Comparison of Deep Learning Models in the Detection of Fraud in Scientific Images

Luan Assis[1], **Isaque Severino**[1], **Gabriel Silva**[1], **Adam Santos**[1]

[1]Faculdade de Sistemas de Informação, Instituto de Geociências e Engenharias,
Universidade Federal do Sul e Sudeste do Pará (Unifesspa)
FL 17, QD 04, LT Especial, Nova Marabá, Marabá, PA, Brasil

luan.assis@unifesspa.edu.br, isaque.severino@unifesspa.edu.br,

gabriel.s_r@unifesspa.edu.br, adamdreyton@unifesspa.edu.br

Abstract. *This paper compares supervised convolutional networks and self-supervised vision foundation models for scientific image forgery detection on the Recod.ai/LUC benchmark. Seven architectures were evaluated under three training-data regimes, yielding 21 configurations at 518×518 resolution. On a fixed test set of 770 images, DINOv2-Large with offline augmentation achieved the best single-run result: 87.79% accuracy, 0.8848 F1-score, 0.9648 ROC-AUC, and 0.9707 PR-AUC. The effect of synthetic forgeries was architecture-dependent: DINOv3-Large improved in both F1 and PR-AUC, whereas DINOv2-Large degraded under the synthetic regime. These results suggest that DINO representations are promising for this benchmark, but they do not establish universal superiority: most configurations were evaluated with a single seed, class-dependent augmentation may introduce shortcuts, and no external validation was performed. Multiseed experiments, balanced ablations, and external validation are left for future work, as the available hardware constrained the number of training runs.*

Resumo. *Este artigo compara redes convolucionais supervisionadas e modelos fundacionais visuais auto-supervisionados na detecção de fraudes em imagens científicas do benchmark Recod.ai/LUC. Sete arquiteturas foram avaliadas em três regimes de dados de treinamento, totalizando 21 configurações na resolução 518×518. Em um conjunto de teste fixo com 770 imagens, o DINOv2-Large com aumento de dados offline obteve o melhor resultado em execução única: 87,79% de acurácia, F1 de 0,8848, ROC-AUC de 0,9648 e PR-AUC de 0,9707. O efeito das falsificações sintéticas variou por arquitetura: o DINOv3-Large melhorou em F1 e PR-AUC, enquanto o DINOv2-Large apresentou degradação no regime sintético. Os resultados sugerem que as representações DINO são promissoras neste benchmark, mas não demonstram superioridade universal, pois a maioria das configurações foi avaliada com uma única semente, o aumento dependente da classe pode introduzir atalhos e não houve validação externa. Experimentos multisseed, ablações balanceadas e validação externa são propostos como trabalhos futuros, pois o hardware disponível limitou novas execuções de treinamento.*

1. Introdução

A integridade de imagens científicas é essencial para a confiabilidade da produção acadêmica, especialmente em áreas biomédicas nas quais imagens de microscopia, Western

blot, lâminas histológicas e culturas celulares funcionam como evidências primárias. Duplicações indevidas, recortes, inserções, montagens ou alterações locais podem modificar a interpretação dos resultados e comprometer conclusões experimentais. Por isso, a detecção automática de manipulações visuais tornou-se uma linha relevante para a ciência forense digital e para a avaliação de integridade científica [Bik et al. 2016, Bucci 2018].

Fraudes por *copy-move* e *splicing* são particularmente desafiadoras em imagens científicas. O primeiro caso reutiliza regiões da própria imagem, preservando ruído, iluminação e textura compatíveis com o restante do conteúdo; o segundo combina regiões de fontes distintas, podendo deixar inconsistências de borda, escala ou estatística de ruído [Fridrich et al. 2003]. Em ambos os casos, a dificuldade aumenta porque imagens biomédicas frequentemente contêm padrões repetitivos, baixo contraste e estruturas naturalmente semelhantes, o que reduz a eficácia da inspeção visual isolada.

Métodos tradicionais de forense digital exploram artefatos estatísticos, compressão, ruído e interpolação, mas tendem a sofrer com a grande variação de aquisição, resolução e modalidade. Redes neurais profundas passaram a ser utilizadas por aprenderem atributos diretamente dos dados, enquanto modelos auto-supervisionados baseados em *Vision Transformers*, como DINOv2 e DINOv3, tornaram-se candidatos promissores por transferirem representações visuais aprendidas em larga escala para domínios especializados [He et al. 2016, Dosovitskiy et al. 2021, Oquab et al. 2023, Siméoni et al. 2025].

Este estudo investiga duas questões: (1) como *baselines* convolucionais supervisionados e modelos DINO performam sob o mesmo benchmark, divisão experimental e protocolo de classificação; e (2) como o aumento de dados *offline* e a inclusão de falsificações sintéticas alteram o desempenho dentro de cada arquitetura. ResNet-50, ResNet-101, DINOv2-Base, DINOv2-Large, DINOv3-Small, DINOv3-Base e DINOv3-Large foram avaliados em três regimes de dados. A contribuição é uma comparação descritiva de 21 configurações em 518×518, através das métricas F1, ROC-AUC e PR-AUC, complementada por análise exploratória de limiar, custo computacional parcial e mapas qualitativos de evidência. Sensibilidade multisseed, comparação controlada de resolução e validação externa são apresentadas como trabalhos futuros. As conclusões são restritas às arquiteturas, ao benchmark, às execuções e aos protocolos avaliados.

