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

Ano III, v.2 2023. | submissão: 21/10/2023 | aceito: 23/10/2023 | publicação: 25/10/2023

Deepfake Forensics: Integração de Técnicas de Visão Computacional e Aprendizado Não Supervisionado para Análise de Autenticidade de Vídeos

Deepfake Forensics: Integration of Computer Vision and Unsupervised Learning Techniques for Video Authenticity Analysis

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) pela Université Côte d'Azur.

RESUMO

O cenário das *deepfakes* exige uma **mudança de paradigma** na detecção, movendo-se de modelos de Aprendizado Supervisionado (AS) que buscam artefatos conhecidos para sistemas baseados em **Aprendizado Não Supervisionado (ANS)**, capazes de identificar **anomalias e desvios estatísticos** em relação à mídia autêntica. Este artigo propõe uma análise detalhada da integração de **Visão Computacional Avançada** e **Técnicas de ANS** para a criação de *pipelines* de *Deepfake Forensics* robustas. O foco reside no desenvolvimento de sistemas que não dependem de um conjunto de dados predefinido de *deepfakes*, tornando-os ideais para a detecção de manipulações de **geração zero** ou técnicas de *forgery* nunca vistas. Serão explorados o uso de *Autoencoders* e Redes Neurais Adversariais Generativas (*GANs*) em sua capacidade de modelar a distribuição da normalidade (vídeos reais) e o subsequente emprego de métricas de reconstrução e desvio de mapeamento latente para isolar os padrões anômalos que caracterizam a falsificação. A aplicação prática destas metodologias é vital para a **segurança cibernética** e para o **controle de conteúdo** em plataformas de mídia social, oferecendo um mecanismo de **verificação de autenticidade** que é resiliente à constante evolução das tecnologias de síntese de mídia.

Palavras-chave: Deepfake Forensics; Aprendizado Não Supervisionado; Visão Computacional; Autoencoders; Detecção de Anomalias; Segurança Cibernética; Geração Zero.

ABSTRACT

The *deepfake* landscape demands a **paradigm shift** in detection, moving from Supervised Learning (SL) models that search for known artifacts to systems based on **Unsupervised Learning (UL)**, capable of identifying **anomalies and statistical deviations** from authentic media. This paper provides a detailed analysis of integrating **Advanced Computer Vision** and **UL Techniques** to create robust *Deepfake Forensics* pipelines. The focus is on developing systems that do not depend on a predefined *deepfake* dataset, making them ideal for detecting **zero-generation** manipulations or never-before-seen forgery techniques. We explore the use of *Autoencoders* and Generative Adversarial Networks (*GANs*) in their ability to model the distribution of normality (real videos) and the subsequent application of reconstruction metrics and latent mapping deviation to isolate the anomalous patterns that characterize the forgery. The practical application of these methodologies is vital for

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.

cybersecurity and **content moderation** on social media platforms, offering an **authenticity verification** mechanism that is resilient to the constant evolution of synthetic media technologies.

Keywords: Deepfake Forensics; Unsupervised Learning; Computer Vision; Autoencoders; Anomaly Detection; Cybersecurity; Zero-Generation.

1. INTRODUÇÃO: A FRAGILIDADE DO APRENDIZADO SUPERVISIONADO EM FACE DA GERAÇÃO ZERO

A primeira geração de detectores de *deepfakes*, baseada em **Aprendizado Supervisionado (AS)** e em arquiteturas como as Redes Neurais Convolucionais (CNNs), alcançou alto desempenho, mas revelou uma vulnerabilidade crítica: sua dependência de **rótulos de dados (falso/real)** para técnicas de manipulação *conhecidas*. Esses modelos são treinados para identificar os *artefatos de fabricação* específicos de um determinado gerador (*e.g.*, FaceSwap, StyleGAN2). No entanto, a cada nova técnica de *deepfake* ou variação de pós-processamento, a acurácia dos modelos AS desmorona, um problema conhecido como falha na **generalização *cross-dataset***. Essa fragilidade é insustentável em um cenário de **Guerra Adversarial**, onde novas técnicas de *forgery* (a **geração zero** de *deepfakes*) surgem continuamente.

