RCMOS – Revista Científica Multidisciplinar O Saber. ISSN: 2675-9128. São Paulo-SP.

Ano III, v.1 2023. | submissão: 10/02/2023 | aceito: 12/03/2023 | publicação: 14/04/2023

Explainable AI (XAI) na Detecção de Deepfakes: Transparência e Interpretação em Modelos de Visão Computacional

Explainable AI (XAI) in Deepfake Detection: Transparency and Interpretation in Computer Vision Models

Matheus de Oliveira Pereira Paula
Bacharelado em Sistemas de Informação pelo Instituto Federal de Educação, Ciência e Tecnologia Fluminense. Mestrando em Data Science and Artificial Intelligence (MSc) na Université Côte d'Azur.

RESUMO

A crescente sofisticação dos *deepfakes* aumentou a urgência de sistemas de detecção baseados em **Aprendizado Profundo** robustos. No entanto, a natureza de "caixa-preta" dos modelos de Visão Computacional, como as Redes Neurais Convolucionais (CNNs) e os Transformers, representa um obstáculo significativo para sua aceitação em domínios críticos, como o **forense e o jurídico**. Este artigo explora a aplicação de técnicas de **Inteligência Artificial Explicável (XAI)** no contexto da detecção de *deepfakes*, investigando como alcançar a transparência e a interpretabilidade dos modelos. Serão discutidas metodologias *post-hoc* e *intrínsecas*, como **CAMs (Class Activation Maps), SHAP e LIME**, analisando sua capacidade de gerar evidências visuais e lógicas sobre o processo de classificação, especificamente, identificando as regiões da imagem ou do vídeo (artefatos) que são determinantes para a decisão de falsidade. O objetivo primário é demonstrar que a integração de XAI é indispensável para construir a **confiança** necessária nos sistemas de detecção, transformando a decisão algorítmica em **prova pericial verificável**, essencial para estabelecer a validade e a admissibilidade dessas tecnologias em tribunais e investigações.

Palavras-chave: XAI; Deepfake; Interpretabilidade; Transparência; Forense Digital; Visão Computacional; Confiança.

ABSTRACT

The increasing sophistication of *deepfakes* has escalated the urgency for robust **Deep Learning** detection systems. However, the "black-box" nature of Computer Vision models, such as Convolutional Neural Networks (CNNs) and Transformers, poses a significant barrier to their acceptance in critical domains like **forensics and law**. This paper explores the application of **Explainable Artificial Intelligence (XAI)** techniques within the context of *deepfake* detection, investigating how model **transparency and interpretability** can be achieved. We

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

discuss *post-hoc* and *intrinsic* methodologies, such as **CAMs (Class Activation Maps), SHAP, and LIME**, and analyze their ability to generate visual and logical evidence of the classification process, specifically identifying the image or video regions (artifacts) that are decisive for the forgery decision. The primary objective is to demonstrate that integrating XAI is indispensable for building the trust needed in detection systems, transforming algorithmic decisions into **verifiable expert evidence**, which is essential for establishing the validity and admissibility of these technologies in courts and investigations.

Keywords: XAI; Deepfake; Interpretability; Transparency; Digital Forensics; Computer Vision; Trust.

1. INTRODUÇÃO: O IMPERATIVO DA EXPLICABILIDADE NA FORENSEDIGITAL

O rápido avanço dos métodos de criação de *deepfakes* estabeleceu um novo patamar de desafios para a sociedade, exigindo contramedidas de **Inteligência Artificial (IA)** igualmente sofisticadas. Modelos de **Visão Computacional**, como as complexas Redes Neurais Convolucionais (CNNs) profundas e os emergentes *Vision Transformers* (ViTs), têm demonstrado alta acurácia na distinção entre mídias autênticas e falsificadas. Contudo, essa alta performance é alcançada à custa da transparência, resultando em uma arquitetura de "caixa-preta" em que a decisão final é um mistério matemático, inacessível até mesmo aos seus desenvolvedores. Esta opacidade é o cerne do problema quando se insere a detecção de *deepfakes* em contextos de alta responsabilidade, como a **investigação forense e o litígio jurídico**, onde a simples declaração de "falso" por um algoritmo é insuficiente para ser considerada prova.

