JNT FACIT BUSINESS AND TECHNOLOGY JOURNAL
ISSN: 2526-4281 - ANO 2026 - MÊS DE MAIO
Ed. 74. Vol. 2 - Págs. 331-348
DOI: 10.5281/zenodo.20396919

CAPES

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

QUALIS A2

REVISÃO SISTEMÁTICA E COMPARAÇÃO DE MODELOS TRANSFORMADORES NA DETECÇÃO DE DEEPFAKE[1]

SYSTEMATIC REVIEW AND COMPARISON OF TRANSFORMER MODELS IN DEEPFAKE DETECTION

Esther Mota REIS
Universidade Estadual do Tocantins (UNITINS)
E-mail: esthermotareis@gmail.com/estherreis@unitins.br
ORCID: http://orcid.org/0009-0002-3118-3009

Alex COELHO
Universidade Estadual do Tocantins (UNITINS)
E-mail: alex.c@unitins.br
ORCID: http://orcid.org/0000-0002-1576-7242

RESUMO

Este estudo realiza uma revisão sistemática, baseada no protocolo PRISMA (Page et al, 2021), e uma análise experimental de modelos baseados em Transformers aplicados à detecção de deepfakes. O mapeamento da literatura indicou a predominância de abordagens baseadas em atenção e métodos híbridos na análise espacial, espaço-temporal e multimodal. Além disso, identificou limitações recorrentes, como a sensibilidade à compressão das mídias e a dificuldade de generalização dos modelos entre diferentes bases de dados. A partir desses achados, o trabalho estruturou uma etapa prática para comparar o desempenho de arquiteturas transformadoras selecionadas sob um protocolo experimental padronizado. Ao integrar a revisão teórica e a análise empírica, o estudo busca fornecer subsídios para o desenvolvimento de soluções mais robustas contra a desinformação digital.

Palavras-chave: Deepfake Detection. Deep Learning. Inteligência Artificial. Modelos Transformadores. Revisão Sistemática. Vision Transformer. Visão Computacional.

ABSTRACT

This study conducts a systematic review, based on the PRISMA protocol (PAGE et al., 2021), and an experimental analysis of Transformer-based models applied to

[1] COMO CITAR: (ABNT): REIS, E. M.; COELHO, A. Revisão Sistemática e Comparação de Modelos Transformadores na Detecção de Deepfake. **JNT Facit Business and Technology Journal.** Qualis A2. ISSN: 2526-4281, Mês de Maio de 2026 - Ed. 74. VOL. 02. Págs.331-348. Disponível: http://revistas.faculdadefacit.edu.br. Acesso em: ___/___/___.

deepfake detection. The literature mapping indicated the predominance of attention-based approaches and hybrid methods in spatial, spatiotemporal, and multimodal analysis. In addition, it identified recurring limitations, such as sensitivity to media compression and the difficulty of model generalization across different datasets. Based on these findings, the work structured a practical stage to compare the performance of selected Transformer architectures under a standardized experimental protocol. By integrating the theoretical review and the empirical analysis, the study aims to provide support for the development of more robust solutions against digital misinformation.

Keywords: Deepfake Detection. Deep Learning. Artificial Intelligence. Transformer Models. Systematic Review. Vision Transformer. Computer Vision.

INTRODUÇÃO

A rápida evolução da Inteligência Artificial (IA) e do Aprendizado Profundo (Deep Learning) revolucionou a geração de conteúdo multimídia sintético. Entre essas inovações, destacam-se as deepfakes, mídias manipuladas com alto grau de realismo produzidas principalmente por meio de arquiteturas como Autoencoders e Redes Adversárias Generativas (GANs) (Goodfellow et al, 2014; Rossler et al, 2019). Embora possuam aplicações legítimas, a facilidade de criação e o realismo dessas mídias impõem sérios riscos associados à desinformação, fraudes digitais e manipulação da opinião pública, tornando a autenticação de conteúdos um desafio crítico para a segurança da informação (Chesney; Citron, 2019).

Diante desse cenário, a detecção de deepfakes consolidou-se como uma área central da Visão Computacional (Verdoliva, 2020), disciplina que utiliza o aprendizado de máquina para extrair padrões e reconhecer anomalias imperceptíveis ao olho humano em imagens e vídeos. Inicialmente, as Redes Neurais Convolucionais (CNNs) dominaram esse campo de detecção (Patel et al, 2023; Edwards et al, 2024). Contudo, a crescente sofisticação das falsificações tem exposto limitações significativas dessas redes, especialmente no que diz respeito à generalização entre diferentes bases de dados (cross-dataset) e à robustez contra compressões e variações de qualidade da mídia (Mirsky; Lee, 2021; Rana et al, 2022).

Para superar essas limitações, arquiteturas baseadas em mecanismos de atenção, em especial os Modelos Transformadores (Transformers), têm ganhado destaque (Vaswani et al, 2017; Dosovitskiy et al, 2021; Liu et al, 2021). Diferentemente de modelos anteriores, os Transformadores convertem informações

visuais em representações vetoriais (embeddings) e processam os dados de forma paralela, utilizando mecanismos de autoatenção para modelar relações globais e dependências de longo alcance, permitindo que o modelo identifique sutis inconsistências contextuais e dinâmicas que escapam a abordagens estritamente locais. Na visão computacional, essa capacidade permite que o modelo identifique sutis inconsistências espaciais, temporais e contextuais deixadas pelos algoritmos de manipulação, oferecendo recursos analíticos superiores para detectar incoerências na dinâmica facial ou textura das imagens (Acheampong et al, 2020; Mohammed; Kora, 2025).

Apesar do notório potencial dos Transformadores, a literatura científica ainda carece de sistematização sobre quais técnicas baseadas nessa arquitetura são mais eficazes e como esses modelos se comportam sob condições experimentais rigorosas e padronizadas. A ausência de avaliações comparativas consistentes dificulta a compreensão real sobre a capacidade de generalização e a robustez das abordagens propostas.

