UNIVERSIDADE FEDERAL DO PARÁ
CAMPUS UNIVERSITÁRIO DE TUCURUÍ
FACULDADE DE ENGENHARIA DE COMPUTAÇÃO

CAMILY FURTADO CANTÃO

# Identificação de Artefatos Visuais em Faces Sintéticas por meio de Transfer Learning e IA Explicável: uma abordagem investigativa com ResNet50 e Grad-CAM

TUCURUÍ
2026

CAMILY FURTADO CANTÃO

# Identificação de Artefatos Visuais em Faces Sintéticas por meio de Transfer Learning e IA Explicável: uma abordagem investigativa com ResNet50 e Grad-CAM

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia de Computação, do Campus Universitário de Tucuruí, da Universidade Federal do Pará, como requisito parcial para obtenção do título de Engenheiro de Computação.

Orientador(a): Iago Lins de Medeiros

TUCURUÍ
2026

CAMILY FURTADO CANTÃO

**Identificação de Artefatos Visuais em Faces Sintéticas por meio de Transfer Learning e IA Explicável: uma abordagem investigativa com ResNet50 e Grad-CAM**

Trabalho de Conclusão de Curso apresentado à Faculdade de Engenharia de Computação, do Campus Universitário de Tucuruí, da Universidade Federal do Pará, como requisito parcial para obtenção do título de Engenheiro de Computação.

Data da aprovação: ____/____/_____

Conceito: _____

**BANCA EXAMINADORA**

_____

Prof. Dr. Iago Lins de Medeiros (FECOMP/UFPA)
Orientador

_____

Prof. Dr. Otávio Noura Teixeira (FECOMP/UFPA)
Membro Interno

_____

MsC. Rafael Veiga Teixeira (FEE-UFPA)
Membro Externo

## AGRADECIMENTOS

A Deus, por nunca ter soltado minha mão, mesmo nos momentos em que eu mal consigo enxergar o caminho à frente.

À minha avó, que hoje não está mais entre nós, mas cuja presença sinto em cada conquista. Ela foi meu maior pilar nos momentos que antecederam o vestibular e tenho a certeza de que ficaria imensamente feliz em ver sua neta se formar engenheira. Esta vitória também é dela.

À minha família, por ser a base de tudo.

Ao meu orientador, Prof. Iago Lins de Medeiros, pela condução cuidadosa desta pesquisa, pela disponibilidade e pela confiança depositada no meu trabalho ao longo de toda essa jornada.

Aos meus amigos David Williams, João Lucas e Rômulo, que foram presentes com paciência, generosidade e dicas valiosas nos momentos em que o caminho parecia mais tortuoso. Tornaram esse processo significativamente mais leve.

Ao Eduardo Miranda, meu namorado, por todo o suporte, pelo colo e pelo apoio incondicional nos momentos em que duvidei da minha própria capacidade. Sua presença fez toda a diferença.

Aos meus colegas de faculdade, que compartilharam comigo os desafios e as alegrias desse percurso e que hoje fazem muita falta, pois hoje, cada um seguiu seu próprio rumo. A saudade que fica é uma prova de que o que vivemos juntos vale muito.

## RESUMO

O desenvolvimento acelerado das tecnologias de Inteligência Artificial generativa levou à criação de conteúdos sintéticos hiper-realistas amplamente acessíveis, consolidando os *deepfakes* como uma ameaça concreta à integridade da informação, à privacidade e à segurança cibernética. Este trabalho investiga o uso de Visão Computacional e Inteligência Artificial Explicável (XAI) para identificar características visuais e artefatos estruturais associados a imagens faciais geradas por IA. Para isso, foi estruturado um *pipeline* metodológico composto por três macroetapas: a curadoria de um conjunto de dados balanceado contendo 10.000 imagens faciais reais e sintéticas, gerenciado pela plataforma Roboflow; o treinamento de um classificador binário baseado na arquitetura ResNet50 utilizando a técnica de Aprendizado por Transferência (*Transfer Learning*), executado na plataforma Kaggle com suporte de GPU NVIDIA Tesla T4; e a aplicação do algoritmo Grad-CAM (*Gradient-weighted Class Activation Mapping*) para gerar mapas de calor que evidenciam as regiões determinantes para as decisões do modelo. Como estudo exploratório preliminar, foi conduzido um experimento com a arquitetura YOLOv8, que revelou alto custo computacional e ausência de mecanismos nativos de interpretabilidade espacial, motivando a adoção da ResNet50 como principal ferramenta metodológica. O classificador final alcançou uma acurácia de 92,30%, um F1-Score de 92,29% e uma AUC-ROC de 0,9758, sem viés significativo entre as classes. Os mapas de calor gerados pelo Grad-CAM revelaram que o modelo direcionou sua atenção, de forma anatomicamente coerente, para a região central da face especificamente os olhos, o nariz e a boca nas imagens sintéticas, regiões consistentes com os artefatos de fusão facial descritos na literatura forense. Os resultados demonstram que a combinação entre Visão Computacional e XAI oferece não apenas elevada capacidade discriminativa, mas também a transparência e a auditabilidade indispensáveis para aplicações de perícia digital, superando as limitações operacionais de modelos puramente classificatórios.


**Palavras-chave:** Deepfakes; ResNet50; Transfer Learning; Inteligência Artificial Explicável; Grad-CAM.

## ABSTRACT

The advanced development of generative Artificial Intelligence technologies has led to the creation of widely accessible hyper-realistic synthetic content, solidifying deepfakes as a concrete threat to information integrity, privacy, and cybersecurity. This work investigates the use of Computer Vision and Explainable Artificial Intelligence (XAI) to identify visual features and structural artifacts associated with AI-generated sound images. To this end, a methodological pipeline was structured, composed of three macro-steps: curation of a balanced dataset of 10,000 real and synthetic material images, managed by the Roboflow platform; training of a binary classifier based on the ResNet50 architecture using Transfer Learning, executed on the Kaggle platform with NVIDIA Tesla T4 GPU support; and application of the Grad-CAM (Gradient-weighted Class Activation Mapping) algorithm to generate heatmaps that externalize the regions crucial for the model's decisions. As a preliminary exploratory study, an experiment was conducted with the YOLOv8 architecture, which revealed high computational cost and a lack of native spatial interpretability mechanisms, motivating the adoption of ResNet50 as the main methodological tool. The final classifier achieved an accuracy of 92.30%, an F1-Score of 92.29%, and an AUC-ROC of 0.9758, with no significant bias between classes. The heat maps generated by Grad-CAM revealed that the model directed its attention, in an anatomically coherent manner, to the central region of the face specifically the eyes, nose, and mouth in the synthetic images, regions consistent with the facial fusion artifacts described in the forensic literature. The results demonstrate that the c-ombination of Computer Vision and XAI offers not only high discriminative capacity but also transparency and auditability indispensable for digital forensics applications, overcoming the operational limitations of purely classificatory models.


Keywords: Deepfakes; ResNet50; Transfer Learning; Explainable Artificial Intelligence; Grad-CAM.

# LISTA DE TABELAS

# LISTA DE FIGURAS

# SUMÁRIO

## INTRODUÇÃO

O avanço acelerado das tecnologias de Inteligência Artificial (IA) tem possibilitado a criação de conteúdos sintéticos cada vez mais realistas, destacando-se os chamados *deepfakes*: imagens, vídeos e áudios gerados ou manipulados por algoritmos de aprendizado profundo (*deep learning*) com o intuito de simular autenticidade. Nos últimos anos, essa tecnologia deixou de ser restrita a laboratórios de pesquisa e tornou-se amplamente acessível ao público geral. Esse cenário impulsionou um crescimento expressivo na produção e disseminação automatizada de mídias falsificadas em redes sociais e plataformas digitais, tornando a distinção entre o real e o simulado um desafio complexo para os sistemas computacionais tradicionais (MUSTAK et al., 2024).

Em contrapartida à velocidade dessa evolução generativa, observa-se uma escassez de ferramentas de código aberto que combinem alta confiabilidade de detecção e viabilidade prática de auditoria pericial (MAWA et al., 2024). Diante desse panorama, o campo da Visão Computacional oferece suporte robusto para o desenvolvimento de classificadores profundos baseados em redes neurais convolucionais (CNNs). Todavia, os modelos tradicionais operam sob a filosofia de "caixa-preta", entregando predições matemáticas acuradas, mas desprovidas de uma justificativa visual estrutural, o que limita sua aplicação em cenários que exigem auditabilidade e confiança pericial (KUMAR et al., 2025).

Essa limitação arquitetônica torna-se ainda mais crítica à medida que as ferramentas generativas evoluem para contornar os métodos tradicionais de análise forense, estabelecendo uma verdadeira corrida armamentista digital (KUMAR et al., 2020; EDWARDS et al., 2024). Pesquisas contemporâneas indicam que a validação de detectores baseados em redes neurais convolucionais densas, como a família *ResNet*, exige não apenas um alto desempenho estatístico isolado, mas uma investigação profunda sobre a capacidade de generalização do modelo perante o escalonamento massivo de dados. Uma vez que a auditoria e a triagem manual de volumes que ultrapassam a escala de milhares de mídias mostram-se cenários operacionalmente inviáveis, a literatura recente tem preconizado a consolidação de *pipelines* automatizados que unam eficiência computacional e interpretabilidade espacial pós-indução. Torna-se imperativo, portanto, documentar a interação prática entre algoritmos de aprendizado

de máquina e a extração de evidências visuais localizadas, fundamentando cientificamente as decisões que ditarão a confiabilidade dos sistemas de segurança cibernética.

Diante disso, esta pesquisa investiga o uso da Visão Computacional aliada ao conceito de Inteligência Artificial Explicável (XAI) para identificar e mapear as características visuais e artefatos estruturais associados a imagens faciais geradas por IA. Utilizando a arquitetura *ResNet50* suportada pela técnica de Aprendizado por Transferência (*Transfer Learning*) como ferramenta metodológica de extração e classificação, o trabalho foca no emprego do algoritmo *Grad-CAM* (*Gradient-weighted Class Activation Mapping*). Dessa forma, o projeto transcende a mera rotulação binária de mídias, buscando correlacionar as regiões de maior ativação espacial da rede com as inconsistências biológicas descritas na literatura forense, fornecendo um mecanismo transparente para auxiliar os processos de análise cibernética e detecção de desinformação.

## JUSTIFICATIVA

A disseminação de *deepfakes* ultrapassa os limites do entretenimento digital, consolidando-se como uma ameaça severa à segurança da informação, à privacidade e à integridade dos ecossistemas digitais (MUSTAK et al., 2024). Sob a ótica científica e tecnológica da Engenharia de Computação, o desafio central reside no processamento e na classificação automática de grandes volumes de dados não estruturados, uma vez que cenários que envolvam dezenas de milhares de imagens inviabilizam qualquer abordagem de auditoria ou triagem pericial manual.

Nesse contexto, este trabalho pauta-se pela necessidade de estabelecer um *pipeline* de Visão Computacional que ofereça alta viabilidade prática e eficiência computacional ao lidar com o escalonamento massivo de dados. A seleção da arquitetura *ResNet50* fundamenta-se em sua consagrada capacidade de extração de características profundas através de *Transfer Learning*, o que mitiga custos proibitivos de infraestrutura e hardware ao reaproveitar de forma eficiente os pesos pré-treinados do banco *ImageNet* (SAMAL et al., 2025). A execução deste modelo na plataforma de nuvem *Kaggle*, com suporte de GPU dedicada, justifica-se pela estabilidade e pelo poder de processamento elástico necessários para gerenciar com segurança o banco de dados definitivo.

Portanto, a pesquisa mostra-se altamente relevante ao preencher uma lacuna prática de engenharia forense: demonstrar como a automação via scripts para curadoria e manipulação de

*datasets* massivos interage com sistemas computacionais baseados em Inteligência Artificial Explicável (XAI). Ao transcender a barreira dos modelos de "caixa-preta", o trabalho fornece uma base científica transparente e auditável via *Grad-CAM*, mapeando as inconsistências visuais das imagens geradas por IA e oferecendo respostas concretas para subsidiar tomadas de decisão no ecossistema de segurança cibernética.

# 1. OBJETIVOS

## 1.1. Objetivo Geral

Investigar, por meio de Visão Computacional e Inteligência Artificial Explicável, as características visuais associadas a imagens faciais geradas por IA, utilizando a arquitetura ResNet50 como ferramenta de classificação e o Grad-CAM como método de análise espacial, com vistas a subsidiar processos de detecção e análise forense de *deepfakes*.