2. Fundamentação e trabalhos correlatos

2.1. Fraudes em imagens científicas

A manipulação de imagens científicas inclui falsificações deliberadas, erros de montagem, duplicação acidental, reutilização indevida de painéis e ajustes visuais excessivos. Estudos sobre integridade de imagens biomédicas indicam que duplicações inapropriadas aparecem em parte relevante da literatura, motivando ferramentas automáticas de triagem para apoiar avaliadores humanos [Bik et al. 2016, Bucci 2018]. Computacionalmente, o problema pode ser tratado como classificação binária, detecção de anomalia, localização de regiões suspeitas ou segmentação da máscara de fraude.

No *copy-move*, uma região é copiada e colada em outra parte da mesma imagem, preservando iluminação, textura e ruído compatíveis com o restante do conteúdo [Fridrich et al. 2003]. No *splicing*, regiões de imagens diferentes são combinadas, podendo gerar inconsistências de borda, escala ou estatística de ruído. Em imagens científicas,

essas manipulações são difíceis porque a própria imagem pode conter estruturas repetitivas e baixo contraste. Assim, o modelo precisa ser sensível a adulterações pequenas, mas robusto para não confundir variações legítimas com fraude.

2.2. Modelos convolucionais

Redes neurais convolucionais representam imagens por filtros locais aplicados em janelas espaciais. Esses filtros compartilham pesos ao longo da imagem e aprendem padrões hierárquicos: camadas iniciais capturam bordas, texturas e contrastes, enquanto camadas profundas combinam esses sinais em estruturas mais abstratas. A ResNet introduziu conexões residuais que facilitam o treinamento de redes profundas, tornando ResNet-50 e ResNet-101 referências fortes para classificação visual [He et al. 2016].

Em forense digital, CNNs podem aprender artefatos associados a manipulações, como transições artificiais, inconsistências de textura, ruído e bordas [Bayar and Stamm 2016, Rao and Ni 2016]. Essa capacidade é positiva porque reduz a dependência de descritores manuais; porém, em bases pequenas ou heterogêneas, CNNs supervisionadas podem aprender atalhos específicos do treinamento. Essa limitação justifica compará-las com modelos pré-treinados em larga escala.

2.3. Transformers de visão e aprendizado auto-supervisionado

Vision Transformers dividem a imagem em *patches* e tratam esses blocos como uma sequência de tokens. O mecanismo de auto-atenção permite relacionar regiões distantes da imagem, favorecendo dependências de longo alcance e contexto global [Dosovitskiy et al. 2021]. Essa propriedade é útil porque a evidência de fraude pode estar em pequenas inconsistências locais, mas sua interpretação depende do padrão visual ao redor.

A família DINO utiliza aprendizado auto-supervisionado em uma estrutura *student-teacher*, na qual uma rede estudante aprende representações consistentes com uma rede professora sem depender de rótulos manuais [Caron et al. 2021]. DINOv2 e DINOv3 buscam atributos visuais transferíveis e mapas densos de características, potencialmente úteis em domínios especializados com poucas anotações [Oquab et al. 2023, Siméoni et al. 2025]. Ao mesmo tempo, esses modelos têm maior custo computacional e exigem ajuste fino conservador.

2.4. Trabalhos correlatos e lacuna investigada

Os trabalhos correlatos podem ser organizados em três grupos. Estudos de integridade científica documentam duplicações e manipulações em publicações biomédicas, reforçando a relevância prática da triagem automática [Bik et al. 2016, Bucci 2018]; porém, nem sempre comparam arquiteturas modernas em um mesmo protocolo. Métodos clássicos e profundos de forense exploram *copy-move*, *splicing* e filtros convolucionais para detectar artefatos locais [Fridrich et al. 2003, Bayar and Stamm 2016, Rao and Ni 2016], mas frequentemente em bases gerais ou cenários específicos de manipulação. Já modelos fundacionais como ViT, DINOv2 e DINOv3 mostram forte transferência visual [Dosovitskiy et al. 2021, Oquab et al. 2023, Siméoni et al. 2025], mas ainda precisam ser avaliados em fraude científica, na qual a classe positiva não representa um objeto, e sim indícios sutis de manipulação. Assim, a lacuna investigada é uma comparação de CNNs

supervisionadas e modelos fundacionais auto-supervisionados no mesmo benchmark e protocolo de classificação, incluindo regimes de dados, custo computacional parcial e mapas qualitativos de evidência. A comparação controlada de resolução permanece como trabalho futuro.

