.4. Usos desinformativos e manipuladores

B.4.1. *Deepfake* de Zelensky pedindo rendição das tropas ucranianas, 2022:

- *Prova dos factos*, no *Público*: https://www.publico.pt/2022/03/17/mundo/noticia/video-zelenskii-anunciar-rendicao-ucrania-verdadeiro-1999129

- Vídeo que foi disseminado nas redes sociais: https://www.youtube.com/watch?v=X17yrEV5sl4

B.5. Usos criminosos e fraudulentos

B.5.1. Relatório da *Alliance to Counter Crime Online*: https://static1.squarespace.com/static/6451b3769ff664542bf67ca9/t/671016a5066d79214960a497/1729107680086/ACCO+-+Deep+Fake+Frauds+Report+-+FINAL+Oct+2024.pdf

B.5.2 No arquivo da *Deutsche Telekom*:

- *A Message from Ella Without Consent*, https://www.youtube.com/watch?v=F4WZ_k0vUDM

B.5.3. Exploração sexual:

- *The most urgent threat of deepfakes isn't politics*: https://www.youtube.com/watch?v=hHHCrf2-x6w

C. Criação de *deepfakes*

C.1. Explicação e demonstração do processo de criação

- *The Deekfake Lab*: https://deepfakelab.theglassroom.org/index-es_ES.html
- *Everybody Dance Now*: https://www.youtube.com/watch?v=PCBTZh41Ris&t=37s

C.2. Exemplos de *deepfakes*

- *This Person Does Not Exist*: https://thispersondoesnotexist.com/ (sempre que se atualiza a página aparece uma nova imagem de uma pessoa que não existe)

C.3. Ferramentas para a criação de *deepfakes*

- Deepface Lab: https://github.com/iperov/DeepFaceLab
- Deepfakes: https://deepfakesweb.com/
- Face Swap: https://faceswap.dev/
- Style-GAN: https://github.com/NVlabs/stylegan

D. Deteção de *deepfakes* com IA

D.1. Diretório de ferramentas

- Detector Tools: https://detectortools.ai/

D.2. Plataformas grátis para deteção de *deepfakes*

- Deepware: https://deepware.ai/
- DeepFake-o-Meter: https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/landing_page
- TrueMedia.org: https://detect.truemedia.org

D.3. Treino em deteção de *deepfakes*

- *Detect Fakes*: https://detectfakes.kellogg.northwestern.edu/

E. Outras técnicas de deteção de *deepfakes*

E1. Marcas de água. Alguns artigos:

- *Big Tech says AI watermarks could curb misinformation, but they're easy to sidestep*: https://www.nbcnews.com/tech/tech-news/watermark-deepfake-solution-ai-misinformation-cant-stop-de-rcna137370
- *Google joins AI watermarking coalition as deepfakes hit mainstream tech platforms*: https://www.nbcnews.com/tech/tech-news/google-joins-ai-watermarking-coalition-deepfakes-hit-mainstream-tech-p-rcna137368
- *First 'certified' deepfake warns viewers not to trust everything they see online*: https://www.thetimes.com/article/first-certified-deepfake-warns-viewers-not-to-trust-everything-they-see-online-kkvctk5kt

E.2. Tecnologia *blockchain*. Alguns vídeos e artigos:

- *How Blockchain Can Be Used To Stop The Proliferation Of Deepfakes*: https://www.youtube.com/watch?v=QfprgikVW0I
- *AI And Blockchain Can Mitigate Fraud Risk Caused By Deepfakes*: https://www.forbes.com/sites/digital-assets/2024/07/06/ai-and-blockchain-synergies-mitigate-risk-of-deepfakes-in-kyc/
- *How blockchain data storage can protect us from deepfakes*: https://www.weforum.org/stories/2023/09/how-blockchain-can-protect-us-again-ai-threats/

- *Blockchain can help combat the threat of deepfakes. Here's how*: https://www.weforum.org/stories/2021/10/how-blockchain-can-help-combat-threat-of-deepfakes/