## 1.2. Objetivos Específicos

- Compilar e estruturar um *dataset* de imagens faciais reais e sintéticas, aplicando técnicas de curadoria e pré-processamento via Roboflow.
- Realizar um experimento exploratório preliminar com YOLOv8 para validação do pipeline de detecção e justificativa da escolha metodológica.
- Implementar *Transfer Learning* com ResNet50 para classificação binária de imagens reais e geradas por IA.
- Avaliar o desempenho do modelo por meio de métricas como Acurácia, F1-*Score* e AUC-ROC.
- Aplicar Grad-CAM para identificar as regiões faciais que mais influenciam a decisão do modelo.
- Relacionar as regiões identificadas pelo Grad-CAM com artefatos visuais descritos na literatura, discutindo suas implicações para análise forense.

## 1.3. Relevância

A relevância deste trabalho manifesta-se, inicialmente, em uma dimensão socioeconômica e de segurança digital premente. Em um cenário em que a Inteligência Artificial generativa democratizou a criação de conteúdos sintéticos hiper-realistas, os *deepfakes* deixaram de ser apenas ferramentas de entretenimento e se consolidaram como ameaças concretas à integridade da informação. A incidência de fraudes financeiras e golpes de engenharia social baseados em falsificação de identidade facial cresceu substancialmente em

mercados globais, tornando urgente o desenvolvimento de contramedidas acessíveis e eficientes para a proteção de usuários e instituições.

Simultâneo a esse impacto social, o problema mostra-se igualmente relevante sob a ótica técnico-científica no campo da Engenharia e da Visão Computacional. O desafio reside na necessidade de automatizar a triagem, o processamento e a validação de grandes volumes de dados não estruturados, visto que analisar conjuntos de dados que ultrapassam a escala de milhares de imagens inviabiliza completamente auditorias manuais. Nesse contexto, investigar como modelos consolidados de aprendizado profundo se comportam diante de demandas massivas é fundamental para determinar o equilíbrio ideal entre a eficiência computacional de processamento e a precisão forense no reconhecimento de anomalias visuais.

### 1.4. Contribuições do Trabalho

Diante dos objetivos delineados, este trabalho oferece contribuições práticas e metodológicas de alto valor para a área de forense digital ao propor o desenvolvimento e a validação de um *pipeline* integrado de Visão Computacional e XAI. A pesquisa avança cientificamente ao afastar-se da abordagem de modelos puramente opacos (*black-box*), estruturando uma metodologia em etapas complementares: primeiramente, a automação e a curadoria digital de um conjunto de dados massivo e balanceado com 10.000 imagens através da plataforma *Roboflow*; em seguida, o processamento e treinamento otimizado da arquitetura *ResNet50* via *Transfer Learning* no ecossistema de nuvem *Kaggle* com aceleração por hardware dedicada; e, por fim, a injeção de interpretabilidade pós-indução por meio do algoritmo *Grad-CAM*. Esse mapeamento multifacetado permite contrastar as predições estatísticas do classificador com as anomalias visuais e os artefatos de fusão facial descritos na literatura, mitigando riscos de memorização (*overfitting*) e fornecendo uma metodologia transparente e auditável para subsidiar a tomada de decisões em segurança cibernética.

Por fim, a pesquisa contribui para a introdução de transparência e auditabilidade em sistemas baseados em inteligência artificial através da aplicação da técnica Grad-CAM no modelo residual. Ao abrir a "caixa-preta" dos modelos de *Deep Learning*, a geração dos mapas de calor prova cientificamente que o classificador toma suas decisões baseando-se em artefatos de fusão faciais reais, descartando vieses ou ruídos de fundo espúrios do conjunto de dados. Como entrega metodológica final, os *pipelines* de código e automação de dados foram documentados e organizados em notebooks de alto desempenho na plataforma Kaggle, servindo

como uma base sólida de consulta, replicação e evolução para futuras investigações sobre a viabilidade de ferramentas forenses de código aberto.

# 2. REFERENCIAL TEÓRICO

## 2.1. Visão Computacional

A Visão Computacional é um dos pilares tecnológicos fundamentais tanto para a criação quanto para a detecção de mídias sintéticas. Segundo Mustapha *et al.* (2025), os objetivos centrais da visão computacional são capacitar as máquinas a perceber e interpretar dados visuais, inspirando-se no funcionamento das redes neurais do cérebro humano. Em sua essência, a área concentra-se primordialmente em duas tarefas cruciais: a detecção de objetos, que busca identificar e localizar computacionalmente classes específicas em fotos (gerando caixas delimitadoras), e a classificação de objetos, que visa atribuir uma classe específica a uma imagem inteira ou a uma região de interesse. Com o avanço do aprendizado profundo (*deep learning*), a combinação de várias estruturas de redes neurais ganhou enorme popularidade para aprimorar o desempenho em diversas aplicações práticas de visão computacional, superando as limitações dos métodos manuais tradicionais.

No contexto específico da análise de imagens e vídeos faciais, Moura (2021) destaca que as tecnologias de visão computacional, aliadas ao aprendizado profundo, oferecem um suporte robusto para a detecção, identificação, reconhecimento e análise minuciosa de corpos e imagens. Uma etapa considerada de extrema importância pela autora é a detecção facial, que consiste no processo de localização exata de uma face dentro de uma imagem para a extração de regiões de interesse e pontos de referência (*landmarks*). Essa capacidade de localizar e analisar traços faciais permitiu que o campo evoluísse de técnicas de extração manual de características para arquiteturas complexas capazes de aprender representações de alto nível. Conforme abordado por Edwards et al. (2024), essa transição histórica na visão computacional foi marcada pelo desenvolvimento inicial de arquiteturas pioneiras como a LeNet-5 na década de 1990, até culminar nos modelos profundos altamente eficientes e paralelizáveis da atualidade.

A aplicação da visão computacional no cenário dos *deepfakes* caracteriza-se como uma verdadeira "corrida armamentista" tecnológica. Sharma *et al.* (2023) argumentam que o surgimento de ferramentas sofisticadas de edição de imagens e de tecnologias de *face swap* (troca de rostos) criou um cenário em que a necessidade de algoritmos avançados de visão computacional para distinguir o real do falso nunca foi tão crucial. Eles ressaltam que o uso de técnicas modernas de aprendizado profundo aumenta não apenas a precisão na detecção, mas também a velocidade de treinamento e a confiabilidade da classificação de imagens faciais.

Por sua vez, Moura (2021) define os *deepfakes* justamente como uma técnica baseada em aprendizado de máquina que proporciona uma ampla variedade de métodos para a troca e manipulação de faces, englobando, em larga escala, a utilização de tecnologias de ponta da visão computacional e do próprio aprendizado profundo. Kumar *et al.* (2020) complementam essa visão ao afirmar que a inteligência artificial forma o núcleo da formulação de *deepfakes*, o que capacita essas mídias fraudulentas a desviar e iludir as técnicas de detecção tradicionais por meio do aprendizado contínuo. Para combater essas falsificações em constante evolução, os pesquisadores de visão computacional recorrem frequentemente ao desenvolvimento de redes de aprendizado profundo personalizadas e técnicas de *Transfer Learning*.

Em suma, a visão computacional não é apenas o domínio que possibilitou o realismo impressionante das manipulações faciais modernas, mas é também a principal ferramenta investigativa forense para combatê-las. A necessidade de modelos com alta precisão levou pesquisadores a combinarem técnicas, como o uso conjunto de classificadores clássicos e detectores modernos, para solucionar os desafios da categorização em cenários complexos. Compreender esses princípios de processamento e interpretação de imagens é o passo inicial para aprofundar-se nas mecânicas de geração e impacto social dos falsos conteúdos digitais.

### 2.2. *Deepfakes*: Conceito, Histórico e Impacto

O termo *Deepfake* surgiu como uma junção das palavras *"deep"* (referente a *deep learning*, ou aprendizado profundo) e *"fake"* (conteúdo digital falso), caracterizando a criação de mídias fraudulentas geradas por Inteligência Artificial. Sathe *et al.* (2024) expandem essa definição, explicando que o termo atua como um guarda-chuva para diversas técnicas de alteração facial baseadas em visão computacional, que podem ser divididas em quatro categorias principais: troca de emoções, síntese total da face, manipulação de traços específicos e a troca de identidade (conhecida como *face swap*). Bipin *et al.* (2023) complementam essa visão ao destacar que, com os rápidos avanços nas redes generativas, as manipulações atingiram níveis de realismo tão altos que se tornam praticamente indistinguíveis ao olho humano.

Historicamente, a primeira aparição de destaque desta tecnologia ocorreu em novembro de 2017, na plataforma de fóruns Reddit. Segundo Moura (2021), um usuário anônimo autointitulado "u/deepfakes" aplicou algoritmos de aprendizado de máquina para trocar os rostos de celebridades e inseri-los em vídeos de conteúdo pornográfico. Edwards *et al.* (2024) observam que as primeiras encarnações das mídias *deepfakes* eram primitivas, frequentemente

associadas a imagens estáticas e de baixa qualidade. No entanto, a rápida evolução no treinamento de modelos de *Deep Learning* e o compartilhamento de algoritmos de código aberto permitiram um salto drástico na resolução e na fluidez dos vídeos, borrando definitivamente as fronteiras entre o que é real e o que é sintético.

As consequências do aprimoramento dessa tecnologia revelaram um impacto massivo e preocupante na sociedade e na política. Bipin et al. (2023) ressaltam que as *deepfakes* foram rapidamente instrumentalizadas para espalhar desinformação, manipular a opinião pública e destruir reputações. Moura (2021) lista exemplos históricos desse abuso político, incluindo vídeos falsificados de Barack Obama, Donald Trump e até mesmo uma alteração do ex-presidente Richard Nixon anunciando o fracasso da missão Apollo 11. Um exemplo geopolítico crítico é citado por Edwards et al. (2024): durante o conflito entre Rússia e Ucrânia em 2022, um vídeo *deepfake* do presidente ucraniano Volodymyr Zelensky foi amplamente distribuído na tentativa de incitar falsamente os cidadãos à rendição. Mustak et al. (2024) discutem o impacto macro desse fenômeno, evidenciando como a desinformação gerada por IA atinge uma escala capaz de ameaçar a estabilidade democrática.

Além da esfera política, os *deepfakes* representam ameaças severas à segurança financeira e à privacidade individual. Kumar *et al.* (2025) relatam um caso alarmante de fraude em que um áudio *deepfake* permitiu que criminosos realizassem um ataque de engenharia social contra uma empresa de energia do Reino Unido, clonando a voz do CEO para autorizar uma transferência ilícita de 220.000 euros. No entanto, a faceta mais sombria da tecnologia ainda reside no entretenimento adulto. Moura (2021) aponta dados alarmantes de que cerca de 96% dos *deepfakes* disponíveis na web possuem natureza pornográfica e, de forma devastadora, 99% desses vídeos têm mulheres como alvo, gerando conteúdos íntimos sem consentimento. Diante dessa realidade, Sharma *et al.* (2025) defendem que o desenvolvimento de classificadores precisos e arquiteturas de detecção baseadas em inteligência artificial deixou de ser um desafio puramente acadêmico para se tornar uma necessidade urgente de segurança global.

### 2.3. Redes Neurais Convolucionais (CNNs)

As **Redes Neurais Convolucionais (CNNs)** representam um dos marcos mais significativos no avanço da visão computacional moderna, sendo a arquitetura de base para a grande maioria dos sistemas de reconhecimento de imagem na atualidade. De acordo com Edwards *et al.* (2024), a fundação dessa arquitetura foi introduzida no final da década de 1980

e popularizada na década de 1990 com o desenvolvimento da LeNet-5, criada inicialmente para resolver problemas de reconhecimento de caracteres numéricos. Esses modelos ganharam enorme notoriedade por sua capacidade de aprender padrões complexos e multidimensionais a partir de grandes conjuntos de dados, utilizando neurônios artificiais que buscam replicar o funcionamento do cérebro humano. Solanke (2022) corrobora essa visão, afirmando que as estruturas das CNNs refletem núcleos internos extremamente complexos, cujas representações visuais se conectam muito bem com os padrões de pensamento humano, estabelecendo as bases para tarefas que vão desde a identificação de objetos até a detecção de anomalias microscópicas em mídias sintéticas.

