

UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO"
FACULDADE DE CIÊNCIAS - CAMPUS BAURU
DEPARTAMENTO DE COMPUTAÇÃO
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

MIGUEL CESAR CORRÊA

**DETECÇÃO DE IMAGENS *DEEPPFAKE*: UM ESTUDO
COMPARATIVO**

BAURU
Novembro/2023

MIGUEL CESAR CORRÊA

DETECÇÃO DE IMAGENS *DEEPPFAKE*: UM ESTUDO COMPARATIVO

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista “Júlio de Mesquita Filho”, Faculdade de Ciências, Campus Bauru.

Orientador: Prof. Dr. Leandro Aparecido Passos Junior

BAURU
Novembro/2023

C824d	Corrêa, Miguel Cesar Detecção de Deepfakes: Um Estudo Comparativo / Miguel Cesar Corrêa. -- Bauru, 2023 29 p. : il., tabs. Trabalho de conclusão de curso (Bacharelado - Ciência da Computação) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências, Bauru Orientador: Leandro Aparecido Passos Junior 1. Aprendizado de Máquina. 2. Detecção de Deepfakes. 3. Inteligência Artificial. I. Título.
-------	--

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências, Bauru. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

Miguel Cesar Corrêa

Detecção de Imagens *deepfake*: um estudo comparativo

Trabalho de Conclusão de Curso do Curso de Bacharelado em Ciência da Computação da Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências, Campus Bauru.

Banca Examinadora

Prof. Dr. Leandro Aparecido Passos Junior

Orientador

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Departamento de Ciência da Computação

Professora Dra. Simone Prado

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Departamento de Ciência da Computação

Professor Dr. Douglas Rodrigues

Universidade Estadual Paulista "Júlio de Mesquita Filho"

Faculdade de Ciências

Departamento de Ciência da Computação

Bauru, _____ de _____ de _____.

Resumo

Com o avanço acelerado da inteligência artificial, os *deepfakes* - vídeos ou imagens manipulados de maneira convincente - emergiram como uma preocupação significativa na era digital. Essas falsificações hiper-realistas têm o potencial de enganar indivíduos, disseminar desinformação e comprometer a autenticidade da informação, representando uma ameaça real à segurança digital e à integridade informativa. Este trabalho aborda a necessidade de desenvolver métodos eficazes para a detecção de *deepfakes*, uma ferramenta essencial na luta contra a desinformação, apresentando os conceitos fundamentais da área. Este trabalho comparou três métodos estado-da-arte de detecção de imagens falsas: Detector de Falsificações com Transformador de Consistência de Identidade, CORE e o Modelo de Detecção de *Deepfake* Ignorante de ID, apresentando resultados satisfatórios, com o método de detecção ignorante de ID apresentando o melhor desempenho.

Palavras-chave: Inteligência Artificial; Deep Learning; Deepfake; Detecção de Deepfake.

Abstract

With the rapid advancement of artificial intelligence, deepfakes – convincingly manipulated videos or images – have emerged as a significant concern in the digital age. These hyper-realistic fakes have the potential to deceive individuals, spread disinformation and compromise the authenticity of information, posing a real threat to digital security and informational integrity. This work addresses the need to develop effective methods for detecting deepfakes, an essential tool in the fight against disinformation, presenting the fundamental concepts of the area. This work compared three state-of-the-art methods for detecting false images: Forgery Detector with Identity Consistency Transformator, CORE and the ID-Unaware Deepfake Detection Model, presenting satisfactory results, with the ID-unaware model having the best performance.

Keywords: Artificial Intelligence; Deep Learning; Deepfake; Deepfake Detection.

Lista de figuras

Figura 1 – Website <i>Deepfakes Web</i>	11
Figura 2 – Imagem <i>deepfake</i> circulada nas redes sociais.	12
Figura 3 – Troca de face aplicada em uma imagem.	14
Figura 4 – Troca de atributos aplicada em uma imagem.	15
Figura 5 – Ventriloquia aplicada em uma imagem.	16
Figura 6 – Exemplo de imagens presentes no conjunto de dados <i>Faceforensics++</i>	19
Figura 7 – Arquitetura do modelo de Dong (2022).	20
Figura 8 – Arquitetura do modelo CORE.	21
Figura 9 – Arquitetura do modelo CORE.	22
Figura 10 – Matriz de Confusão do modelo de DONG et al.(2022).	24
Figura 11 – Matriz de Confusão do modelo de NI et al.(2022).	25
Figura 12 – Matriz de Confusão do modelo de DONG et al. (2023)	26