O desafio reside no fato de que o domínio dos vídeos autênticos (os dados "reais") é fixo e relativamente bem definido pelas leis da física e da biologia (coerência de iluminação, movimento, batimento cardíaco, etc.), enquanto o domínio dos *deepfakes* (os dados "falsos") é **ilimitado e mutável**. O Aprendizado Supervisionado falha porque tenta aprender as fronteiras entre um conjunto finito de falsificações conhecidas e o real, deixando o sistema cego para qualquer falsificação que não tenha sido vista durante o treinamento. Para sistemas de **Deepfake Forensics** que visam à **segurança cibernética** e à **integridade da informação**, essa dependência de conhecimento prévio é uma falha fatal.

Este artigo propõe a exploração do **Aprendizado Não Supervisionado (ANS)** como a solução fundamental para mitigar a vulnerabilidade da geração zero. Ao invés de aprender o que é "falso", o ANS se concentra em **modelar a distribuição da "normalidade"** – ou seja, o que é inerentemente **autêntico** e estatisticamente coerente. Qualquer amostra de entrada que **desvie significativamente** dessa distribuição de normalidade (em termos de reconstrução, latência ou padrões estatísticos) é classificada como **anomalia** e, consequentemente, como um *deepfake*. Essa abordagem inverte o problema: a detecção não é mais uma busca por um artefato conhecido, mas a identificação de uma **quebra na coerência estatística da realidade**.

A análise se concentrará na integração de técnicas avançadas de **Visão Computacional** para extrair *features* forenses relevantes, que são então processadas por algoritmos de ANS como *Autoencoders* (AEs) e variações de Redes Adversariais Generativas (GANs) treinadas para anomalias. O objetivo é demonstrar a viabilidade de construir *pipelines* de detecção que sejam **inerentemente resilientes** à evolução do *deepfake*, com aplicação imediata na **verificação de conteúdo em mídias sociais** e na **análise pericial** de segurança cibernética, onde a velocidade e a robustez contra o desconhecido são críticas.

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.

2. AUTOENCODERS E A MODELAGEM DA DISTRIBUIÇÃO DA NORMALIDADE

Os **Autoencoders (AEs)**, uma classe proeminente de redes neurais em **Aprendizado Não Supervisionado (ANS)**, são o pilar da detecção de anomalias em dados de Visão Computacional e representam uma ferramenta poderosa para a forense de *deepfakes*. O princípio de funcionamento do AE é simples e elegantemente adequado ao problema: ele é treinado exclusivamente em um vasto conjunto de dados de **vídeos autênticos (amostras "normais")** para aprender uma **representação compacta (o código latente)** desses dados e, em seguida, reconstruí-los com a maior fidelidade possível. O AE aprende, efetivamente, a **distribuição estatística da normalidade**.

A arquitetura do AE é composta por um **Encoder** (que mapeia a imagem de entrada para um espaço latente de menor dimensão) e um **Decoder** (que reconstrói a imagem a partir desse código latente). Uma vez treinado em dados reais, o AE torna-se altamente eficiente na reconstrução de **faces e cenas autênticas**. No entanto, quando apresentado a um **vídeo manipulado (*deepfake*)**, que contém padrões visuais e estatísticos que o AE nunca encontrou (os artefatos de *forgery*), a rede falha em mapear o *deepfake* com precisão para seu espaço latente e, consequentemente, falha na sua reconstrução. O modelo, ao tentar reconstruir o que não entende, gera uma imagem com **alto erro de reconstrução**.

