

Detecção de *Deepfakes* Baseada em Metadados Forenses Utilizando Perceptron de Múltiplas Camadas

Claudio Caueh O. Xavier

Estudante de Engenharia de Software
Universidade Federal Rural do Semi-Árido (UFERSA)
Rio Grande do Norte, Brasil
cqueholiveira@gmail.com

Rosana C. B. Rego

Departamento de Engenharias e Tecnologia
Universidade Federal Rural do Semi-Árido (UFERSA)
Rio Grande do Norte, Brasil
rosana.rego@ufersa.edu.br

Abstract—As tecnologias de síntese de mídias falsas, denominadas *deepfakes*, constituem um dos principais desafios contemporâneos à segurança digital. Com o avanço das ferramentas de manipulação de vídeo e áudio, a distinção entre conteúdo autêntico e sintetizado artificialmente tornou-se progressivamente complexa. Este trabalho propõe uma abordagem baseada em metadados forenses: em vez de realizar análise em nível de *pixel*, o método proposto opera sobre características numéricas abstratas da mídia, tais como inconsistências na sincronização audiovisual e anomalias de compressão. Utilizou-se um *dataset* científico contendo essas métricas forenses, sobre o qual foi desenvolvida uma Rede Neural Artificial do tipo *Multilayer Perceptron* (MLP). O modelo proposto demonstrou capacidade de classificação binária entre mídias autênticas e *deepfakes*. Os resultados confirmam que as assinaturas forenses estruturais presentes em *deepfakes* constituem características discriminativas suficientes para sistemas de detecção computacionalmente eficientes, demandando recursos substancialmente inferiores quando comparados a abordagens baseadas puramente em visão computacional profunda. O modelo alcançou acurácia de 100% no conjunto de teste, com ausência total de falsos positivos e falsos negativos.

Index Terms—*deepfake*, forense digital, redes neurais artificiais, aprendizado de máquina, *multilayer perceptron*, interpretabilidade

I. INTRODUÇÃO E REVISÃO DE LITERATURA

A. O Problema dos *Deepfakes* na Sociedade

A geração automatizada de conteúdo multimídia por meio de inteligência artificial (IA) proporcionou significativos avanços em diversos domínios, contudo, introduziu desafios substanciais à autenticidade digital: os *deepfakes* [1]. Essa tecnologia, frequentemente fundamentada em *Generative Adversarial Networks* (GANs), possibilita a substituição de faces e a síntese de voz em imagens e vídeos com elevado grau de realismo visual e acústico.

Esse cenário configura-se como um desafio crítico para pesquisadores de segurança da informação, uma vez que os *deepfakes* podem ser empregados para disseminação de desinformação, fraudes financeiras e difamação de indivíduos públicos. O problema central reside na democratização das ferramentas de síntese, cuja barreira técnica para produção de *deepfakes* de alta qualidade foi substancialmente reduzida,

possibilitando que usuários sem conhecimento especializado tenham acesso a sistemas avançados de geração de faces sintéticas.

O crescimento exponencial dessas mídias sintéticas demanda métodos de detecção cada vez mais eficientes e precisos. Inicialmente, as abordagens de identificação de *deepfakes* baseavam-se em inspeção visual manual, todavia, as imagens sintéticas contemporâneas apresentam qualidade perceptual próxima à de mídias autênticas [2]. Consequentemente, a utilização de IA para detectar artefatos gerados por outras arquiteturas de *machine learning* consolidou-se como abordagem predominante na literatura científica. Entretanto, grande parte desses métodos fundamenta-se em análise intensiva de características em nível de *pixel*, demandando infraestrutura computacional de alto desempenho e processos de treinamento extensos, com duração de múltiplas horas ou dias. Essa elevada demanda computacional limita sua aplicabilidade em sistemas escaláveis de detecção automatizada, como plataformas de redes sociais.

B. Proposta Focada em Metadados Forenses

As diversas técnicas de síntese de mídias audiovisuais artificiais geram diferentes assinaturas forenses imperceptíveis à percepção humana direta. A literatura científica evidencia que os processos de geração de *deepfakes* apresentam significativa heterogeneidade metodológica [4]. Essa premissa tem direcionado pesquisadores a uma mudança de paradigma: em vez de analisar o conteúdo visual em nível de *pixel*, investigar metadados forenses e características estatísticas secundárias.

Essa abordagem motiva o emprego de atributos multidimensionais, tais como os indicadores forenses utilizados no presente estudo [4]. Diversos estudos demonstram que a detecção pode ser realizada de forma mais eficiente fundamentando-se nessas anomalias sutis de compressão e inconsistências estruturais. Uma vez que os algoritmos geradores de mídias sintéticas otimizam prioritariamente a fidelidade visual, aspectos secundários do processamento, como estabilidade de compressão em regiões de fundo e sincronia labial, frequentemente apresentam artefatos detectáveis estatisticamente.