Lista de tabelas

Tabela 1 – Resultados obtidos para os modelos testados	24
--	----

Sumário

1	INTRODUÇÃO	10
1.1	Problema	10
1.2	Justificativa	11
1.3	Objetivos	11
1.3.1	Objetivo Geral	11
1.3.2	Objetivos Específicos	12
1.4	Organização	12
2	FUNDAMENTAÇÃO TEÓRICA	13
2.1	Inteligência Artificial	13
2.2	Redes Neurais Artificiais	13
2.2.1	Aprendizado Profundo	13
2.3	<i>Deepfakes</i>	14
2.3.1	Tipos de <i>Deepfakes</i>	14
2.3.1.1	Troca de Face (<i>face-swap</i>)	14
2.3.1.2	Troca de Atributos Faciais	15
2.3.1.3	Ventriloquia (<i>puppet-master</i>)	15
2.4	Métodos de Geração de <i>Deepfakes</i>	15
2.4.1	<i>FaceSwap</i>	15
2.4.2	<i>Face2Face</i>	16
2.4.3	<i>Deepfake</i>	16
2.4.4	<i>NeuralTextures</i>	16
3	METODOLOGIA	18
3.1	Ferramentas	18
3.1.1	<i>Python</i>	18
3.1.2	<i>Hardware</i> Utilizado	19
3.2	Conjunto de Dados	19
3.3	Métodos de Detecção de Falsificação	20
3.3.1	Detectando Falsificações com Transformador de Consistência de Identidade	20
3.3.2	CORE: Aprendizado de Representação Consistente para Detecção de Falsificação de Rosto	21
3.3.3	Modelo de Detecção de Deepfake Ignorante de ID	21
3.4	Métricas	22
4	RESULTADOS	24

4.1	Resultados Obtidos	24
4.2	Discussão	25
5	CONCLUSÃO	27
	REFERÊNCIAS	28

1 Introdução

O surgimento de tecnologia *deepfake* trouxe consigo novos desafios no campo de forense multimídia. *Deepfakes* são conteúdos multimídia criados com a ajuda de algoritmos de Inteligência Artificial (IA) que imitam pessoas reais e eventos (CHAWLA, 2019). Essa tecnologia se provou útil na criação de fraudes convincentes que podem enganar o olho humano e ter um impacto significativo na sociedade. No entanto, também apresenta graves ameaças à segurança, já que *deepfakes* podem ser usados para propagar desinformação e propaganda política (WESTERLUND, 2019).

Esse trabalho tem como objetivo realizar um estudo comparativo sobre técnicas de detecção de *deepfake* em imagens. O objetivo principal é desenvolver uma compreensão profunda do estado da arte atual em técnicas de detecção de *deepfake* e explorar sua eficácia no reconhecimento de imagens forjadas. O estudo também examinará os desafios e limitações dos métodos existentes e identificará áreas para pesquisas futuras, uma pesquisa necessária observando a facilidade de geração de conteúdos do tipo. (PASSOS et al., 2023)

A pesquisa proposta tem implicações significativas para várias áreas, incluindo cibersegurança, forense de imagens e jornalismo. Os resultados do estudo podem ser usados para o desenvolvimento de técnicas eficazes de detecção de *deepfake* em imagens, reduzindo assim os riscos associados à essa tecnologia.

1.1 Problema

O avanço tecnológico na criação de *deepfakes* tem levantado preocupações sobre a autenticidade e confiabilidade de imagens em várias áreas, como segurança, justiça, jornalismo e entretenimento (BORGES; MARTINS; CALADO, 2019). Detectar *deepfakes* é um problema complexo, pois é uma área em constante evolução na qual as técnicas de manipulação estão se tornando cada vez mais sofisticadas.

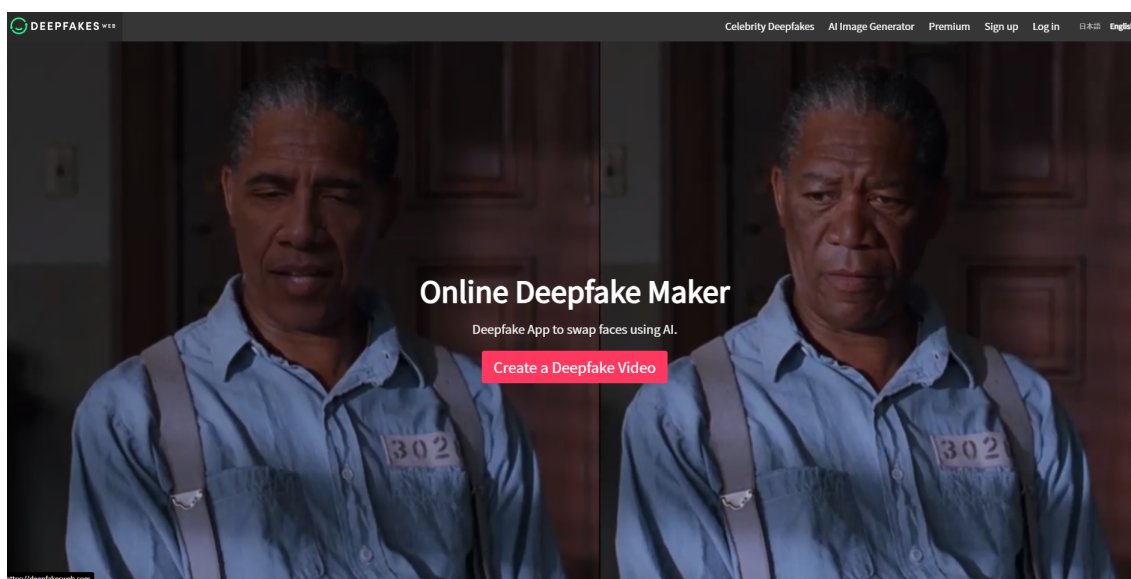
O problema central abordado neste trabalho é: quais técnicas de detecção de *deepfake* em imagens baseadas em aprendizado profundo (*deep learning*) são mais eficazes e eficientes? Essa monografia se concentrará apenas em imagens *deepfake*, e não outros tipos de mídia, como vídeos ou áudios, expandindo o trabalho inicial desenvolvido por Passos et al. (2023), abordando técnicas mais modernas relacionadas ao mesmo problema. É uma pesquisa viável, visto que a detecção de *deepfakes* é um problema em ascensão que tem atraído interesse em diversas áreas. Além disso, existem diversas técnicas de detecção baseadas em aprendizado profundo em desenvolvimento que podem ser analisadas e comparadas. A relevância dessa pesquisa é clara, visto os usos maliciosos em que essa tecnologia pode ser empregada.