Para compreender a eficácia das CNNs, é necessário analisar a sua mecânica de funcionamento, que se divide primordialmente em duas fases: a extração de características e a classificação. Kumar *et al.* (2025) explicam que, na primeira fase, a rede aprende a representação das feições da imagem por meio de uma hierarquia de camadas, eliminando a necessidade de extração manual de recursos. Moura (2021) detalha que essas redes são compostas tipicamente por camadas de convolução, de agrupamento (*pooling*) e camadas totalmente conectadas (*fully connected*). Bipin et al. (2023) aprofundam esse conceito ao descrever que as camadas convolucionais aplicam filtros espaciais que, em níveis superficiais, capturam características de baixo nível (como bordas e texturas) e, em camadas mais profundas, identificam padrões altamente complexos e anomalias deixadas pelos algoritmos geradores de *deepfakes*. Após a convolução, as camadas de *pooling* (como o *Max Pooling*) são aplicadas para reduzir a dimensionalidade espacial dos mapas de características, retendo as informações mais proeminentes e descartando dados redundantes, o que diminui o peso computacional da rede. Por fim, uma camada de achatamento (*flatten*) converte esses mapas em um vetor unidimensional que alimenta as camadas totalmente conectadas, responsáveis por cruzar os dados extraídos e realizar a classificação binária final.

No domínio específico da detecção forense de *deepfakes*, as CNNs provaram ser ferramentas excepcionalmente eficientes. Mustapha *et al.* (2025) argumentam que as camadas profundas das CNNs capacitam o modelo a distinguir entre itens visualmente muito semelhantes ao adquirir representações hierárquicas dos dados, tornando-as ideais para tarefas de classificação de granulação fina, como diferenciar uma pele real de uma textura gerada artificialmente. No entanto, a literatura técnica também aponta limitações importantes associadas a essas redes. Edwards *et al.* (2024) destacam que as implementações mais tradicionais de CNNs podem apresentar fraquezas na extração simultânea de características

locais e globais, dificultando o aprendizado de relações espaciais precisas entre os componentes do rosto. Uma das críticas recai justamente sobre o design da própria camada de *max pooling*, que, segundo análises levantadas por Edwards *et al.* (2024), pode resultar na perda de informações espaciais vitais e de detalhes de baixo nível que, de outra forma, seriam fundamentais para expor falsificações extremamente sutis.

Para mitigar essas vulnerabilidades arquitetônicas, a comunidade científica passou a desenvolver CNNs progressivamente mais densas, profundas e estruturadas como as famílias VGG e ResNet, ou a adotar abordagens híbridas. No âmbito do desenvolvimento experimental desta pesquisa, o uso prático de CNNs evoluiu a partir de modelos-piloto mais simples, os quais tendem a sofrer com os efeitos do *overfitting* ao memorizarem ruídos de conjuntos de dados reduzidos para o emprego de arquiteturas robustas suportadas pela técnica de *Transfer Learning*. A aplicação de uma rede madura e profunda como a ResNet50 (implementada na etapa principal do projeto) fornece a capacidade discriminativa necessária para mitigar o elevado nível de realismo presente nas mídias manipuladas (*deepfakes*), conseguindo reter as informações contextuais e espaciais complexas sem comprometer a eficiência computacional durante o processo de treinamento.

### 2.4. Arquitetura ResNet50

Do ponto de vista estrutural e operacional, Sathe *et al.* (2020) destacam que a arquitetura da ResNet50 é composta por blocos residuais profundos e densamente conectados. Embora a adição de múltiplas camadas de extração de características exija maior capacidade de memória e complexidade computacional, essa profundidade confere à ResNet50 uma vantagem significativa: a capacidade de identificar *deepfakes* com maior precisão do que redes mais rasas, como a VGG-19, por conseguir mapear anomalias hierárquicas de maior complexidade. Complementarmente, Kumar *et al.* (2020) detalham o funcionamento da extração de características nessa rede. Segundo os autores, as representações extraídas fluem até uma camada de agrupamento global (*Average Pooling Layer*), na qual a dimensionalidade é consolidada em um vetor de 2048 dimensões, formando um mapeamento semântico de alto nível antes das etapas finais de classificação.

A base desse funcionamento reside no mecanismo de aprendizado residual, elemento central da arquitetura ResNet50. Conforme ilustrado na Figura 1, cada bloco residual recebe uma entrada x e, em vez de aprender diretamente o mapeamento desejado $h(x)$ aprende a

função residual $F(x) = h(x) - x$. A saída do bloco é então obtida pela soma $F(x) + x$, viabilizada pela conexão de atalho (*skip connection*) que contorna as camadas convolucionais. Esse mecanismo garante que o gradiente flua diretamente pelas camadas mais profundas da rede durante o treinamento, mitigando o problema do desvanecimento do gradiente e permitindo que a ResNet50 alcance profundidades que redes convencionais não conseguem otimizar de forma eficaz (HE et al., 2016).

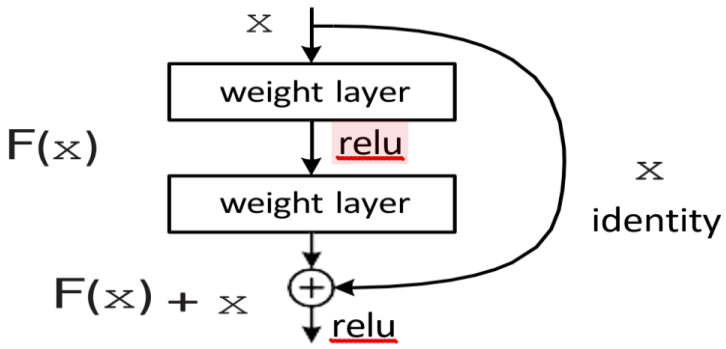


**Figura 1 – Mecanismo de um Bloco Residual com *Skip Connection* (por peso/convolução/ReLU e a linha curva da identidade pulando para a soma $f(x) + x$. Fonte: Adaptado de He et al. (2016).**

Na literatura acadêmica focada na detecção forense de mídias sintéticas, o desempenho da ResNet50 varia significativamente dependendo de como a sua camada final (*head*) é ajustado. Estudos conduzidos por Sharma *et al.* (2022) e Kumar *et al.* (2022) evidenciaram que, quando utilizada em sua forma padrão, sem otimizações adicionais, a ResNet50 atinge uma acurácia modesta, na faixa de 53% a 54%. No entanto, Kumar *et al.* (2025) demonstraram que o verdadeiro poder da arquitetura se revela quando ela é utilizada como extratora de características acoplada a técnicas de *Transfer Learning*: ao integrar as 2048 características da ResNet50 com um classificador de Regressão Logística, a acurácia salta drasticamente para 88,47%. Em experimentos híbridos semelhantes, Sathe et al. (2020) também registraram que a ResNet superou arquiteturas mais simples, atingindo 74% de acurácia de forma independente.

Esse aprendizado é orientado pela minimização contínua de uma função de perda, métrica matemática que quantifica o erro entre as predições do modelo e os rótulos verdadeiros do conjunto de dados. Por se tratar de um problema de classificação binária no qual cada imagem pertence exclusivamente à classe *real* ou à classe *fake* a função adotada neste trabalho é a *Binary Cross-Entropy* (BCE), amplamente consolidada na literatura de detecção de deepfakes para esse tipo de tarefa. Formalmente, a BCE é definida conforme a Equação 1:

$$L = -\frac{1}{N}\sum_{i=1}^{N} \left[ y_i \, log \, log \left( \hat{y}_i \right) + (1 - y_i) \, log \, log \left( 1 - \hat{y}_i \right) \right] \qquad (Equação\ 1)$$

onde $N$ representa o número total de amostras do lote de treinamento, $y_i$ corresponde ao rótulo verdadeiro da amostra $i$ (assumindo valor 0 para imagens reais e 1 para imagens sintéticas) e $\hat{y}_i$ é a probabilidade predita pelo modelo para essa mesma amostra. A cada época de treinamento, o otimizador Adam ajusta iterativamente os pesos da rede no sentido oposto ao gradiente dessa função, reduzindo progressivamente o erro de classificação e conduzindo o modelo à convergência comportamento evidenciado pela queda acentuada da curva de *loss* ao longo das 20 épocas de treinamento, conforme ilustrado na **Figura 2.**

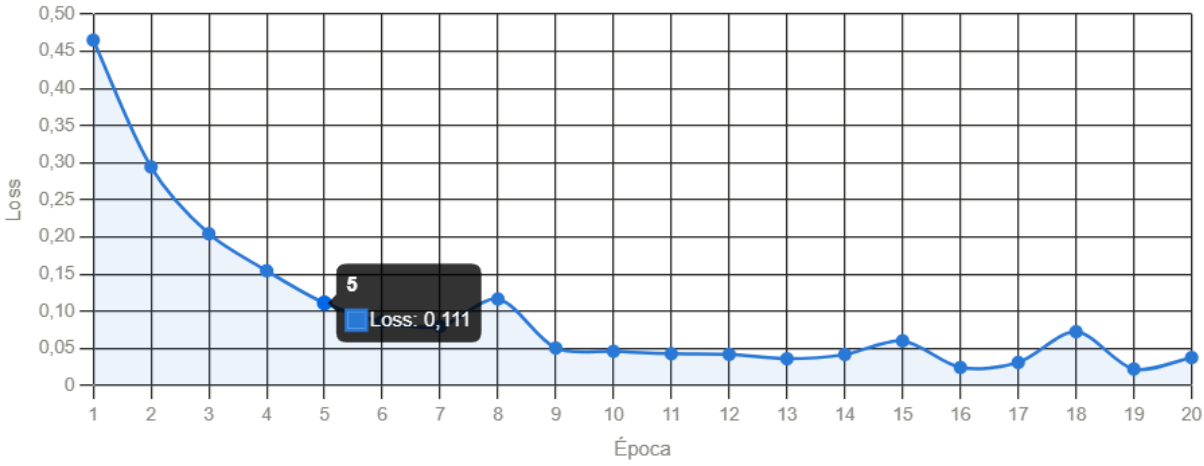


**Figura 2 - Curvas de Aprendizado de Perda (*Loss*) e Acurácia da *ResNet50*. Fonte: autoria própria, 2026.**

## 2.5. *Datasets* para Detecção de *Deepfake* e Métricas de Avaliação

A eficácia de qualquer modelo de aprendizado profundo na detecção de *deepfakes* está intrinsecamente ligada à qualidade, diversidade e volume dos dados utilizados durante o seu treinamento. Edwards *et al.* (2024) argumentam que o acesso a *datasets*, isto é, repositórios estruturados de amostras utilizados para treinar e avaliar modelos de inteligência artificial, que forneçam uma representação justa da diversidade demográfica (idade, gênero e etnia) e das mais recentes técnicas de falsificação é um dos desafios mais críticos da área. A ausência de dados representativos frequentemente resulta em modelos enviesados e com baixa capacidade de generalização em cenários do mundo real.

Nesse contexto, a literatura padrão-ouro consolidou um conjunto de *benchmarks* padrões de referência utilizados para comparar e avaliar o desempenho de diferentes sistemas que moldaram o desenvolvimento da área. Pesquisadores como Rössler *et al.* (2019) foram pioneiros ao introduzir o *FaceForensics++* (FF++), um banco de dados com milhares de vídeos manipulados por abordagens como *Face2Face* e *FaceSwap*, técnicas de troca e animação facial baseadas em aprendizado profundo, que se tornou a base para testes de inúmeras arquiteturas. Posteriormente, Li et al. (2020) propuseram o *Celeb-DF* (v2), um conjunto de dados de alta fidelidade desenvolvido para solucionar as falhas visuais óbvias dos bancos anteriores, fornecendo vídeos com artefatos de falsificação minimizados. Por sua vez, Dolhansky *et al.* (2020) lançaram o *Deepfake Detection Challenge* (DFDC), um *benchmark* colaborativo patrocinado por grandes empresas de tecnologia, introduzindo ampla variabilidade de iluminação, oclusões e cenários complexos para simular condições do mundo real.

Apesar da relevância histórica desses *benchmarks* clássicos, a literatura recente voltada à forense digital tem preconizado a adoção de bancos de dados massivos de imagens estáticas, mais adequados a pipelines de classificação binária. Um exemplo notório é o conjunto *140k Real and Fake Faces*, que contrapõe 70.000 faces autênticas extraídas da plataforma *Flickr* a 70.000 imagens sintéticas geradas pelo *StyleGAN*, algoritmo generativo de síntese facial desenvolvido pela NVIDIA capaz de produzir rostos humanos fotorrealistas. Fundamentada nessa necessidade de dispor de dados altamente balanceados e maduros para garantir a confiabilidade estatística dos classificadores, a etapa de curadoria deste Trabalho de Conclusão de Curso foi estruturada.