A utilização dessas características tabuladas demanda apenas uma fração mínima de recursos computacionais quando comparada aos métodos tradicionais de visão computacional. Portanto, o objetivo fundamental e explícito desta pesquisa consiste em apresentar uma abordagem preditiva de baixo custo computacional baseada em metadados forenses para a detecção de *deepfakes*. Pretende-se estruturar e validar um algoritmo *Multilayer Perceptron* (MLP) capaz de operar sobre características numéricas abstratas e gerar previsões em escala de milissegundos.

C. Trabalhos Relacionados

A literatura científica evidencia que os processos de geração de *deepfakes* apresentam significativa heterogeneidade metodológica [4]. Trabalhos recentes indicam uma mudança de paradigma: em vez de focar puramente na inspeção de conteúdo em nível de *pixel*, diversos estudos demonstram que a detecção e classificação podem ser realizadas de forma mais eficiente empregando metadados e atributos multidimensionais fundamentados em anomalias sutis de compressão [4]. O método proposto neste artigo avança nessa discussão ao explorar uma arquitetura leve operando estritamente sobre características de inconsistência estrutural e sincronia, visando superar a elevada demanda de recursos das abordagens tradicionais de visão computacional.

II. CARACTERÍSTICAS DA MODELAGEM EXPERIMENTAL

A. A Base de Dados Original

Para garantir uma avaliação científica rigorosa e reprodutível, os dados utilizados no escopo desta pesquisa pertencem ao *dataset* público internacional intitulado “*Deepfake and Synthetic Media Detection*” [3]. Esse repositório, amplamente utilizado para fins de pesquisa acadêmica, contém atributos e características representativas de cenários reais, refletindo as assinaturas forenses presentes tanto em mídias sintéticas (classe *Fake*) quanto em mídias autênticas não manipuladas (classe *Real*).

O *dataset* não contém arquivos multimídia brutos, mas sim características numéricas pré-extraídas e tabuladas para cada registro. Após a etapa de análise exploratória de dados e auditoria de interpretabilidade, o modelo foi otimizado para operar estritamente sobre três atributos forenses essenciais:

- 1) Índice de dessincronização labial: métrica que quantifica a discrepância temporal entre os movimentos labiais e o sinal de áudio correspondente (*lip sync score*).
- 2) Pontuação de artefatos visuais: métrica que avalia a presença de anomalias e artefatos de compressão em regiões de fundo da imagem (*visual artifacts score*).
- 3) Pontuação de inconsistência de iluminação: métrica que detecta padrões atípicos ou incoerentes de iluminação decorrentes de manipulações digitais (*lighting inconsistency score*).

Para evitar que o modelo memorizasse informações em vez de aprender padrões reais, identificadores que causam *Data Leakage* foram estritamente removidos.

B. A Padronização e Divisão do Processo

Em seguida, as três características numéricas contínuas selecionadas foram submetidas a padronização estatística. Esta etapa é fundamental para prevenir que atributos com escalas numéricas distintas introduzam viés no processo de otimização dos pesos sinápticos. Utilizando a biblioteca *Scikit-learn* [6], o método *StandardScaler* normalizou cada característica para apresentar média $\mu = 0$ e desvio padrão $\sigma = 1$.

Por fim, os dados foram submetidos ao particionamento estratificado. Utilizou-se a função `train_test_split` para separar 80% das amostras para o conjunto de treinamento, e os 20% remanescentes constituíram o conjunto de teste (*test set*), preservando a proporção original das classes mediante estratificação [6].

C. Proposição da Arquitetura e Modelagem Matemática

A arquitetura proposta consiste em um modelo *Multilayer Perceptron* (MLP) adaptado para dados de baixa dimensionalidade (3 características de entrada). De natureza *feedforward*, a rede executa propagação unidirecional da camada de entrada até a camada de saída. A arquitetura modular proposta pode ser observada na Figura 1. Para a otimização dos pesos sinápticos, empregou-se o algoritmo *Adaptive Moment Estimation* (Adam). A função de perda adotada foi a entropia cruzada binária (*Binary Cross-Entropy*), matematicamente definida como:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \quad (1)$$

onde N é o número de amostras, y_i é a classe verdadeira (0 ou 1) e $\hat{y}_i$ é a probabilidade predita pelo modelo.