1.2 Justificativa

Esse trabalho tem como objetivo abordar um problema atual e relevante para a sociedade. Com o avanço da tecnologia de geração de imagens falsas, qualquer pessoa com acesso a um computador é capaz de gerar arquivos de mídia indistinguíveis de imagens reais. (WESTERLUND, 2019).

A Figura 1 é retirada de um *website* que permite a geração de conteúdos *deepfake* diretamente de navegadores *web*.

Figura 1 – Website Deepfakes Web



Fonte: Adaptado de *Tecnoblog*, (2023).

No mês de abril de 2023, a Figura 2 foi circulada nas redes sociais e em veículos de notícia. A imagem mostra o Papa Francisco com um casaco atípico. Após averiguação, foi constatado que se tratava de uma imagem falsa. Levando isso em consideração, um estudo comparativo das técnicas atuais de detecção de imagens forjadas pode ajudar a apontar os melhores métodos atuais para detecção de imagens falsas e orientar o desenvolvimento de técnicas de detecção mais eficazes e eficientes baseadas em aprendizado profundo para combater os *deepfakes* em imagens, indicando as abordagens mais eficientes e com melhor resultado.

1.3 Objetivos

O objetivo geral e objetivos específicos deste trabalho estão apresentados a seguir.

1.3.1 Objetivo Geral

Este trabalho teve como objetivo comparar métodos de detecção de *deepfake* em imagens utilizando técnicas de *deep learning*.

Figura 2 – Imagem *deepfake* circulada nas redes sociais.



Fonte: CBS News, (2023).

1.3.2 Objetivos Específicos

- Realizar uma revisão bibliográfica sobre *deep learning* e as técnicas de aprendizado de máquina utilizadas para detecção de *deepfakes*;
- Identificar os conjuntos de dados de *deepfake* disponíveis e selecionar os conjuntos a serem utilizados;
- Implementar os métodos de detecção de *deepfake* selecionados;
- Treinar os métodos escolhidos utilizando o conjunto de dados selecionado;
- Comparar e analisar os resultados obtidos pelos diferentes métodos de detecção;

1.4 Organização

Este trabalho está organizado desta maneira:

- **Capítulo 2:** Fundamentação teórica apresentando os conceitos de IA, Aprendizado de Máquina, Redes Neurais Artificiais. O capítulo também introduz o conceito de *deepfakes*, suas categorias e principais métodos de geração.
- **Capítulo 3:** Apresenta as ferramentas utilizadas para o desenvolvimento deste trabalho, o conjunto de dados utilizado e os métodos testados.
- **Capítulo 4:** Expõe os resultados obtidos e a discussão destes.
- **Capítulo 5:** Apresenta a conclusão do trabalho e caminhos para serem explorados futuramente.

2 Fundamentação Teórica

Neste capítulo, apresenta-se a base teórica na qual este trabalho se apoia, abordando os tópicos de estudo.

2.1 Inteligência Artificial

A inteligência artificial (IA) é um campo da ciência da computação dedicado à criação de sistemas capazes de realizar tarefas que, tradicionalmente, exigiriam a inteligência humana. Isso inclui o processamento de linguagem natural, reconhecimento de padrões, aprendizado de máquina e raciocínio lógico (BODEN, 1996). Uma sub-área da IA, conhecida como Aprendizado de Máquinas, tem como seu objetivo desenvolver algoritmos que permitem que máquinas aprendam a partir de dados, identifiquem padrões, tomem decisões e melhorem suas capacidades com o passar do tempo, sem a necessidade de intervenção humana. Desde seu surgimento, a IA e o aprendizado de máquinas transformaram diversos setores, desde a automação industrial até a assistência pessoal virtual, e continua a expandir suas fronteiras.

2.2 Redes Neurais Artificiais

As redes neurais artificiais são algoritmos para atividades cognitivas, como aprendizado e otimização, que são baseadas no conceito do sistema nervoso de neurônios. Uma rede neural artificial, de maneira geral, recebe dados em seus nós (ou neurônios) de entrada, que são processados e propagados pelos nós das camadas seguintes, resultando em uma saída. Essa saída é então comparada com um resultado esperado e, com isso, é encontrado um valor de perda (*loss*), que é utilizado para o ajuste dos pesos das conexões. Este processo se repete até que se atinja um número pré-determinado de repetições ou alguma determinada métrica seja atingida. (MÜLLER; REINHARDT; STRICKLAND, 1995)

2.2.1 Aprendizado Profundo

O aprendizado profundo, ou *deep learning*, é uma subcategoria avançada do aprendizado de máquina, caracterizado pelo uso de redes neurais artificiais com várias camadas ocultas que simulam o funcionamento do cérebro humano para processar dados e criar padrões de reconhecimento em grande escala. Essas redes neurais são compostas por nós, comparados aos neurônios, interligados de forma a propagar informações e realizar transformações nos dados de entrada. Essa técnica permite que a máquina melhore seu desempenho na realização de tarefas como visão computacional, processamento de linguagem natural e análise preditiva, ao aprender com grandes volumes de dados. O aprendizado profundo revolucionou a capacidade

das máquinas de interpretar o mundo, facilitando avanços significativos em campos como reconhecimento facial, tradução automática e condução autônoma.