Em um ambiente legal, a prova pericial deve ser **verificável, replicável e, acima de tudo, logicamente justificada**. A confiabilidade de um sistema de detecção de *deepfake* é colocada em xeque se o perito não puder apresentar ao juiz, de forma clara e intuitiva, *quais* as características do vídeo levaram o modelo a classificar o conteúdo como forjado. Se a CNN está focando em um artefato de compressão, em vez de uma inconsistência na geometria facial, a decisão pode ser errônea, mas a opacidade do modelo impede essa distinção crítica. Portanto, a **Inteligência Artificial Explicável (XAI)**surge não apenas como um aprimoramento, mas como um **imperativo metodológico** para garantir a **admissibilidade e a validade** dos resultados da detecção de *deepfakes* em qualquer processo que exija um padrão rigoroso de prova.

O objetivo central deste artigo é fornecer um arcabouço teórico e prático para a integração de técnicas de XAI na pipeline de detecção de *deepfakes*. Analisaremos como as principais metodologias de explicabilidade podem ser adaptadas para o domínio forense, transformando a saída binária do detector (*real* ou *fake*) em um **relatório pericial com evidências visuais e ponderadas**. A relevância deste tema reside na necessidade de construir uma ponte entre a

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

capacidade preditiva da IA avançada e a **exigência de justificação racional** dos sistemas legais, promovendo um nível de **confiança e rastreabilidade** que é atualmente inexistente nos modelos opacos tradicionais.

A pesquisa se aprofundará na diferenciação entre as técnicas de explicabilidade *agnósticas ao modelo* e as *específicas*, avaliando sua aplicabilidade na identificação de artefatos de *deepfake* que variam em sutileza e localização (desde micro-artefatos de pixel até inconsistências de iluminação global). A finalidade é robustecer a detecção, garantindo que o modelo não esteja "trapaceando" ao aprender correlações espúrias (como marcas d'água de conjuntos de dados de treinamento ou metadados) em vez das verdadeiras assinaturas de manipulação. Em última análise, a adoção de XAI transcende a mera interpretação algorítmica; é a base para a construção de um sistema de defesa digital **responsável e eticamente fundamentado**.

2. A NATUREZA DA "CAIXA-PRETA" E O DESAFIO DA CONFIABILIDADE FORENSE

A ascensão do **Aprendizado Profundo** em tarefas de Visão Computacional foi marcada por uma mudança de paradigma: ganhos exponenciais em acurácia foram obtidos às custas da **interpretabilidade**. Modelos de detecção de *deepfakes*, como as CNNs profundas (e.g., ResNet, XceptionNet) e os ViTs, são compostos por milhões ou até bilhões de parâmetros dispostos em camadas complexas, onde as transformações não lineares tornam o mapeamento da entrada (o vídeo) para a saída (a classificação) intrinsecamente opaco. Essa opacidade, conhecida como o problema da **"caixa-preta"**, gera uma desconfiança fundamental em qualquer ambiente onde a falha do sistema possa ter consequências graves e irreversíveis, sendo o sistema jurídico o exemplo mais proeminente.

No contexto forense, a **confiabilidade** de uma evidência digital depende de sua **rastreabilidade e justificação**. Quando um perito apresenta um laudo alegando que um vídeo é um *deepfake*, essa conclusão não pode se basear unicamente no resultado de 98% de probabilidade de falsidade emitido por um algoritmo desconhecido. O juiz, os advogados e o júri precisam entender **o *porquê***; é necessário que o modelo aponte, com precisão espacial e temporal, as características (ou a ausência delas) que corroboram a tese da manipulação, transformando a predição probabilística em **evidência causal**. A ausência dessa justificação explícita impede o contraditório e a crítica técnica da prova, violando princípios básicos do devido processo legal.

