# Sobel filter and linear classification for deepfake analysis of faces

**Fernanda G. Tamanaka**[1], **Carlos E. Thomaz**[1]

[1]Departamento de Engenharia Elétrica – Centro Universitário FEI (FEI)
Avenida Humberto Alencar Castelo Branco — São Bernardo do Campo — SP — Brasil

fegoyo@fei.edu.br, cet@fei.edu.br

***Abstract.*** *Through artificial intelligence (AI) techniques, deepfakes are created. Currently, almost six years after this topic's popularization, some people have been using AI algorithms to change, for example, face. In this scenario, this article proposes to apply Sobel filter, together with multivariate statistical learning techniques, to identify the most discriminating characteristics. Complementarily, we aim to analyze the areas that humans have visually identified as relevant to detect a deepfake. The results show that the statistical model (PCA plus MLDA) combined with Sobel filter correctly classified most of the images, highlighting regions that discriminate a deepfake.*

***Resumo.*** *Por meio de técnicas de inteligência artificial (IA) são criadas deepfakes. Atualmente, quase seis anos após a popularização desse tema, algumas pessoas estão utilizando algoritmos de IA para alterar, por exemplo, o rosto. Neste cenário, este artigo tem como proposta aplicar o filtro Sobel, em conjunto com técnicas multivariadas de aprendizagem estatística, para a identificação das características mais discriminantes. Complementarmente, visamos analisar as áreas que humanos visualmente identificaram como relevantes para detectar uma deepfake. Os resultados mostram que o modelo estatístico (PCA mais MLDA) combinado com o filtro Sobel classificou corretamente a maioria das imagens, realçando as regiões que discriminam uma deepfake.*

## 1. Introdução

Por meio de técnicas de inteligência artificial (IA) são criadas deepfakes (deep learning + fakes), sendo que estas podem ser por áudio, vídeo e/ou imagem [Kietzmann et al. 2020]. Com a popularização das deepfakes no fórum do Reddit, em meados de 2017 e ínicio de 2018, diversas empresas (exemplo: Google e Facebook) demonstraram preocupação com esse tema, ao ponto de lançarem campeonatos (com banco de dados públicos) em que o melhor modelo que predizia a detecção de deepfakes era vencedor [Mitra et al. 2021]. Com isso, recentemente, aumentaram exponencialmente os trabalhos que visam compreender como detectar essas deepfakes [Rana et al. 2022].

Segundo [Kietzmann et al. 2020], existem sete tipos de deepfakes, sendo estes categorizados em imagens, áudios, vídeos e áudio combinado com vídeo. Todavia, essa classificação pode variar de autor para autor, tal como [Juefei-Xu et al. 2022, Tolosana et al. 2020] que focaram apenas na manipulação facial, seja por vídeo ou por imagem. Considerando o trabalho de [Kietzmann et al. 2020], esses tipos de deepfakes contemplam 1 (uma) técnica de deepfake para imagem (Face and/or body-swapping),

2 (duas) para áudios (Voice-swapping e Text-to-Speech), 3 (três) para vídeos (Face-swapping, Face-morphing, Full-body puppetry) e 1 (uma) para áudios com vídeos (Lip-syncing).

Neste contexto, a literatura afim aponta distintas técnicas para identificar uma deepfake, porém resultam da observação das técnicas citadas anteriormente de forma isolada [Chen et al. 2020, Demir and Ciftci 2021] ou combinada de duas ou mais técnicas [Tariq et al. 2021], seja analisando somente a face [Tolosana et al. 2021] ou combinando análise das emoções com áudio [Mittal et al. 2020]. Em se tratando de um tema recente, ainda não há um consenso tanto para categorizar os métodos de detecção de deepfakes, como para identificar os tipos de deepfakes.

O presente trabalho tem como base a classificação feita por [Masood et al. 2022], mediante que os autores descreveram os critérios usados para classificar cada um dos nove tipos de detecção: 1) Inconsistência; 2) Ambientais; 3) Comportamentais; 4) Forencis; 5) Sincronização; 6) Fisiológicas; 7) Coerência; 8) Classificação e 9) Detecção de anomalia. Dentre essas características, o foco deste artigo será em características comportamentais (ex. Expressões faciais) e ambientais (ex. Iluminação), para analisar quais são as regiões que identificam uma deepfake em quadros de vídeos de faces frontais [Tolosana et al. 2021, Demir and Ciftci 2021, Gerstner and Farid 2022].