2.3 *Deepfakes*

Deepfakes são resultados de uma técnica sofisticada de síntese de imagem e vídeo baseada em IA, especificamente no aprendizado profundo, que permite a criação ou manipulação de conteúdo audiovisual com alto grau de realismo. Utilizando redes neurais como as GANs (*Generative Adversarial Networks*), *deepfakes* podem substituir rostos, replicar vozes e modificar expressões faciais em vídeos, gerando representações quase indistinguíveis da realidade. Eles se dividem em vários tipos, incluindo troca de rostos, onde a face de uma pessoa é sobreposta à de outra; simulação de voz, reproduzindo a fala de alguém; e manipulação de ações. Enquanto têm aplicações legítimas em entretenimento e educação, *deepfakes* também levantam preocupações éticas e legais, particularmente em relação à desinformação, à privacidade e à segurança, dada a dificuldade em discernir entre o que é real e o que é fabricado.

A seguir, são apresentadas as principais categorias de falsificações.

2.3.1 Tipos de *Deepfakes*

Deepfakes podem ser categorizadas, de maneira geral, em três principais categorias:

2.3.1.1 Troca de Face (*face-swap*)

Este é o método mais comum de manipulação de imagens. Nesta técnica, a face de uma pessoa é sobreposta em outra pessoa, mantendo as expressões originais. A Figura 3 demonstra o resultado da troca de face aplicada na face do ex-presidente americano Bill Clinton, que teve sua face substituída pelo rosto do ator Nicolas Cage.

Figura 3 – Troca de face aplicada em uma imagem.



Fonte: Adaptado de Korshunova et al. (2017).

2.3.1.2 Troca de Atributos Faciais

A técnica de substituição de atributos faciais consiste na troca de atributos de uma foto original, como a cor do cabelo, idade ou gênero. A Figura 4 apresenta os resultados de diversos métodos de troca de atributos faciais.

Figura 4 – Troca de atributos aplicada em uma imagem.



Fonte: Adaptado de Choi et al. (2017).

2.3.1.3 Ventriloquia (*puppet-master*)

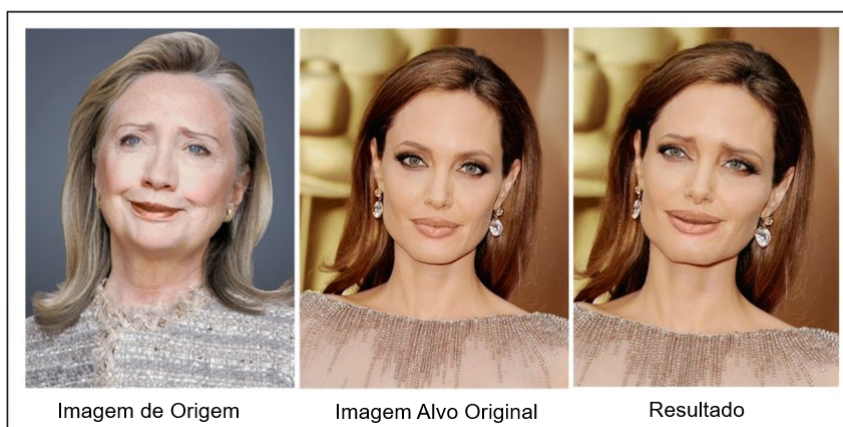
Os *deepfakes* de ventriloquia manipulam a imagem final para seguir os movimentos da face, cabeça e corpo da imagem original. A Figura 5 exemplifica este método, onde as expressões faciais da Hillary Clinton foram aplicadas em uma imagem da Angelina Jolie. É uma técnica com utilidades práticas nos campos de animação e jogos eletrônicos, mas também é utilizada para a geração de desinformação política.

2.4 Métodos de Geração de *Deepfakes*

2.4.1 FaceSwap

Este método é uma abordagem gráfica para a transferência de uma face de um vídeo origem para o vídeo destino, se encaixando na categoria de falsificações "troca de face". Ele detecta pontos de referência faciais para extrair a região facial, e com isso, gera um modelo tridimensional da face utilizando *blendshapes*, que são expressões faciais e movimentos pré-determinados. Com o modelo pronto, a face do vídeo de origem é então projetada no vídeo destino, utilizando os pontos de referência para o alinhamento. Por fim, correções de cor são

Figura 5 – Ventriloquia aplicada em uma imagem.



Fonte: Adaptado de Masood et al. (2022).

aplicadas no resultado final, de modo que a nova face tenha a mesma tonalidade que os *pixels* ao seu redor. (RÖSSLER et al., 2019)

2.4.2 Face2Face

Face2Face é um método de ventriloquia. Este modelo combina reconstrução tridimensional com técnicas de renderização baseadas em imagem. Primeiramente, o modelo constrói um modelo tridimensional detalhado da face objetivo, capturando suas expressões e iluminação, e em seguida, transfere as expressões do vídeo de origem para o vídeo final mantendo a identidade do indivíduo no vídeo objetivo. Isso é feito aplicando as expressões capturadas do vídeo de origem para o modelo tridimensional do vídeo objetivo. (RÖSSLER et al., 2019)