Nesse contexto, a relevância da investigação não está apenas em verificar se modelos transformadores alcançam bons resultados em bases específicas, mas em compreender até que ponto esses resultados se sustentam diante de mudanças de domínio. Em aplicações reais, conteúdos manipulados podem apresentar diferentes níveis de compressão, iluminação, resolução, enquadramento facial e técnicas de síntese, o que torna insuficiente avaliar os modelos apenas em ambientes controlados. Assim, estudos que combinem revisão sistemática e experimentação prática contribuem para aproximar a análise acadêmica dos desafios encontrados em cenários concretos de verificação de autenticidade digital.

Para preencher essa lacuna, este estudo tem como objetivo identificar as principais técnicas de detecção de deepfakes baseadas em Transformers por meio de uma revisão sistemática da literatura, fundamentada nas diretrizes do protocolo PRISMA (PAGE et al., 2021). Complementarmente, a pesquisa conduz um estudo experimental estruturado para reproduzir, analisar e comparar o desempenho dos principais modelos transformadores (em cenários de avaliação interna e cross-dataset), oferecendo subsídios e evidências empíricas para o desenvolvimento de soluções forenses mais robustas no combate à desinformação.

METODOLOGIA

A presente pesquisa caracteriza-se como aplicada, exploratória e descritiva, conduzida com uma abordagem quali-quantitativa. O estudo foi estruturado em duas

etapas complementares fundamentais: a realização de uma Revisão Sistemática da Literatura (RSL) para identificar o estado da arte, seguida por um estudo experimental focado na reprodução e comparação de modelos de detecção.

Fluxo Metodológico

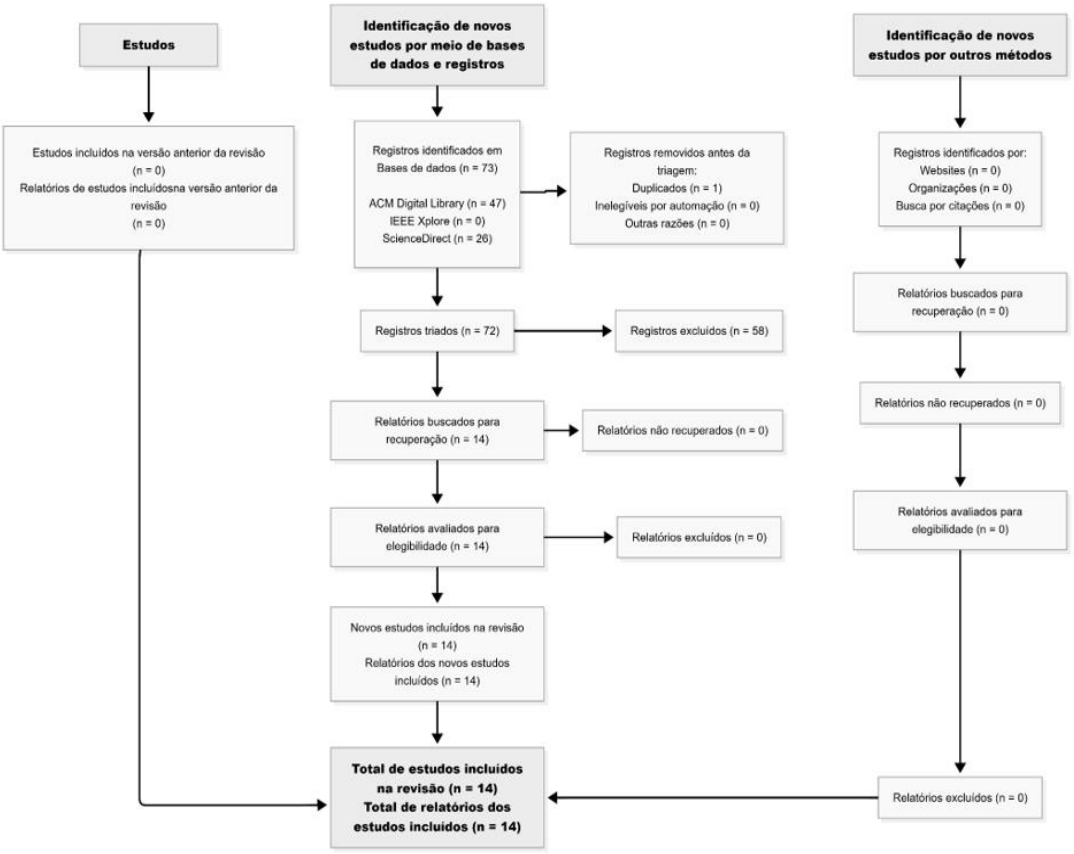
O desenvolvimento da pesquisa obedeceu a etapas sequenciais para assegurar transparência e reprodutibilidade. Inicialmente, definiu-se o problema e estruturou-se o protocolo da revisão com base na estratégia PICOC: a População (mídias sintéticas e *deepfakes*), a Intervenção (técnicas de detecção), a Comparação (diferentes modelos), os *Outcomes* (métricas de desempenho e limitações) e o Contexto (literatura de IA e Visão Computacional).

A triagem ocorreu em fases: busca nas bases, remoção de duplicidades, leitura de títulos e resumos e, posteriormente, análise integral dos textos para extração estruturada de dados. A etapa final do fluxo constituiu a fase prática, na qual os modelos baseados em Transformers mais recorrentes na literatura foram reproduzidos sob um protocolo experimental padronizado, permitindo cruzar os dados teóricos levantados com evidências empíricas.

Mapeamento Sistemático de Trabalhos e Análise

O levantamento dos estudos primários seguiu as diretrizes do protocolo PRISMA 2020 (PAGE et al., 2021), como apresenta a figura 1. As buscas foram executadas nas bases ACM Digital Library, IEEE Xplore e Science Direct, considerando artigos de acesso aberto publicados a partir de 2022.

Figura 1: Fluxograma PRISMA 2020 do processo de identificação, triagem, elegibilidade e inclusão dos estudos.



Fonte: Elaboração própria (2026).