Apêndice 2 – Testes no Deepware, Deepfake-O-Meter e TrueMedia.org

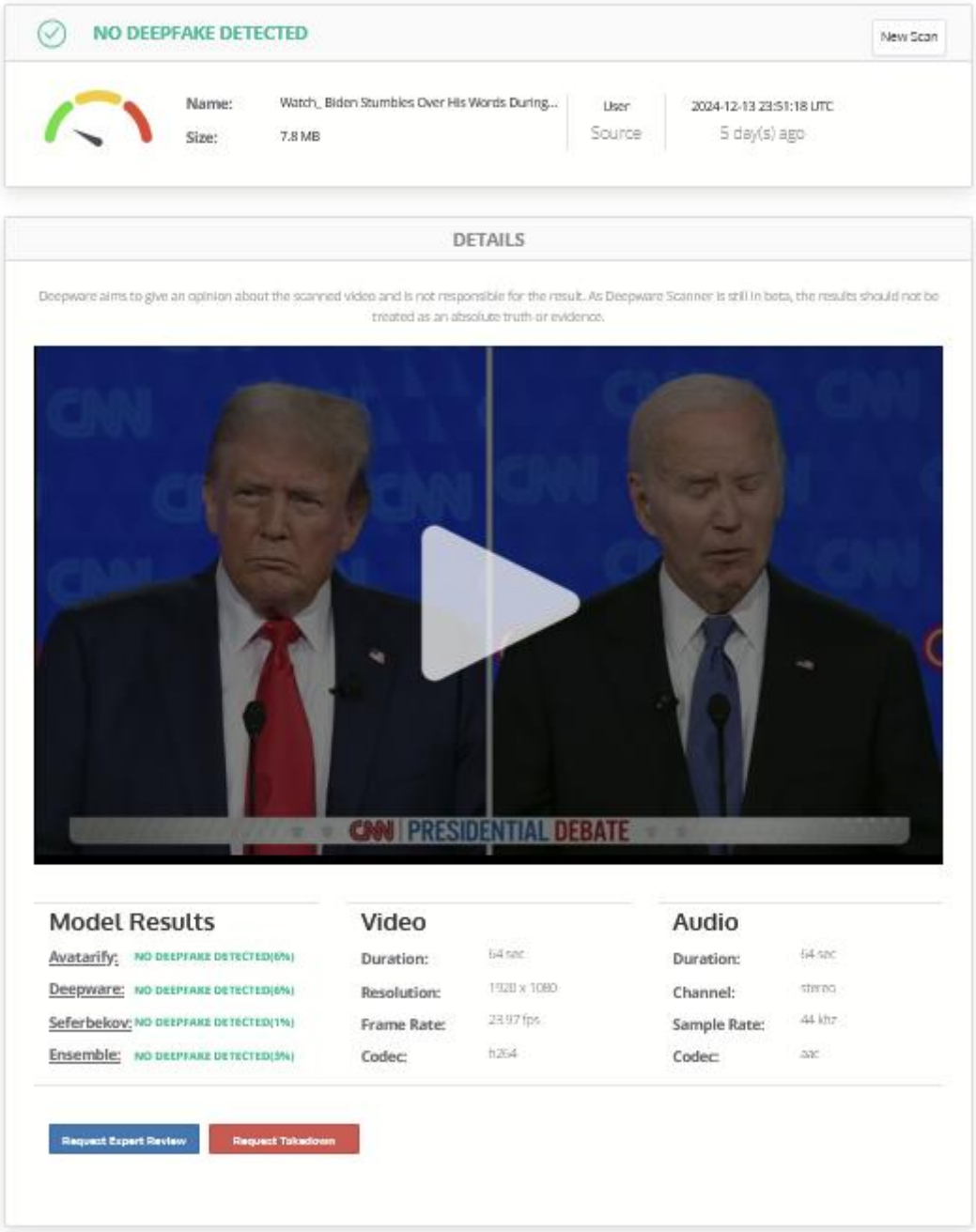

Figura 1- Análise de vídeo autêntico no Deepware

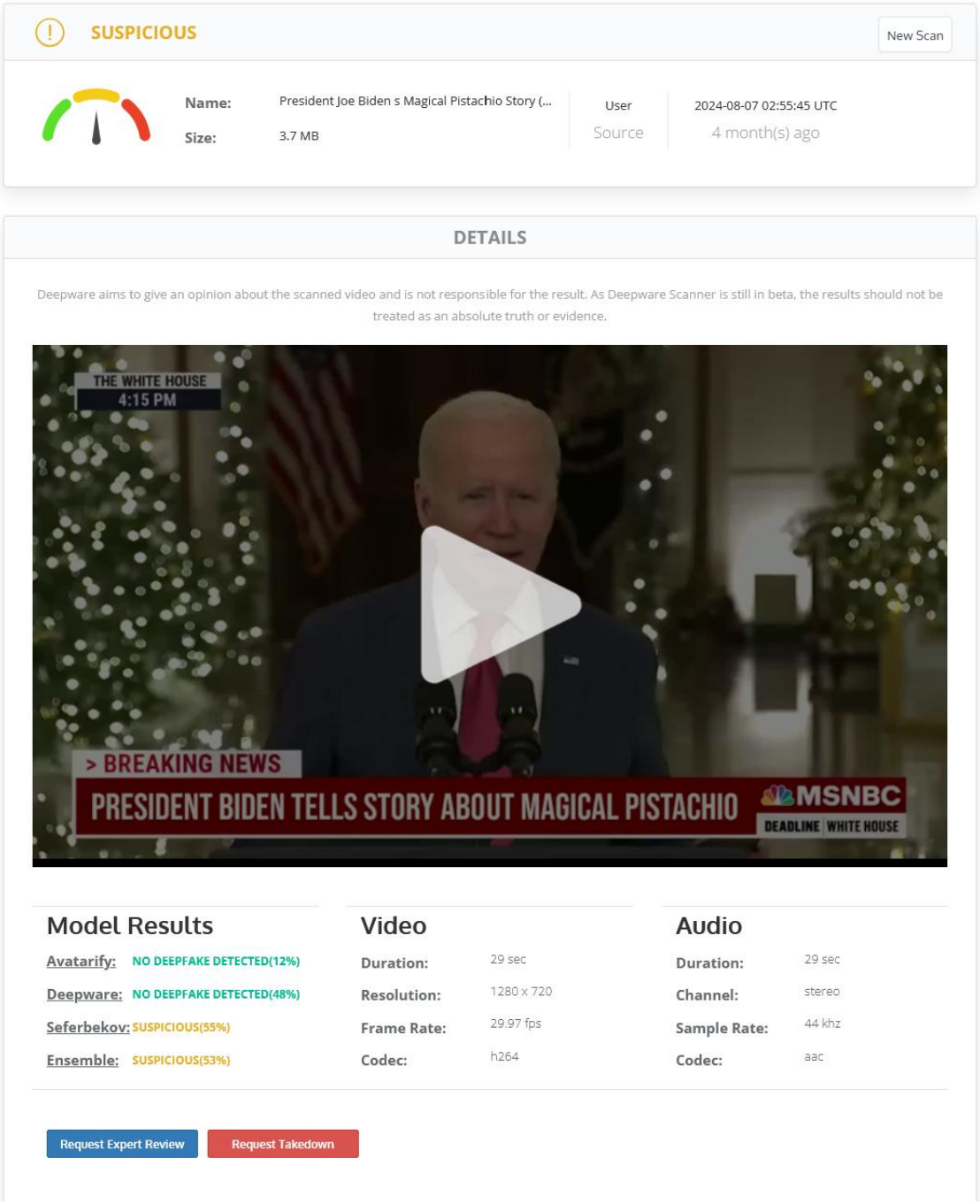

Figura 2- Análise de *deepfake* no Deepware

Detector	Result	Details
AltFreezing (2023)	12.2% AI-Generated Likelihood	View detail
DSP-FWA (2019)	27.5% AI-Generated Likelihood	View detail
FTCN (2021)	1.4% AI-Generated Likelihood	View detail
LIPINC (2024)	100.0% AI-Generated Likelihood	View detail
LSDA (2024)	13.4% AI-Generated Likelihood	View detail
SBI (2022)	4.5% AI-Generated Likelihood	View detail
WAV2LIP-STA (2022)	7.8% AI-Generated Likelihood	View detail

Figura 3- Análise de vídeo autêntico no Deepfake-O-Meter

Detector	Result	Details
AltFreezing (2023)	18.3% AI-Generated Likelihood	View detail
DSP-FWA (2019)	97.5% AI-Generated Likelihood	View detail
FTCN (2021)	1.1% AI-Generated Likelihood	View detail
LIPINC (2024)	100.0% AI-Generated Likelihood	View detail
LSDA (2024)	48.8% AI-Generated Likelihood	View detail
SBI (2022)	17.6% AI-Generated Likelihood	View detail
WAV2LIP-STA (2022)	89.1% AI-Generated Likelihood	View detail