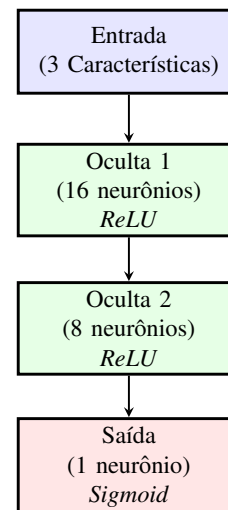


Fig. 1. Diagrama visual do sequenciamento arquitetural da rede neural *Multilayer Perceptron*.

Para capturar relações não lineares nos dados, empregou-se a função de ativação *Rectified Linear Unit* (ReLU) nas

camadas ocultas. A equação que define a função *ReLU* é dada por:

$$f(x) = \max(0, x) \quad (2)$$

Devido à sua simplicidade computacional e às suas propriedades de gradiente, essa função otimiza o processo de treinamento ao mitigar o problema de desvanecimento do gradiente (*vanishing gradient*). A saída da última camada oculta é processada por uma camada densa final que utiliza a função de ativação sigmoide (*Sigmoid*), matematicamente expressa por:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

Esta função mapeia os valores de saída para o intervalo contínuo $[0, 1]$, possibilitando a interpretação probabilística do resultado, onde valores próximos a 1 indicam classe *Fake* e valores próximos a 0 indicam classe *Real*.

III. ANÁLISE DE SIMULAÇÕES E RESULTADOS

A. Treinamento e Convergência

O processo de treinamento da rede foi configurado para 100 épocas. O modelo apresentou convergência ao longo das iterações, as métricas de desempenho convergiram. A arquitetura simplificada, operando sobre apenas três características de entrada, aliada ao emprego de regularização *L2*, permitiu que a rede capturasse os padrões discriminativos de forma robusta, sem apresentar instabilidades numéricas, conforme mostra a Figura 2.

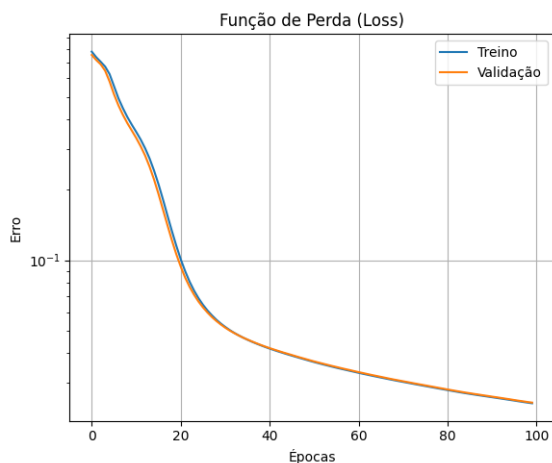


Fig. 2. Convergência contínua das curvas de Perda durante o processo de treinamento, evidenciando a ausência de sobreajuste.

Analisando a evolução do processo de treinamento ilustrada nas Figuras 2 e 3, observa-se comportamento de convergência altamente estável. Em paralelo, a Figura 3 evidencia que a acurácia do modelo partiu de valor inicial moderado e apresentou crescimento acentuado, atingindo acurácia de 1,0 antes da décima época. A manutenção dessa métrica em valor unitário até o término do treinamento confirma que os metadados

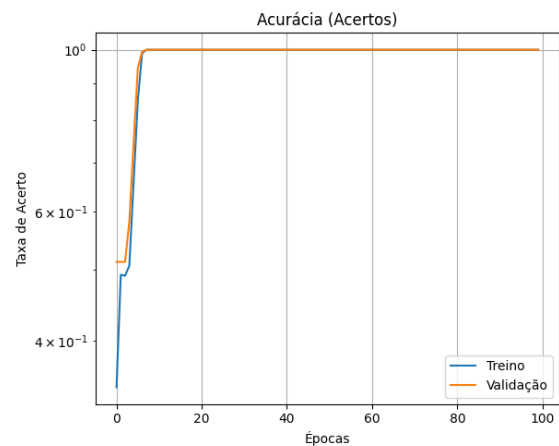


Fig. 3. Evolução em escala logarítmica da acurácia, comprovando a estabilização rápida e consistente dos acertos.

forenses selecionados constituem características discriminativas suficientemente representativas, permitindo convergência rápida e estável do modelo, dispensando ciclos extensos de otimização.

B. Desempenho no Conjunto de Testes

Para a avaliação final de desempenho, empregou-se o conjunto de teste previamente isolado (20% do *dataset* total). A matriz de confusão é apresentada na Figura 4.