Adicionalmente, visando reduzir os custos computacionais e trabalhar com um número limitado de amostras, este artigo tem como proposta aplicar o filtro Sobel, em conjunto com técnicas multivariadas de aprendizagem estatística, para a identificação das características mais discriminantes das deepfakes. Complementarmente, visa analisar de forma inédita as áreas que os humanos identificaram como relevantes para detectar uma deepfake, por meio do rastreamento ocular (em inglês *eye-tracking*) .

## 2. Materiais e Métodos

Nesta seção, descrevemos a metodologia aplicada em relação às bases de dados de vídeos estudadas nesta pesquisa (Celeb-DF) de [Li et al. 2020] e (Faceforencis) de [Rössler et al. 2018], mencionando (brevemente) estas bases. Além disso, são descritas as análises estatísticas e o método de rastreamento ocular implementado.

### 2.1. Bases de dados

A escolha das bases de dados (Celeb-DF [Li et al. 2020] e Faceforencis++ [Rossler et al. 2019]) se deu por três fatores: 1) os vídeos são de fácil acesso ao público (mediante preenchimento de formulário), 2) são bases antigas para o estudo das deepfakes e 3) ambas possuem diversidade de vídeos tanto para homens como para mulheres.

#### 2.1.1. Celeb-DF

Celeb-DF é uma base de dados pública [1] com vídeos extraídos do Youtube, de 59 distintas celebridades com diferentes nacionalidades. Ao todo, são 6.229 vídeos, com tempo médio de 13 segundos, divididos entre 590 reais e 5.639 fakes. Não foi detalhado pelos criadores do Celeb-DF, [Li et al. 2020], qual foi a técnica utilizada para criar os deepfakes.

[1] Site do celeb-df: https://github.com/yuezunli/celeb-deepfakeforensics

### 2.1.2. Faceforencis++

Extraído do Youtube, Faceforencis++ é uma evolução do Faceforencis [Rössler et al. 2018] e é uma base de dados pública [2], composta por quatro técnicas de criação de conteúdos fakes (Deepfakes, Faceswap, Face2face e Neural Textures), com diferentes vídeos de distintas etnias. Ao todo são 5.000 vídeos divididos entre 1.000 reais e 4.000 fakes. Diferentemente do Celeb-DF, os autores do Faceforencis++ não colocaram em seu trabalho o tempo médio dos vídeos.

## 2.2. Análise estatística

Para esta análise experimental, foram utilizados 200 (duzentos) frames, 100 frames para cada base de dados. Cada frame frontal da face tem resolução de 224 x 224 e inclui os movimentos naturais da face (abertura e fechamento dos olhos, além dos movimentos da boca) e a posição da câmera.

Esta amostra de imagens, uma para cada pessoa, é útil para analisar se o movimento facial, identificado por alguns pontos específicos, é uma característica discriminante à identificação de uma deepfake. Ao todo, são 100 (cem) imagens reais e 100 imagens fakes. Estas faces foram detectadas usando a técnica Single Shot Scale-invariant Face Detector (S3fd) [Zhang et al. 2017] e alinhadas usando a técnica Face Alignment Network (FAN) [Bulat and Tzimiropoulos 2017].

Em cada uma dessas 200 imagens, foram aplicadas as técnicas de Análise de Componentes Principais (PCA) [Wold et al. 1987] e Máxima Discriminância Linear (MLDA [Thomaz et al. 2006] para identificar se as imagens foram classificadas corretamente, por meio da redução de dimensionalidade e das características faciais que discriminam uma imagem real de uma fake, em se tratando de um problema com poucas amostras. Todavia, antes de aplicar essas técnicas estatísticas, foi aplicado o filtro Sobel [Kanopoulos et al. 1988], para detectar os contornos das imagens e, assim, identificar (pela diferença entre os pixels) as características predominantes das imagens reais e das fakes com e sem filtro.

## 2.3. Percepção visual humana

O método de rastreamento ocular explorado é uma adaptação do trabalho de [Caporusso et al. 2020] e a amostra é composta por 36 participantes voluntários que tiveram que assinar o termo de consentimento livre e esclarecido (TCLE). A análise foi com base nos participantes que tiveram taxa de aquisição do sinal acima de 80%.