2.5. Aumento de dados e falsificações sintéticas

O aumento de dados melhora a robustez diante de variações de posição, orientação, iluminação e aparência [Shorten and Khoshgoftaar 2019]. Em detecção de fraude, por outro lado, transformações fortes ou assimétricas podem criar atalhos artificiais para o classificador. Falsificações sintéticas também podem ampliar a classe positiva, mas seu benefício depende do realismo e da proximidade estatística em relação às fraudes reais. Assim, comparar cenários com e sem falsificações sintéticas permite verificar se esses exemplos aumentam a generalização ou introduzem desvio de distribuição.

3. Materiais e métodos

3.1. Dataset e organização experimental

O estudo utiliza o benchmark *Recod.ai/LUC – Scientific Image Forgery Detection*, voltado à detecção de falsificações em imagens científicas e biomédicas [Recod.ai and LUC 2025], aqui tratado como classificação binária em nível de imagem: cada amostra é autêntica ou forjada. O conjunto real contém 5.128 imagens (2.377 autênticas e 2.751 forjadas), divididas de forma estratificada em 70% / 15% / 15%, o que produz 3.588 imagens de treino, 770 de validação e 770 de teste, este último com 357 autênticas e 413 forjadas. A divisão precede a geração de novos exemplos e agrupa as amostras pelo nome-base da imagem-fonte, mantendo na mesma partição as versões autêntica e forjada de uma mesma origem. Imagens geradas por aumento de dados ou por síntese entram apenas no treino: no cenário *Sem aumento* ele contém as 3.588 imagens reais e, nos cenários com aumento, 868 imagens geradas o elevam a 4.456 amostras, enquanto validação e teste permanecem com 770 imagens reais cada. Assim, validação e teste medem a generalização para amostras reais não transformadas artificialmente.

3.2. Cenários e falsificações sintéticas

Foram definidos três cenários: *Sem aumento*, apenas com imagens originais; *Aumento offline*, que acrescenta aumento de dados *offline* sem falsificações sintéticas; e *Aumento com sintéticos*, que soma exemplos forjados sinteticamente. A separação permite medir o efeito do aumento convencional e o efeito adicional dos sintéticos.

O aumento *offline* aplica transformações conservadoras às imagens autênticas, com inversão horizontal, rotação de até 5°, deslocamentos de até 2%, brilho entre 0,95 e 1,05, *zoom* de até 3% e deslocamento de canais de até 5 níveis. Às imagens forjadas aplicam-se transformações mais agressivas, com inversões horizontal e vertical, rotação de até 20°, deslocamentos de até 10%, brilho entre 0,8 e 1,2, *zoom* de até 15%, cisalhamento de 0,1 e deslocamento de canais de até 15 níveis. O preenchimento de bordas é reflexivo e as transformações geométricas das imagens forjadas são replicadas nas máscaras disponíveis. No cenário *Aumento com sintéticos*, acrescentam-se falsificações geradas a partir de imagens autênticas por *copy-move* (30%), *splicing* (25%), *inpainting* (20%), injeção de ruído (15%) e manipulação local de brilho (10%).

Como as transformações *offline* não têm a mesma distribuição entre as classes, o protocolo pode induzir atalhos correlacionados ao rótulo. Na ausência de uma ablação balanceada, o efeito do aumento deve ser lido como específico desse protocolo assimétrico, sem atribuição causal isolada aos dados sintéticos; a comparação com transformações equivalentes nas duas classes exige novas execuções e fica como trabalho futuro.

3.3. Modelos avaliados

Foram avaliadas duas famílias de arquiteturas. A primeira é composta pelas CNNs supervisionadas ResNet-50 e ResNet-101, usadas como *baselines* e inicializadas com pesos ImageNet, com o *backbone* seguido de uma cabeça MLP binária. A segunda reúne cinco modelos auto-supervisionados baseados em *Vision Transformers*, a saber, DINOv2-Base, DINOv2-Large, DINOv3-Small, DINOv3-Base e DINOv3-Large, nos quais o *backbone* pré-treinado é adaptado por uma cabeça de classificação. Os pesos correspondem aos identificadores `facebook/dinov2-base` e `facebook/dinov2-large` no Hugging Face e, para o DINOv3, `facebook/dinov3-vits16-pretrain-lvd1689m`, `facebook/dinov3-vitb16-pretrain-lvd1689m` e `facebook/dinov3-vitl16-pretrain-lvd1689m`, estes últimos em repositórios de acesso controlado, que exigem autenticação para reprodução exata.

Os códigos de preparação dos dados, divisão das amostras, treinamento, avaliação e geração dos mapas de evidência estão disponíveis em um repositório[1], que documenta dependências, comandos de reprodução e a localização externa do dataset e dos *checkpoints*.

3.4. Pré-processamento, treinamento e métricas

Define-se como protocolo principal a configuração de referência do estudo, composta pela resolução 518×518, pelo limiar de decisão 0,4 e pelas 21 combinações de modelo e cenário treinadas com semente única, sobre a qual se baseia a comparação central da Tabela 2. As variações em 384×384 e no limiar 0,5 são tratadas como análises de sensibilidade sobre esse protocolo, e não como protocolos alternativos independentes.