Para a captura dos trabalhos, definiu-se a seguinte string de busca: *("deepfake detection*" OR *"synthetic media detection"*) AND (*"detection" OR "identification"*) AND (*"transformer-based model"*). Os critérios de inclusão consideraram estudos que abordassem diretamente a detecção de deepfakes, com descrição clara das técnicas, conjuntos de dados e resultados obtidos. Foram priorizados trabalhos baseados em Transformers ou arquiteturas híbridas com mecanismos de atenção, enquanto estudos não transformadores foram mantidos apenas quando contribuíam para contextualizar o estado da arte e comparar abordagens correlatas. Trabalhos puramente teóricos ou focados exclusivamente na geração de mídias foram excluídos. Adicionalmente, os trabalhos incluídos passaram por uma avaliação de qualidade metodológica que pontuou a clareza dos objetivos, delineamento experimental e detalhamento das métricas de avaliação.

A avaliação de qualidade teve como finalidade verificar não apenas a pertinência temática dos estudos, mas também a consistência das informações necessárias para comparação entre eles. Dessa forma, foram observados aspectos como a clareza do objetivo, a descrição da técnica de detecção empregada, a identificação dos conjuntos de dados, a explicitação das métricas de desempenho e a

apresentação dos resultados. Esse procedimento foi importante para reduzir a inclusão de estudos com baixa aderência metodológica e para assegurar que os artigos selecionados apresentassem informações suficientes para subsidiar a etapa comparativa da pesquisa.

Ferramentas e Estrutura Utilizada

Para garantir a organização e o rigor em todas as fases da pesquisa, adotou-se um conjunto específico de ferramentas:

1. Parsifal (PARSIFAL, 2026): Utilizado para o gerenciamento da revisão sistemática, elaboração do protocolo, aplicação dos critérios de inclusão/exclusão e extração padronizada das informações dos estudos;
2. NotebookLM (GOOGLE, 2026): Empregado como ferramenta de apoio durante a triagem primária, facilitando a síntese de conteúdos e a identificação preliminar dos elementos centrais (como objetivos e métricas) dos artigos lidos;
3. Google Colab (GOOGLE, 2026): Utilizado como o ambiente de desenvolvimento para a fase experimental. A plataforma hospedou a implementação computacional, execução dos modelos Transformers selecionados e a extração das métricas de desempenho para análise comparativa.

RESULTADOS

A busca inicial nas bases de dados recuperou 73 estudos (47 da ACM Digital Library e 26 da Science Direct), evidenciando que a IEEE Xplore não retornou resultados alinhados aos critérios neste recorte. Após a aplicação dos critérios de inclusão/exclusão e a remoção de duplicidades via protocolo PRISMA, 14 artigos primários compuseram o escopo final desta revisão sistemática.

Esse resultado indica que a estratégia de busca adotada apresentou maior retorno inicial na ACM Digital Library, base que concentrou 64,38% dos registros recuperados. Entretanto, a quantidade inicial de estudos não correspondeu necessariamente à maior aderência ao recorte da pesquisa, uma vez que parte significativa dos trabalhos recuperados foi excluída após a aplicação dos critérios de seleção. A Science Direct, embora tenha retornado menor volume inicial de estudos, apresentou contribuição expressiva para o conjunto final de artigos incluídos, o que sugere maior alinhamento proporcional com o tema da detecção de deepfakes baseada em arquiteturas transformadoras. Já a ausência de resultados na IEEE Xplore pode estar relacionada à especificidade da string utilizada, especialmente pela

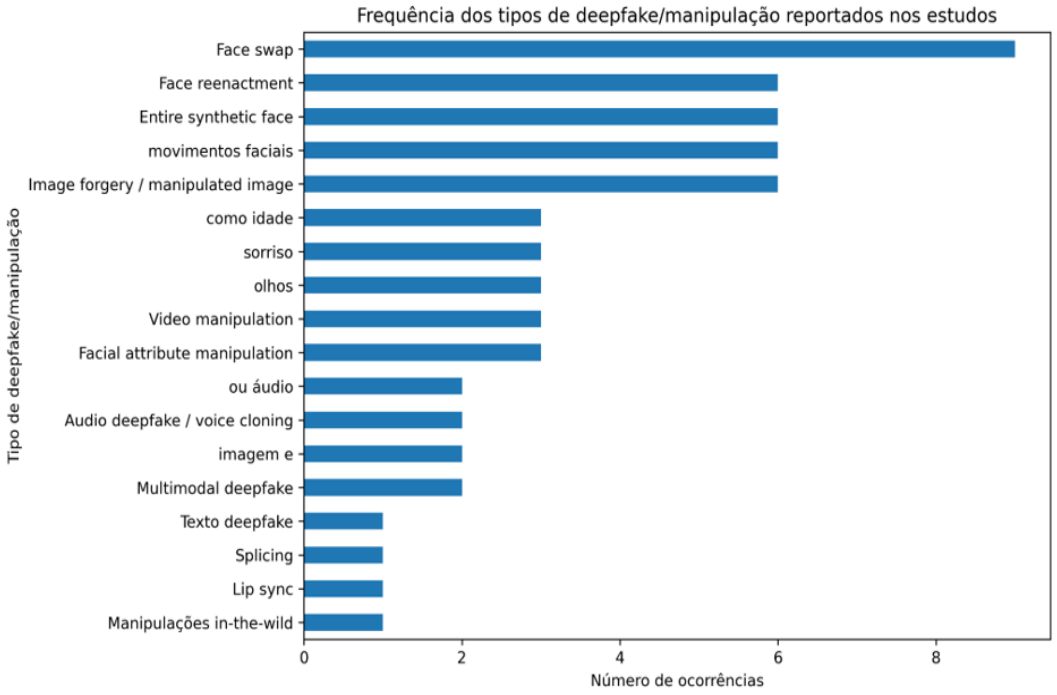
presença do termo “transformer-based model”, que restringiu a recuperação a trabalhos explicitamente associados a essa abordagem.