2.4.3 Deepfake

Deepfake é outro método de troca de face. O método utiliza um par de *autoencoders* com um *encoder* compartilhado, onde cada *autoencoder* é treinado para reconstruir as faces de origem e alvo. O processo de troca de face envolve a aplicação do *encoder* e *decoder* da face de origem na face alvo, efetivamente as trocando. A saída deste processo é então mesclada com o restante da imagem. Este método requer o treinamento para cada par de vídeos, resultando em um processo que consome mais tempo, mas efetivo em criar trocas de face realistas. (RÖSSLER et al., 2019)

2.4.4 NeuralTextures

Este método é uma abordagem avançada de ventriloquismo. O modelo foca em aprender a textura da pessoa alvo, que é usada em conjunto de uma rede de renderização. Essa abordagem, particularmente, modifica as expressões faciais da boca enquanto mantém a região dos olhos inalterada, o que resulta em maior realismo no produto final. Este método necessita

da geometria da face capturada para o treinamento. É conhecido por resultados altamente detalhados e realísticos. ([RÖSSLER et al., 2019](#))

3 Metodologia

A princípio, para o desenvolvimento do trabalho, foi feito um levantamento bibliográfico abrangente, explorando conceitos de IA, aprendizado profundo e o surgimento e evolução de *deepfakes*. Em seguida, foram determinados os algoritmos a serem utilizados no projeto.

Feito isso, foi selecionado o conjunto de dados a ser utilizado no projeto. Os algoritmos foram então aplicados sobre esse conjunto de dados e os resultados obtidos foram analisados e apresentados.

3.1 Ferramentas

Nesta seção estão descritas as ferramentas utilizadas para o desenvolvimento do trabalho.

3.1.1 *Python*

Python é uma linguagem de programação de alto nível, interpretada e de propósito geral, conhecida pela sua simplicidade. *Python* suporta diversos paradigmas de programação, incluindo programação orientada a objetos, imperativa, funcional e procedural.

Pela sua clareza e facilidade de aprendizado, *Python* tornou-se uma das linguagens mais populares para iniciantes, ao mesmo tempo que é robusta o suficiente para ser utilizada em algumas das maiores plataformas do mundo e em áreas complexas. No passado, diversas linguagens foram utilizadas no desenvolvimento e expansão da área de IA, mas *Python* tem visto um grande aumento de popularidade para essas aplicações, assim como para outras áreas como análise de dados, e desenvolvimento web ([RASCHKA; PATTERSON; NOLET, 2020](#)).

Para realização do trabalho, as seguintes bibliotecas do *Python* foram utilizadas:

- *Pytorch*
- *OpenCV*
- *Scipy*
- *Numpy*
- *Timm*

3.1.2 Hardware Utilizado

O *hardware* utilizado para o desenvolvimento do trabalho, com o treinamento e execução dos testes foram:

- Processador Intel I5-12400F com 12 *threads* e frequência de *clock* de 2.5 GHz;
- Placa de Vídeo GeForce com 12 Gb e VRam GDDR5;
- 16 Gb de memória RAM DDR4 com 3200 MHz de frequência;
- Sistema Operacional *Arch Linux*;

3.2 Conjunto de Dados

O conjunto de dados *FaceForensics++* é uma base de dados com mais de 1.8 milhão imagens manipuladas. Essas imagens foram geradas através de 1000 vídeos retirados do *YouTube*, que foram manipulados com os métodos *Deepfake*, *Faceswap*, *Face2Face* e *NeuralTextures*. O conjunto também fornece *benchmarks* de performance de diversos métodos de detecção aplicados sobre a base de dados (RÖSSLER et al., 2019). Para a realização deste trabalho, foram extraídos 24 mil *frames* do conjunto de dados, dos quais 16800 foram utilizados para o treinamento e 7200 para teste.

Figura 6 – Exemplo de imagens presentes no conjunto de dados *Faceforensics++*.



Fonte: Rössler et al. (2019).

3.3 Métodos de Detecção de Falsificação

Para o desenvolvimento do estudo, o processo de seleção dos algoritmos de detecção foi feito com base em critérios como as tecnologias empregadas, a recenticidade da publicação e a relevância no meio acadêmico e profissional, evidenciada pelo número de citações. Essa abordagem levou a escolha de algoritmos que estão no fronte da pesquisa em IA. Dessa forma, o trabalho buscou não apenas aplicar o que há de mais eficaz em termos de detecção, mas também contribuir de forma significativa com aplicações que reflitam o estado da arte e promovam avanços no campo dos *deepfakes*. A seguir estão descritos os métodos selecionados.

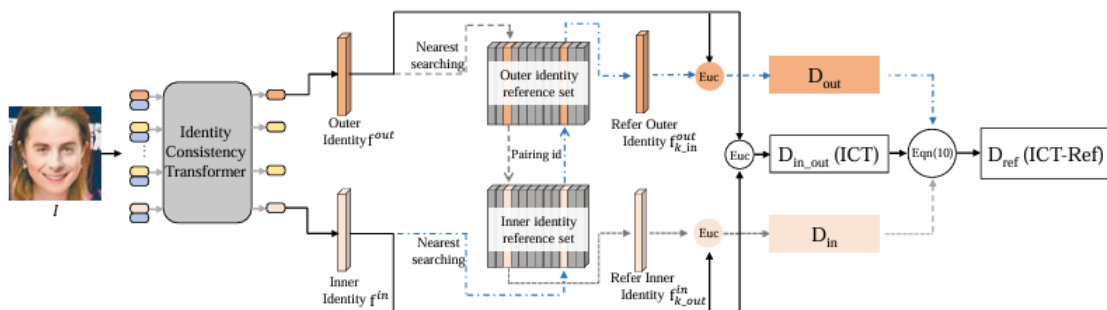
3.3.1 Detectando Falsificações com Transformador de Consistência de Identidade