As imagens são redimensionadas para 518×518 pixels no protocolo principal, resolução compatível com o *patch size* 14 do DINOv2 (grade de 37×37 *patches*) e que preserva detalhes relevantes para manipulações sutis; como as CNNs são frequentemente avaliadas em resoluções menores, mantiveram-se também pesos avaliados em 384×384 para análise de sensibilidade. No treino, uma etapa *online* comum às duas classes aplica inversão horizontal com probabilidade 0,5, inversão vertical com probabilidade 0,2, rotação de até 10° e *ColorJitter* com brilho e contraste de 0,1, saturação de 0,05 e matiz de 0,03; validação e teste usam apenas redimensionamento e normalização pelos parâmetros do ImageNet.

O treinamento ocorre em duas fases: a primeira treina somente a cabeça de classificação, com *backbone* congelado, e a segunda realiza o *fine-tuning* completo com taxas distintas para *backbone* e cabeça, *warmup*, agendamento por cosseno, *early stopping* por F1 de validação, FP16 e entropia cruzada ponderada para favorecer a classe forjada. O *pipeline* fixa o limiar 0,4 para priorizar a redução de falsos negativos e usa 0,5 apenas como análise de sensibilidade; trata-se de um ponto operacional definido a priori no código, e

[1] `https://github.com/LuanAssis01/Comparison-of-Deep-Learning-Models-in-the-Detection-of-Fraud-in-Scientific-Images`

não de um limiar universalmente ótimo ou otimizado no teste, escolha compatível com cenários de custo assimétrico [Sheng and Ling 2006] e com a literatura que mostra que 0,5 não é necessariamente ótimo para F1 [Lipton et al. 2014]. Os hiperparâmetros, obtidos por busca em grade ou aleatória e mantidos fixos nos treinos completos, constam na Tabela 1; em todos os modelos a taxa do *backbone* é menor que a da cabeça, com ajustes mais conservadores nos modelos DINO de maior porte.

Tabela 1. Hiperparâmetros do treinamento. E_1/E_2: épocas por fase; η_{c1}: taxa da cabeça na fase 1; η_{b2} e η_{c2}: taxas do *backbone* e da cabeça na fase 2; LS: *label smoothing*; w_1: peso da classe forjada.

Modelo	E_1/E_2	η_{c1}	η_{b2}	η_{c2}	**Warmup**	**WD**	**LS**	w_1
ResNet-50	8/40	$6{,}67\times10^{-4}$	$6{,}80\times10^{-5}$	$5{,}55\times10^{-4}$	3	$3{,}63\times10^{-5}$	0,134	1,447
ResNet-101	8/40	$6{,}00\times10^{-4}$	$5{,}00\times10^{-5}$	$3{,}00\times10^{-4}$	3	$1{,}00\times10^{-3}$	0,050	1,500
DINOv2-Base	12/35	$3{,}20\times10^{-4}$	$2{,}89\times10^{-6}$	$2{,}56\times10^{-4}$	4	$6{,}62\times10^{-3}$	0,042	1,451
DINOv2-Large	12/35	$3{,}00\times10^{-4}$	$1{,}00\times10^{-6}$	$2{,}00\times10^{-4}$	5	$2{,}00\times10^{-3}$	0,050	1,500
DINOv3-Small	10/35	$1{,}11\times10^{-4}$	$4{,}00\times10^{-6}$	$9{,}07\times10^{-5}$	3	$3{,}64\times10^{-3}$	0,047	1,424
DINOv3-Base	12/35	$3{,}07\times10^{-4}$	$4{,}00\times10^{-6}$	$2{,}70\times10^{-4}$	4	$3{,}00\times10^{-3}$	0,023	0,987
DINOv3-Large	12/35	$5{,}72\times10^{-5}$	$3{,}25\times10^{-6}$	$2{,}00\times10^{-4}$	5	$4{,}17\times10^{-3}$	0,003	1,964

Todos os modelos usam AdamW, agendamento por cosseno, FP16, acumulação de gradiente em 2 passos e *early stopping* com paciência de 12 épocas. O lote físico é 4 nas ResNet e nos modelos Base/Small (efetivo 8) e 2 nos modelos Large (efetivo 4). A execução principal usa a semente 42 em uma NVIDIA RTX 3060 de 12 GB; o script `src/summarize_seed_study.py` define as sementes 42, 123, 777, 2025 e 3407 para um estudo de robustez que não foi concluído, de modo que nenhuma média multisseed é usada aqui. A limitação de hardware impediu a repetição sistemática das 21 configurações nesta versão.