Durante a metodologia experimental, os voluntários tiveram que visualizar vinte imagens (10 reais e 10 fakes) que foram compostas por frames de faces frontais de vídeos. Ao todo, foram escolhidos (das 200 imagens mencionadas anteriormente); doze imagens (6 reais e 6 fakes) da base Celeb-DF e oito imagens (4 reais e 4 fakes) da base Faceforencis++, que foram estatisticamente menos discriminantes. Cada imagem, colocada na tela do Tobii TX300, teve uma resolução de 940x940 pixels e foram de faces detectadas usando a técnica S3fd [Zhang et al. 2017] e faces alinhadas usando a técnica FAN [Bulat and Tzimiropoulos 2017].

[2] Site do faceforencis++: https://github.com/ondyari/FaceForensics

O voluntário teve 5 segundos para analisar a imagem e mais 3 segundos para dizer se o que acabou de ver era real ou fake. O voluntário apertou a barra de espaço, uma única vez, apenas para pular a introdução sobre os resultados experimentais. Para passar automaticamente para a próxima imagem, foi definido um tempo fixo de visualização, assim que finalizasse os 3 segundos da pergunta anterior.

## 3. Experimentos, Resultados e Discussões

Nesta seção descrevemos experimentos, a análise dos resultados nas bases de dados: Celeb-DF e Faceforencis++, além de fazer uma breve discussão desses resultados.

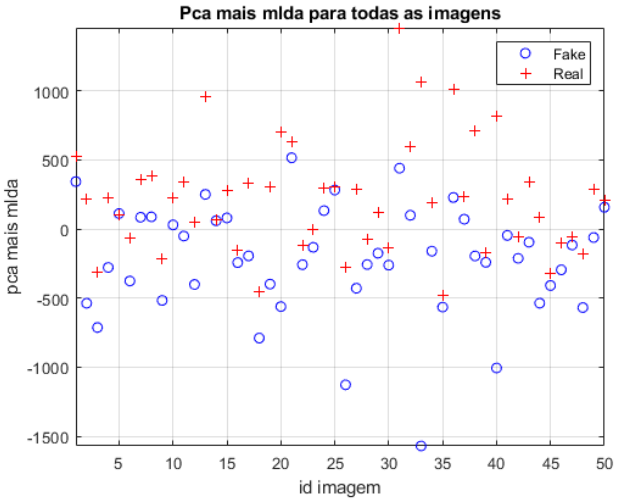


**Figura 1. Separação linear dos dados Celeb-DF com tons de cinza.**

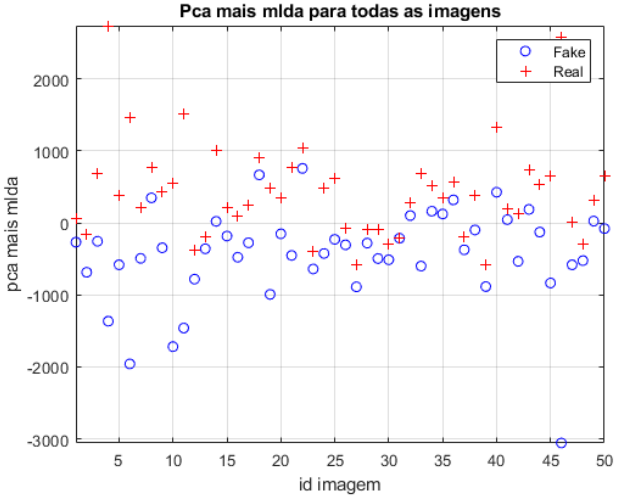


**Figura 2. Separação linear dos dados Faceforencis++ com tons de cinza.**

### 3.1. Análise das bases de dados (estatística e eye-tracking)

As Figuras 1 a 4 apresentam os resultados experimentais das características discriminantes das 200 imagens, 100 para cada base de dados, mencionadas na seção 2. Como o estudo trabalha com uma amostra de poucas imagens, foram aplicadas as técnicas de PCA [Wold et al. 1987], para reduzir a dimensionalidade dos dados; e de MLDA