A partir desse referencial, superou-se a limitação inicial de um experimento piloto estruturado com apenas 965 imagens, cuja dimensão reduzida induziu a rede a um cenário de *overfitting,* fenômeno em que o modelo memoriza os padrões específicos do conjunto de treinamento em detrimento da capacidade de generalizar para dados novos, resultando em uma acurácia *baseline* ilusória de 99,5%. Para mitigar essa vulnerabilidade, consolidou-se o conjunto denominado *Teste_TCC_10k*, obtido por meio da plataforma *Roboflow*, ferramenta de gerenciamento e curadoria de dados visuais para aprendizado de máquina. Este *dataset* forneceu um universo de 10.000 imagens rigorosamente balanceadas entre as classes *real* e *fake*, submetidas a rotinas de pré-processamento para redimensionamento fixo de 224×224 pixels, além da aplicação de técnicas de aumento de dados *data augmentation*, como o espelhamento horizontal aleatório (*Random Horizontal Flip*), durante a fase de treinamento. Esse *pipeline* de engenharia de dados assegurou a robustez necessária para evitar a memorização de ruídos

espúrios e expandir a capacidade de generalização do modelo final, conforme mostrado na **figura 3.**

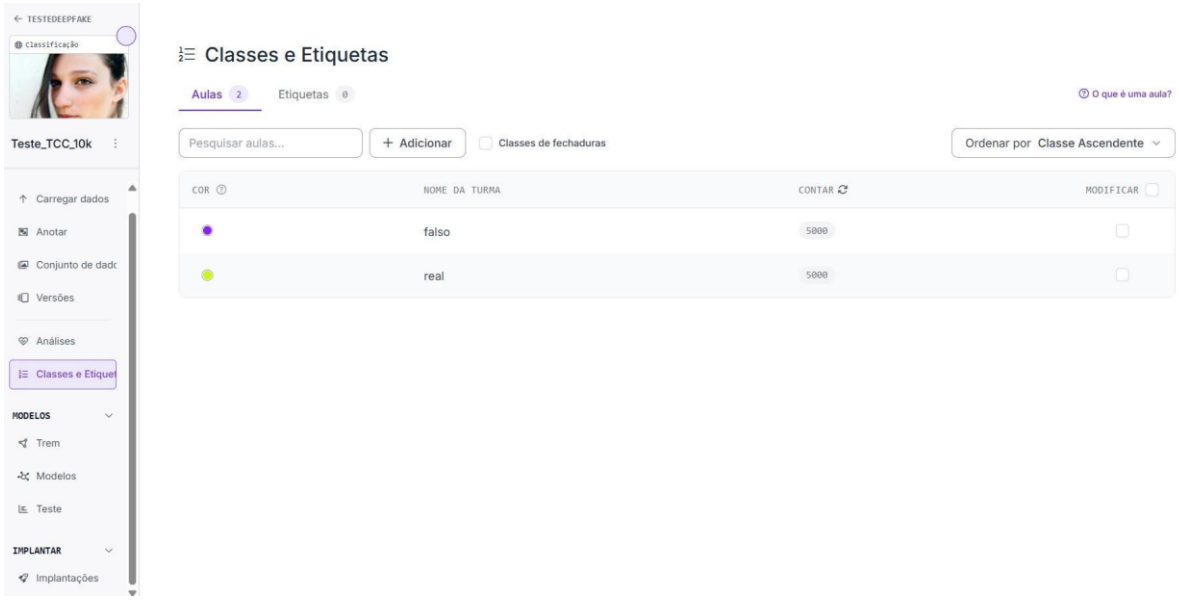


**Figura 3 – Distribuição e Curadoria do Dataset no Painel da Plataforma Roboflow. Fonte: autoria própria, 2026.**

Diante da robustez do conjunto de dados construído, tornou-se igualmente necessário definir critérios de avaliação à altura do rigor metodológico adotado. Edwards *et al.* (2024) e Moura (2021) advertem que o uso exclusivo da acurácia, proporção geral de previsões corretas em relação ao total de amostras analisadas como métrica de sucesso, não é suficiente para atestar a verdadeira confiabilidade de um classificador pericial, especialmente em dados desbalanceados ou em cenários do mundo real. Dessa forma, a avaliação técnica de estado da arte baseia-se na extração de valores da Matriz de Confusão, tabela que mapeia os acertos e erros do modelo por classe, para calcular um conjunto de métricas mais profundas. A Precisão (*Precision*) avalia a exatidão das predições positivas, sendo crucial quando o custo de um falso positivo é alto; o *Recall* mede a sensibilidade do modelo em detectar todos os *deepfakes* que realmente existiam no lote; e o *F1-Score* atua como a média harmônica entre Precisão e *Recall*, consolidando o poder geral do classificador em uma única métrica equilibrada. Adicionalmente, a métrica AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*), ou Área Sob a Curva de Característica de Operação do Receptor, é o padrão máximo para medir a capacidade do modelo de distinguir entre as classes independentemente do limiar de decisão escolhido, onde um valor igual a 1,00 representa a perfeição discriminativa.

O experimento principal desta pesquisa aplicou esse exato rigor forense após a execução de 20 épocas de treinamento, ciclos completos de exposição do modelo ao conjunto de dados de treino. Os ensaios computacionais documentam que o classificador obteve, junto ao conjunto de teste, uma acurácia de 92,30%, amparada por taxas consistentes de Precisão e *Recall* de 92,40% e 92,30%, respectivamente. A estabilidade metodológica foi ratificada por meio de um *F1-Score* de 92,29%; contudo, o principal indicador de validação científica do projeto residiu na AUC-ROC, que atingiu a marca de 0,9758, comprovando elevada capacidade discriminativa do modelo perante amostras inéditas e não estruturadas. Adicionalmente, a análise detalhada da Matriz de Confusão revelou um controle rigoroso de vieses, estabelecendo margens de erro de 10,41% para a taxa de falsos negativos (classe *fake*) e 5,10% para falsos positivos (classe *real*), atestando a eficácia e a maturidade da aplicação de redes neurais convolucionais profundas na esfera da forense digital.

### 2.6. Ambientes de Desenvolvimento e Ecossistemas de Nuvem

O treinamento de redes neurais convolucionais profundas demanda um poder computacional massivo que, frequentemente, ultrapassa as capacidades de processamento de hardware locais. Para solucionar esse obstáculo e viabilizar a execução dos algoritmos de visão computacional, o desenvolvimento prático deste trabalho apoiou-se no uso de ecossistemas de dados baseados em computação em nuvem. Conceitualmente, esses ecossistemas compreendem ambientes digitais integrados que permitem escalar análises e processar volumes massivos de dados de forma elástica, segura e distribuída, eliminando a necessidade de infraestrutura local.

Para preencher a lacuna de engenharia deste projeto, a infraestrutura computacional foi empregada de forma estratégica em duas plataformas de nuvens distintas, adequando-se à fase de maturação do modelo e à respectiva volumetria de dados.

- **Google Colaboratory (Colab):** Adotado na fase preliminar da investigação devido à sua viabilidade para prototipagem rápida e validação de scripts em linguagem Python. Operando sobre uma instância com suporte à aceleração por *hardware* via *GPU NVIDIA Tesla T4*, esse ambiente abrigou o primeiro experimento (Modelo 1), focado na homologação técnica do *pipeline* de *Transfer Learning* com o conjunto piloto de 965 imagens. Sequencialmente, conduziu-se um ensaio de estresse, submetendo a mesma plataforma a um conjunto massivo de 10.000 imagens, atuando como um cenário de calibração para avaliar a estabilidade do algoritmo e identificar potenciais vieses sistemáticos (*biases*) de amostragem.

- **Plataforma Kaggle:** Selecionada como o ecossistema definitivo para a execução e consolidação do experimento principal (Modelo 2). A transição justificou-se pela necessidade de um ambiente de processamento escalável e de alta estabilidade para manipular com segurança o banco de dados final. Apoiada pela aceleração de *hardware* de uma GPU NVIDIA Tesla T4 dedicada, a arquitetura realizou a ingestão e o processamento intensivo das 10.000 imagens faciais devidamente distribuídas em subconjuntos isolados de treino, validação e teste, convergindo de forma robusta em um tempo total de aproximadamente 1 hora e 30 minutos.

## 2.7. Métricas de Avaliação de Desempenho

A escolha de arquiteturas profundas, como a ResNet50 e o YOLOv8, bem como a utilização de ecossistemas de nuvem elásticos, ambientes de processamento distribuído em nuvem que permitem escalar recursos computacionais de forma dinâmica conforme a demanda para o processamento massivo de dados, levanta a necessidade de uma avaliação de desempenho igualmente rigorosa. Na literatura de forense digital, a avaliação de detectores de *deepfakes* historicamente sofre com a ausência de uma abordagem de medição estruturada, o que compromete a comparabilidade entre estudos e a confiabilidade dos resultados reportados.

Nesse cenário, pesquisadores alertam que a dependência exclusiva da métrica de acurácia pode ser ilusória, mascarando o real comportamento do modelo, especialmente ao lidar com dados desbalanceados ou com cenários do mundo real. Para contornar essa limitação e atestar a verdadeira confiabilidade da rede, o estado da arte exige o cruzamento de múltiplas métricas extraídas a partir da Matriz de Confusão, tabela que mapeia os acertos e erros do classificador em quatro quadrantes: Verdadeiros Positivos (*True Positives* TP), casos corretamente identificados como *fake*; Verdadeiros Negativos (*True Negatives* TN), casos corretamente identificados como *real*; Falsos Positivos (*False Positives* FP), imagens reais erroneamente classificadas como falsas; e Falsos Negativos (*False Negatives* FN), *deepfakes* que passaram despercebidos pelo modelo. A partir desse mapeamento detalhado, torna-se possível compreender a real eficácia do classificador em distinguir entre faces originais e sintéticas, pautando o desempenho computacional em métricas fundamentais como a Acurácia, a Precisão, o Recall, o F1-Score e a AUC-ROC — além do mAP (mean Average Precision), utilizado na avaliação de modelos da família YOLO. As definições detalhadas e as fórmulas de

cada uma dessas métricas são apresentadas na Seção 3.4.2, no contexto da metodologia experimental.

A evolução no emprego dessas métricas quantitativas marca a maturidade metodológica alcançada ao longo do desenvolvimento desta pesquisa. No primeiro experimento (Modelo 1), executado em ambiente preliminar sob um conjunto reduzido de 965 imagens, obteve-se uma acurácia baseline de 99,5%. Todavia, conforme aponta a literatura forense, a avaliação isolada desse indicador, sem o suporte de métricas complementares de dispersão e sensibilidade como o *F1-Score* e a área sob a curva ROC (AUC-ROC), mascara uma forte probabilidade de *overfitting*. Sob essa condição, o classificador tende a memorizar as peculiaridades estáticas do conjunto amostral reduzido em detrimento do aprendizado e da generalização dos padrões globais de falsificação biológica.

Para sanar essa limitação e atestar a verdadeira generalização da rede, o Modelo 2 (ResNet50 10k), treinado na plataforma Kaggle com 10.000 imagens, foi submetido a uma avaliação exaustiva. Os resultados provaram a excelência do classificador perante um volume massivo de dados novos: o modelo não apenas atingiu uma acurácia realista de 92,30%, como sustentou esse desempenho com uma precisão de 92,40%, *recall* de 92,30% e F1-*Score* de 92,29%

O grande destaque metodológico foi a métrica de AUC-ROC alcançando 0,9758, o que comprova uma altíssima capacidade de discriminação entre as faces. Aliadas a uma Matriz de Confusão balanceada (apresentando taxas de erro próximas e aceitáveis, de 10,41% para a classe fake e 5,10% para a classe real), essas métricas solidificam a arquitetura proposta como uma ferramenta robusta e transparente, superando a avaliação isolada e garantindo a confiabilidade técnica da monografia. A condução do aprendizado operou de forma contínua ao longo das épocas de treinamento sob a gerência do otimizador *Adam*, cuja aplicação é amplamente chancelada pela literatura de detecção digital por proporcionar estabilidade na minimização da função de perda. Essa consolidação de infraestrutura e engenharia pavimentou o caminho para a extração analítica das métricas e curvas de aprendizado que fundamentam a validação deste projeto.