Figura 4- Análise de vídeo *deepfake* no Deepfake-O-Meter

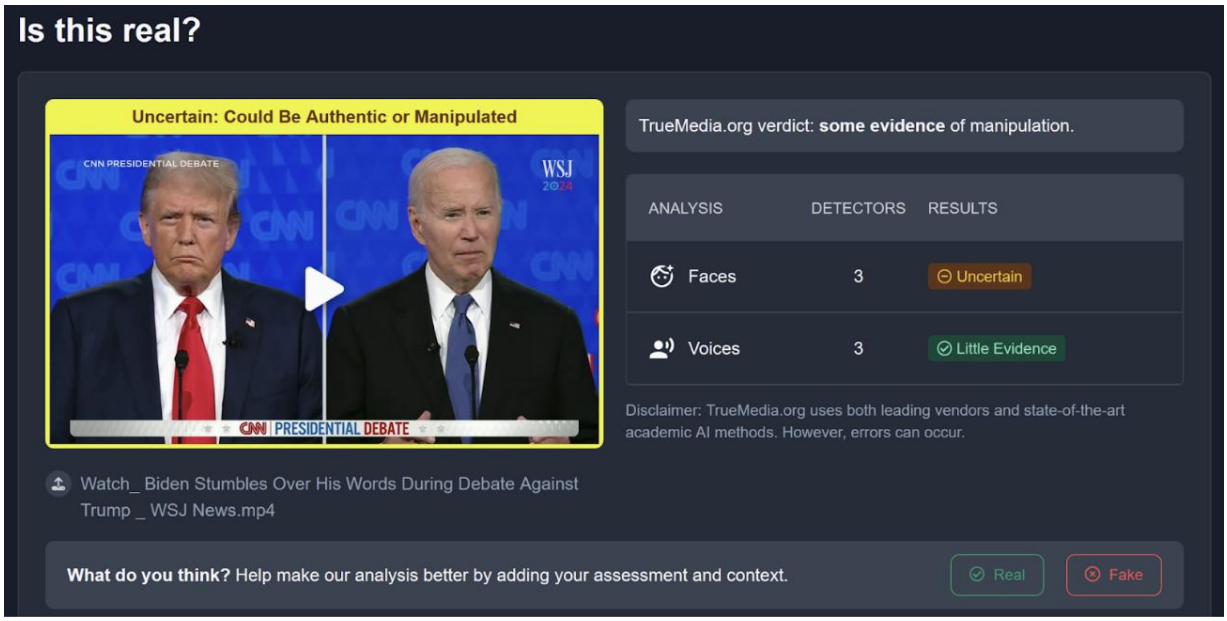

Figura 5- Análise de vídeo autêntico no TrueMedia.org

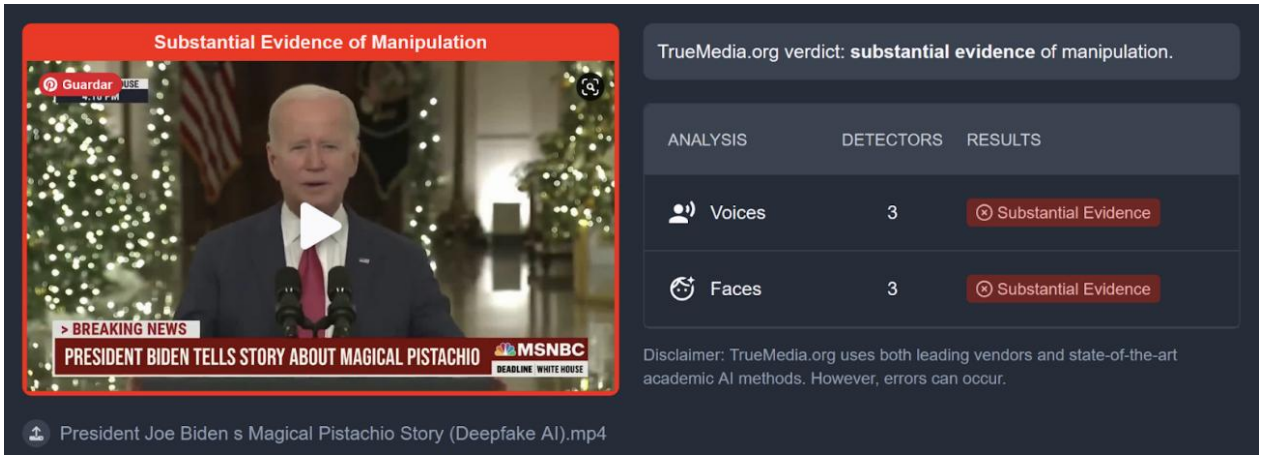

Figura 6- Análise de *deepfake* no TrueMedia.org