O desafio da caixa-preta na detecção de *deepfakes* é agravado pela própria natureza adversarial do problema. Sem XAI, é impossível garantir que o modelo não esteja focando em **artefatos espúrios** que se correlacionam acidentalmente com a classe "falso" no conjunto de dados de treinamento. Exemplos disso incluem modelos que aprendem a identificar a **marca d'água invisível** de um *codec* ou as **bordas da caixa delimitadora** usada na fase de *face swapping*, em vez da falha na textura da pele ou nas sombras. Tais correlações são frágeis e quebram sob

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

mínima variação, tornando o modelo inútil na generalização e totalmente inadmissível como prova confiável, pois a causa da classificação não é o *deepfake*, mas um artefato de produção do *dataset*.

Portanto, a implementação de XAI na detecção de *deepfakes* é uma **necessidade ético-legal** que transcende a mera curiosidade acadêmica. Ela é a única via para mitigar a opacidade, permitindo que a comunidade forense e o judiciário possam auditar o processo decisório do algoritmo. Ao transformar a caixa-preta em uma **caixa de vidro**, onde as saliências da decisão são visíveis, o XAI não apenas valida a alta acurácia dos modelos de Aprendizado Profundo, mas também eleva sua decisão de uma sugestão técnica para uma **prova pericial de alto valor probatório**, essencial para enfrentar a desinformação na era da mídia sintética.

3. TÉCNICAS DE XAI *POST-HOC*: MAPAS DE ATIVAÇÃO E ATRIBUIÇÃO DE RECURSOS

As técnicas de **XAI *post-hoc*** (pós-análise) são o grupo mais comum de metodologias de explicabilidade, pois podem ser aplicadas a qualquer modelo de detecção de *deepfake* já treinado, sem a necessidade de modificar sua arquitetura interna. Essas técnicas se concentram em medir a **contribuição e a influência** de cada *feature* de entrada (pixels ou regiões de pixels) na predição final do modelo. No domínio da Visão Computacional, duas categorias principais dominam: os **Mapas de Ativação de Classe (CAMs)** e os métodos de **Atribuição de Recursos (*Feature Attribution*)**.

Os **Class Activation Maps (CAMs)** e suas variações (como Grad-CAM e Grad-CAM++) são ferramentas visuais que geram um mapa de calor (heatmap) sobre a imagem de entrada, destacando as regiões mais relevantes para a classificação. Em um contexto de detecção de *deepfake*, um Grad-CAM bem-sucedido deve focar nas áreas que contêm os artefatos da manipulação, como a **linha de fusão da face, inconsistências nos olhos/boca, ou discrepâncias na iluminação e na textura da pele**. Se o modelo está corretamente focado, o mapa de calor deve se concentrar nos pontos de *forgery*. Esta visualização é extremamente valiosa para a forense, pois transforma uma decisão numérica abstrata em uma **evidência visual intuitiva e espacialmente localizada**, permitindo ao perito identificar se o modelo está detectando a falha real da manipulação ou um ruído irrelevante do fundo.

Em paralelo, os métodos de **Atribuição de Recursos**, como **SHAP (SHapley Additive exPlanations)** e **LIME (Local Interpretable Model-agnostic Explanations)**, fornecem uma **justificativa mais numérica e causal**. O SHAP, baseado na teoria dos jogos de Shapley, calcula o valor de contribuição marginal de cada *pixel* ou *patch* de imagem para a probabilidade de falsidade. Ele fornece um conjunto de **valores de Shapley** que quantificam o peso exato de cada região na classificação final, garantindo que a explicação seja **globalmente consistente e localmente fiel**. O LIME, por sua vez, constrói um modelo linear interpretável localmente

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

(próximo ao ponto de dados que está sendo explicado) para aproximar o comportamento complexo do modelo de caixa-preta.

A aplicabilidade desses métodos *post-hoc* na detecção de *deepfakes* reside na sua capacidade de **auditar e validar o viés** do modelo. Se um SHAP ou um Grad-CAM consistentemente atribui alta importância a áreas fora da região facial manipulada (e.g., a marca d'água no canto do vídeo), isso indica que o modelo sofreu *overfitting* a um artefato de *dataset* espúrio. A limitação desses métodos é que fornecem apenas uma explicação local (válida apenas para o vídeo em análise, e não para o comportamento geral do modelo) e podem ser **computacionalmente caros** (como o SHAP), o que impacta a análise em tempo real de vídeos longos. Contudo, para a elaboração de laudos periciais *offline*, a riqueza e a profundidade dessas explicações causais são um recurso inestimável, estabelecendo a base para a prova técnica em tribunal.