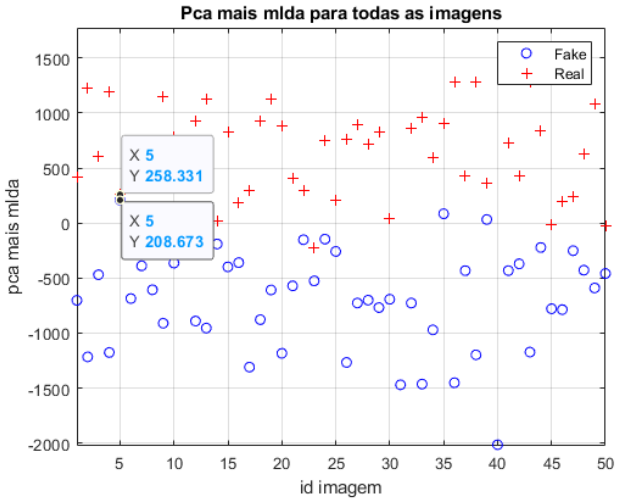


**Figura 3. Separação linear dos dados Celeb-DF com filtro Sobel.**

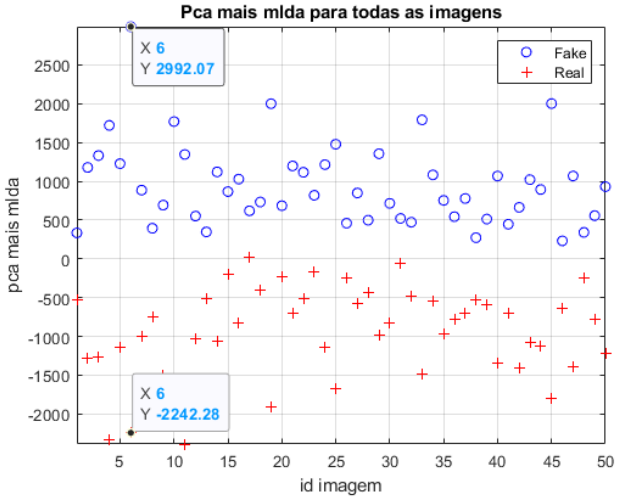


**Figura 4. Separação linear dos dados Faceforencis++ com filtro Sobel**

[Thomaz et al. 2006] para maximizar a separação linear amostral e encontrar os valores discriminantes e, assim, identificar quais são as características mais relevantes para distinguir uma imagem real de uma fake.

Para o cálculo do MLDA, foram utilizadas todas componentes principais com autovalor não-nulo do PCA. Nos gráficos das Figuras 1 a 4, o eixo ”x”representa as imagens (50 reais e 50 fakes) e o eixo ”y”os valores do PCA mais MLDA. Com isso, para descobrir se uma característica discriminante é predominante, basta identificar a variação no eixo ”y”, no qual quanto maior for a diferença entre as imagens fakes e reais, maior será a discriminância.

As diferenças gráficas entre as Figuras 1 e 2 (ambas sem o filtro Sobel) e as Figuras 3 e 4 (ambas com o filtro Sobel) são notáveis. É possível verificar que o filtro Sobel auxilia na separação das imagens. Para calcular a acurácia, nós usamos o método de validação cruzada [Kohavi et al. 1995], separando (para cada base de dados) 20 (vinte) imagens para teste e 80 (oitenta) imagens para treino, com k-fold igual a 5 (cinco), como ilustrado

nas Tabelas 1 e 2.

De acordo com as Tabelas 1 e 2, é possível verificar visualmente que o classificador linear teve dificuldade em discriminar as imagens com tons de cinza, com uma acurácia inferior a 70% para o Celeb-DF e inferior a 85% para Faceforencis++. No entanto, com o filtro Sobel a acurácia foi superior a 90%, para a base Celeb-DF, e de 100% para o Faceforencis++.