Este método propõe uma nova técnica de detecção de falsificações, chamado "Transformador de Consistência de Identidade" (*Identity Consistency Transformer*), se concentrando em informações de identidade. O método funciona procurando inconsistências entre as regiões internas e externas de um rosto, e pode ser aprimorado com informações de identidades adicionais, o que o torna ideal para detectar falsificações relacionadas a celebridades. (DONG et al., 2022).

O modelo foi originalmente treinado no conjunto de dados *MS-Celeb-1M* (GUO et al., 2016), que consiste de 10 milhões de imagens de faces, com mais de 1 milhão de identidades. O método gerou novas imagens trocando as regiões internas de pares de faces contidas no conjunto de dados, gerando um conjunto de testes diversos sem a necessidade de *deepfakes* reais. O modelo obteve uma média de desempenho de 94,36%.

A Figura 7 demonstra a arquitetura do modelo proposto. A implementação deste método está disponível em (ICT..., 2022).

Figura 7 – Arquitetura do modelo de Dong (2022).



Fonte: Dong et al. (2022).

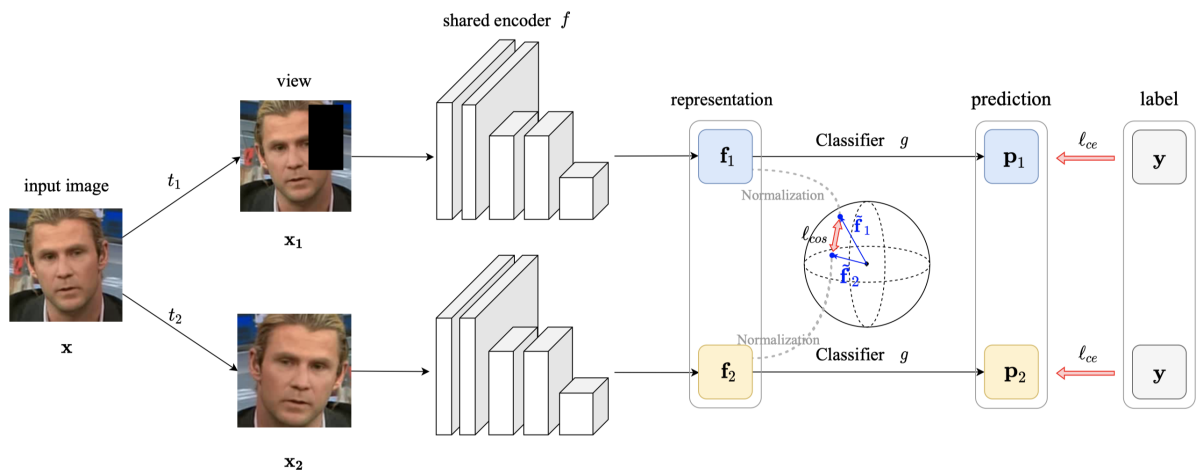
3.3.2 CORE: Aprendizado de Representação Consistente para Detecção de Falsificação de Rosto

A técnica proposta por NI et al. (2022) apresenta um método de detecção que visa mitigar o problema de sobre-ajuste. Isso é feito aplicando manipulações aleatórias em pares em uma imagem de entrada, gerando diferentes visões. Essas visões são processadas por um *encoder* compartilhado para extrair os pontos de interesse e são depois classificadas. (NI et al., 2022)

O método utiliza *encoders* compartilhados para processar os pares de imagens gerados, extraíndo os pontos de interesse e reduzindo a dimensionalidade dos dados. As saídas dos *encoders* são processadas por um classificador, responsável por prever se a face de entrada é real ou falsa.

A Figura 8 apresenta a arquitetura do modelo proposto. Sua implementação está disponível em (CORE..., 2023).

Figura 8 – Arquitetura do modelo CORE.



Fonte: Ni et al. (2022).

3.3.3 Modelo de Detecção de Deepfake Ignorante de ID

Este método trás a tona um problema com as técnicas de detecção de *deepfakes*, conhecido como "vazamento de identidade implícita", onde os classificadores binários aprendem, sem intenção, representações de identidade a partir das imagens. Isso pode levar a queda de desempenho quando aplicado a métodos não vistos anteriormente, devido ao sobre-ajuste.

Para resolver isso, é proposto o Modelo de Detecção de Deepfake Ignorante de ID (*ID-Unaware Deepfake Detection Model*). O modelo contém um módulo de detecção de artefatos que busca áreas locais de artefatos ao invés de características de identidade globais. O método

$$F - Score = \frac{2PR}{P + R} \quad (3.4)$$

Onde:

- VP - Verdadeiro Positivo, a imagem é corretamente indicada como falsa;
- VN - Verdadeiro Negativo, a imagem é corretamente indicada como real;
- FP - Falso Positivo, a imagem é erroneamente indicada como falsa;
- FN - Falso Negativo, a imagem é erroneamente indicada como real;
- P - Precisão;
- R - Revocação;

4 Resultados

Neste capítulo, detalham-se as métricas adotadas para a análise dos resultados, bem como os resultados alcançados.