4. TÉCNICAS INTRÍNSECAS E A TRANSIÇÃO PARA MODELOS INERENTEMENTE INTERPRETÁVEIS

Embora os métodos *post-hoc* sejam essenciais para a auditoria de modelos existentes, a tendência de pesquisa em XAI aponta para a criação de **modelos inerentemente interpretáveis** ou **técnicas intrínsecas**. Estas abordagens visam incorporar o mecanismo de explicabilidade diretamente na arquitetura da rede neural, garantindo que a decisão final seja logicamente rastreável por *design*, e não por análise retrospectiva. A busca por modelos intrinsecamente interpretáveis é uma resposta direta à fragilidade inerente dos métodos *post-hoc*, cuja explicação é sempre uma aproximação e pode ser enganosa se o modelo subjacente for extremamente complexo.

No domínio da Visão Computacional para detecção de *deepfakes*, isso se manifesta no uso de arquiteturas que aprendem explicitamente a separar a representação semântica da **dos artefatos**. Um exemplo teórico é o uso de **Modelos Baseados em Protótipos**, onde a classificação não é feita por uma fronteira de decisão opaca, mas pela similaridade com "protótipos" visuais armazenados em uma camada latente, representando exemplos canônicos de faces reais e faces falsas. A explicação para a classificação de um vídeo falso seria, então, a simples apresentação do protótipo mais similar de *deepfake* ao qual o vídeo se assemelha, fornecendo uma **justificativa por analogia** altamente intuitiva para o ser humano.

A maior contribuição para a interpretabilidade intrínseca na detecção de *deepfakes* vem, ironicamente, da arquitetura **Transformer** e seu mecanismo de **auto-atenção**. O mapa de atenção gerado naturalmente pelo Transformer, que indica como o modelo pondera diferentes *patches* da imagem para construir sua representação, pode ser diretamente interpretado como um **mapa de relevância**. Embora não seja uma explicação causal completa no sentido do SHAP, o mapa de atenção é uma ferramenta intrínseca que revela onde o modelo está

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

"olhando". Se o modelo de detecção está focado em anomalias de iluminação ou nas bordas da máscara facial aplicada no *deepfake*, a matriz de atenção refletirá essa concentração de peso.

A grande vantagem dos modelos intrinsecamente interpretáveis é a **confiança inerente** que eles proporcionam, uma vez que a explicação não é um subproduto, mas parte integrante do processo de inferência. Isso simplifica o *deployment* em ambientes forenses, pois o laudo pericial pode se basear em um processo que é, por definição, transparente. Contudo, o design dessas arquiteturas é um desafio, pois a restrição de interpretabilidade pode, paradoxalmente, **limitar a capacidade preditiva** do modelo, forçando uma troca de desempenho por transparência. O campo de pesquisa busca incessantemente mitigar essa troca, desenvolvendo modelos que sejam simultaneamente **altamente acurados e perfeitamente rastreáveis** em sua lógica decisória.

5. XAI NO COMBATE AO VIÉS E NA VALIDAÇÃO DA GENERALIZAÇÃO

O uso estratégico da **Inteligência Artificial Explicável (XAI)** na detecção de *deepfakes* transcende a mera justificação de uma única classificação; torna-se uma ferramenta crítica para a validação da capacidade de **generalização** e o **combate ao viés algorítmico**. Como discutido em estudos comparativos, a principal fragilidade dos modelos de detecção de *deepfake* é sua tendência a **falhar em dados não vistos** (*cross-dataset generalization*), o que frequentemente é um sintoma de que o modelo aprendeu correlações espúrias nos dados de treinamento. O XAI fornece o microscópio necessário para identificar e corrigir esse comportamento indesejado.