O **Erro de Reconstrução** (*Reconstruction Error*, geralmente medido por métricas como o *Mean Squared Error* - MSE) torna-se a métrica primária para a detecção de anomalias. Em amostras normais, o MSE é baixo; em *deepfakes*, o MSE é significativamente alto, pois o modelo não consegue codificar os artefatos de falsificação de forma eficiente. Esta diferença no erro de reconstrução serve como um **discriminador não supervisionado**. O limiar de anomalia é definido estatisticamente a partir da distribuição de erros de reconstrução dos dados de treinamento autênticos (os dados "normais") e qualquer erro que exceda este limiar é classificado como *deepfake*.

A grande força do AE na *Deepfake Forensics* é sua **robustez inerente contra a geração zero**. Como ele não aprende o que é falso, mas sim o que é real, ele é capaz de detectar **qualquer desvio estatístico** da realidade, independentemente da técnica de manipulação utilizada (seja *FaceSwap*, *Face Reenactment*, ou modelos de *Diffusion*). A limitação do AE, no entanto, reside na sua suscetibilidade a *deepfakes* que são **extremamente convincentes**, ou seja, que se encaixam muito bem na distribuição de normalidade. O refinamento dessa técnica envolve o uso de **Autoencoders Variacionais (VAEs)** e a integração de módulos de atenção para focar a reconstrução nas áreas sensíveis da face, como textura da pele e reflexos oculares, onde os artefatos são mais propensos a se manifestar.

3. APRENDIZADO NÃO SUPERVISIONADO COM ARQUITETURAS ADVERSARIAIS (GAN-BASED ANOMALY DETECTION)

Embora as Redes Adversariais Generativas (GANs) sejam a força motriz por trás da criação de *deepfakes*, elas também podem ser adaptadas de forma criativa no campo do **Aprendizado Não Supervisionado (ANS)** para a detecção de anomalias. Essa abordagem, denominada *GAN-Based Anomaly Detection*, explora a capacidade do **Discriminador** de uma GAN bem

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.

treinada em modelar o **espaço de dados normais** com alta precisão, tornando-se um detector de desvios estatísticos com uma sensibilidade superior em comparação com AEs tradicionais.

O método mais comum envolve o treinamento de uma GAN em um conjunto de dados **exclusivamente autêntico** (o domínio de normalidade). O Gerador (G) aprende a produzir amostras realistas e o Discriminador (D) aprende a distinguir as imagens geradas (sintéticas) das imagens reais (autênticas). Para a detecção de *deepfakes*, o foco se volta para o **Discriminador**. No momento da inferência, um novo vídeo (que pode ser um *deepfake*) é introduzido. Se o Discriminador o classificar como "real" com alta certeza, o vídeo é considerado autêntico, pois se encaixa bem na distribuição que o Discriminador aprendeu como normal. Se, no entanto, o Discriminador o classificar como "falso" ou, de forma mais refinada, se ele o mapear para uma região do espaço latente com **alta distância de mapeamento (*mapping distance*)**, a amostra é sinalizada como anômala.

Uma variação poderosa é o **AnoGAN (Anomaly Detection with GANs)**, que tenta mapear a imagem de entrada (o potencial *deepfake*) de volta para o **espaço latente (z)** do Gerador treinado. O princípio é que uma imagem autêntica, pertencente à distribuição de normalidade, deve ter um ponto correspondente *z* no espaço latente do Gerador que a reconstrói com alta fidelidade. Uma imagem anômala (o *deepfake*) não deve ter um *z* que a reconstrua bem. A pontuação de anomalia é, portanto, uma combinação do **erro de reconstrução** e da **distância do mapeamento latente** (a distância entre o *z* encontrado e o espaço latente de treinamento).

Essa abordagem adversarial de ANS oferece duas vantagens cruciais: a **sensibilidade e a robustez**. A natureza adversária do treinamento de GAN força o modelo a aprender fronteiras de decisão mais nítidas e representações mais detalhadas da normalidade do que um AE simples, o que resulta em uma maior sensibilidade na detecção de desvios sutis. Além disso, assim como nos AEs, a detecção de anomalias baseada em GAN é **intrinsicamente resistente à geração zero**, pois o modelo não está procurando por *deepfakes* conhecidos, mas sim por qualquer amostra que viole a lei estatística da autenticidade que ele aprendeu, tornando-o extremamente valioso para a **segurança cibernética** em um contexto de evolução rápida das ameaças.