A avaliação usa acurácia, precisão, *recall*, especificidade, F1-score, ROC-AUC e PR-AUC nas 21 configurações, sendo a PR-AUC especialmente informativa quando o desempenho sobre a classe positiva é central [Davis and Goadrich 2006, Saito and Rehmsmeier 2015]. Como falsos negativos correspondem a imagens forjadas classificadas como autênticas, a escolha do melhor modelo considera sensibilidade, especificidade, F1 e separabilidade global [Chicco and Jurman 2021]. Como etapa complementar, foram gerados mapas qualitativos de evidência: Grad-CAM da última camada convolucional nas ResNet [Selvaraju et al. 2017] e mapas de atenção do token CLS para os *patches* nos modelos DINO, com oclusão como alternativa independente da arquitetura. Havendo máscara real, calcula-se IoU@0,50 entre o mapa binarizado e a máscara redimensionada. A etapa não converte o classificador em segmentador; indica apenas se a decisão se associa a regiões visualmente plausíveis.

4. Resultados e discussão

4.1. Comparação geral entre arquiteturas

A Tabela 2 apresenta os resultados atualizados do protocolo principal em 518×518, usando o mesmo conjunto de teste de 770 imagens. Os valores de acurácia, precisão, *recall* e especificidade estão em porcentagem. A especificidade foi calculada a partir das mesmas predições de teste, considerando a distribuição de 357 imagens autênticas e 413 forjadas.

Tabela 2. Desempenho dos modelos avaliados no protocolo principal 518×518.

Modelo	Cenário	Acc.	Prec.	Recall	Esp.	F1	ROC-AUC	PR-AUC
ResNet-50	Sem aumento	78,44	74,17	91,77	63,02	0,8203	0,8769	0,8834
ResNet-101	Sem aumento	80,52	78,16	88,38	71,43	0,8295	0,8930	0,9158
DINOv2-Base	Sem aumento	84,94	83,98	88,86	80,41	0,8635	0,9492	0,9616
DINOv2-Large	Sem aumento	85,58	85,28	88,38	82,34	0,8680	0,9534	0,9652
DINOv3-Small	Sem aumento	79,48	77,07	87,89	69,75	0,8213	0,8907	0,9126
DINOv3-Base	Sem aumento	82,21	81,36	86,68	77,04	0,8394	0,9230	0,9420
DINOv3-Large	Sem aumento	86,75	86,59	89,10	84,03	0,8783	0,9497	0,9606
ResNet-50	Aumento offline	80,26	77,94	88,14	71,14	0,8273	0,8888	0,9107
ResNet-101	Aumento offline	81,30	81,35	84,50	77,60	0,8290	0,8924	0,9180
DINOv2-Base	Aumento offline	84,55	84,51	87,17	81,52	0,8582	0,9433	0,9579
DINOv2-Large	**Aumento offline**	**87,79**	**89,58**	**87,41**	**88,23**	**0,8848**	**0,9648**	**0,9707**
DINOv3-Small	Aumento offline	77,27	77,67	80,87	73,11	0,7924	0,8891	0,9108
DINOv3-Base	Aumento offline	80,52	84,33	78,21	83,19	0,8116	0,9232	0,9400
DINOv3-Large	Aumento offline	85,45	84,76	88,86	81,51	0,8676	0,9489	0,9559
ResNet-50	Aum. + sintéticos	75,84	77,89	76,76	74,78	0,7732	0,8727	0,8928
ResNet-101	Aum. + sintéticos	80,26	78,56	86,92	72,56	0,8253	0,8990	0,9159
DINOv2-Base	Aum. + sintéticos	84,81	82,46	91,04	77,60	0,8654	0,9463	0,9600
DINOv2-Large	Aum. + sintéticos	85,97	86,75	87,17	84,58	0,8696	0,9580	0,9686
DINOv3-Small	Aum. + sintéticos	79,22	76,97	87,41	69,75	0,8186	0,8889	0,9074
DINOv3-Base	Aum. + sintéticos	80,91	83,08	80,87	80,96	0,8196	0,9199	0,9410
DINOv3-Large	Aum. + sintéticos	87,14	87,56	88,62	85,43	0,8809	0,9558	0,9626

O melhor resultado foi obtido pelo DINOv2-Large com aumento *offline*. Em relação ao mesmo modelo sem aumento, a precisão passou de 85,28% para 89,58% e a especificidade de 82,34% para 88,23%, enquanto o *recall* variou de 88,38% para 87,41%. O aumento do F1, de 0,8680 para 0,8848, está associado principalmente à redução de falsos positivos, com pequena redução da sensibilidade.

Os modelos DINO de maior porte também apresentaram os maiores valores de ROC-AUC e PR-AUC. Entretanto, DINOv3-Small e DINOv3-Base não superaram as ResNet em todas as configurações. Os resultados permitem afirmar que DINOv2-Large e DINOv3-Large foram os modelos mais competitivos neste protocolo, mas não que toda a família DINO seja superior. Além disso, o experimento não separa os efeitos da arquitetura, do número de parâmetros e do pré-treinamento; por isso, não é possível atribuir o desempenho a um único desses fatores.