4.1 Resultados Obtidos

A tabela 1 apresenta os resultados obtidos com base nas métricas descritas na seção 3.4. Pode-se observar um resultado superior do método proposto por Dong et al. (2023), que utiliza um módulo de detecção de artefatos para reduzir a superespecialização.

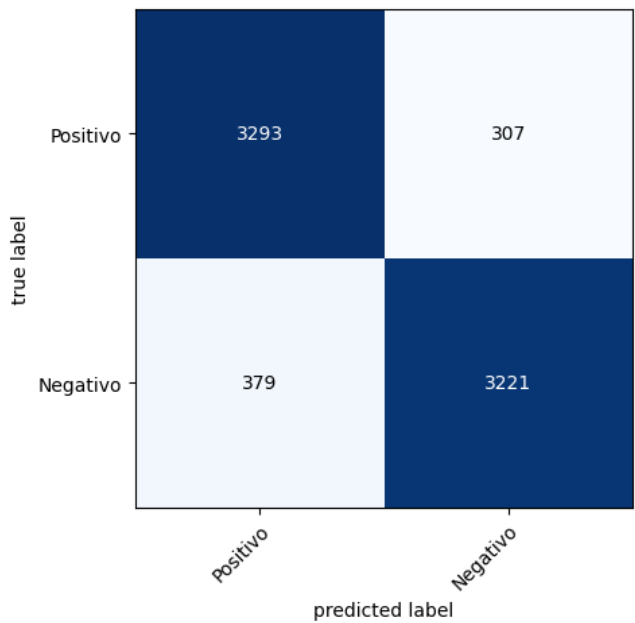
Tabela 1 – Resultados obtidos para os modelos testados

Modelo	Acurácia (%)	Precisão (%)	Revocação (%)	F-Score (%)
(DONG et al., 2022)	90,47	89,68	91,47	90,57
(NI et al., 2022)	93,79	92,39	95,06	93,70
(DONG et al., 2023)	97,47	95,91	98,99	98,44

Fonte: Elaborado pelo autor.

A Figura 10 apresenta a matriz de confusão do modelo de Dong et. al (2022), é possível notar uma quantidade maior de falsos positivos e falsos negativos em comparação aos outros métodos.

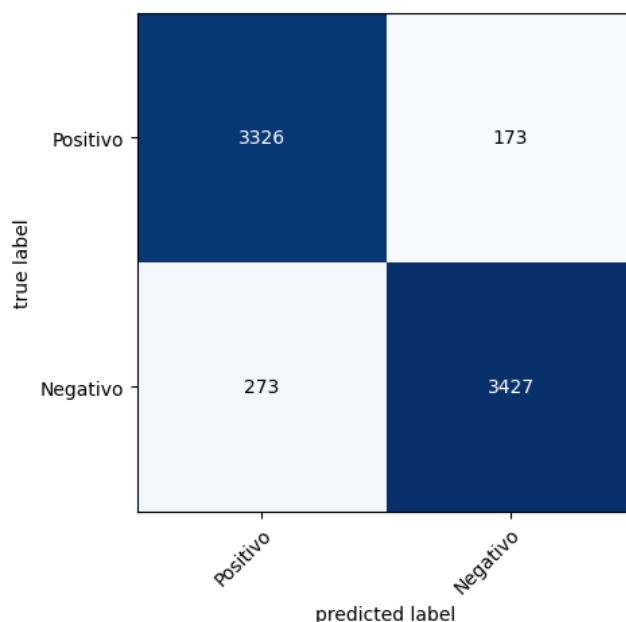
Figura 10 – Matriz de Confusão do modelo de DONG et al.(2022).



Fonte: Elaborado pelo autor.

A Figura 11 exibe a matriz de confusão do modelo de Ni et. al (2022), que teve um desempenho melhor que o método anterior, mas ainda apresenta uma quantidade significativa de falsos positivos e falsos negativos.

Figura 11 – Matriz de Confusão do modelo de NI et al.(2022).



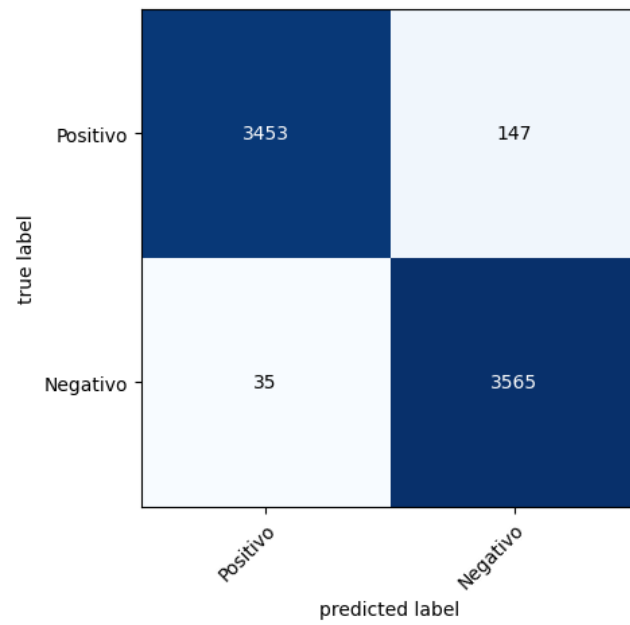
Fonte: Elaborado pelo autor.