Ao aplicar técnicas como Grad-CAM ou SHAP em larga escala sobre um conjunto de dados de validação, os peritos podem realizar uma **auditoria de *feature***. Se o modelo, em todos os vídeos classificados como "falso", estiver consistentemente focando nos metadados ou em um *artefato de compressão* uniforme, isso revela um **viés de *dataset***. O detector não está aprendendo a manipular o conceito de falsidade facial, mas sim a detectar a impressão digital do processo de criação do conjunto de dados de treinamento. Essa descoberta, facilitada pelo XAI, permite aos pesquisadores **retreinar o modelo** utilizando técnicas de **data augmentation** mais robustas (como a introdução de diferentes níveis de compressão) ou forçar o modelo a ignorar áreas com artefatos espúrios por meio de **máscaras de atenção** (*attention masking*).

A validação da **capacidade de generalização** também se beneficia imensamente do XAI. Quando um modelo treinado em um *dataset* (ex: FaceForensics++) apresenta uma queda de acurácia em um *dataset* não visto (ex: CelebDF-V2), o XAI pode diagnosticar a causa dessa falha. Se as explicações XAI dos vídeos "falsos" no *dataset* original se concentram em um artefato de baixa frequência, mas o modelo não consegue produzir um *heatmap* coerente para os vídeos falsos do novo *dataset*, a evidência é clara: o *deepfake* do novo conjunto de dados não possui o mesmo artefato, e o modelo antigo não conseguiu generalizar para uma **característica de manipulação semântica** mais complexa.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

O uso contínuo de XAI na pipeline de desenvolvimento, portanto, transforma-se em um **mecanismo de *feedback* e correção**. Ele garante que os modelos de detecção sejam treinados para focar nas **invariantes de falsidade** – as falhas lógicas e físicas que são difíceis de eliminar para qualquer gerador, como a inconsistência na iluminação global ou o erro no mapeamento da geometria 3D da face – em vez de artefatos de implementação facilmente eliminávels. Essa abordagem não apenas aumenta a **robustez** contra *deepfakes* não vistos, mas também fundamenta a **confiabilidade científica** do modelo.

6. IMPLICAÇÕES FORENSES E JURÍDICAS DA TRANSPARÊNCIA DO MODELO

A transição da detecção opaca para a **Inteligência Artificial Explicável (XAI)** possui profundas e transformadoras implicações para as esferas forense e jurídica. A principal barreira para a adoção de sistemas de IA em tribunais não é a falta de acurácia, mas a incapacidade de atender ao **padrão de prova científica**, que exige que a metodologia seja auditável, compreensível e passível de refutação por um perito da parte adversa. O XAI é a ferramenta que viabiliza essa admissibilidade, transformando a decisão do algoritmo em **prova pericial robusta**.

Em um processo legal, a **explicação do modelo** se torna o próprio **laudo pericial**. Um mapa de calor Grad-CAM, por exemplo, pode ser anexado ao laudo para demonstrar visualmente que o modelo baseou sua classificação de "falso" na falha de interpolação ao redor do maxilar (uma área comum de artefatos) e não em uma mancha aleatória no fundo do vídeo. Da mesma forma, os valores de Shapley podem ser usados para quantificar a **magnitude do artefato de falsidade** em termos da probabilidade de manipulação, fornecendo uma medida numérica da certeza da decisão que é transparente e logicamente fundamentada. A ausência dessa explicitação deixa a prova vulnerável a ser classificada como "evidência de caixa-preta" e, portanto, inadmissível.

A transparência gerada pelo XAI também facilita o processo de **contraditório e o *Daubert Standard*** (em jurisdições que o utilizam), onde a evidência científica deve ser testável, revisada por pares e ter uma taxa de erro conhecida. Com a XAI, o advogado da parte contrária pode analisar se o modelo está focado em artefatos espúrios ou irrelevantes, permitindo a **crítica técnica** da prova apresentada. Essa capacidade de auditar o viés do algoritmo é essencial para garantir a **imparcialidade** do sistema de justiça. Sem a explicabilidade, a única forma de refutação é alegar que o modelo "simplesmente errou", o que não é uma crítica científica válida.