4.2. Influência do limiar de classificação

O limiar 0,4 foi adotado no pipeline como decisão operacional de triagem. Nesse contexto, um falso negativo deixa uma imagem potencialmente forjada sem revisão, enquanto um falso positivo encaminha uma imagem autêntica para inspeção humana adicional. A Figura 1 mostra esse compromisso: em algumas configurações, a redução do número de falsos negativos é acompanhada por diminuição da precisão ou do F1. O valor 0,4 não é apresentado como limiar universal; em uma aplicação externa, ele deve ser escolhido somente com dados de validação e de acordo com o custo atribuído a cada tipo de erro.

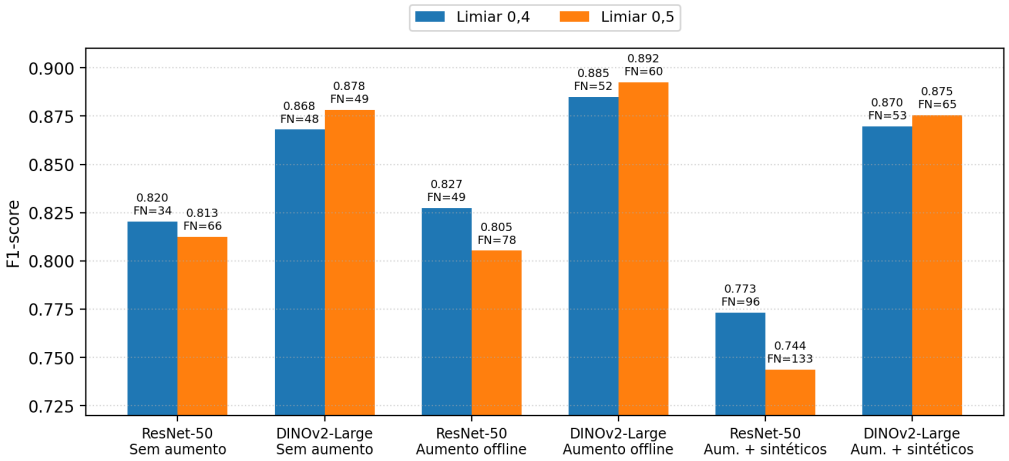


Figura 1. Comparação entre limiares 0,4 e 0,5 em 518 × 518. As anotações indicam o F1-score e o número de falsos negativos (FN).

4.3. Efeito dos cenários de dados

Na execução única, a inclusão de falsificações sintéticas apresentou resultados diferentes entre os modelos. Em comparação com o aumento *offline*, o F1 do DINOv3-Large aumentou 0,0133 e a PR-AUC aumentou 0,0067. No DINOv2-Large, o F1 diminuiu 0,0152 e a PR-AUC diminuiu 0,0021. No DINOv3-Small, o F1 aumentou 0,0262, mas a PR-AUC diminuiu 0,0034.

F1 e PR-AUC descrevem aspectos distintos do desempenho. O F1 resume as decisões produzidas por um limiar específico, enquanto a PR-AUC avalia a ordenação das amostras ao longo de vários limiares. Assim, o aumento de F1 acompanhado de redução da PR-AUC no DINOv3-Small indica melhora no ponto de corte adotado, mas não na ordenação global das amostras.

Essas diferenças não permitem isolar o efeito das imagens sintéticas. As transformações *offline* foram mais intensas para a classe forjada do que para a classe autêntica, incluindo maiores amplitudes de rotação, deslocamento, brilho e *zoom*. Esse procedimento pode produzir artefatos de pré-processamento associados ao rótulo, permitindo que o modelo reconheça diferenças entre as transformações em vez de características próprias da fraude. Como cada configuração principal foi executada uma única vez, os resultados desta seção são descritivos. Uma avaliação causal exigiria transformações equivalentes nas duas classes e repetições com diferentes sementes.

4.4. Sensibilidade à resolução de entrada

Os checkpoints disponíveis em 384 × 384 e 518 × 518 foram obtidos em configurações que podem diferir quanto ao regime de dados e a outras condições de treinamento. Portanto, os valores registrados não permitem atribuir diferenças de desempenho apenas à resolução. Para realizar essa comparação, será necessário manter arquitetura, regime de dados, divisão, sementes, hiperparâmetros e limiar fixos.

4.5. Custo de treinamento, inferência e viabilidade prática

O custo computacional foi avaliado parcialmente. Em 518 × 518, o DINOv2-Large exigiu cerca de 10h40min de treinamento e 43s para avaliar o teste; o DINOv3-Large exigiu

9h12min e 39s. A ResNet-50 possui cerca de 24,7 milhões de parâmetros e latência próxima de 7,2–7,7 ms, enquanto o DINOv2-Large possui 305,5 milhões de parâmetros e latência de aproximadamente 62,2 ms. Esses valores mostram o compromisso entre desempenho e custo, mas não permitem comparar todas as arquiteturas enquanto memória de pico, tempo por época e latência padronizada não forem reportados para cada modelo.