Por fim, a Figura 12 exibe a matriz de confusão do modelo de Dong et. al (2023), que apresentou o melhor desempenho dentre os métodos estudados, com uma quantidade de falsos negativos comparável a da Figura 11, mas com uma quantidade de falsos positivos muito inferior.

4.2 Discussão

Em uma comparação direta, pode-se notar que o modelo ignorante de ID de Dong et al. (2023) obteve um melhor desempenho que as demais abordagens, devido a introdução do seu mecanismo de detecção desenvolvido com o objetivo de reduzir a dependência de artefatos globais das imagens falsificadas, indicando um caminho a ser seguido para trabalhos futuros, expandindo as técnicas de detecção de artefatos locais e reduzindo ainda mais o problema de vazamento implícito de identidades. A abordagem de Ni et al. (2022), demonstrou o segundo melhor desempenho dentre as três técnicas comparadas neste trabalho. Esta é outra técnica que buscou mitigar o problema de sobre-ajuste, mas sua técnica de expansão das imagens de não obteve resultados tão positivos como o método de Dong et al. (2023).

Figura 12 – Matriz de Confusão do modelo de DONG et al. (2023)



Fonte: Elaborado pelo autor.

5 Conclusão

Com uma sociedade cada vez mais conectada, com imagens da vida pessoal de toda a população disponíveis publicamente, o advento da tecnologia de *deepfakes* trás grandes preocupações para a sociedade.

Este trabalho realizou a comparação de três métodos estado-da-arte de detecção de imagens forjadas utilizando conceitos de inteligência artificial, aprendizado de máquina e aprendizado profundo.

Os resultados obtidos demonstram uma alta taxa de detecção de imagens falsas, principalmente com o método que emprega técnicas contra a superespecialização. Ainda assim, há espaço para evolução. É possível fazer uma analogia à corridas armamentistas, onde quando um lado evolui, o outro lado gera inovações de contraponto. É necessário o constante avanço nas tecnologias de detecção de *deepfakes*, para combater as inovações contínuas na geração das imagens falsas.

Trabalhos futuros podem abordar diferentes métodos de detecção e conjuntos de dados mais expansivos, podendo explorar a possibilidade de combinação de métodos e técnicas distintas para melhor desempenho.

Referências

- BODEN, M. A. *Artificial intelligence*. [S.l.]: Elsevier, 1996.
- BORGES, L.; MARTINS, B.; CALADO, P. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *Journal of Data and Information Quality (JDIQ)*, ACM New York, NY, USA, v. 11, n. 3, p. 1–26, 2019.
- CHAWLA, R. Deepfakes: How a pervert shook the world. *International Journal of Advance Research and Development*, v. 4, n. 6, p. 4–8, 2019.
- CORE: Consistent representation learning for face forgery detection, CVPRW 22. 2023. Disponível em: <<<https://github.com/niyunsheng/CORE>>>. Acesso em 15 de Outubro de 2023.
- DONG, S.; WANG, J.; JI, R.; LIANG, J.; FAN, H.; GE, Z. *Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization*. 2023.
- DONG, X.; BAO, J.; CHEN, D.; ZHANG, T.; ZHANG, W.; YU, N.; CHEN, D.; WEN, F.; GUO, B. Protecting celebrities from deepfake with identity consistency transformer. *arXiv preprint arXiv:2203.01318*, 2022.
- GUO, Y.; ZHANG, L.; HU, Y.; HE, X.; GAO, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. *CoRR*, abs/1607.08221, 2016. Disponível em: <<http://arxiv.org/abs/1607.08221>>.
- HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, Academy & Industry Research Collaboration Center (AIRCC), v. 5, n. 2, p. 1, 2015.
- ICT DeepFake Detection, CVPR2022. 2022. Disponível em: <<https://github.com/LightDXY/ICT_DeepFake>>. Acesso em 15 de Outubro de 2023.
- MEGVII-RESEARCH/CADDM: Official implementation of ID-unaware Deepfake Detection Model. 2023. Disponível em: <<<https://github.com/megvii-research/CADDM>>>. Acesso em 15 de Outubro de 2023.
- MÜLLER, B.; REINHARDT, J.; STRICKLAND, M. T. *Neural networks: an introduction*. [S.l.]: Springer Science & Business Media, 1995.
- NI, Y.; MENG, D.; YU, C.; QUAN, C.; REN, D.; ZHAO, Y. Core: Consistent representation learning for face forgery detection. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. [S.l.: s.n.], 2022. p. 12–21.
- PASSOS, L. A.; JODAS, D.; COSTA, K. A. P. da; JÚNIOR, L. A. S.; RODRIGUES, D.; SER, J. D.; CAMACHO, D.; PAPA, J. P. *A Review of Deep Learning-based Approaches for Deepfake Content Detection*. 2023.
- RASCHKA, S.; PATTERSON, J.; NOLET, C. Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence. *Information, MDPI*, v. 11, n. 4, p. 193, 2020.

RÖSSLER, A.; COZZOLINO, D.; VERDOLIVA, L.; RIESS, C.; THIES, J.; NIESSNER, M. Faceforensics++: Learning to detect manipulated facial images. *CoRR*, abs/1901.08971, 2019. Disponível em: <<http://arxiv.org/abs/1901.08971>>.

WESTERLUND, M. The emergence of deepfake technology: A review. *Technology innovation management review*, v. 9, n. 11, 2019.