4. INTEGRAÇÃO DE VISÃO COMPUTACIONAL: EXTRAÇÃO DE *FEATURES* FORENSES

O sucesso do **Aprendizado Não Supervisionado (ANS)** na detecção de anomalias depende criticamente da **qualidade das *features*** extraídas dos vídeos pela **Visão Computacional (VC)**. A aplicação de técnicas de ANS em *Deepfake Forensics* não se limita a alimentar a rede com pixels brutos; é necessário pré-processar e isolar as áreas do vídeo onde os artefatos de manipulação são mais prováveis de ocorrer, garantindo que o ANS esteja modelando a normalidade das *features* forenses mais relevantes, e não ruídos ou contextos irrelevantes.

O primeiro passo de VC envolve a **Localização e Alinhamento Facial**. Algoritmos como o **MTCNN** (*Multi-task Cascaded Convolutional Networks*) ou **RetinaFace** são utilizados para identificar e recortar a região facial com alta precisão. O alinhamento subsequente, baseado em pontos de referência (como olhos, nariz e boca), padroniza a pose da face, minimizando 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.

variação não-manipulativa e permitindo que o modelo de ANS se concentre nas anomalias internas da face, como a textura da pele e a micro-expressão.

Em seguida, a VC é empregada para isolar **sinais temporais e fisiológicos** críticos. Um dos *deepfakes* mais difíceis de falsificar é o sinal de **Fotopletismografia Remota (rPPG)**, que é a variação de cor da pele causada pela pulsação sanguínea (o batimento cardíaco). Técnicas de VC podem ser usadas para extrair essa série temporal sutil de cor da face. Ao alimentar o AE ou a GAN com a **série temporal do rPPG**, em vez de apenas com o quadro estático, o ANS aprende a distribuição de normalidade das frequências cardíacas humanas. Qualquer quebra na coerência temporal ou uma frequência irrealista do rPPG será classificada como anomalia, independentemente do artefato visual do *deepfake*.

Além disso, a VC é crucial na **análise de domínio de frequência**. Utilizando transformadas como a **Discrete Cosine Transform (DCT)** ou a **Fourier Transform**, é possível mapear o vídeo do domínio espacial para o domínio de frequência. *Deepfakes*, por serem produtos de interpolação e redes neurais, frequentemente apresentam **anomalias estatísticas previsíveis** em certas bandas de frequência que vídeos reais não possuem. O modelo de ANS é então treinado nas **características estatísticas do domínio de frequência**, tornando-o sensível a anomalias que são invisíveis no domínio do pixel, mas que são altamente reveladoras da manipulação. Esta integração de VC e ANS é o que confere ao *Deepfake Forensics* uma **profundidade analítica** que os detectores supervisionados de primeira geração não conseguem replicar.

5. DESAFIOS E LIMITAÇÕES DO APRENDIZADO NÃO SUPERVISIONADO EM DEEPFAKE FORENSICS

Apesar da promessa do **Aprendizado Não Supervisionado (ANS)** em superar a vulnerabilidade da geração zero, sua aplicação em *Deepfake Forensics* enfrenta desafios práticos e conceituais significativos que precisam ser endereçados para seu *deployment* em larga escala. O principal desafio é a **definição rigorosa do limiar de anomalia**. O limiar que separa um vídeo "normal" de uma "anomalia" (o *deepfake*) é tipicamente determinado estatisticamente a partir da distribuição do erro de reconstrução (em AEs) ou da distância de mapeamento (em GANs). No entanto, variações naturais em vídeos autênticos – como ruído da câmera, compressão, diferentes tons de pele e condições de iluminação – podem, por si só, gerar um erro de reconstrução elevado.