Resultados da Revisão Sistemática

Os artigos selecionados foram submetidos a uma rigorosa avaliação de qualidade metodológica, na qual a grande maioria obteve pontuação máxima (7.0), demonstrando clareza nos objetivos, conjuntos de dados e protocolos experimentais adotados. A análise cronológica evidenciou uma concentração de publicações no ano de 2025, indicando um interesse crescente e recente por arquiteturas sofisticadas para este fim.

Em relação ao escopo das investigações, as abordagens focam majoritariamente em manipulações como face swap e face *reenactment* aplicadas a imagens e vídeos (Figura 2). As métricas de desempenho predominantes nos estudos são Acurácia, AUC/AUROC, Recall e F1-score, sendo os métodos validados frequentemente por meio de comparações com baselines, estudos de ablação (*ablation study*) e testes entre bases distintas (*cross-dataset*).

Figura 2: Frequência dos tipos de deepfake e manipulação reportados nos estudos.



Fonte: Elaboração própria (2026).

A Figura 2 evidencia os principais tipos de deepfake e manipulação abordados nos estudos analisados. Entre eles, destacam-se categorias recorrentes como face swap, face reenactment, manipulações de imagem estática e vídeo, além de deepfakes multimodais. A diversidade de modalidades identificadas, que abrange em menor

frequência até manipulações específicas como lip sync, clonagem de voz e splicing, demonstra que a literatura não se concentra em um único tipo de falsificação. Esse cenário reforça a complexidade do problema investigado e evidencia a necessidade de adotar modelos arquiteturais avançados, capazes de processar contextos múltiplos para lidar com os variados cenários de fraude digital.

Outro aspecto relevante observado na revisão refere-se à diversidade dos tipos de evidência explorados pelos estudos. Enquanto algumas abordagens se concentram em características visuais e espaciais, como textura, bordas, regiões faciais e artefatos de manipulação, outras buscam incorporar relações temporais, padrões comportamentais ou informações multimodais. Essa diversidade demonstra que a detecção de deepfakes não depende de um único tipo de indício, mas da combinação de diferentes sinais capazes de revelar inconsistências produzidas durante o processo de geração sintética. Nesse sentido, arquiteturas baseadas em atenção tornam-se especialmente relevantes, pois permitem ao modelo relacionar regiões distintas da imagem ou do vídeo e identificar padrões distribuídos que podem não ser evidentes em análises locais.

Desenvolvimento Comparativo

A análise das propostas da literatura (tabela 1) demonstra uma clara predominância de famílias arquiteturais baseadas em *Vision Transformer* (ViT), *DeiT* e *TimeSformer*, além de propostas multimodais e abordagens híbridas (como a união de EfficientNet com EVA Transformer).

Tabela 1: Síntese comparativa dos estudos selecionados quanto às arquiteturas, mídias, datasets, métricas e elegibilidade para a etapa prática.

Estudo	Família arquitetural	Tipo de mídia	Dataset(s)	Métricas	Apto p/ etapa prática?
Realistic Facial Deep Fakes Detection Through Self-Supervised Features Generated by a Self-Distilled Vision Transformer	ViT / Vision Transformer	Imagem	Deepfake Detection Challenge Dataset (DFDC)	AUC; F1-score	Sim
Voice-Face Homogeneity Tells Deepfake	Transformer multimodal	Multimodal	DFDC; FakeAVCeleb; DeepfakeTIMIT; VoxCeleb2	AUC; Acurácia; F1-score; Precisão; Recall/Sensibilidade	Sim

Self-supervised Bidirectional Synchronization Estimation for Multimodal Deepfake Detection with Short-term Dependency	Transformer temporal / multimodal	Multimodal	FakeAVCeleb; KoDF	AP / Average Precision; AUC	Sim
A Multi-Grained Parallel Spatio-Temporal Learning Architecture for Deepfake Video Detection	Híbrido com atenção espaço-temporal	Vídeo	FF++ (HQ/LQ); CelebDF; DFDC; FaceShifter; DeeperForensics	AUC; Acurácia	Sim
DeiTFake: Deepfake detection model using DeiT multi-stage training	DeiT	Imagem	OpenForensics (face-cropped); Celeb-DF v2; FaceForensics++; OpenForensics Test-Dev	AUROC; Acurácia; F1-score	Sim
Robust cross-dataset deepfake detection with multitask self-supervised learning	Híbrido (EfficientNet + EVA Transformer)	Imagem, Vídeo	FF++; DFD; CelebDF-v2; DFDC; DFDCp; FFIW; DFF	AUC	Sim
A deep fake detection approach for cyber security threat based on deep learning and diffusion-osmosis model	Não transformer (U-Net + GCN)	Imagem	Celeb-HQ; Celeb-DF; DFDC; FaceForensics++; CelebA	AUC; Acurácia; F1-score; Precisão; Recall/Sensibilidade	Não
Customized Convolutional Neural Network for Accurate Detection of Deep Fake Images in Video Collections	Não transformer (CNN)	Imagem, Vídeo	Deep Fake Detection Challenge (DFDC)	AUC; Acurácia	Não
Integrating perceptual quality analysis and caption-based features for robust deepfake video detection	Transformer multimodal / híbrido	Multimodal, Vídeo	FaceForensics++ (FF++); DFDC; Celeb-DF v2	Acurácia; F1-score	Sim
Real-Time Deepfake Detection via Gaze and Blink Patterns: A	TimeSformer / híbrido	Vídeo	FaceForensics++; CelebDF-V2; DFDC; FakeAVCeleb	Acurácia; F1-score; Precisão; Recall/Sensibilidade	Sim