4.6. Mapas de evidência e interpretabilidade

As Figuras 2 e 3 apresentam exemplos reais para os quais há máscara da região manipulada. Nas CNNs, o Grad-CAM indica as regiões que mais contribuíram para a classe forjada; nos modelos DINO, a atenção do token CLS indica os *patches* mais relacionados à decisão. Esses mapas não medem o desempenho da classificação e não correspondem a uma segmentação da fraude [Selvaraju et al. 2017]. Eles mostram somente onde o classificador concentrou evidência ao produzir sua resposta.

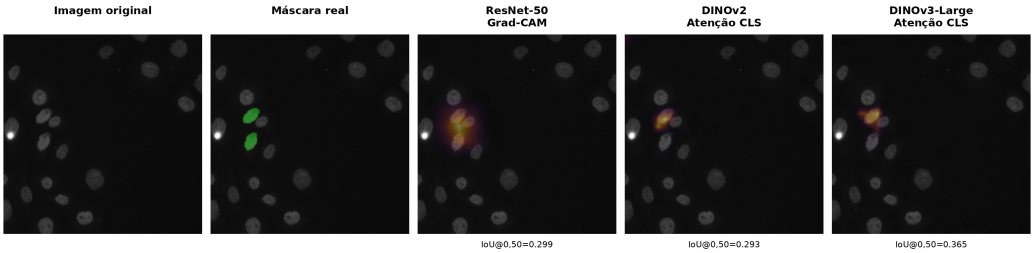


Figura 2. Exemplo favorável de mapas de evidência em imagem forjada com máscara disponível, comparando imagem original, máscara real, Grad-CAM e atenção CLS. Os mapas indicam regiões relevantes para a decisão, mas não uma segmentação precisa.

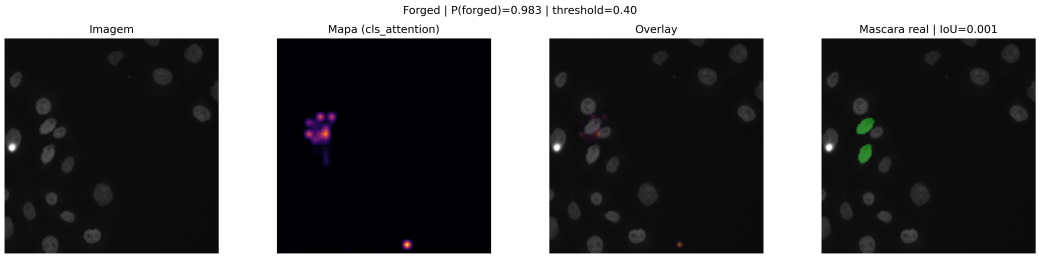


Figura 3. Caso-limite do DINOv2-Large em 518×518: a amostra é classificada corretamente como forjada, mas a atenção apresenta baixa sobreposição com a máscara real (IoU@0,50 = 0,001).

A Figura 2 apresenta um caso em que as regiões destacadas coincidem parcialmente com a área manipulada, sugerindo que a classificação utilizou características relacionadas à fraude. Na Figura 3, a imagem foi classificada corretamente como forjada, mas a sobreposição com a máscara foi de apenas 0,001. Nesse caso, o modelo pode ter utilizado outras características da imagem para decidir. A comparação mostra que uma classificação correta não garante a localização da região adulterada.

O uso pretendido dos classificadores é apoiar a triagem, e não determinar automaticamente a ocorrência de má conduta. Uma predição positiva pode priorizar uma imagem para inspeção humana, mas não identifica autoria, intenção ou gravidade da manipulação.

Do mesmo modo, os mapas de evidência servem como explicação qualitativa da decisão e não constituem prova de adulteração.

5. Limitações e trabalhos futuros

5.1. Limitações experimentais

Cada uma das 21 configurações principais foi avaliada uma vez, com a semente 42. Dessa forma, não foi possível estimar a variação dos resultados entre execuções, nem calcular média, desvio-padrão ou intervalo de confiança. O repositório contém uma rotina de agregação para cinco sementes, mas o conjunto correspondente de treinamentos ainda não foi concluído e validado sob o mesmo protocolo. Por esse motivo, as diferenças apresentadas devem ser interpretadas como resultados das execuções realizadas, e não como estimativas da estabilidade de cada arquitetura.

As transformações *offline* também diferem entre as classes. Como discutido na Seção 4.3, essa assimetria pode produzir artefatos de pré-processamento associados ao rótulo. Assim, os experimentos atuais não separam o benefício de adicionar dados do possível aprendizado desses artefatos. Além disso, a análise de resolução não foi controlada, pois os checkpoints de 384×384 e 518×518 podem diferir em outras condições de treinamento.

Os modelos foram treinados para classificar a imagem inteira, e não para segmentar a região adulterada. Grad-CAM, atenção CLS e oclusão são usados apenas para examinar quais regiões influenciaram a decisão. Como não há quantidade suficiente de máscaras e métricas agregadas para todas as configurações, a localização permanece qualitativa. A generalização dos resultados também não foi testada fora do benchmark Recod.ai/LUC, e a comparação de eficiência é parcial porque memória de pico, tempo por época e latência padronizada não estão disponíveis para todos os modelos.