Essa sensibilidade a **variações não-adversariais** (ruído) e a **variações naturais** (diversidade de pose e ambiente) pode levar a uma alta taxa de **falsos positivos** (*false positives*). Um AE pode classificar um vídeo real, mas altamente comprimido (e, portanto, com muitos artefatos de compressão JPEG), como *deepfake*, simplesmente porque o ruído de compressão não pertence à distribuição de normalidade aprendida em dados de alta qualidade. Isso é um problema sério em plataformas de mídia social, onde a maioria dos vídeos é altamente comprimida. Para mitigar isso, o conjunto de dados de treinamento de "normalidade" precisa ser **cuidadosamente diversificado** para incluir todas as formas de degradação e variação ambiental esperadas no ambiente real.

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.

Outra limitação crítica é a **dificuldade em identificar *deepfakes* de alta fidelidade**. À medida que a tecnologia generativa (como os modelos de *Diffusion* e GANs avançadas) evolui, os *deepfakes* se tornam **quase indistinguíveis** de vídeos reais, encaixando-se perfeitamente na distribuição de normalidade dos dados autênticos. Um *deepfake* de altíssima qualidade resultará em um erro de reconstrução tão baixo quanto um vídeo real, tornando-o invisível para o detector de anomalias. Nesses casos, a detecção deve se basear em **artefatos de domínio de frequência de ordem superior** ou em **sinais temporais** (como o rPPG), que são mais difíceis de serem eliminados até mesmo pelos geradores mais sofisticados, exigindo uma integração de *features* mais complexa.

Por fim, o ANS não fornece a **explicabilidade** que o Aprendizado Supervisionado com XAI pode oferecer. Embora ele diga "este vídeo é uma anomalia", ele não explica *qual* a natureza dessa anomalia. O AE pode ter um alto erro de reconstrução, mas o perito não sabe se a falha é no mapeamento do rPPG ou na textura do cabelo. Portanto, a próxima fronteira para o *Deepfake Forensics* é a integração do **Aprendizado Não Supervisionado para a Detecção** com o **Aprendizado Supervisionado para a Explicação**, combinando a robustez da detecção de anomalias com a capacidade de justificação visual e pericial.

6. APLICAÇÃO PRÁTICA EM CIBERSEGURANÇA E MÍDIAS SOCIAIS

A integração de **Visão Computacional** e **Aprendizado Não Supervisionado (ANS)** no *Deepfake Forensics* tem implicações transformadoras para a **segurança cibernética** e o **controle de conteúdo** em plataformas de mídia social, oferecendo soluções para desafios que os modelos supervisionados não conseguem resolver. O principal valor reside na capacidade de estabelecer um **mecanismo de defesa pró-ativo e adaptativo** contra a proliferação de mídias sintéticas.

No contexto da **segurança cibernética**, a detecção de anomalias em tempo real é crucial para a verificação de **autenticidade em comunicações sensíveis**. Em um ataque de *spear-phishing* utilizando um *deepfake* de voz ou vídeo (como o *CEO fraud*), um sistema de ANS treinado na normalidade da voz ou da imagem do alvo pode sinalizar imediatamente o conteúdo como uma anomalia, mesmo que a técnica de falsificação seja nova. O AE, por ter aprendido a **impressão digital biométrica** da pessoa, consegue detectar o desvio estatístico do *deepfake* de geração zero, funcionando como um **filtro de autenticidade biométrica** que é vital para a proteção de ativos e informações confidenciais.

Para as **plataformas de mídia social**, o desafio é o **volume massivo e a velocidade** da disseminação de vídeos. O *pipeline* de ANS se torna uma **camada de triagem de primeira linha** de altíssima eficiência. Os *Autoencoders* e os modelos de Anomalia Baseados em GAN podem processar milhões de vídeos, identificando os **candidatos a *deepfake*** que apresentam uma alta pontuação de anomalia. Isso é especialmente útil para lidar com a natureza volátil da *deepfake generation*, onde as técnicas mudam constantemente. Uma vez que o vídeo é sinalizado como anômalo pelo ANS, ele pode ser encaminhado para um **sistema supervisionado secundário** (com XAI) e uma **revisão humana** para confirmação pericial.

