Combate à Falsificação Digital: Um Modelo de Detecção de Imagens DeepFake com Aprendizado Profundo

Vinicius N. Lopes, Tadeu Januario, Vitor G. Balsanello
Diego Kreutz, Dionatan R. Schmidt, Eliezer Flores, Elder Rodrigues

[1]PPGES- Programa de Pós Graduação em Engenharia de Software
Universidade Federal do Pampa (UNIPAMPA) - Alegrete, RS - Brasil

{viniciusnunez,tadeujenuario,vitorbalsanello}.aluno@unipampa.edu.br
{kreutz,dionatanschmidt,eliezerflores,elderrodrigues}@unipampa.edu.br

Abstract. *DeepFake detection is a central challenge in digital security. This work presents a pipeline based on the fine-tuning of EfficientNet-B0, employing progressive data scaling, threshold optimization via the F1-score, and synthetic diversification using OmniGen. Evaluated on the DFDC Preview and OpenForensics datasets (179,000 images), the method achieved 98% accuracy, a threshold of 0.3495, and precision/recall of 0.98, demonstrating high effectiveness and stability.*

Resumo. *A detecção de DeepFakes é um desafio central da segurança digital. Este trabalho apresenta um pipeline baseado no fine-tuning da EfficientNet-B0, com escalonamento progressivo de dados, otimização do threshold via F1-score e diversificação sintética com OmniGen. Avaliado nas bases DFDC Preview e OpenForensics (179 mil imagens), o método atingiu 98% de acurácia, threshold de 0.3495 e precisão/recall de 0.98, evidenciando alta eficácia e estabilidade.*

1. Introdução

A evolução das técnicas de inteligência artificial, especialmente os *DeepFakes* [Altuncu et al. 2024], possibilita a criação de mídias sintéticas realistas que representam sérias ameaças à segurança digital [Rana et al. 2022]. Sua detecção automatizada é um campo ativo de pesquisa, com métodos baseados principalmente em redes neurais convolucionais (CNNs).

Conjuntos de dados de larga escala, como *DFDC Preview* [Dolhansky et al. 2019] e *OpenForensics* [Le et al. 2021], oferecem cenários amplos e realistas para a avaliação de modelos, servindo como *benchmarks* consolidados na detecção de *DeepFakes*. Entre as arquiteturas de destaque, a EfficientNet [Tan and Le 2019] apresenta um bom equilíbrio entre acurácia e custo computacional, tornando-se uma candidata promissora para essa tarefa.

Neste trabalho, emprega-se a *EfficientNet-B0* com aprendizado progressivo, utilizando subconjuntos *Mini*, *Middle*, *Big* e *Full* (totalizando mais de 179 mil imagens), aliados a um processo de *fine-tuning* em três fases e técnicas de diversificação de dados (*data augmentation* e *OmniGen*). Os experimentos demandaram mais de 50 horas de processamento em GPU (NVIDIA RTX 5070 Ti), executados em ambiente Docker configurado com CUDA Toolkit 12.2 e cuDNN 8.9.5. As bibliotecas utilizadas incluíram TensorFlow 2.13.0, Keras 2.13.1, Scikit-learn 1.3.0 e Matplotlib 3.7.2.

Adotamos *fine-tuning* progressivo (Mini→Full), *data augmentation* e calibração de limiar via F1-score, alcançando desempenho estável em DFDC Preview e OpenForensics. A *EfficientNet-B0* foi escolhida por equilibrar acurácia e custo computacional (modelo leve, adequado a 224×224) e pela eficiência em transferência em cenários de artefatos texturais e locais, nos quais CNNs compactas permanecem competitivas, ao contrário dos *Transformers*, que exigem mais dados e cálculo. A **Seção 2** descreve dados e treino, a **Seção 3** resultados, e a **Seção 4** limitações, próximos passos e implicações éticas.

2. Materiais e Métodos

2.1. Bases de Dados e Estratégia de Escalonamento

Foram utilizadas as bases públicas *DFDC Preview*[Dolhansky et al. 2019] e *OpenForensics*[Le et al. 2021] com escalonamento incremental com quatro subconjuntos: *Mini* (testes rápidos), *Middle* (diversidade facial), *Big* (idades/poses variadas) e *Full* (cenário de produção), todos com divisão de 80% (treino) e 20% (validação) e classes balanceadas.

Os conjuntos de dados foram balanceados entre exemplos reais e falsos. O conjunto *Mini* apresentou 534 imagens de cada classe (50,0% reais), *Middle* teve 2.978 reais e 2.539 falsas (54,0% reais), e Big contou com 5.413 reais e 5.492 falsas (49,6% reais). O conjunto *Full* de treinamento possuía 70.001 imagens por classe (50,0% reais), enquanto o *Full* de validação incluiu 19.787 reais e 19.641 falsas (50,7% reais).