**Tabela 1. Acurácia por filtro (Celeb-DF)**

| K-fold | Sem Sobel | Com Sobel |
|---|---|---|
| 1 | 70,00% | 95,00% |
| 2 | 71,25% | 97,50% |
| 3 | 70,00% | 96,25% |
| 4 | 65,00% | 93,75% |
| 5 | 70,00% | 97,75% |
| Acurácia Média | 69,25% $\pm$ 2,44% | 96,00% $\pm$ 1,63% |

**Tabela 2. Acurácia por filtro (Faceforencis++)**

| K-fold | Sem Sobel | Com Sobel |
|---|---|---|
| 1 | 83,75% | 100,00% |
| 2 | 83,75% | 100,00% |
| 3 | 78,75% | 100,00% |
| 4 | 82,50% | 100,00% |
| 5 | 83,75% | 100,00% |
| Acurácia Média | 82,50% $\pm$ 2,17% | 100,00% |

De modo geral, o método combinado PCA mais MLDA com o filtro Sobel classificou bem essas imagens, porém (em alguns pontos) ocorreu o oposto disto, especialmente na base de dados Celeb-DF. Neste caso, tiveram poucas características discriminantes que permitiram classificar a imagem corretamente, independentemente do gênero. Com isso, dos pontos ilustrados nas Figuras 3 e 4, serão apresentados dois casos de cada base de dados, destacados nas Figuras 5 a 8 que ou tiveram características menos discriminantes (Figuras 5 e 6) ou tiveram características muito discriminantes (Figuras 7 e 8).

Na Figura 3, é possível ver a proximidade de dois pontos: 5 (mulher real) e 5 (mulher fake), sendo que desses o 5 (mulher real) foi classificado corretamente, porém de forma muito próxima a sua imagem fake. Com isso, nota-se que (neste caso) não existe uma característica discriminante predominante, pelo filtro Sobel, que a diferencie de uma imagem fake. Já na Figura 4, referente a base de dados Faceforencis++, é possível ver que há uma larga discriminância entre as imagens. Isto se deve ao fato de apresentarem uma maior densidade do filtro Sobel, como ilustrado nas Figuras 7 e 8, permitindo inferir que a iluminação ajuda na discriminação das características e, por consequência, na classificação das imagens. Tal resultado vai ao encontro dos achados de [Gerstner and Farid 2022].

Nas Figuras 5 e 6, as principais características a serem observadas são: boca, olhos, nariz e sobrancelhas, porém com uma predominância do filtro de Sobel nas imagens

**Figura 5. Celeb-DF com Sobel - Fake.**

**Figura 6. Celeb-DF com Sobel - Real.**

**Figura 7. Faceforencis++ com Sobel - Fake.**

**Figura 8. Faceforencis++ com Sobel - Real.**

reais. Já nas Figuras 7 e 8, as principais características a serem observadas são: boca, olhos (ênfase na pupila), nariz, sobrancelhas, outras regiões da face (exceto a orelha) e a iluminação, porém com uma predominância do filtro de Sobel nas imagens reais, assim como os resultados das Figuras 5 e 6.

Tais achados, tanto da análise da base de dados do Celeb-DF como do Faceforencis++, concordam com os trabalhos recentes de [Nguyen and Derakhshani 2020, Tolosana et al. 2021, Demir and Ciftci 2021, Gerstner and Farid 2022] na identificação das regiões que auxiliam na detecção de uma deepfake.

A diferença entre as imagens do Celeb-DF (Figuras 5 e 6) e do Faceforencis++ (Figuras 7 e 8) é o grau de iluminação das imagens, sendo que como as imagens do Faceforencis++ apresentam uma maior intensidade de iluminação, as mesmas ilustram com maiores detalhes as características discriminantes e, por isto, analisando somente o contorno da imagem, é possível identificar que há presença de informações que permitem distinguir imagens reais das fakes.

A acurácia média da percepção visual humana foi de 66,53% $\pm$ 10,87%, bem inferior ao algoritmo. Para visualizar comparativamente os resultados do eye-tracking com o Sobel foram escolhidas quatro imagens que representam as figuras com menor acurácia, tanto para imagens reais como para as fakes. Estas imagens são apresentadas nas Figuras 9, 12, 15 e 18; em conjunto com suas imagens com filtros aplicados (Sobel e tons de cinza).