Para sintetizar a sequência operacional descrita e evidenciar a integração entre as etapas de tratamento de dados, processamento em nuvem e validação forense, o fluxo metodológico completo é mapeado de forma sequencial na **Figura 4.**

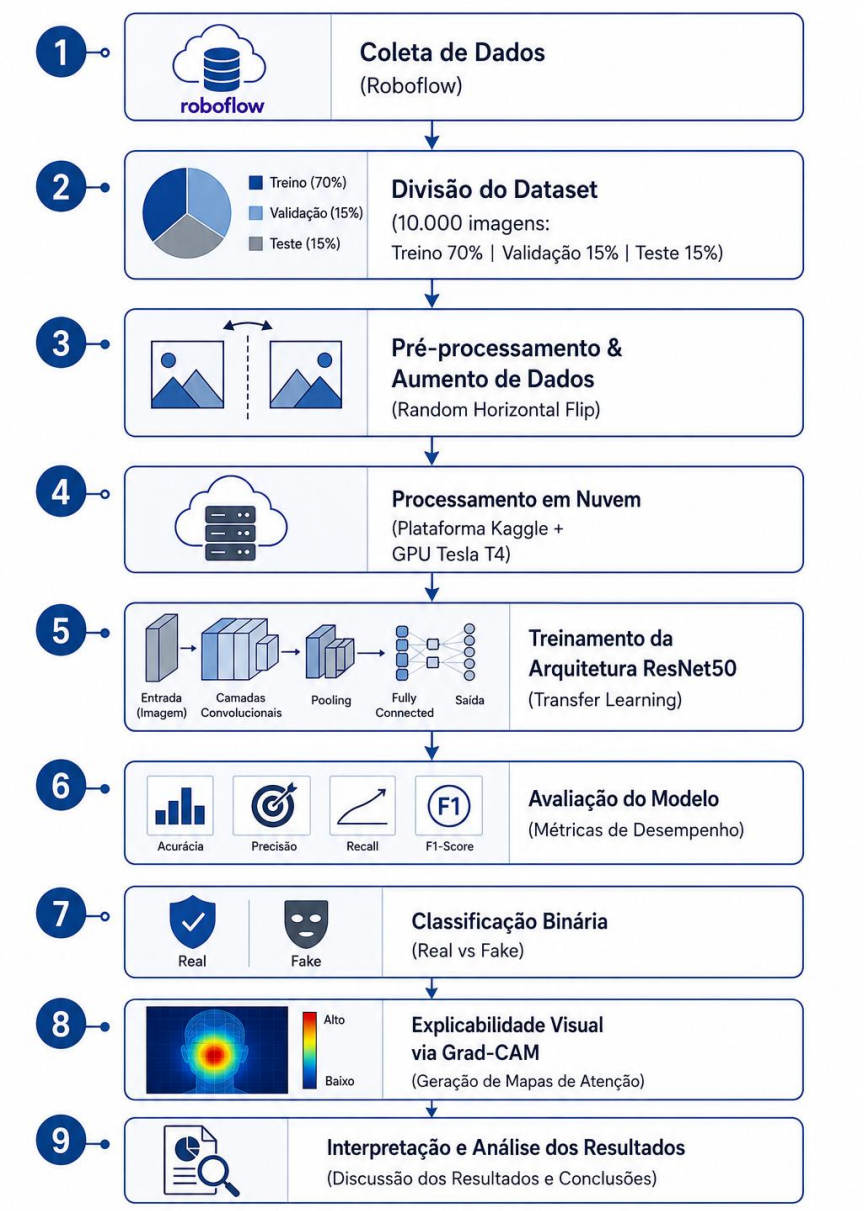


**Figura 4 – Fluxograma Geral do Pipeline de Engenharia e Processamento de Dados.**
**Fonte: autoria própria, 2026.**

## 2.8 Trabalho Anterior Correlato

Em trabalho anterior desenvolvido no âmbito da mesma instituição, Cantão e Medeiros (2025) realizaram uma revisão crítica da literatura sobre a detecção visual de *deepfakes* a olho nu, investigando de que forma a percepção humana pode atuar como ferramenta complementar aos detectores automáticos. O estudo sistematizou os principais indícios visuais recorrentes em vídeos sintéticos como ausência de piscadas naturais, fluidez exagerada de movimentos, textura de pele homogênea e inconsistências entre figura e fundo e os aplicou na análise de um vídeo gerado pela plataforma Google Veo 3. Os achados evidenciaram que, mesmo diante de

conteúdos de alta qualidade estética, artefatos visuais ainda são detectáveis por observadores atentos. Esse trabalho fundamenta e complementa a presente pesquisa: enquanto aquele explora os limites da percepção humana, esta propõe uma abordagem computacional automatizada, utilizando Transfer Learning com ResNet50 e IA Explicável por meio do Grad-CAM para identificar e visualizar os mesmos tipos de artefatos de forma sistemática e auditável.

## 3. METODOLOGIA

O presente capítulo descreve de forma minuciosa o pipeline metodológico e de engenharia de software desenvolvido para a concepção do sistema de detecção de mídias sintéticas faciais. Afastando-se de abordagens puramente comparativas de arquiteturas, a abordagem experimental deste trabalho assume um caráter investigativo e pericial, estruturado em três macroetapas sequenciais: a engenharia e curadoria do conjunto de dados; o protocolo de indução e classificação por meio de uma rede neural convolucional profunda; e, por fim, o arcabouço forense de injeção de interpretabilidade visual pós-indução. Todo o *pipeline* foi planejado para garantir a reprodutibilidade científica e assegurar a auditabilidade das decisões tomadas pelo modelo preditivo.

### 3.1. Estudo Exploratório Preliminar (*Pipeline* YOLOv8)

Como etapa preliminar, foi conduzido um experimento exploratório utilizando a arquitetura *YOLOv8* na variante de classificação de imagens (*yolov8n-cls*), com o objetivo de validar o *pipeline* de detecção e embasar de forma empírica a escolha metodológica para a etapa principal do trabalho.

Foram realizados dois experimentos com o *YOLOv8*, ambos executados no ambiente *Google Colab*. O primeiro utilizou um subconjunto reduzido de 965 imagens extraídas do *dataset 140k Real and Fake Faces (Kaggle)*, servindo como prova de conceito para verificar a viabilidade inicial do pipeline. O segundo experimento foi conduzido sobre o *dataset* completo *Teste_TCC_10k*, com 10.000 imagens, permitindo avaliar o comportamento do modelo em maior escala. Em ambos os experimentos exploratórios, foram adotados os mesmos hiperparâmetros: 20 épocas de treinamento, tamanho de lote (*batch size*) de 32 e resolução de entrada de 224 × 224 pixels. O treinamento do experimento em larga escala demandou aproximadamente 5 horas de processamento no *Colab*.

Embora o *YOLOv8* tenha apresentado métricas competitivas com AUC-ROC de 0,9934 e AP de 0,9939, o elevado custo computacional e o tempo de treinamento (5 horas), aliados à ausência de mecanismos nativos de interpretabilidade espacial pós-indução, motivaram a

adoção da arquitetura *ResNet50* associada ao algoritmo *Grad-CAM* como o arcabouço principal para a investigação forense proposta neste trabalho.

## 3.2. Coleta e Preparação dos Dados (Etapa Principal)

A etapa principal da metodologia concentra-se na estruturação do insumo fundamental para o aprendizado de máquina adaptativo. Esta fase compreende a seleção criteriosa das amostras, a organização das classes e as transformações lineares aplicadas no espaço vetorial das imagens.

### 3.2.1. Construção do *Dataset*

Para a consolidação desta pesquisa, utilizou-se o conjunto de dados definitivo denominado *Teste_TCC_10k*, centralizado e gerenciado por meio da plataforma de curadoria digital *Roboflow*. O volume total do banco de dados compreende exatamente 10.000 imagens estáticas de rostos humanos, distribuídas de forma rigorosamente paritária: 5.000 imagens correspondentes a faces autênticas (*Real*) e 5.000 imagens correspondentes a faces geradas artificialmente (*Fake*) por algoritmos generativos de ponta. O processo de organização e rotulagem das classes foi automatizado via código, associando metadados binários estruturados a cada imagem para permitir a indexação eficiente pelas rotinas de carregamento de dados (*data loaders*).

### 3.2.2. Divisão dos Dados (Data Split)

Buscando garantir a lisura estatística e o isolamento completo de dados para a validação do classificador, o universo de 10.000 imagens foi particionado estrategicamente em três subconjuntos mutuamente exclusivos, adotando a proporção matemática de 70% para treinamento, 15% para validação e 15% para teste. A distribuição quantitativa e o papel pericial das imagens foram configurados da seguinte forma:

- **Conjunto de Treinamento (7.000 imagens):** Destinado exclusivamente à alimentação das camadas convolucionais e à retropropagação do erro para a atualização dinâmica dos pesos sinápticos do modelo;
- **Conjunto de Validação (1.500 imagens):** Isolado para o monitoramento contínuo da convergência e do comportamento da função de perda ao final de cada época, atuando como o balizador para identificar sinais prematuros de superajuste (*overfitting*);

- **Conjunto de Teste (1.500 imagens):** Reservado de forma estrita para a avaliação final do classificador. Este lote permaneceu completamente inédito para a rede durante todo o ciclo de aprendizado, servindo para extrair as métricas reais de generalização e fornecer as amostras que seriam submetidas à auditoria visual.

### 3.2.3. Pré-processamento e Aumento de Dados (*Data Augmentation*)

Como as imagens originais do ecossistema apresentavam resoluções heterogêneas, implementou-se um *pipeline* de pré-processamento padronizado utilizando a biblioteca *PyTorch*. Primeiramente, aplicou-se um algoritmo de redimensionamento espacial fixo para transpor todas as matrizes de pixels para a dimensão de 224 × 224 pixels. Essa resolução atende à restrição dimensional nativa da arquitetura de extração de características adotada no projeto. Sequencialmente, os valores numéricos dos canais de cores (RGB) foram normalizados com base nos parâmetros estatísticos de média e desvio padrão do banco *ImageNet*, acelerando a velocidade de convergência do algoritmo.

Por fim, visando expandir artificialmente a variabilidade do cenário e injetar robustez ao modelo, o lote destinado exclusivamente ao treinamento foi submetido à técnica de aumento de dados (*data augmentation*) em tempo de execução. Aplicou-se a transformação linear de espelhamento horizontal aleatório (*Random Horizontal Flip*) com probabilidade de **50%**. Essa operação altera a orientação espacial da face sem modificar o rótulo biológico da imagem, forçando as camadas convolucionais a aprenderem invariâncias estruturais profundas e impedindo que o classificador memorize o posicionamento ou ruídos de fundo espúrios das fotos originais.

## 3.3. Treinamento do Modelo

Este estágio descreve o *pipeline* de engenharia de software e o arranjo algorítmico concebidos para a consolidação do classificador inteligente. O processo documenta desde o acoplamento estrutural da arquitetura profunda utilizada até a calibração fina dos hiperparâmetros que regeram a otimização matemática do sistema.

### 3.3.1. Arquitetura ResNet50 como Ferramenta de Extração

Como espinha dorsal (*backbone*) para a extração de características profundas e padrões de alta complexidade nas imagens faciais, o *pipeline* incorporou a arquitetura *ResNet50*. A

escolha dessa rede neural convolucional profunda fundamenta-se na presença de conexões residuais (*skip connections*), um mecanismo de engenharia que resolve o problema do desaparecimento do gradiente em redes de grande profundidade, permitindo o fluxo direto do sinal de aprendizado e preservando a integridade das informações ao longo das 50 camadas da estrutura.

No escopo desta investigação, a *ResNet50* não atua como o objeto de estudo analítico, mas sim como uma ferramenta metodológica de suporte. Em vez de inicializar o treinamento do zero o que exigiria um tempo de processamento proibitivo, adotou-se o paradigma de *Transfer Learning*. A arquitetura foi importada contendo os pesos sinápticos previamente otimizados no banco de dados massivo *ImageNet* (*IMAGENET1K_V1*). Essa abordagem confere ao modelo a capacidade de reaproveitar um conhecimento prévio robusto sobre formas, texturas, geometrias e descontinuidades espaciais, direcionando essa inteligência computacional especificamente para o escopo de detecção de anomalias sintéticas.

### 3.3.2. Customização e Isolamento de Camadas

Para adequar a ferramenta originalmente projetada para classificar 1.000 categorias de objetos gerais à especificidade forense deste trabalho, implementou-se uma rotina de modificação estrutural na rede via código. O procedimento técnico consistiu em congelar os parâmetros e pesos de todas as camadas convolucionais iniciais e intermediárias da *ResNet50*, blindando o conhecimento prévio obtido no *ImageNet* e forçando essas camadas a atuarem estritamente como um extrator de características fixo.

Na sequência, a camada final totalmente conectada (*Fully Connected Layer*) original da rede foi removida. Em seu lugar, acoplou-se uma nova estrutura de classificação binária customizada. Essa nova camada foi programada para receber o vetor denso de 2.048 dimensões gerado após a operação de agrupamento global (*Global Average Pooling*) e mapeá-lo diretamente em uma saída linear simplificada, correspondente aos dois neurônios do nosso problema: classe 0 (*Fake*) e classe 1 (*Real*).