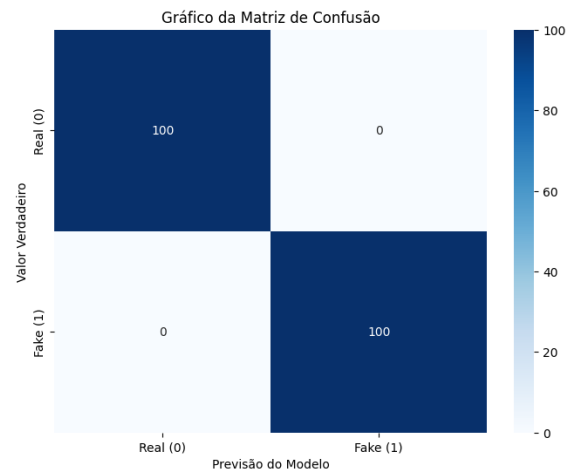


Fig. 4. Matriz de Confusão evidenciando desempenho na classificação do *dataset* de teste (ausência total de falsos positivos e falsos negativos).

O modelo atingiu acurácia de 100% no conjunto de teste. Das amostras apresentadas, tanto *Real* quanto *Fake*, a rede foi capaz de classificá-las corretamente sem registrar falsos positivos ou falsos negativos. Este resultado ratifica a eficácia da arquitetura simplificada proposta para o *dataset* em questão. Ao discutir esses resultados em face aos trabalhos relacionados, evidencia-se que a estratégia de operar sobre assinaturas

forenses pré-extraídas superou as abordagens tradicionais ao atingir desempenho máximo demandando apenas uma fração mínima de recursos computacionais. O alto nível de acerto comprova que inconsistências na sincronização audiovisual e anomalias de compressão são características discriminativas suficientes e plenas para a identificação de mídias sintetizadas.

C. Interpretabilidade Computacional (XAI)

Para elucidar o processo de decisão da Rede Neural e garantir transparência interpretativa, aplicou-se o método de Inteligência Artificial Explicável *SHapley Additive exPlanations (SHAP)* [7].

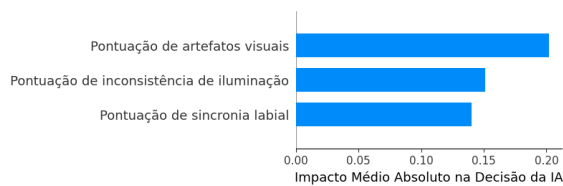


Fig. 5. Impacto médio absoluto de cada característica na decisão final do modelo (*SHAP Bar Plot*).

A Figura 5 revela que a pontuação de artefatos visuais constitui o atributo de maior importância para a classificação de mídias como *Fake*, seguida pela inconsistência de iluminação e, finalmente, pela dessincronização labial. O modelo extraiu hierarquia de importância consistente com o conhecimento forense esperado, corroborando a relevância das características selecionadas para detecção de mídias sintéticas.

IV. CONCLUSÕES E TRABALHOS FUTUROS

A arquitetura *MLP*, após otimizações baseadas em análise de explicabilidade (*SHAP*), demonstrou desempenho robusto com 100% de acurácia no *dataset* avaliado. A classificação operando exclusivamente sobre três metadados forenses demonstrou que é viável substituir a análise massiva em nível de *pixel*. Contudo, devem-se pontuar limitações metodológicas e ameaças à validade do estudo. A obtenção de 100% de acurácia e a ausência total de erros foram viabilizadas pelo uso de métricas numéricas pré-extraídas e tabuladas, o que abstrai o elevado custo computacional que seria inevitavelmente exigido para a extração forense dessas características a partir de mídias brutas em tempo real. Além disso, tal perfeição no conjunto de teste pode indicar uma dependência do modelo às assinaturas específicas da base de dados utilizada. Portanto, como perspectiva futura, propõe-se a aplicação do modelo em dados provenientes do tráfego real de usuários em plataformas de redes sociais e serviços de compartilhamento de vídeo.

DECLARAÇÃO DE USO DE INTELIGÊNCIA ARTIFICIAL

Os autores declaram que utilizaram ferramentas de IA, incluindo Gemini, exclusivamente para auxiliar na melhoria da linguagem, estruturação do código das figuras e legibilidade deste manuscrito. O uso dessas ferramentas foi realizado sob supervisão humana, e os autores revisaram e editaram

cuidadosamente o conteúdo, assumindo total responsabilidade pelo texto final publicado.

REFERENCES

- [1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.
- [2] T. T. Nguyen *et al.*, "Deep learning for deepfakes creation and detection: A survey," *Computer Vision and Image Understanding*, vol. 196, 2020.
- [3] P. Dhokane, "Deepfake and Synthetic Media Detection Dataset," Kaggle.
- [4] Z. Yan *et al.*, "DF40: Toward Next-Generation Deepfake Detection," *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.
- [5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
- [6] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [7] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [8] Keras, "tf.keras.Sequential - TensorFlow API Documentation." [Online]. Disponível: https://www.tensorflow.org/api_docs/python/tf/keras/Sequential.