**Tabela 3. Acurácia das imagens menos discriminantes por cada método avaliado (Celeb-DF)**

| Figura | Sem Sobel | Com Sobel | Eye-tracking |
|---|---|---|---|
| F5 | Errou | Errou | 47,22% |
| F22 | Acertou | Acertou | 58,33% |
| F24 | Errou | Acertou | 66,67% |
| F35 | Acertou | Errou | 77,78% |
| F39 | Acertou | Errou | 41,67% |
| F44 | Acertou | Acertou | 72,22% |
| R14 | Acertou | Acertou | 75,00% |
| R16 | Errou | Acertou | 69,44% |
| R23 | Errou | Errou | 72,22% |
| R30 | Errou | Acertou | 86,11% |
| R45 | Errou | Errou | 50,00% |
| R50 | Acertou | Errou | 75,00% |

Para fins de comparação de desempenho dos três métodos de detecção estudados neste artigo: 1) análise estatística com filtro Sobel, 2) análise estatística com tons de cinza e 3) análise da atenção visual humana; são apresentadas as Tabelas 3 e 4 com a acurácia de todas as imagens utilizadas no experimento do eye-tracking, ou seja, doze para a Tabela 3 (Celeb-DF) e oito para a Tabela 4 (Faceforencis++). Para o cálculo da acurácia de cada imagem, com e sem o filtro Sobel, foi utilizada a técnica leave-one-out [Fukunaga and Hummels 1989], obtendo uma resposta binária para cada imagem. Desta forma, ”Acertou”significa que o algoritmo classificou corretamente a imagem e ”Errou”, o contrário.

**Figura 9. Celeb-DF com eye-tracking - Fake.**

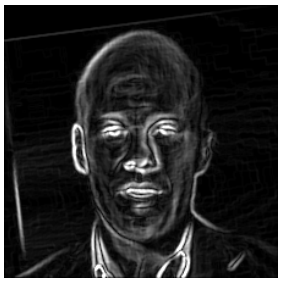

**Figura 10. Celeb-DF com Sobel - Fake.**

**Figura 11. Celeb-DF com Tons de cinza - Fake.**

Na Tabela 3 é possível ver que as imagens em tons de cinza tiveram um desempenho inferior para classificar imagens reais. Ao passo que com o filtro Sobel, a acurácia foi a mesma, independentemente se a imagem é real ou fake. Além disso, no eye-tracking, nota-se que os humanos tiveram mais dificuldade em identificar uma imagem fake de uma real, principalmente a imagem ”F39”que é ilustrada na Figura 9. Em relação às imagens reais, a imagem ”R45”(Figura 12) foi a que os humanos tiveram maior dificuldade em classificar se é fake ou real.

Na Tabela 4, de forma análoga, as imagens em tons de cinza tiveram um desempe-

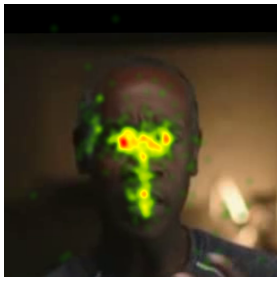

**Figura 12. Celeb-DF com eye-tracking - Real.**

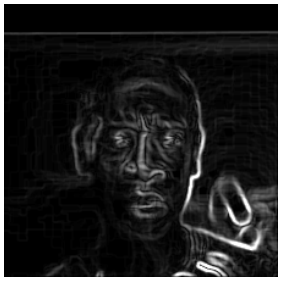

**Figura 13. Celeb-DF com Sobel - Real.**

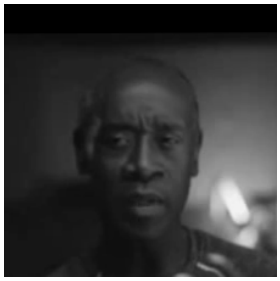

**Figura 14. Celeb-DF com Tons de cinza - Real.**

**Tabela 4. Acurácia das imagens menos discriminantes por cada método avaliado (Faceforencis++)**

| Figura | Sem Sobel | Com Sobel | Eye-tracking |
|---|---|---|---|
| F13 | Acertou | Acertou | 72,22% |
| F26 | Acertou | Acertou | 91,67% |
| F38 | Acertou | Acertou | 97,22% |
| F46 | Errou | Acertou | 75,00% |
| R15 | Acertou | Acertou | 41,67% |
| R17 | Acertou | Acertou | 55,56% |
| R23 | Errou | Acertou | 50,00% |
| R31 | Errou | Acertou | 55,56% |

nho inferior para classificar imagens reais. Ao passo que com o filtro Sobel, o classificador acertou todas as imagens. Além disso, no eye-tracking, nota-se que os humanos tiveram mais dificuldade em identificar uma imagem real de uma fake, principalmente, a imagem ”R15”que é ilustrada na Figura 18. Em relação às imagens fakes, a imagem ”F13”(Figura 15) foi a que os humanos tiveram mais dificuldade em classificar se é fake ou real.