### 3.3.3. Configuração dos Hiperparâmetros e Otimização

A calibração dos hiperparâmetros foi planejada para assegurar a estabilidade matemática do algoritmo e mitigar o risco de divergência do gradiente. O ciclo de aprendizagem operou sob as seguintes diretrizes periciais**:**

- Função de Perda (*Loss Function*): Utilizou-se a Entropia Cruzada Binária (*Binary Cross-Entropy Loss*), algoritmo ideal para problemas de classificação segregada, encarregado de mensurar a discrepância quantitativa entre as probabilidades preditas pela rede e os rótulos biológicos reais das imagens faciais;
- Algoritmo Otimizador: Aplicou-se o otimizador *Adam* (*Adaptive Moment Estimation*). A escolha justifica-se por sua eficiência computacional e baixa demanda de memória, calculando taxas de aprendizado adaptativas para diferentes parâmetros a partir do monitoramento dos momentos dos gradientes;
- Taxa de Aprendizado (*Learning Rate*): Fixada no valor constante de 0,001, garantindo passos de ajuste finos o suficiente para uma convergência suave sem causar oscilações caóticas no gradiente de perda;
- Tamanho do Lote (*Batch Size*): Configurado em 32 imagens por lote, otimizando o uso dos registradores de memória de vídeo da GPU e estabilizando a estimativa do gradiente a cada iteração de treino.

### 3.3.4. Processo de Execução e Infraestrutura de Nuvem

O pipeline de treinamento foi executado continuamente ao longo de 20 épocas. Visando contornar as limitações de hardware local, o processamento foi hospedado integralmente no ecossistema de nuvem da plataforma *Kaggle*, utilizando a aceleração gráfica de uma GPU *NVIDIA Tesla T4* dedicada.

Durante cada época, o conjunto de treinamento alimentou a rede em lotes sequenciais; ao término de cada ciclo completo de rotação dos dados, o classificador foi submetido ao conjunto de validação isolado para extração das curvas de perda e acurácia. O *pipeline* convergiu de forma madura em um tempo total de processamento de aproximadamente 1 hora e 30 minutos, gerando o arquivo de pesos binários consolidado e pronto para a etapa de auditoria e aplicação da inteligência explicável.

## 3.4. Avaliação e Interpretabilidade

Esta etapa final da metodologia estabelece critérios científicos e os mecanismos algorítmicos empregados para auditar a capacidade de generalização do modelo e expor os critérios visuais que foram utilizados pela rede neural em suas respectivas tomadas de decisão.

### 3.4.1. Protocolo Experimental e Isolamento de Teste

Para garantir a total isenção estatística na validação do sistema, o modelo foi submetido ao conjunto de teste isolado composto por 1.500 imagens inéditas (750 reais e 750 sintéticas). Este lote funcionou como uma simulação de ambiente de produção real, impossibilitando qualquer vazamento de dados (*data leakage*). O protocolo consistiu em alimentar a rede com essas amostras para que o classificador binário realizasse as inferências independentes, gerando as matrizes de confusão que subsidiam as análises estatísticas.

### 3.4.2. Métricas de Desempenho Quantitativo

A validação quantitativa da capacidade preditiva do detector foi estruturada a partir de cinco métricas consolidadas na literatura de aprendizado de máquina, calculadas a partir das relações de Verdadeiros Positivos (TP), Verdadeiros Negativos (TN), Falsos Positivos (FP) e Falsos Negativos (FN) extraídas da matriz de confusão:

- **Acurácia:** Mede a proporção geral de previsões corretas (reais e falsas) em relação ao total de amostras analisadas no lote de teste;
- **mAP (mean Average Precision):** Muito utilizada na avaliação de modelos da família YOLO, calcula a precisão média em diferentes limiares de interseção, sendo empregada na avaliação do experimento exploratório com o YOLOv8 (Seção 3.1);
- **Precisão:** Avalia a exatidão das predições positivas, calculando a proporção de imagens rotuladas como sintéticas que de fato eram *deepfakes* legítimos;
- **Recall:** Quantifica a sensibilidade do classificador em mapear e identificar corretamente as mídias fraudulentas existentes, minimizando a incidência de falsos negativos;
- **F1-Score:** Atua como a média harmônica entre a Precisão e o *Recall*, fornecendo um indicador unificado e equilibrado sobre a eficácia geral do classificador.

### 3.4.3. Engenharia de Explicabilidade Visual via Grad-CAM

Com o propósito de mitigar a opacidade inerente ao modelo convolucional profundo ("caixa-preta") e fornecer uma ferramenta transparente de suporte à auditoria pericial, implementou-se o algoritmo *Grad-CAM* (*Gradient-weighted Class Activation Mapping*). Em termos de engenharia de software, o script foi programado para interceptar o fluxo de retropropagação e capturar os gradientes de erro calculados especificamente na última camada

convolucional da *ResNet50*, logo antes da redução espacial promovida pelo agrupamento global.

Os gradientes capturados foram submetidos a uma operação de agrupamento de importância por canal, gerando um peso ponderado para cada mapa de ativação. Na sequência, aplicou-se uma combinação linear ponderada seguida por uma função de retificação linear (ReLU), isolando apenas as características que influenciaram positivamente a decisão de classificar a imagem dentro do seu respectivo rótulo predito.

O resultado matemático dessa operação gerou uma matriz bidimensional de intensidade, a qual foi redimensionada por interpolação bilinear para corresponder à resolução original de 224×224 pixels da face examinada. Por fim, essa matriz foi sobreposta à imagem facial na forma de um mapa de calor em espectro pseudocor, onde tonalidades quentes indicam as regiões de máxima atenção da rede. Essa abordagem permite correlacionar diretamente as áreas de convergência computacional com os artefatos de renderização e fusão facial descritos na literatura, transformando a resposta matemática em uma evidência visual auditável para o ecossistema de forense digital.

# 4. RESULTADOS E DISCUSSÕES

## 4.1. Resultados do Experimento Exploratório com YOLOv8

O experimento exploratório com a arquitetura *YOLOv8* foi conduzido em duas etapas sobre o conjunto de dados *Teste_TCC_10k*, ambas executadas no ambiente de nuvem *Google Colab*. O objetivo fundamental dessa fase preliminar não residia na obtenção do modelo definitivo do trabalho, mas sim na validação do *pipeline* de detecção e no fornecimento de subsídios empíricos para balizar a escolha metodológica da etapa principal.

Os resultados obtidos mostraram-se expressivos sob a ótica quantitativa. A matriz de confusão revelou que, sobre o conjunto de teste isolado, o modelo classificou corretamente 460 imagens da classe *Fake* e 494 da classe *Real*, registrando uma taxa residual de apenas 30 falsos negativos e 16 falsos positivos. A curva ROC apresentou uma Área Sob a Curva (**AUC**) de **0,9934**, enquanto a curva de Precisão-*Recall* registrou uma Precisão Média (**AP** - *Average Precision*) de **0,9939**, ambas indicativas de uma altíssima capacidade discriminativa entre as distribuições das classes.

No entanto, a despeito do elevado desempenho quantitativo, dois fatores técnicos limitaram a adoção do *YOLOv8* como o modelo principal da pesquisa:

1. **O Custo Computacional:** O treinamento em larga escala demandou aproximadamente 5 horas de processamento analítico no ambiente *Colab*, evidenciando uma alta exigência de recursos de hardware e tempo de execução desfavorável.
2. **A Opacidade Algorítmica:** O fator mais determinante para os objetivos deste trabalho foi a ausência de mecanismos nativos de interpretabilidade espacial, característica essencial para a análise das regiões faciais associadas a imagens geradas por IA, que constitui o cerne da investigação proposta.

Esses fatores mitigaram a viabilidade do *framework* para a continuidade do projeto e motivaram a adoção da arquitetura *ResNet50* como o modelo principal, dada a sua compatibilidade nativa com técnicas de Inteligência Artificial Explicável (XAI), em especial o algoritmo *Grad-CAM*.

## 4.2. Resultados da Classificação com ResNet50

O modelo principal da pesquisa foi treinado por meio do paradigma de *Transfer Learning* sobre a arquitetura profunda *ResNet50*, incorporando os pesos pré-treinados do banco *ImageNet* (IMAGENET1K_V1). O ciclo de indução estendeu-se por 20 épocas contínuas no ecossistema de nuvem *Kaggle*, sob a aceleração de hardware de uma GPU *NVIDIA Tesla T4*, demandando um tempo total de processamento de aproximadamente 1 hora e 30 minutos.

### 4.2.1. Curvas de Aprendizado

As curvas de acurácia e perda (*Loss*) por época evidenciam um processo de aprendizado estável, consistente e termodinamicamente equilibrado. A acurácia do conjunto de treinamento expandiu-se progressivamente, atingindo patamares superiores a 98% nas épocas finais do ciclo. Em contrapartida, a acurácia de validação estabilizou-se de forma assintótica na faixa compreendida entre 91% e 93%, demonstrando uma sólida capacidade de generalização do modelo perante dados não vistos durante a otimização dos parâmetros.

A função de perda manifestou uma trajetória de queda contínua ao longo das épocas, exibindo oscilações pontuais e controladas decorrentes do uso de lotes reduzidos (*mini-*

*batches*), sem apresentar qualquer indício de divergência matemática ou de superajuste (*overfitting*). Os resultados descritos podem ser observados na **figura 5:**

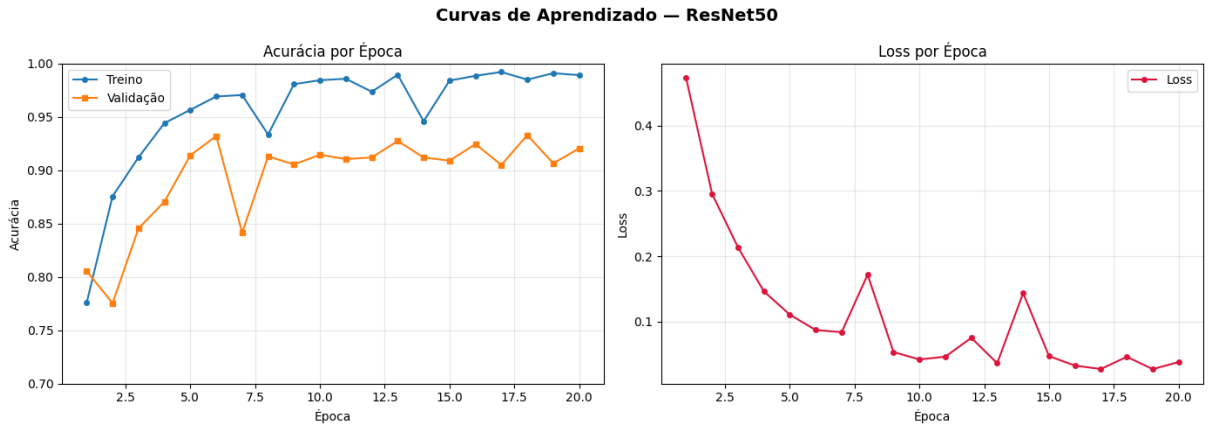


**Figura 5 - Curvas de Aprendizado de Perda (*Loss*) e Acurácia da *ResNet50*.**
**Fonte: autoria própria, 2026.**

### 4.2.2. Métricas de Desempenho Quantitativo

A avaliação estatística realizada sobre o conjunto de teste estritamente isolado produziu uma matriz de resultados homogênea e robusta, conforme sistematizado na Tabela 1.

**Tabela 1 – Métricas de desempenho**

| **Tabela - MÉTRICAS DE DESEMPENHO DO MODELO RESNET50** | | |
|---|---|---|
| **Métrica** | **Resultado** | **Interpretação** |
| **Acurácia** | 92,30% | Alta taxa de acerto geral no conjunto de teste |
| **Precisão (*Weighted*)** | 92,40% | Baixa taxa de falsos positivos entre as classes |
| **Recall (*Weighted*)** | 92,30% | Alta capacidade de recuperar amostras corretas |
| **F1-Score (*Weighted*)** | 92,29% | Equilíbrio entre Precisão e Recall |
| **AUC-ROC** | 0,9758 | Excelente capacidade discriminativa entre as classes |

**Fonte: Elaborada pela autora, 2026**.

O índice AUC-ROC de 0,9758 merece um destaque pericial especial. Enquanto um classificador puramente aleatório apresentaria uma área sob a curva de 0,50 e um classificador teórico perfeito atingiria 1,00, o valor obtido nesta pesquisa comprova a altíssima capacidade discriminativa do pipeline para segregar imagens reais daquelas geradas artificialmente por IA, operando com confiabilidade mesmo em cenários de baixa confiança estatística na predição linear.