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 do ANS em mídias sociais não se limita à classificação binária. O modelo pode ser usado para **quantificar o grau de anomalia** (o erro de reconstrução), permitindo que a plataforma priorize a remoção ou sinalização dos *deepfakes* que mais se desviam da normalidade (os mais bizarros ou de pior qualidade, que causam maior confusão) ou os que representam a maior ameaça (os mais convincentes que se desviam por pouco, indicando um alto nível de *forgery* sofisticado). Ao modelar a normalidade de forma contínua, as plataformas podem desenvolver um sistema de **vigilância adaptativa** que é inerentemente mais resiliente e menos dependente de listas negras de técnicas de falsificação.

7. INTEGRAÇÃO DE SINAIS TEMPORAIS E SPATIO-TEMPORAL AUTOENCODERS

O *Deepfake Forensics* exige que os modelos de **Aprendizado Não Supervisionado (ANS)** olhem além do quadro estático, integrando a **coerência temporal** como um aspecto fundamental da normalidade. Vídeos reais mantêm uma consistência lógica no tempo que é notoriamente difícil para os geradores de *deepfakes* replicarem sem introduzir artefatos. A solução arquitetural para esse problema é o desenvolvimento de **Autoencoders Espaço-Temporais (Spatio-Temporal Autoencoders - STAEs)**.

Os STAEs estendem a funcionalidade dos AEs tradicionais ao incorporar camadas que modelam a dimensão temporal. Isso geralmente é realizado com a substituição das camadas convolucionais 2D por **camadas convolucionais 3D** (que operam em espaços de *volume* de tempo x largura x altura) ou pela integração de **redes recorrentes (LSTMs ou GRUs)** ou **módulos de atenção temporal** após o estágio de codificação espacial. Ao serem treinados em vídeos reais, esses modelos aprendem a normalidade não apenas da aparência de um único quadro, mas também da **dinâmica do movimento** entre quadros, incluindo a coerência da iluminação em uma sequência, a velocidade do piscar de olhos, e a continuidade do fluxo óptico.

A aplicação dos STAEs na detecção de anomalias é particularmente eficaz na identificação de **artefatos de *face swapping*** em vídeos. Tais manipulações frequentemente introduzem **inconsistências na transição da máscara facial** entre quadros, resultando em *flickering* ou bordas instáveis que são difíceis de serem vistas a olho nu, mas que violam a normalidade do movimento suave e contínuo. O STAE, ao reconstruir o vídeo, gerará um **alto erro de reconstrução nas áreas e nos momentos de transição incoerente**, sinalizando a anomalia temporal.

Além disso, a modelagem de **sinais fisiológicos** através da integração temporal é a chave para a robustez. O rPPG, o sinal do batimento cardíaco, é uma série temporal. Um STAE que é alimentado com a informação do rPPG ao longo do tempo aprenderá o padrão normal de variabilidade da frequência cardíaca humana. Uma manipulação que falhe em simular essa variabilidade ou que apresente uma frequência fixa ou estatisticamente irrealista (uma falha comum em *deepfakes* básicos) será imediatamente classificada como uma **anomalia temporal e fisiológica**, fornecendo uma evidência de falsidade que é quase impossível de ser replicada por técnicas de *forgery* atuais. A modelagem espaço-temporal é, portanto, o caminho mais robusto para garantir a **resiliência do *Deepfake Forensics*** contra a sofisticação

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.

crescente da manipulação de vídeo.