**Figura 15. Facefo-rencis++ com eye-tracking - Fake.**

**Figura 16. Facefo-rencis++ com So-bel - Fake.**

**Figura 17. Facefo-rencis++ com Tons de cinza - Fake.**

Analisando as imagens com os resultados do eye-tracking (Figuras 9, 12, 15 e 18) é possível notar uma semelhança entre elas, no que se refere as áreas de atenção visual. Essas áreas são, em ordem decrescente de importância, considerando o mapa de calor (em que a cor vermelha representa maior quantidade de fixação e a cor verde o oposto): olhos,

**Figura 18. Faceforencis++ com eye-tracking - Real.**

**Figura 19. Faceforencis++ com Sobel - Real.**

**Figura 20. Faceforencis++ com Tons de cinza - Real.**

boca, nariz, orelha, outras regiões da face e sobrancelha. Tais regiões, especialmente, olhos, boca e nariz; por serem áreas que compreendem a região central do rosto, apresentam informações relevantes para identificar uma pessoa e, consequentemente, distinguir se uma imagem é real ou fake [Li et al. 2021, Wang et al. 2022]

Somente na base Faceforencis++, as imagens que os humanos mais erraram (Figuras 15 e 18) foram as imagens que o algoritmo teve facilidade em detectar se é uma deepfake, ou seja, o algoritmo identificou características que permitiram distinguir uma imagem deepfake de uma real, enquanto que para os humanos estas imagens apresentaram características menos discriminantes. Já na base Celeb-DF (Figuras 9 e 12) os humanos acertaram mais que o algoritmo. Tanto no eye-tracking como o filtro Sobel identificaram a orelha como uma característica relevante para distinguir as imagens reais das fakes. Tal achado vai ao encontro de [Agarwal and Farid 2021].

Nas imagens com tons de cinza (Figuras 11, 14, 17 e 20), é possível visualizar que, diferentemente do filtro Sobel que destaca os contornos da imagem, não há um destaque para uma região específica, o que torna difícil a classificação correta das imagens. Já nas imagens com o filtro Sobel (Figuras 10, 13, 16 e 19) destacam-se as seguintes regiões: olhos, sobrancelha, nariz, boca, outras regiões da face, iluminação e orelha; diferente das imagens com o filtro Sobel (figuras 5, 6, 7 e 8) que não destacam a orelha, tendo em vista que o cabelo impede a sua classificação.

Na atenção visual, por ora, não se analisou a influência da iluminação na identificação das imagens, porém não se descarta a possibilidade de, em uma próxima versão do estudo, analisar não somente a iluminação, mas também se o nível da dilatação pupilar está relacionado com a acurácia.

À critério de curiosidade, na literatura mais recente de detecção de deepfakes, usando as bases Celeb-df e Faceforencis++, observa-se que os autores [Mohiuddin et al. 2023] criaram um modelo de machine learning que usa a técnica de seleção das características hierárquicas, para comparar a acurácia desse modelo em cada base de dados, seja Celeb-DF ou Faceforencis++, em relação a outras literaturas. [Mohiuddin et al. 2023] utilizaram 1.416 frames reais e 9.262 fakes para Celeb-DF. Já para o Faceforencis, os autores contaram com 3.294 frames reais e 3.176 fakes. Em ambas as bases de dados, todos os frames foram divididos em treino, validação e teste. Ao longo do trabalho, [Mohiuddin et al. 2023] descobriram a acurácia de 99,35% e 99,16% para ambas as bases de dados, respectivamente. Diferentemente dessa literatura, pode

observar, portanto, que a principal vantagem deste estudo é a pesquisa com um conjunto limitado de amostras e com acurácias encontradas no Sobel semelhantes a esse modelo.