Transformer Framework					
Constructing an Arabic Deepfake Detection Dataset from Government Sources and LLM-Generated Text	Transformer multimodal	Texto	Dataset árabe customizado (46.712 artigos)	Outra	Sim
GAN-ViT-CMFD: A novel framework integrating generative adversarial networks and vision transformers for enhanced copy-move forgery detection and classification with spectral clustering	Híbrido (GAN + ViT + CNN)	Imagem	CoMoFoD; CASIA; MICC-F220	Acurácia; F1-score; Precisão; Recall/Sensibilidade	Sim
An adaptive dual-domain feature representation method for enhanced deep forgery detection	Transformer não especificado	Imagem	FaceForensics++ (c23/c40); Celeb-DeepFake-v2 (CDF); DFDC	AUC; Acurácia	Sim
Deepfake detection: Enhancing performance with spatiotemporal texture and deep learning feature fusion	Não transformer (3D CNN com atenção)	Vídeo	Celeb-DF; FaceForensics++; DeepfakeTIMIT; FaceShifter	AUC; Acurácia; Standard Deviation (estabilidade)	Não

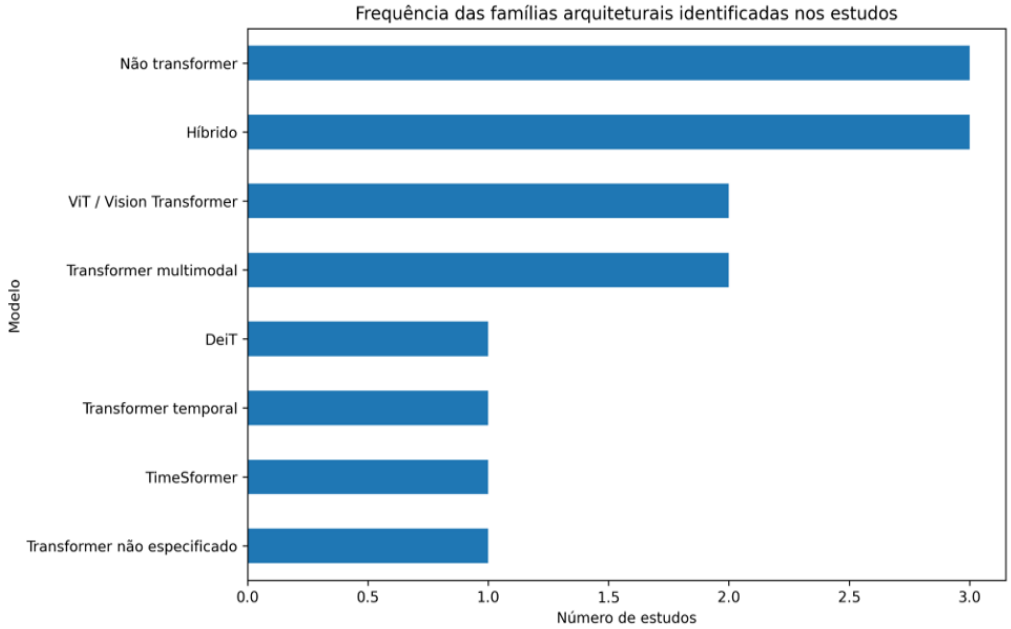
Fonte: Elaboração própria (2026).

A presença de estudos classificados como “não transformer” na Tabela 1 não descaracteriza o recorte da pesquisa, uma vez que esses trabalhos foram mantidos com finalidade de contextualização do estado da arte e comparação indireta com abordagens correlatas. Embora não tenham composto a etapa prática, eles permitem observar que a literatura ainda combina modelos transformadores, híbridos e arquiteturas tradicionais de aprendizado profundo na tentativa de enfrentar diferentes tipos de manipulação. Essa coexistência evidencia que o campo da detecção de deepfakes ainda se encontra em fase de consolidação metodológica, com

diferentes estratégias sendo exploradas conforme o tipo de mídia, o dataset utilizado e o cenário experimental adotado.

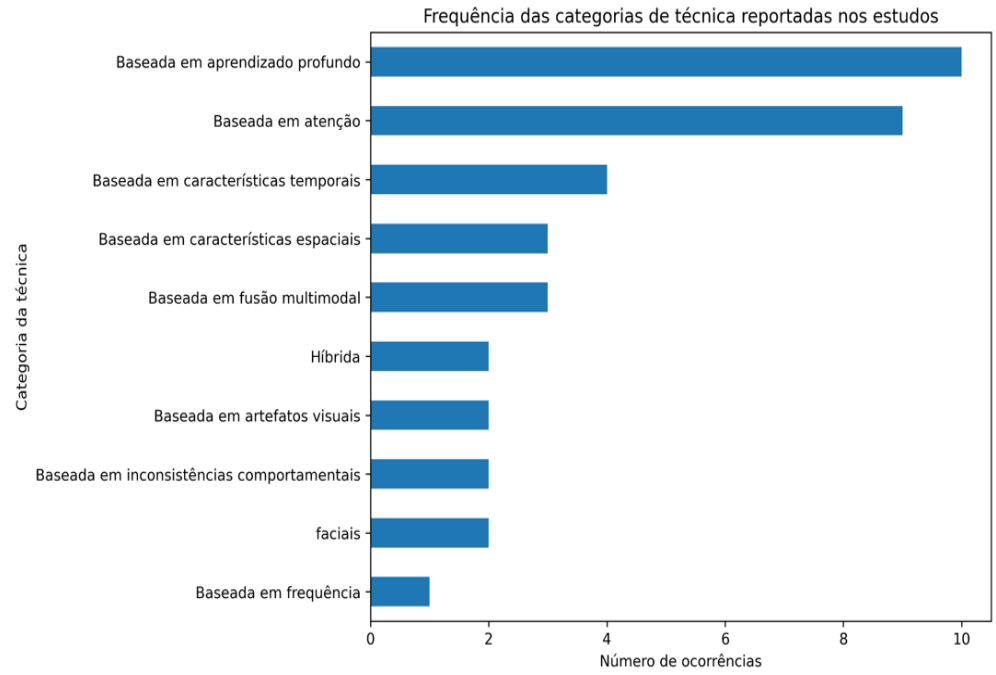
Observou-se que a literatura recente combina frequentemente essas arquiteturas (Figura 3) com mecanismos de extração de características espaciais e temporais para identificar inconsistências sutis (Figura 4).

Figura 3: Frequência das famílias arquiteturais identificadas nos estudos selecionados.