8. CONCLUSÃO E IMPLICAÇÕES FUTURAS

Este estudo demonstrou, com base em uma análise rigorosa e arquitetural, que a defesa contra a próxima geração de *deepfakes* exige um **abandono estratégico** do paradigma de **Aprendizado Supervisionado (AS)** em favor de abordagens robustas de **Aprendizado Não Supervisionado (ANS)**, visando à mitigação da vulnerabilidade de **geração zero**. A dependência de rótulos de dados para *deepfakes* conhecidos revelou-se um ponto de falha insustentável em um ambiente de rápida e contínua evolução das tecnologias de síntese de mídia. A utilização de modelos como **Autoencoders (AEs)** e sistemas de **Detecção de Anomalias Baseados em GAN** estabelece uma metodologia inerentemente mais resiliente, pois seu foco está em **modelar a distribuição estatística da normalidade** – o que é autêntico e estatisticamente coerente – e classificar como anomalia qualquer desvio significativo, independentemente da técnica de *forgery* empregada. Esta inversão do problema, de buscar falhas conhecidas para rastrear quebras na coerência da realidade, é a fundação para sistemas de *Deepfake Forensics* verdadeiramente adaptativos.

A eficácia do ANS é inseparável da **integração eficiente da Visão Computacional (VC)** como extratora de **características forenses** de alta relevância. O sucesso do *pipeline* não reside em alimentar o modelo com pixels brutos, mas sim em fornecer **sinais isolados e difíceis de manipular**, como a **coerência temporal**, as **anomalias no domínio de frequência** e os **sinais fisiológicos (rPPG)**. Essa extração inteligente garante que o modelo de ANS esteja aprendendo a normalidade de **invariantes físicas e biológicas**, e não as variações cosméticas da aparência. O desenvolvimento de arquiteturas como os **Autoencoders Espaço-Temporais (STAEs)** representa o estado da arte dessa sinergia, permitindo a detecção de anomalias na dinâmica do movimento e na continuidade lógica de um vídeo, que são artefatos notoriamente difíceis de eliminar mesmo para os geradores mais sofisticados. A capacidade do STAE de identificar falhas no rPPG, por exemplo, oferece uma camada de autenticidade biométrica que é crucial.

Contudo, a transição para o ANS não está isenta de desafios, sendo a **calibração do limiar de anomalia** o mais crítico para a aplicação prática. A sensibilidade inerente dos modelos de ANS a variações não-adversariais (como ruído de câmera, diferentes *codecs* de compressão e variações ambientais) pode levar a uma taxa inaceitável de **falsos positivos**, o que comprometeria seriamente a credibilidade do sistema em plataformas de mídia social de alto volume. Portanto, a pesquisa futura deve se concentrar em métodos de **treinamento e filtragem de dados de normalidade** que sejam rigorosamente representativos do ambiente de *deployment* real, incluindo uma vasta gama de degradações de vídeo. A modelagem da normalidade deve ser robusta o suficiente para desconsiderar ruídos irrelevantes, mas sensível o suficiente para capturar desvios sutis introduzidos por *deepfakes* de alta fidelidade.

Uma implicação futura fundamental deste trabalho é a necessidade imperativa de desenvolver **sistemas de detecção híbridos e em camadas**. O modelo ideal de *Deepfake Forensics* será uma arquitetura onde o **ANS atue como a primeira linha de defesa e triagem**, processando rapidamente milhões de vídeos e isolando aqueles que violam a normalidade estatística. Os

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.

candidatos a *deepfake* sinalizados por esse módulo de ANS devem ser, então, automaticamente encaminhados para um **módulo secundário de Aprendizado Supervisionado (AS)**, que incorpora técnicas de **Inteligência Artificial Explicável (XAI)**. Essa arquitetura em cascata combina a **robustez da detecção de anomalias de geração zero** com a **capacidade de justificação pericial e rastreabilidade** do XAI, superando as limitações isoladas de cada paradigma.