A separação por fonte depende da convenção de que versões relacionadas compartilhem o mesmo nome-base. O código verifica interseções entre os identificadores de treino, validação e teste e mantém amostras aumentadas ou sintéticas somente no treino; ainda assim, a garantia depende da correção dos nomes e metadados fornecidos pelo dataset.

5.2. Limitações computacionais e trabalhos futuros

Os experimentos foram executados em uma NVIDIA RTX 3060 de 12 GB. Esse recurso limitou a realização, dentro do período do estudo, de novas repetições e avaliações que exigiriam o treinamento completo das 21 configurações. A primeira extensão prevista é repetir os treinamentos com as cinco sementes já definidas no repositório, registrando os resultados de cada execução e calculando medidas de dispersão.

Também será necessário repetir os cenários de aumento com as mesmas transformações nas duas classes. Essa ablação permitirá verificar se as diferenças observadas decorrem do acréscimo de imagens ou de artefatos associados ao pré-processamento. Outros trabalhos futuros incluem escolher o limiar exclusivamente no conjunto de validação; comparar 384×384 e 518×518 mantendo as demais condições fixas; avaliar os modelos em outro benchmark; e registrar memória de pico, tempo de treinamento e latência pelo mesmo procedimento em todas as arquiteturas. A localização poderá ser avaliada quantitativamente quando estiver disponível um conjunto suficiente de máscaras anotadas. Esses experimentos não são usados para sustentar as conclusões atuais.

6. Conclusão

Sob o benchmark Recod.ai/LUC e os protocolos avaliados, os modelos DINO de maior porte alcançaram os melhores resultados de classificação. Na execução principal, o DINOv2-Large com aumento *offline* obteve 87,79% de acurácia, F1 de 0,8848, ROC-AUC de 0,9648 e PR-AUC de 0,9707.

Em comparação com o aumento *offline*, a inclusão de falsificações sintéticas elevou o F1 do DINOv3-Large em 0,0133 e sua PR-AUC em 0,0067. No DINOv2-Large, o F1 diminuiu 0,0152 e a PR-AUC diminuiu 0,0021. No DINOv3-Small, o F1 aumentou 0,0262, enquanto a PR-AUC diminuiu 0,0034. Portanto, os dados sintéticos não produziram melhora consistente entre os modelos e as métricas avaliadas.

Os resultados indicam que modelos fundacionais podem apoiar a triagem de imagens científicas, desde que a decisão final permaneça sob revisão humana. As conclusões se restringem ao benchmark, às arquiteturas e às execuções avaliadas. Repetições com múltiplas sementes, transformações balanceadas, comparação controlada de resolução, avaliação externa e análise quantitativa da localização são necessárias para verificar a estabilidade e a generalização dos resultados.

Referências

Bik, E. M., Casadevall, A., and Fang, F. C. (2016). The prevalence of inappropriate image duplication in biomedical research publications. *mBio*, 7(3):e00809–16.

Bucci, E. M. (2018). Automatic detection of image manipulations in the biomedical literature. *Cell Death & Disease*, 9(3):400.

Fridrich, J., Soukal, D., and Lukas, J. (2003). Detection of copy-move forgery in digital images. In *Proceedings of Digital Forensic Research Workshop*.

Bayar, B. and Stamm, M. C. (2016). Universal image manipulation detection with a constrained convolutional layer. In *ACM IH&MMSec*, pages 5–10.

Rao, Y. and Ni, J. (2016). Deep learning for splicing and copy-move forgery detection. In *IEEE WIFS*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*, pages 770–778.

Dosovitskiy, A. et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Caron, M. et al. (2021). Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660.

Oquab, M. et al. (2023). DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Siméoni, O. et al. (2025). DINOv3. *arXiv preprint arXiv:2508.10104*.

Recod.ai and LUC (2025). Recod.ai/LUC – Scientific Image Forgery Detection. Dataset documentation.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6:60.

Chicco, D. and Jurman, G. (2021). Advantages of MCC over F1 and accuracy in binary classification. *BMC Genomics*, 21:6.

Selvaraju, R. R. et al. (2017). Grad-CAM: Visual explanations via gradient-based localization. In *ICCV*, pages 618–626.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *ICML*, pages 233–240.

Saito, T. and Rehmsmeier, M. (2015). Precision-recall plots are more informative than ROC plots for imbalanced data. *PLOS ONE*, 10(3):e0118432.

Lipton, Z. C., Elkan, C., and Narayanaswamy, B. (2014). Thresholding classifiers to maximize F1 score. *arXiv preprint arXiv:1402.1892*.

Sheng, V. S. and Ling, C. X. (2006). Thresholding for making classifiers cost-sensitive. In *AAAI*, pages 476–481.