Fonte: Elaboração própria (2026).

Figura 4: Frequência das categorias de técnica reportadas nos estudos selecionados.



Fonte: Elaboração própria (2026).

A análise das propostas da literatura (Tabela 1) e das famílias arquiteturais mais frequentes (Figura 3) demonstra uma clara predominância de modelos baseados em Vision Transformer (ViT), DeiT e TimeSformer, além de propostas multimodais e híbridas. Complementarmente, a Figura 4 evidencia que a literatura recente tem priorizado abordagens ancoradas em aprendizado profundo e mecanismos de atenção, combinando frequentemente essas arquiteturas com métodos de extração de características espaciais e temporais para identificar inconsistências sutis nas mídias manipuladas.

Além disso, constatou-se uma forte heterogeneidade nos datasets empregados (como FaceForensics++, DFDC e Celeb-DF) e nos protocolos de validação, o que dificulta a comparação direta e justa do desempenho empírico entre os diferentes trabalhos publicados.

Discussão de Resultados da Revisão Sistemática

Os achados da literatura confirmam que os Transformers superam limitações inerentes às redes estritamente convolucionais (CNNs) graças à sua capacidade de modelar dependências globais, integrar contextos e representar padrões complexos sem depender de engenharia manual de atributos. Como os vestígios de falsificação (*deepfakes*) frequentemente se manifestam de forma distribuída na imagem ou no vídeo, essa análise global se mostra essencial.

Contudo, a literatura reporta gargalos críticos recorrentes: a alta dependência de conjuntos de dados específicos, a sensibilidade à compressão da mídia e a dificuldade de generalização entre bases de dados distintas (*cross-dataset*). Essa lacuna, aliada à ausência de uniformidade nos protocolos experimentais teóricos, fundamenta e justifica a necessidade da etapa experimental padronizada conduzida neste estudo.

ETAPA EXPERIMENTAL

Com base nos achados da revisão sistemática, foi conduzida uma etapa experimental comparativa para avaliar o desempenho de três arquiteturas representativas do estado da arte na detecção de deepfakes em imagens: Vision Transformer (ViT), Data-efficient Image Transformer (DeiT) e uma arquitetura híbrida (EfficientNet + EVA Transformer).

O experimento consistiu em uma tarefa de classificação binária (imagens reais vs. manipuladas). O treinamento, validação e teste interno utilizaram um subconjunto do FaceForensics++, com o conjunto de teste composto por 280 imagens

perfeitamente balanceadas. Para avaliar a capacidade de generalização e robustez (cross-dataset), os modelos treinados foram testados em um subconjunto do Celeb-DF contendo 3000 amostras (1500 reais e 1500 falsas). Os modelos foram submetidos a um protocolo padronizado, incluindo imagens redimensionadas para 224×224 pixels, 10 épocas de treinamento, função de perda Cross-Entropy e otimizador AdamW. Os desempenhos alcançados estão sintetizados na Tabela 2.

Tabela 2: Desempenho dos modelos no teste interno (FaceForensics++) e no teste cruzado (Celeb-DF).

Modelo	Cenário	Acurácia	Precisão	Recall	F1-score	AUC/ROC	Loss
ViT	Teste interno	0,9964	0,9929	1,0000	0,9964	1,0000	0,0090
DeiT	Teste interno	0,9893	0,9928	0,9857	0,9892	0,9985	0,0372
Efficient Net + EVA	Teste interno	1,0000	1,0000	1,0000	1,0000	1,0000	0,0164
ViT	Cross-dataset	0,5150	0,5159	0,4873	0,5012	0,5280	1,4615
DeiT	Cross-dataset	0,5020	0,5048	0,2100	0,2966	0,5077	2,0226
Efficient Net + EVA	Cross-dataset	0,5930	0,7254	0,2993	0,4238	0,6888	0,8163

Fonte: Elaboração própria (2026).

A comparação entre os dois cenários experimentais evidencia uma diferença expressiva entre desempenho interno e capacidade de generalização. No teste interno, realizado no mesmo domínio de dados do treinamento, os três modelos apresentaram resultados próximos ao desempenho perfeito, com AUC/ROC igual ou superior a 0,9985. Entretanto, no teste cross-dataset com o Celeb-DF, observou-se uma redução significativa em todas as arquiteturas avaliadas. O ViT, por exemplo, passou de AUC/ROC 1,0000 para 0,5280; o DeiT reduziu de 0,9985 para 0,5077; e o EfficientNet + EVA, embora tenha mantido o melhor desempenho relativo, caiu de 1,0000 para 0,6888. Esse comportamento demonstra que a avaliação interna, isoladamente, pode superestimar a robustez dos modelos, uma vez que o aprendizado pode estar fortemente associado às características específicas da base de treinamento.

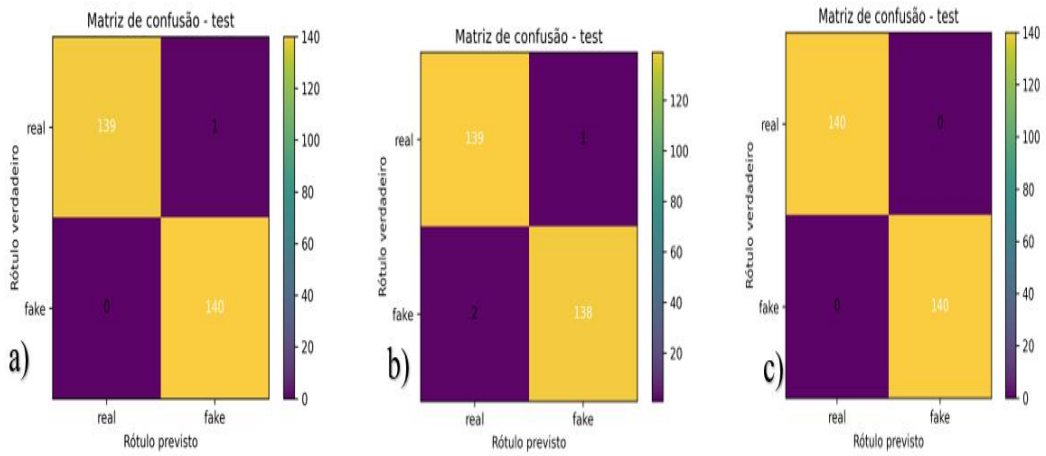
A análise conjunta das métricas também revela diferenças importantes entre os modelos. No cenário cross-dataset, o EfficientNet + EVA apresentou a maior precisão entre as arquiteturas avaliadas, com valor de 0,7254, indicando que, quando o modelo classificou uma amostra como falsa, houve maior probabilidade de acerto em comparação aos demais. Contudo, seu recall foi de apenas 0,2993, o que