### 4.2.3. Matriz de Confusão

A análise fina dos quadrantes da Matriz de Confusão revelou que a rede classificou corretamente 439 imagens sintéticas (*Fake*) e 484 imagens autênticas (*Real*) dentro do universo do lote de teste. O sistema registrou 51 falsos negativos e 26 falsos positivos.

A partir desses valores absolutos, computou-se que as taxas de erro por classe foram de 10,41% para as mídias *Fake* e de apenas 5,10% para as fotos *Real*. Esses percentuais demonstram a ausência de viés (*bias*) significativo entre as categorias analisadas, confirmando que a engenharia do modelo não favorece sistematicamente nenhuma das classes e mantém a estabilidade da auditoria, conforme detalhado na **figura 6**.

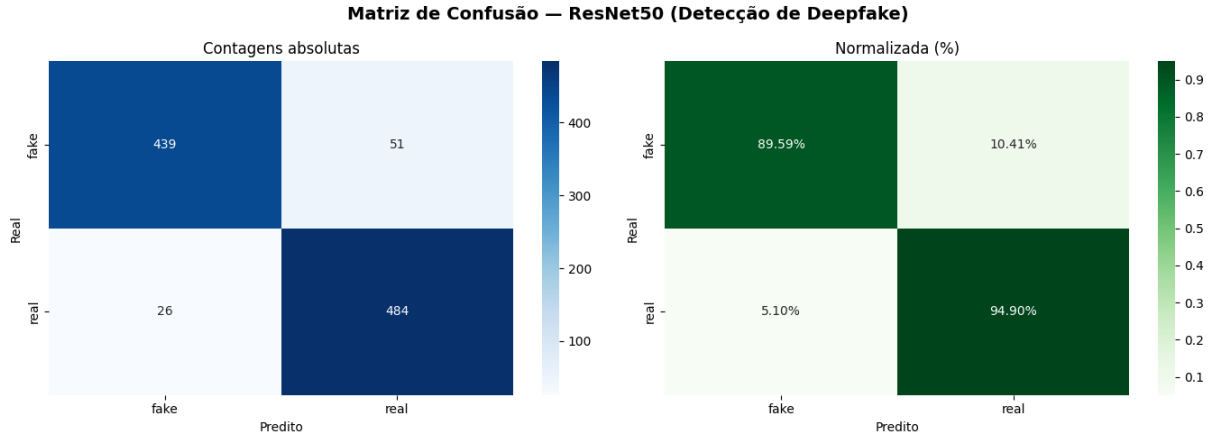


**Figura 6 - Matriz de Confusão do Modelo ResNet50 no Lote de Teste.**
**Fonte: autoria própria, 2026.**

## 4.3. Análise do Grad-CAM e Regiões Faciais

O algoritmo *Grad-CAM* (*Gradient-weighted Class Activation Mapping*) foi aplicado sobre a arquitetura *ResNet50* pós-treinamento com o objetivo analítico de identificar quais

regiões faciais exerceram maior influência vetorial nas decisões de classificação do modelo. A técnica atua interceptando os gradientes da última camada convolucional densa para gerar mapas de calor sobrepostos às matrizes de pixels originais. Nesse mapeamento de relevância, tonalidades cromáticas quentes como o vermelho e o laranja indicam zonas de máxima ativação e peso decisório, enquanto tonalidades frias como o azul e o verde denotam baixa contribuição para a predição linear da rede.

### 4.3.1 Padrões Observados nas Imagens *Fake*

Nas amostras classificadas como sintéticas, os mapas de calor gerados revelaram uma ativação espacial predominantemente concentrada na região central do rosto, abrangendo de forma densa o alinhamento dos olhos, do nariz e da cavidade bucal, conforme ilustrado na **Figura 7**.


**Figura 7 - Mapas de Calor do Grad-CAM em Amostras Classificadas como *Fake*.**
**Fonte: autoria própria, 2026.**

Esse padrão de convergência computacional mostra-se altamente consistente com o que a literatura forense descreve como as zonas de maior incidência de artefatos estruturais em imagens geradas por redes adversariais generativas (GANs) e modelos de difusão. Inconsistências microtexturais na pele, assimetrias pupilares sutis, descontinuidades de borda ao redor dos lábios e irregularidades na geometria do septo nasal são falhas biológicas frequentemente introduzidas pelos processos de síntese e reconstrução artificial.

A concentração da atenção do modelo nessas regiões específicas atesta que a *ResNet50* não memorizou padrões espúrios, ruídos de fundo ou correlações superficiais do *dataset*, mas

aprendeu a isolar assinaturas morfológicas genuinamente associadas ao processo de falsificação de rostos. Esse fenômeno confere estrita validade científica e robustez ao pipeline de engenharia desenvolvido.

### 4.3.2 Implicações para a Análise Forense Digital

Os resultados viabilizados pelo *Grad-CAM* oferecem uma contribuição metodológica que transcende a limitação operacional da classificação binária tradicional. Ao externalizar e tornar visível o *locus* exato onde o modelo identifica os indícios de fraude computacional, o sistema deixa de atuar como uma "caixa-preta" inacessível e consolida-se como um mecanismo auditável de suporte à tomada de decisões periciais, como esquematizado na correlação visual da **Figura 8**.

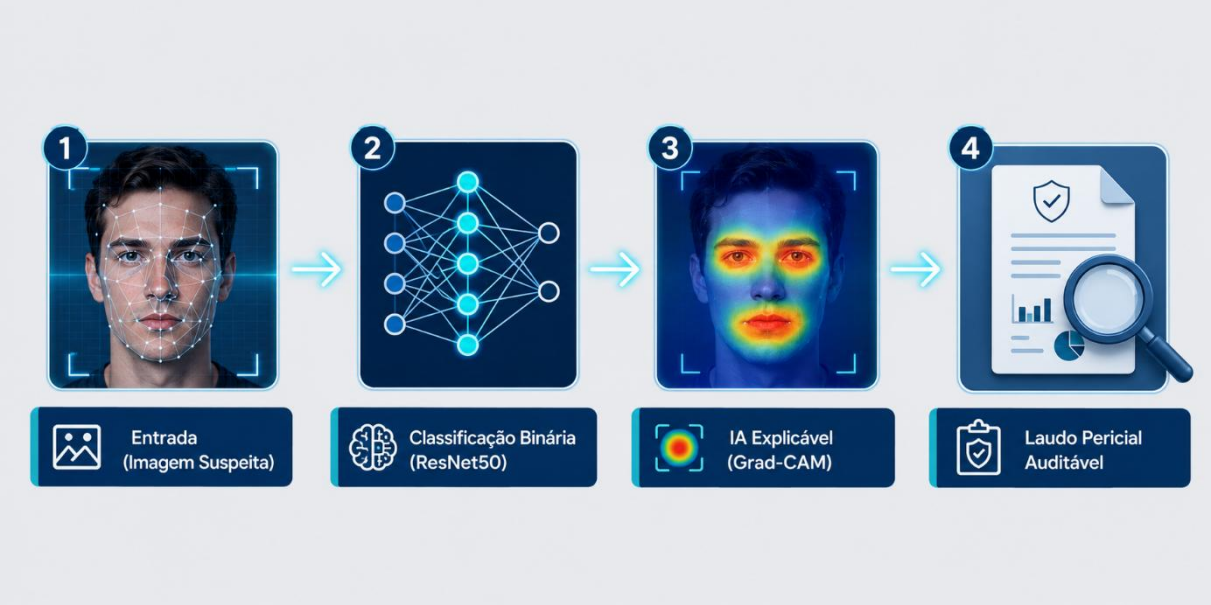


**Figura 8 - Pipeline de Análise Forense Digital com *Grad-CAM*: da entrada da imagem suspeita ao laudo pericial auditável. Fonte: autoria própria, 2026.**

Em contextos práticos de segurança cibernética tais como o combate à desinformação em redes digitais, sistemas de verificação de identidade ou perícia forense criminal, a capacidade técnica de apontar as regiões faciais específicas que sustentam a suspeita de falsificação representa um avanço disruptivo. Essa abordagem fornece a um perito humano a evidência visual necessária para fundamentar laudos técnicos, superando os detectores convencionais que emitem veredictos probabilísticos sem qualquer justificativa interpretável.

Dessa forma, a engenharia de explicabilidade cumpre um papel duplo e crucial neste trabalho: valida a coerência do aprendizado da rede perante o fenômeno investigado e

demonstra o potencial prático da IA Explicável (XAI) como ferramenta científica de apoio à segurança da informação.

# 5. DISCUSSÃO GERAL

Os resultados obtidos ao longo dos experimentos conduzidos neste trabalho permitem consolidar uma análise abrangente e integrada sobre o uso de Visão Computacional e XAI na identificação de características visuais associadas a imagens faciais geradas por IA.

O percurso metodológico adotado partindo de um experimento exploratório com a arquitetura *YOLOv8* até a aplicação do algoritmo *Grad-CAM* sobre a *ResNet50* revelou-se não apenas tecnicamente coerente, mas também cientificamente relevante. Cada etapa contribuiu de forma distinta para a construção do conhecimento: o experimento com o *YOLOv8* validou a viabilidade do ecossistema de dados e evidenciou os limites práticos de uma abordagem puramente classificatória, enquanto a *ResNet50* com *Transfer Learning* demonstrou que é possível alcançar alto desempenho discriminativo com custo computacional razoável e, sobretudo, com capacidade de interpretação profunda dos resultados.

## 5.1. Sobre o Desempenho dos Modelos

O *YOLOv8* apresentou métricas quantitativas ligeiramente superiores com *AUC-ROC* de 0,9934 frente a 0,9758 da *ResNet50*, porém ao custo de aproximadamente 5 horas de treinamento e sem oferecer mecanismos nativos de interpretabilidade espacial. A *ResNet50*, por sua vez, atingiu acurácia de 92,30% e *AUC-ROC* de 0,9758 em cerca de 1 hora e 30 minutos de treinamento, registrando um *F1-Score* de 92,29% e ausência de viés significativo entre as classes.

Essa relação entre desempenho estatístico e custo-benefício computacional reforça a adequação da *ResNet50* como ferramenta metodológica ideal para os objetivos deste trabalho. Em cenários reais de segurança cibernética, onde novas variantes de falsificações emergem diariamente e exigem o recondicionamento massivo e frequente de bancos de dados, a otimização do tempo de treinamento e o menor consumo de infraestrutura em nuvem tornam-se critérios de viabilidade técnica determinantes para o mercado.

É importante destacar que a diferença de aproximadamente 1,76% na *AUC-ROC* entre os dois modelos não representa uma superioridade determinante do *YOLOv8* ambas as arquiteturas operam em faixas de excelência discriminativa. O que de fato diferencia as

abordagens é a dimensão interpretativa: enquanto o *YOLOv8* entrega um veredito numérico isolado, a *ResNet50* combinada ao *Grad-CAM* entrega um veredito cientificamente explicado.

## 5.2. Sobre a Contribuição do Grad-CAM

A aplicação do *Grad-CAM* representa a contribuição central deste trabalho. Os mapas de calor gerados demonstraram que o modelo aprendeu a direcionar sua atenção para as regiões anatomicamente coerentes com os artefatos introduzidos por algoritmos de síntese facial especialmente o alinhamento dos olhos, do nariz e da boca nas imagens *Fake*, enquanto distribuiu a ativação de forma difusa e uniforme nas imagens *Real*, refletindo com precisão a ausência de manipulação localizada.

Esse resultado estabelece implicações práticas que transcendem a limitação operacional da classificação binária tradicional. Em contextos de análise forense digital, a capacidade técnica de indicar graficamente onde uma imagem apresenta indícios de manipulação e não apenas se ela é manipulada representa um avanço qualitativo significativo para a área.

Na engenharia forense, a ausência de uma justificativa visual anula a utilidade prática do modelo para subsidiar auditorias ou compor laudos periciais em juízo. Um sistema que apenas classifica oferece uma resposta probabilística; um sistema que classifica e explica, conforme estruturado neste *pipeline*, oferece uma evidência técnica auditável.

## 5.3. Sobre as Limitações e Perspectivas Futuras

Este trabalho reconhece limitações metodológicas que abrem caminho para investigações futuras. O conjunto de dados utilizado, embora criteriosamente curado, é composto majoritariamente por imagens geradas por redes adversariais generativas (GANs), o que pode limitar a generalização direta do modelo para *deepfakes* produzidos por arquiteturas generativas de outra natureza, como os modelos de difusão. Além disso, os experimentos foram conduzidos em um ambiente laboratorial controlado, com imagens em resolução fixa e padronizada, condição que pode diferir de cenários forenses reais do cotidiano, onde os arquivos frequentemente apresentam severas variações de qualidade, compressão de rede, iluminação e oclusão facial parcial.