2.2. Pré-processamento e Aumento de Dados

As imagens foram redimensionadas para 224×224 e normalizadas em [0,1]. Aplicou-se *data augmentation* (espelhamento, rotação e *zoom* ±10%) para simular variações de ângulo e pose. O *OmniGen* gerou amostras sintéticas com variações de expressão, iluminação e manipulações quase-adversariais, ampliando o espaço de características e tornando o modelo mais robusto a *DeepFakes* complexos.

2.3. Arquitetura do Modelo e Treinamento Progressivo

O modelo adotado foi a *EfficientNet-B0* [Tan and Le 2019], pré-treinado no *ImageNet* [Deng et al. 2009] ou seja, já treinada em milhões de imagens para reconhecer padrões visuais, aproveitando esse conhecimento sem treinar do zero e ajustada para classificação binária com uma cabeça personalizada composta por *Global Average Pooling* e três camadas densas (512, 256 e 1024 unidades) com ReLU e *dropout* (0.4, 0.4, 0.5), finalizando com *sigmoide*. Esta configuração permite extração de características complexas de padrões sutis de manipulação.

O treinamento seguiu três fases progressivas de *fine-tuning*. Na **Fase 1**, com a base convolucional congelada, apenas as camadas densas foram treinadas (5 épocas, $\eta = 1 \times 10^{-3}$, *batch size* = 32), resultando em desempenho aleatório (acurácia $A \approx 0{,}50$, AUC $\approx 0{,}50$) com viés acentuado para a classe *"fake"* ($R_{\text{real}} \approx 0$, $R_{\text{fake}} \approx 1$). Na **Fase 2**, o descongelamento parcial das últimas 20-40 camadas (10 épocas, $\eta = 1 \times 10^{-3}$) trouxe leve melhora ($A \approx 0{,}55$, R_{real} entre $0{,}01$ e $0{,}19$), mas com viés persistente e capacidade discriminativa limitada (AUC $< 0{,}60$). Na **Fase 3**, o *fine-tuning* completo (30-50 épocas, $\eta = 1 \times 10^{-5}$) com *early stopping* e redução adaptativa da taxa de aprendizado permitiu o ajuste simultâneo de θ_{conv} (parâmetros convolucionais da *EfficientNet-B0*) e θ_{denso} (parâmetros das camadas densas adicionadas), capturando padrões sutis de manipulação e atingindo robustez ($A = 0{,}98$, AUC $\approx 1{,}00$) com equilíbrio entre precisão e *recall*.

Além disso, otimizou-se o *threshold* τ que define a classificação (probabilidades acima de τ são falsas; abaixo, autênticas). O valor padrão $\tau = 0.5$ mostrou-se subótimo devido a variações sutis na distribuição das saídas. Para equilibrar precisão e *recall*, recalibrou-se τ maximizando o *F1-score* via curva *Precision–Recall*, ajustando o *threshold* às particularidades dos dados. Essa estratégia equilibra falsos positivos / negativos, com valores distintos de τ^* obtidos: Mini ($0{,}1591$), Middle ($0{,}0731$) e Big/Full ($0{,}3495$), refletindo a dependência do *threshold* com a distribuição dos dados e presença de ruído.

3. Resultados e Discussão

O impacto da liberação gradual das camadas da *EfficientNet-B0* foi avaliado por métricas como acurácia (A), precisão (P), *recall* (R), *F1-score* ($F1$) e (*AUC*), que medem acertos, exatidão, cobertura, equilíbrio entre precisão e *recall*, e a distinção entre classes.

3.1. Desempenho por Fase de Treinamento

Na **Fase 1**, com base congelada (θ_{conv} da *EfficientNet-B0* fixos), desempenho aleatório (acurácia $A \approx 0{,}50$, AUC $\approx 0{,}50$), com viés para *fake* ($R_{\text{real}} \approx 0$, $R_{\text{fake}} \approx 1$), mostrando limitação das *features* genéricas. Na **Fase 2**, descongelamento parcial ($L = 20$ a 40 camadas) trouxe leve melhora (acurácia $A \approx 0{,}55$, R_{real} entre $0{,}01$ e $0{,}19$), mas com capacidade discriminativa restrita (AUC $< 0{,}60$). Na **Fase 3**, com *fine-tuning* completo atingiu robustez (acurácia $A = 0{,}98$, AUC $\approx 1{,}00$ em $n = 179.430$ imagens) com equilíbrio entre precisão (P), *recall* (R) em ambas classes, sendo a otimização de τ^* via *F1-score* crucial para calibrar a sensibilidade, conforme Tabela 1.

Tabela 1. Resultados por fase de treinamento.