Além disso, a XAI é crucial para a atribuição de responsabilidade e para a **ética da IA**. Ao justificar a classificação, a explicação XAI pode, em tese, ajudar a identificar o **tipo de artefato** deixado por um gerador específico (GAN, Autoencoder, Diffusion), auxiliando na rastreabilidade e na origem da manipulação. Em um futuro onde a legislação de IA será mais rigorosa, a capacidade de um sistema de detecção de *deepfake* de **auto-explicar-se** será um **requisito normativo**, e não apenas uma funcionalidade desejável.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

7. INTEGRAÇÃO DE XAI EM AMBIENTES DE *DEPLOYMENT* EM TEMPO REAL

Embora as técnicas de **Inteligência Artificial Explicável (XAI)** sejam academicamente robustas, sua integração em ambientes de detecção de *deepfake* em **tempo real** (como plataformas de *streaming* ou redes sociais) apresenta desafios computacionais significativos. A maioria dos métodos de explicabilidade *post-hoc* mais informativos, como SHAP e LIME, exige um **alto custo computacional**, pois eles dependem da avaliação de múltiplas perturbações da entrada ou da execução de um grande número de iterações para estimar a contribuição de cada *feature*. Essa latência adicional é muitas vezes incompatível com a exigência de processar dezenas ou centenas de quadros por segundo.

O desafio reside na busca por um *trade-off* otimizado entre a **profundidade da explicação** e a **velocidade de inferência**. Para aplicações em tempo real, as técnicas **intrinsecamente interpretáveis** e as variações otimizadas de **CAMs** se tornam as mais viáveis. O uso dos mapas de atenção gerados naturalmente pelos **Transformers** (ViTs) pode ser explorado como uma explicação "quase gratuita" em termos de latência. Uma vez que o cálculo da atenção é inerente à inferência do Transformer, a visualização da matriz de atenção pode ser extraída com custo computacional adicional mínimo, oferecendo uma **explicação espacialmente intuitiva** sem degradar significativamente a taxa de quadros por segundo.

Outra estratégia de otimização para *deployment* em larga escala é a **Explicabilidade por Amostragem Temporal**. Em vez de gerar um mapa XAI para cada quadro do vídeo, o sistema pode ser configurado para gerar explicações apenas em **quadros-chave** (por exemplo, a cada 30 ou 50 quadros) ou somente quando a probabilidade de falsidade ultrapassa um **limite de incerteza**. Essa amostragem reduz o custo computacional total e concentra o esforço de explicabilidade nos momentos mais críticos da manipulação. Isso permite que a detecção *core* (a classificação binária) continue a ser executada em alta velocidade, enquanto a **camada de *auditabilidade*** (XAI) é ativada seletivamente.

A pesquisa futura em XAI para *deepfakes* em tempo real deve focar no desenvolvimento de **modelos *post-hoc* leves** e **aproximativos**. Isso inclui o treinamento de **redes neurais explicativas** menores que aprendem a prever o mapa de calor de um modelo de caixa-preta maior (conhecimento de destilação para XAI) ou o uso de técnicas de **interrogações adversariais** otimizadas para gerar explicações em tempo constante. A integração bem-sucedida do XAI em ambientes de alta velocidade não apenas valida a decisão algorítmica para fins forenses, mas também melhora a **experiência do usuário e do moderador de conteúdo**, permitindo-lhes compreender rapidamente a origem da manipulação e tomar decisões informadas sobre a remoção ou sinalização do material.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

8. CONCLUSÃO E IMPLICAÇÕES FUTURAS

A presente análise confirmou que a **Inteligência Artificial Explicável (XAI)** é o elo indispensável que conecta a alta acurácia dos modelos de Aprendizado Profundo de detecção de *deepfakes* com as rigorosas exigências de **confiança, transparência e justificação causal** dos domínios forense e jurídico. A simples performance preditiva, inerente às arquiteturas de caixa-preta como CNNs e Transformers, não é suficiente para a admissibilidade em um tribunal, onde a prova deve ser auditável e refutável. O XAI, por meio de técnicas como **Grad-CAM e SHAP**, transforma a classificação binária opaca em um **laudo pericial detalhado**, apresentando evidências visuais e quantitativas sobre as regiões e *features* específicas que motivaram a decisão de falsidade.