demonstra dificuldade em recuperar todas as amostras manipuladas. Esse comportamento sugere um modelo mais conservador na classificação da classe fake, reduzindo falsos positivos, mas ainda produzindo quantidade elevada de falsos negativos. O DeiT apresentou o comportamento mais crítico nesse aspecto, com recall de 0,2100 no teste cruzado, revelando forte tendência a classificar imagens manipuladas como reais.

Os resultados do teste interno com o FaceForensics++ evidenciaram um aprendizado quase perfeito. Como pode ser observado na expressiva concentração de amostras nas diagonais principais das matrizes (Figura 5), a incidência de falsos positivos e falsos negativos foi praticamente nula durante a validação interna. O destaque visual fica para a arquitetura híbrida (EfficientNet + EVA), que não cometeu nenhum erro de classificação.

A arquitetura híbrida (EfficientNet + EVA) obteve o melhor desempenho, alcançando acurácia, precisão, recall, F1-score e AUC de 1,0000, não cometendo nenhum erro de classificação. O ViT e o DeiT também registraram desempenhos altíssimos, com acurácias de 0,9964 e 0,9893, respectivamente, comprovando a excelente capacidade dessas arquiteturas em discriminar conteúdos no domínio em que foram treinadas.

Figura 5: Matrizes de confusão dos modelos avaliados no teste interno com o FaceForensics++. a) ViT; b)DeiT; c) EfficientNet + EVA.

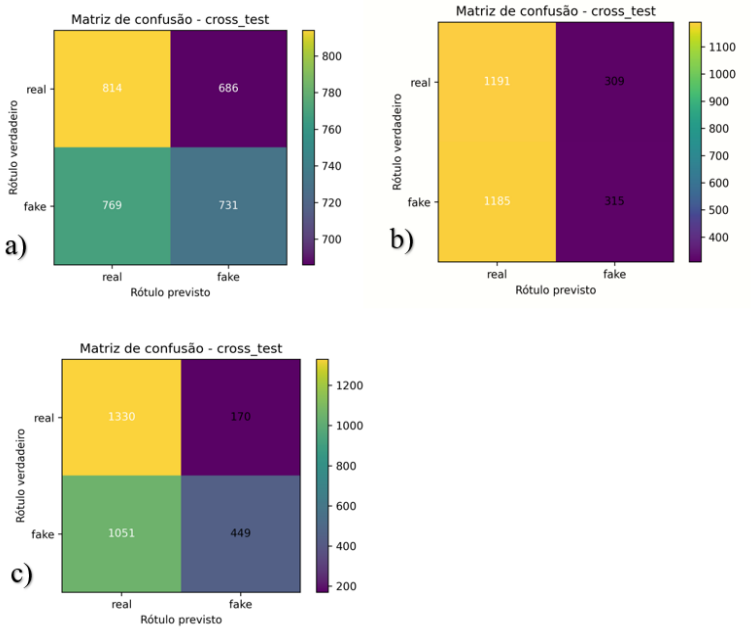


Fonte: Elaboração própria (2026).

Contudo, o cenário sofreu uma degradação expressiva na avaliação cross-dataset com o Celeb-DF (Figura 6), confirmando o principal gargalo apontado pela RSL: a dificuldade de generalização entre bases de dados distintas. O ViT e o DeiT caíram para acurácias muito próximas ao acaso (0,5150 e 0,5020). O DeiT demonstrou uma vulnerabilidade particular, apresentando forte viés para a classe

real e um recall de apenas 0,2100 para a classe manipulada. A arquitetura híbrida, combinando extração convolucional e representação baseada em Transformer, demonstrou maior robustez relativa no teste externo (Acurácia de 0,5930), mas ainda distante de um nível ideal de segurança.

Figura 6: Matrizes de confusão dos modelos avaliados no teste cruzado com o Celeb-DF. a) ViT; b)DeiT; c) EfficientNet + EVA.



Fonte: Elaboração própria (2026).

Os resultados evidenciam que a queda de desempenho não ocorreu de forma pontual, mas se manifestou de maneira consistente entre os três modelos avaliados. O ViT e o DeiT, embora tenham alcançado AUC/ROC praticamente perfeita no teste interno, apresentaram redução expressiva quando submetidos ao Celeb-DF, aproximando-se de um comportamento aleatório na avaliação externa. Esse resultado sugere que tais modelos podem ter aprendido padrões específicos do FaceForensics++, como artefatos visuais, distribuição das imagens e características próprias do processo de manipulação presente na base de treinamento. A arquitetura EfficientNet + EVA apresentou a menor degradação relativa entre os modelos, mantendo AUC/ROC de 0,6888 no teste cross-dataset. Esse comportamento indica que a combinação entre extração convolucional e representação baseada em Transformer pode contribuir para maior robustez diante de mudanças de domínio. Ainda assim, o baixo recall observado para a classe fake demonstra que o modelo continuou deixando de identificar parte significativa das imagens manipuladas, o que limita sua aplicação prática em contextos sensíveis de verificação de autenticidade digital.