Fase	τ	A	P/R real	P/R fake	n
Base Congelada	–	0.50	– / 0.00	– / 1.00	1.068
Parcial	–	0.55	0.80 / 0.01	0.55 / 1.00	1.103
Fase 1	0.1591	0.80	0.73 / 0.93	0.90 / 0.67	2.181
Fase 2	0.0731	0.77	0.71 / 0.91	0.88 / 0.64	7.885
Fase 3	0.3495	0.98	0.98 / 0.98	0.98 / 0.99	179.430

A Tabela 1 mostra que *features* genéricas são insuficientes (Fase 1); o descongelamento parcial traz ganhos limitados e enviesados (Fase 2); já o ajuste completo é essencial para capturar padrões sutis e equilibrar as métricas (Fase 3).

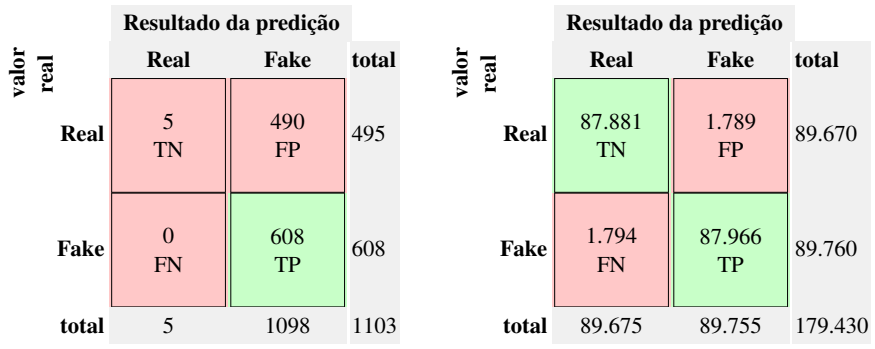
3.2. Matriz de Confusão

A Figura 1 apresenta os acertos (*TP*, *TN*) e erros (*FP*, *FN*) do classificador, sendo os *FN* os mais críticos por aceitarem falsificações como legítimas. O equilíbrio entre essas taxas é obtido pela calibração de τ^* via *F1-score*. A matriz (a) apresenta os resultados com $\tau = 0.5$ e a (b) com *threshold* ajustado ($\tau = 0.3495$).

4. Conclusão

Foi apresentada uma abordagem para detecção de *DeepFakes* com *EfficientNet-B0* e *fine-tuning* progressivo, alcançando 98% de acurácia *intra-dataset*[1]. O treinamento em fa-

[1]Código-fonte: `https://github.com/AI-Horizon-Labs/DeepFake-Inspector`



(a) Base parcialmente descongelada (Acurácia: 55%)

(b) Fine-tuning completo (Acurácia: 98%)

Figura 1. Matrizes de confusão: pior caso vs. melhor caso

ses, aliado ao pré-processamento, ao aumento de dados com *OmniGen* e ao ajuste do *threshold*, mostrou-se essencial para a robustez e a generalização do modelo.

Apesar do desempenho, as análises permaneceram restritas ao cenário *intra-dataset*. A validação *cross-dataset*, estudos ablativos, replicações com múltiplas sementes e testes em vídeos são etapas necessárias para consolidar os resultados e avaliar a sensibilidade do modelo a artefatos (compressão, iluminação, desfoque).

Considerações éticas: Sistemas de detecção devem visar aplicações benéficas como moderação de conteúdo, verificação de identidade e perícia digital, evitando usos indevidos para vigilância massiva ou violações de privacidade. Falsos negativos/positivos podem ter consequências graves, exigindo auditorias rigorosas, transparência sobre limitações, e *frameworks* que assegurem responsabilidade ética e legal. Modelos devem ser avaliados quanto a possíveis vieses discriminatórios em dados desbalanceados.

Agradecimentos. A pesquisa contou com apoio da CAPES (Código de Financiamento 001) e da FAPERGS (termos de outorga 24/2551-0001368-7 e 24/2551-0000726-1).

Referências

Altuncu, E., Franqueira, V. N. L., and Li, S. (2024). Deepfake: definitions, performance metrics and standards, datasets, and a meta-review. *Frontiers in Big Data*, 7:1400024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *IEEE CCVPR*, pages 248–255. IEEE.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., and Ferrer, C. C. (2019). The deepfake detection challenge (DFDC) preview dataset. `https://arxiv.org/abs/1910.08854`. Acesso em: 19 ago. 2025.

Le, T.-N., Nguyen, H. H., Yamagishi, J., and Echizen, I. (2021). Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild. In *IEEE/CVF ICCV*, pages 10117–10127.

Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, 10:25494–25513.

Tan, M. and Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *36th ICML*, volume 97, pages 6105–6114.