O principal *insight* metodológico reside no reconhecimento de que o XAI não é um mero acessório, mas uma **ferramenta de validação e diagnóstico** que aborda a maior falha da detecção de *deepfakes*: a **má generalização decorrente** do *overfitting* a artefatos espúrios do *dataset*. Ao auditar o modelo com XAI em larga escala, os desenvolvedores podem garantir que o algoritmo está focando nas **invariantes de autenticidade** (coerência física, iluminação, geometria) em vez de artefatos de compressão ou metadados de baixo nível. A integração da **interpretabilidade intrínseca** através da análise dos mapas de autoatenção dos Vision Transformers (ViTs) representa o caminho mais promissor para alcançar a transparência com o mínimo de sobrecarga computacional, equilibrando a necessidade de alta acurácia com a demanda por rastreabilidade.

As implicações futuras deste estudo são vastas e delineiam uma agenda de pesquisa tripla. Em primeiro lugar, o desenvolvimento de **modelos híbridos XAI-otimizados** é prioritário, com foco na criação de métodos post-hoc que sejam computacionalmente eficientes e possam ser integrados a *pipelines* de detecção em tempo real sem degradação da latência. Técnicas de **destilação de conhecimento de explicabilidade** – onde um modelo menor e rápido aprende a imitar as explicações de um modelo de caixa-preta grande – serão cruciais para este fim.

Em segundo lugar, a pesquisa deve se concentrar na criação de **métricas de interpretabilidade forense**. Atualmente, a interpretabilidade é avaliada subjetivamente ou por métricas genéricas; para o domínio forense, é vital criar **escalas de confiança XAI** que quantifiquem o grau de validade da explicação, permitindo que o sistema jurídico tenha um padrão objetivo para aceitar ou rejeitar a prova algorítmica. Essa padronização é fundamental para a futura **regulamentação da IA** em contextos de segurança e justiça.

Por fim, a XAI deve ser expandida para o domínio **multimodal**, oferecendo explicações coerentes que integrem evidências de manipulação visual, auditiva e fisiológica. Ao justificar por que o áudio não corresponde ao movimento labial, ou por que o sinal de rPPG é inconsistente, a explicação XAI se torna mais robusta e difícil de ser contestada. A **transparência e a interpretabilidade** não são mais luxos; são as bases sobre as quais deve ser

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.

construída a próxima geração de defesa contra a desinformação, garantindo que a **Inteligência Artificial seja uma aliada da justiça e não uma fonte de opacidade e desconfiança**.

REFERÊNCIAS

Livros e Artigos

1. BACH, Sebastian et al. On Pixel-Wise Explanations for Non-Linear Classifier Decisionsby Layer-Wise Relevance Propagation. PLoS ONE, v. 10, n. 7, p. e0130140, 2015.
2. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: Advances in NeuralInformation Processing Systems (NIPS), 2014.
3. KINGMA, Diederik P.; BA, Jimmy. Adam: A Method for Stochastic Optimization. In: International Conference on Learning Representations (ICLR), 2015.
4. RIBEIRO, Marco Tulio; SINGH, Sameer; GUEST RINARD, Carlos. "Why Should I TrustYou?": Explaining the Predictions of Any Classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016.
5. SUNDARARAJAN, Mukund; TALY, Ankur; YAN, Vijay. Axiomatic Attribution forDeep Networks. In: International Conference on Machine Learning (ICML), 2017.
6. VASWANI, Ashish et al. Attention Is All You Need. In: Advances in Neural InformationProcessing Systems (NIPS), 2017.
7. SELVARAIU, Ramprasaath R. et al. Grad-CAM: Visual Explanations from DeepNetworks via Gradient-based Localization. In: IEEE International Conference on Computer Vision (ICCV), 2017.
8. LUNDBERG, Scott M.; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. In: Advances in Neural Information Processing Systems (NIPS), 2017.
9. FONG, Ruth; VEDANTAM, Shreya; LIM, Joo Hwee. Interpretable Deep Learning forImage Analysis. arXiv:1901.07746, 2019.
10. ROSSLER, Andreas et al. FaceForensics++: Learning to Detect Manipulated Facial Images. In: International Conference on Computer Vision (ICCV), 2019.

Este é um artigo publicado em acesso aberto (Open Access) sob a licença Creative Commons Attribution, que permite uso, distribuição e reprodução em qualquer meio, sem restrições desde que o trabalho original seja corretamente citado.