Em síntese, os resultados obtidos confirmam a relevância dos modelos transformadores e híbridos para a detecção de deepfakes, mas também demonstram que o desempenho elevado em testes internos não garante robustez em cenários externos. A queda observada no cross-dataset reforça uma das principais limitações identificadas na revisão sistemática: a dificuldade de generalização entre bases de dados distintas. Em cenários reais, nos quais as mídias podem apresentar diferentes níveis de compressão, qualidade, origem e técnica de manipulação, modelos com alto desempenho em um único dataset podem não manter a mesma confiabilidade. Dessa forma, a principal contribuição empírica do estudo está em demonstrar, sob protocolo padronizado, que a comparação entre arquiteturas deve considerar não apenas métricas internas elevadas, mas também sua estabilidade em bases externas e sua capacidade de manter desempenho consistente diante de mudanças de domínio, qualidade visual e padrões de manipulação.

CONSIDERAÇÕES FINAIS

A principal contribuição do estudo está em demonstrar que a comparação entre modelos transformadores deve considerar não apenas métricas elevadas em bases internas, mas também sua capacidade de generalização em bases externas, condição mais próxima dos desafios encontrados em aplicações reais de detecção de deepfakes.

O presente estudo cumpriu seu objetivo ao mapear as principais técnicas de detecção de deepfakes e validá-las empiricamente sob protocolo padronizado. A revisão sistemática confirmou que a adoção de Transformers e modelos híbridos representa uma mudança de paradigma, permitindo que a detecção utilize o contexto global, dependências espaciais e representações multimodais para identificar falsificações que passariam despercebidas por métodos estritamente convolucionais.

A fase experimental reforçou a superioridade analítica dessas arquiteturas em domínios fechados, mas expôs criticamente sua fragilidade em avaliações cross-dataset. Constatou-se que a generalização e a sensibilidade a diferentes processos de síntese continuam sendo os maiores desafios da área forense digital.

Para superar essas barreiras, trabalhos futuros devem ampliar os testes para o formato de vídeo contínuo, explorando dinâmicas temporais, metodologias de aprendizado autossupervisionado e técnicas de adaptação de domínio para garantir a integridade da informação face a um cenário de ameaças cada vez mais sofisticado.

REFERÊNCIAS

ACHEAMPONG, F. A. *et al*. Transformer Models for Text-based Emotion Detection: A Review of BERT-based Approaches. **Artificial Intelligence Review**, v. 54, p. 5789–5829, 2021. Disponível em: https://doi.org/10.1007/s10462-021-09958-2. Acesso em: 1 mar. 2026.

CHESNEY, R.; CITRON, D. K. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. **Foreign Affairs**, v. 98, n. 1, p. 147–155, 2019. Disponível em: https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war. Acesso em: 1 mar. 2026.

DOSOVITSKIY, A. *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. In: **International Conference on Learning Representations (ICLR)**, 2021. Disponível em: https://openreview.net/forum?id=YicbFdNTTy. Acesso em: 1 mar. 2026.

EDWARDS, P. *et al*. Review of deepfake techniques: architecture, detection, and datasets. **IEEE Access**, v. 12, p. 154718–154742, 2024. Disponível em: https://doi.org/10.1109/ACCESS.2024.3477257. Acesso em: 8 mar. 2026.

GOODFELLOW, I. *et al*. Generative adversarial nets. In: **Advances in Neural Information Processing Systems**. 2014. p. 2672–2680. Disponível em: https://papers.nips.cc/paper/5423-generative-adversarial-nets. Acesso em: 8 mar. 2026.

GOOGLE. **Google Colaboratory**. Disponível em: https://colab.research.google.com/. Acesso em: 11 abr. 2026.

GOOGLE. **NotebookLM**. Disponível em: https://notebooklm.google.com/. Acesso em: 11 mar. 2026.

LIU, Z. *et al*. Swin transformer: Hierarchical vision transformer using shifted windows. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision**, 2021. Disponível em: https://doi.org/10.1109/ICCV48922.2021.00986. Acesso em: 24 mar. 2026.

MIRSKY, Y.; LEE, W. The creation and detection of deepfakes: A survey. **ACM Computing Surveys (CSUR)**, v. 54, n. 1, p. 1–41, 2021. Disponível em: https://doi.org/10.1145/3425780. Acesso em: 24 mar. 2026.

MOHAMMED, A.; KORA, R. A Comprehensive Overview and Analysis of Large Language Models: Trends and Challenges. **IEEE Access**, v. 13, p. 1-15, 2025. Disponível em: https://doi.org/10.1109/ACCESS.2025.3573955. Acesso em: 24 mar. 2026.

PAGE, M. J. *et al*. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. **BMJ**, v. 372, n. 71, 2021. Disponível em: https://doi.org/10.1136/bmj.n71.

PARSIFAL. **Parsifal**: a collaborative tool for systematic literature reviews. Disponível em: https://parsif.al/. Acesso em: 10 mar. 2026.

PATEL, Y. *et al*. Deepfake generation and detection: case study and challenges. **IEEE Access**, v. 11, p. 135200-135215, 2023. Disponível em: https://doi.org/10.1109/ACCESS.2023.3342107. Acesso em: 1 abr. 2026.

RANA, M. S. *et al*. Deepfake Detection: A Systematic Literature Review. **IEEE Access**, v. 10, p. 25494–25513, 2022. Disponível em: https://doi.org/10.1109/ACCESS.2022.3154404. Acesso em: 8 mar. 2026.

ROSSLER, A. *et al*. Faceforensics++: Learning to detect manipulated facial images. In: **Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)**. Seoul: [s. n.], 2019. p. 1–11. Disponível em: https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html. Acesso em: 1 abr. 2026.

VASWANI, A. *et al*. Attention is all you need. **Advances in Neural Information Processing Systems**, v. 30, p. 5998-6008, 2017. Disponível em: https://papers.nips.cc/paper/7181-attention-is-all-you-need. Acesso em: 1 abr. 2026.

VERDOLIVA, L. Media forensics and deepfakes: An overview. **IEEE Journal of Selected Topics in Signal Processing**, v. 14, n. 5, p. 910–932, 2020. Disponível em: https://doi.org/10.1109/JSTSP.2020.3002101. Acesso em: 5 abr. 2026.