## 4. Conclusão

Os resultados experimentais mostraram quantitativa e qualitativamente que o modelo estatístico linear (PCA mais MLDA) combinado com o filtro Sobel classificou corretamente a maioria das imagens, realçando as regiões que discriminam uma deepfake com uma acurácia média superior a 95%, independentemente da base de dados. Além disso, mesmo com poucas amostras, os achados deste artigo estão de acordo com o que foi apresentado pela literatura afim, ou seja, que as características faciais discriminantes para diferenciar um frame real de um fake são: olhos (ênfase na pupila), nariz, boca, orelha, sobrancelha, outras regiões da face e iluminação.

Na atenção visual, os humanos mostraram uma habilidade inferior ao algoritmo na detecção de imagens fakes, para a base Celeb-DF, e o contrário para a base Faceforencis++. Analisando somente o filtro Sobel é possível inferir que a iluminação talvez seja a característica mais discriminante para classificação dessas imagens.

Por fim, observa-se que as vantagens são a redução dos custos computacionais e o estudo com um número limitado de amostras. Ao passo que, a limitação consiste em não explorar a dilatação pupilar na análise da percepção visual e, por isso, como trabalhos futuros, espera-se estender esses resultados para analisar também se o nível da dilatação pupilar está relacionado com a acurácia da percepção humana.

## Agradecimentos

Este estudo foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de financiamento 001.

## Referências

Agarwal, S. and Farid, H. (2021). Detecting deep-fake videos from aural and oral dynamics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 981–989.

Bulat, A. and Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030.

Caporusso, N., Zhang, K., and Carlson, G. (2020). Using eye-tracking to study the authenticity of images produced by generative adversarial networks. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pages 1–6. IEEE.

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., and Khoury, E. (2020). Generalization of audio deepfake detection. In *Proc. Odyssey 2020 The Speaker and Language Recognition Workshop*, pages 132–137.

Demir, I. and Ciftci, U. A. (2021). Where do deep fakes look? synthetic face detection via gaze tracking. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–11.

Fukunaga, K. and Hummels, D. M. (1989). Leave-one-out procedures for nonparametric error estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(4):421–423.

Gerstner, C. R. and Farid, H. (2022). Detecting real-time deep-fake videos using active illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 53–60.

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., and Liu, Y. (2022). Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision*, 130(7):1678–1734.

Kanopoulos, N., Vasanthavada, N., and Baker, R. L. (1988). Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367.

Kietzmann, J., Lee, L. W., McCarthy, I. P., and Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons*, 63(2):135–146.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Li, M., Liu, B., Hu, Y., Zhang, L., and Wang, S. (2021). Deepfake detection using robust spatial and temporal features from facial landmarks. In *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, pages 1–6. IEEE.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.

Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., and Malik, H. (2022). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, pages 1–53.

Mitra, A., Mohanty, S. P., Corcoran, P., and Kougianos, E. (2021). A machine learning based approach for deepfake detection in social media through key video frame extraction. *SN Computer Science*, 2(2):1–18.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., and Manocha, D. (2020). Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832.

Mohiuddin, S., Sheikh, K. H., Malakar, S., Velásquez, J. D., and Sarkar, R. (2023). A hierarchical feature selection strategy for deepfake video detection. *Neural Computing and Applications*, 35(13):9363–9380.

Nguyen, H. M. and Derakhshani, R. (2020). Eyebrow recognition for identifying deepfake videos. In *2020 international conference of the biometrics special interest group (BIOSIG)*, pages 1–5. IEEE.

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

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2018). Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Tariq, S., Jeon, S., and Woo, S. S. (2021). Am i a real or fake celebrity? measuring commercial face recognition web apis under deepfake impersonation attack. *arXiv preprint arXiv:2103.00847*.

Thomaz, C. E., Kitani, E. C., and Gillies, D. F. (2006). A maximum uncertainty lda-based approach for limited sample size problems—with application to face recognition. *Journal of the Brazilian Computer Society*, 12(2):7–18.

Tolosana, R., Romero-Tapiador, S., Fierrez, J., and Vera-Rodriguez, R. (2021). Deepfakes evolution: Analysis of facial regions and fake detection performance. In *International Conference on Pattern Recognition*, pages 442–456. Springer.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., and Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148.

Wang, G., Jiang, Q., Jin, X., and Cui, X. (2022). Ffr_fd: Effective and fast detection of deepfakes via feature point defects. *Information Sciences*, 596:472–488.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., and Li, S. Z. (2017). S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201.