Como perspectivas futuras para a continuidade da pesquisa, destacam-se:

- A expansão do *dataset* para incluir imagens geradas de forma heterogênea por diferentes arquiteturas de ponta;
- A investigação e acoplamento de técnicas complementares de XAI, como o *SHAP* (*SHapley Additive exPlanations*) e o *LIME* (*Local Interpretable Model-agnostic Explanations*), para fins de comparação com os mapas do *Grad-CAM*;
- A integração do *pipeline* de engenharia desenvolvido em uma ferramenta de análise forense modular e com interface amigável, tornando-a acessível a profissionais do direito e peritos não especialistas em Inteligência Artificial.

### 5.4. Considerações Finais da Seção

Em síntese, os resultados obtidos demonstram que a combinação entre Visão Computacional e Inteligência Artificial Explicável não apenas se mostra tecnicamente viável para a detecção de mídias sintéticas, mas oferece uma camada de transparência indispensável para aplicações modernas que exigem auditabilidade e justificativa de decisões automatizadas.

A *ResNet50*, utilizada como instrumento metodológico de engenharia, e o *Grad-CAM*, atuando como lente interpretativa pericial, compõem juntos um *pipeline* que vai muito além da detecção ele investiga, revela e comunica de forma explícita as marcas invisíveis deixadas pelos algoritmos de IA na síntese de rostos humanos.

## 6. CONCLUSÃO

O presente trabalho investigou o uso de Visão Computacional e Inteligência Artificial Explicável na identificação de características visuais associadas a imagens faciais geradas por algoritmos de síntese artificial, com foco na aplicação do *Grad-CAM* como mecanismo de interpretabilidade espacial sobre um modelo *ResNet50* treinado via *Transfer Learning*.

O percurso percorrido demonstrou que é possível construir um pipeline técnico robusto, interpretável e cientificamente fundamentado para a análise de *deepfakes* faciais, utilizando recursos computacionais acessíveis e ferramentas consolidadas na literatura de Visão Computacional.

O experimento exploratório com *YOLOv8* cumpriu seu papel de validar o ecossistema de dados e o *pipeline* de detecção, registrando AUC-ROC de 0,9934 e AP de 0,9939 métricas expressivas que, no entanto, não foram acompanhadas de capacidade interpretativa. Essa limitação fundamentou a escolha da *ResNet50* como arquitetura principal, não por

superioridade métrica isolada, mas por sua compatibilidade nativa com técnicas de XAI e por oferecer uma relação custo-benefício computacional superior em cenários de atualização contínua de modelos.

A *ResNet50* treinada sobre o *dataset* Teste_TCC_10k atingiu acurácia de 92,30%, AUC-ROC de 0,9758 e F1-*Score* de 92,29%, com ausência de viés significativo entre as classes *real* e *fake*. Mais do que os números, o resultado determinante do trabalho foi obtido pela aplicação do *Grad-CAM*: os mapas de calor gerados confirmaram que o modelo direcionou sua atenção às regiões anatomicamente coerentes com os artefatos de síntese facial olhos, nariz e boca nas imagens *fake*, enquanto distribuiu a ativação de forma difusa nas imagens *real*. Esse padrão demonstra que o modelo aprendeu representações genuínas do fenômeno investigado, e não decorou padrões espúrios do *dataset*.

A contribuição central deste trabalho reside, portanto, na demonstração de que a IA Explicável não é um recurso acessório em sistemas de detecção de *deepfakes* ela é um componente estrutural indispensável para qualquer aplicação que aspire à auditabilidade técnica e à validade pericial. Um classificador que apenas emite um veredito probabilístico é insuficiente para subsidiar laudos, auditorias ou decisões jurídicas. Um *pipeline* que classifica, localiza e comunica visualmente os indícios de manipulação representa um avanço qualitativo que aproxima a Inteligência Artificial das demandas reais da engenharia forense.

As limitações identificadas como a composição do *dataset* majoritariamente por imagens geradas por *GANs* e a padronização das condições experimentais não comprometem a validade dos resultados obtidos, mas apontam direções concretas para a continuidade da pesquisa. A expansão para *datasets* heterogêneos, o acoplamento de técnicas complementares de XAI como SHAP e LIME, e o desenvolvimento de uma ferramenta forense com interface acessível a profissionais não especialistas em IA constituem desdobramentos naturais e promissores deste trabalho.

Por fim, este trabalho reafirma que a detecção de *deepfakes* não se resolve apenas com modelos mais precisos; resolve-se com modelos mais transparentes. Em um cenário em que imagens sintéticas ameaçam a integridade da informação, da identidade e da justiça, tornar a Inteligência Artificial explicável não é uma escolha técnica: é uma responsabilidade científica.

## REFERÊNCIAS

BHAT, N. A.; GIRI, K. J. DeepFake detection in images and videos: a survey on models, datasets, and evaluation metrics. *Multimedia Tools and Applications*, v. 85, n. 233, 2026. Disponível em: https://doi.org/10.1007/s11042-026-21365-9.

BUDATI, J. L. Narayana; JADAM, A. Bharathi; MALLEBOYINA, R. Explainable AI for Deepfake Detection: A Grad-CAM Approach to Video Forensics. In: *INTERNATIONAL CONFERENCE FOR EMERGING TECHNOLOGY (INCET)*, 6., 2025, Belgaum, India. Proceedings [...]. Belgaum: IEEE, 2025. p. 1-7. DOI: 10.1109/INCET64471.2025.11140952.

CANTÃO, Camily; MEDEIROS, Iago. Deepfakes além do algoritmo: uma revisão sobre indícios visuais identificáveis a olho nu na era da inteligência artificial generativa. In: *ESCOLA REGIONAL DE INFORMÁTICA NORTE 2 (ERIN 2)*, 18., 2025, Macapá/AP. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025. p. 61-66. DOI: https://doi.org/10.5753/erin.2025.16056.

DOLHANSKY, Brian; HOWES, Russ; PFLAUM, Ben; BARAM, Nicole; FERRER, Cristian Canton. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv preprint*, 2020.

EDWARDS, Peter; NEBEL, Jean-Christophe; GREENHILL, Darrel; LIANG, Xing. A Review of Deepfake Techniques: Architecture, Detection, and Datasets. *IEEE Access*, 2024.

GOODFELLOW, Ian J.; POUGET-ABADIE, Jean; MIRZA, Mehdi; XU, Bing; WARDE-FARLEY, David; OZAIR, Sherjil; COURVILLE, Aaron; BENGIO, Yoshua. Generative Adversarial Nets. *Advances in Neural Information Processing Systems*, p. 2672-2680, 2014. Disponível em: https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

HARMA, Vaishali; SINGH, Neetu; PRASAD, Rahul. Assessing YOLOv8 as a Classifier for Detection of Synthetically Generated Facial Imagery. In: *INTERNATIONAL CONFERENCE ON SIGNAL PROCESSING AND INTEGRATED NETWORKS (SPIN)*, 11., 2024, Noida, India. Proceedings [...]. Noida: IEEE, 2024. p. 99-104. DOI: 10.1109/SPIN60856.2024.10511616.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep residual learning for image recognition. In: *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, p. 770-778, 2016. Disponível em: https://ieeexplore.ieee.org/document/7780459.

KHAN, Abdullah Ayub; LAGHARI, Asif Ali; INAM, Syed Azeem; ULLAH, Sajid; SHAHZAD, Muhammad; SYED, Darakhshan. A survey on multimedia-enabled deepfake

detection. *Intelligent Decision Technologies*, 2025. Disponível em: https://link.springer.com/article/10.1007/s10791-025-09550-0.

KUMAR, Niteesh; P., Pranav; NIRNEY, Vishal; V., Geetha. Deepfake Image Detection using CNNs and Transfer Learning. S.l.: s.n., 2020.

KUMAR, S. S. *et al*. Explainable Deepfake Detection: Integrating Grad-CAM with CNN-RNN Architecture for Real-Time Forensic Analysis. *TechRxiv*, 2 dez. 2025. DOI: 10.36227/techrxiv.176471850.05490312.

KROIß, Lukas; RESCHKE, Johannes. Deepfake Detection of Face Images based on a CNN. *arXiv*, 2025. Disponível em: https://arxiv.org/abs/2503.11389.

LI, Yuezun; YANG, Xin; SUN, Pu; QI, Honggang; LYU, Siwei. Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. p. 3207-3216, 2020.

MANDAL, Unmesh; SETUA, Sanjit Kumar; SARMA, Sayan Sen. Deepfakes and beyond in the era of AI-generated disinformation: a systematic literature review of emerging technologies, challenges, and policy recommendations. *Journal of Visual Communication and Image Representation*, v. 117, p. 104748, 2026. DOI: 10.1016/j.jvcir.2026.104748. Disponível em: https://www.sciencedirect.com/science/article/pii/S104732032600043X.

MAWA, Jannatul; KABIR, Md. Humayun. Study on Deepfake Face Detection using Transfer Learning Approach. *International Journal of Computer Applications*, v. 186, n. 37, p. 44-48, Aug. 2024. DOI: 10.5120/ijca2024923948.

MOURA, Camila Steffane Fernandes Teixeira de. *Detecção de deepfakes a partir de técnicas de visão computacional e aprendizado de máquina*. 2021. Dissertação (Mestrado em Ciência da Computação) - Instituto de Computação, Universidade Estadual de Campinas, Campinas, 2021.

MUSTAK, M. *et al*. Deepfakes and beyond in the era of AI-generated disinformation: A systematic literature review. *Computers in Human Behavior Reports*, v. 16, art. 100538, 2024. DOI: 10.1016/j.chbr.2024.100538.

PR, B.; BIJU, A. K.; M, B.; SHYAM, B.; ADITHYA, C. A. DeepFake Detection Using CNN Models with Explainable AI Techniques. In: *INTERNATIONAL CONFERENCE ON ADVANCES IN COMPUTING, COMMUNICATION, EMBEDDED AND SECURE SYSTEMS (ACCESS)*, 4., 2025, Ernakulam, India. Proceedings [...]. Ernakulam: IEEE, 2025. p. 341-349. DOI: 10.1109/ACCESS65134.2025.11135663.

RÖSSLER, Andreas; COZZOLINO, Davide; VERDOLIVA, Luisa; RIESS, Christian; THIES, Justus; NIEBNER, Matthias. FaceForensics++: Learning to detect manipulated facial images. In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. p. 1-11, 2019.

SAMAL, Debasish; NAGPAL, Dimple; AGRAWAL, Prateek; MADAAN, Vishu. DeFakeNet: A ResNet50V2-Based Deep Learning Model for Deepfake Detection and Classification. *Journal of Innovative Image Processing*, v. 7, n. 4, p. 1356-1373, nov. 2025. DOI: 10.36548/jiip.2025.4.015.

SATHE, Pranav; SABANE, Varad; UNDALE, Chaitanya; UTTARKAR, Atharva; CHAVHAN, Vaibhav; SABLE, Nilesh P.; YENKIKAR, Anuradha. Deepfake Image Detection Using YOLOv8. In: *IEEE PUNE SECTION INTERNATIONAL CONFERENCE (PuneCon)*, 2024, Pune, India. Proceedings [...]. Pune: IEEE, 2024. p. 1-5. DOI: 10.1109/PuneCon63413.2024.10895706.

SHARMA, J.; SHARMA, S.; KUMAR, V.; HUSSEIN, H. S.; ALSHAZLY, H. Deepfakes classification of faces using convolutional neural networks. *Traitement du Signal*, v. 39, n. 3, p. 1027-1037, 2022.

SHARMA, S. South korean woman loses £40k in elon musk romance scam involving deepfake video. *The Independent*, 2024. Disponível em: https://www.independent.co.uk/asia/east-asia/elon-musk-romance-scam-dupes-south-korean-b2533764.html.

SOLANKE, A. A.; BIASIOTTI, M. A. Artificial Intelligence in Digital Forensics: Evaluating, Standardizing and Optimizing Digital Evidence Mining Techniques. *Künstliche Intelligenz*, v. 36, n. 2, p. 143-161, 2022. DOI: 10.1007/s13218-022-00763-9.

SOUDY, A. H.; SAYED, O.; TAG-ELSER, H. *et al*. Deepfake detection using convolutional vision transformers and convolutional neural networks. *Neural Computing and Applications*, v. 36, p. 19759-19775, 2024. DOI: 10.1007/s00521-024-10181-7.