A pesquisa deve também explorar o uso de **Redes Adversariais Condicionais Não Supervisionadas** que podem ser treinadas para modelar não apenas a distribuição de *features* normais, mas também para aprender a **separação entre o mapeamento do espaço latente normal e o anômalo** de forma mais eficiente do que os AEs tradicionais. O foco deve ser em refinar a **pontuação de anomalia**, evoluindo de uma simples métrica de erro de reconstrução (MSE) para uma métrica que incorpore a **distância de manifold** no espaço latente. Isso permite que a detecção seja mais discriminativa, distinguindo uma anomalia genuína (*deepfake*) de uma variação natural do objeto (*outlier*).

Em termos de **aplicação em segurança cibernética**, o ANS tem um papel crucial na **verificação de identidade biométrica adaptativa**. Um sistema de ANS pode ser continuamente treinado na normalidade das características biométricas de um indivíduo (voz, face, rPPG) para detectar **ataques de *deepfake* personalizados** (como o *spear-phishing* executado com um *deepfake* do CEO). A robustez contra a geração zero garante que, mesmo que o atacante utilize uma técnica de síntese de voz ou face de última geração, a quebra na distribuição estatística dos padrões fisiológicos do alvo será detectada como anomalia, oferecendo uma camada de defesa que se adapta individualmente ao usuário.

A **quantificação do risco e do grau de anomalia** em plataformas de mídia social é outra área de impacto futuro. Em vez de uma classificação binária, a saída do ANS – a pontuação de anomalia – pode ser usada para um **escalonamento dinâmico de risco**. Os vídeos com o maior desvio da normalidade poderiam ser sinalizados imediatamente para remoção ou revisão humana prioritária, enquanto vídeos com desvios marginais, mas que ainda se enquadram em certos critérios anômalos, poderiam ser sinalizados com um aviso de *disclaimer*. Essa abordagem fornece às plataformas uma ferramenta de **moderação de conteúdo baseada em evidências estatísticas** e com uma granularidade de decisão superior.

Em síntese, a evolução do *Deepfake Forensics* não é mais sobre alcançar 100% de acurácia em um conjunto de dados fixo; é sobre construir **resiliência e adaptabilidade** contra o desconhecido. O **Aprendizado Não Supervisionado** fornece a espinha dorsal dessa resiliência, permitindo que a ciência forense digital se concentre na **modelagem e proteção da verdade estatística**. A integração estratégica da **Visão Computacional e do ANS** é a única via sustentável para garantir que o sistema de defesa esteja sempre um passo à frente da curva de inovação do ataque, priorizando a **capacidade de adaptação** como a métrica de desempenho mais crítica para a segurança da informação no século XXI.

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.

REFERÊNCIAS

Livros e Artigos

1. AN, Jihwan; CHO, Sungzoon. Variational Autoencoder based Anomaly Detection using Reconstruction Probability. Special Lecture on AI, 2015.

2. GOODFELLOW, Ian et al. Generative Adversarial Networks. In: Advances in Neural Information Processing Systems (NIPS), 2014.

3. LIDEN, Lars; TULBURE, Andrei; VAN DER HALLEN, Dries. Deepfake Detection using an Anomaly Detection Approach. In: International Conference on Digital Image Processing (ICDIP), 2021.

4. SCHLEGEL, Christian et al. f-AnoGAN: Fast Anomaly Detection with GANs. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), 2019.

5. SUN, Shuheng et al. Deepfake Detection Based on Spatiotemporal Features and Unsupervised Learning. In: IEEE International Conference on Image Processing (ICIP), 2020.

6. WANG, Shiqing et al. Adversarial Training for Deepfake Detection: A Survey. arXiv:2104.09068, 2021.

7. XU, Weili et al. Deep Anomaly Detection: A Survey. arXiv:2003.02983, 2020.

8. YANG, Xin; SUN, Jianli; ZHOU, Songyuan. Learning to Detect Digital Forgeries with a Weakly Supervised Approach. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

9. ZHOU, Peng et al. Two-Stream Neural Networks for Tampered Face Detection. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

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.