# Evaluation of Projected Gradient Descent in Attacks against Age Estimation Models

**Marcelo Nascimento Oliveira Soares**[1], **Leonardo Barbosa Oliveira**[2], **Antonio João Gonçalves Azambuja**[3], **Jean Phelipe de Oliveira Lima**[3,4]

[1]Instituto de Informática — Universidade Federal de Goiás (UFG)
Goiânia — GO — Brasil

[2]Departamento de Ciência da Computação — Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte — MG — Brasil

[3]Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos — SP — Brasil

[4]Universidade Federal de Mato Grosso (UFMT)
Várzea Grande — MT — Brasil

marcelo.soares@discente.ufg.br, leob@dcc.ufmg.br,

azambuja@ita.br, jean.lima@ga.ita.br

***Abstract.*** *The global landscape demands secure age verification mechanisms to protect against fraud and threats to minors in digital environments. However, the use of Artificial Intelligence in this context is vulnerable to Adversarial Machine Learning. In this work, we investigate the security of these models against transfer attacks. We employ Projected Gradient Descent on a surrogate model and transfer the perturbation to a system that simulates production environments, validating the vulnerability of biometric architectures and highlighting the need for robust security measures to prevent bypass in age verification platforms.*

***Resumo.*** *O cenário mundial exige mecanismos seguros de verificação etária para a proteção contra fraudes e ameaças contra menores em ambientes digitais. No entanto, o uso de Inteligência Artificial neste contexto é vulnerável ao Aprendizado de Máquina Adversarial. Neste trabalho, investigamos a segurança desses modelos frente a ataques de transferência. Utilizamos o Projected Gradient Descent em um modelo substituto (surrogate) e transferimos a perturbação para um sistema que simula ambientes de produção, validando a vulnerabilidade das arquiteturas biométricas e evidenciando a necessidade de medidas de segurança robustas para evitar o bypass em plataformas de verificação etária.*

## 1. Introdução

No cenário digital, a proteção de menores tem sido regulamentada por diferentes estatutos legais, que passaram a exigir que as plataformas *online* implementem mecanismos de verificação biométrica para a proteção dos usuários [Sousa et al. 2025]. A implementação de biometria no ecossistema *Roblox*, por exemplo, evidencia a necessidade de modelos de estimativa etária para assegurar a conformidade legal e para proteger os usuários.

No entanto, a adoção de inteligência artificial como um método biométrico é frequentemente ameaçada sob a ótica do Aprendizado de Máquina Adversarial (AML) [Biggio and Roli 2018]. A literatura documenta que modelos de aprendizado profundo podem ser vulneráveis a diversos ataques conhecidos [Qiao et al. 2026]. No referido contexto, este trabalho investiga a integridade de modelos no de verificação etária. Nesse sentido, a pesquisa demonstrou que o ataque pode não apenas fraudar a classificação etária, mas também degradar a confiança nos sistemas de reconhecimento facial e ameaçar a eficácia da proteção de crianças e adolescentes na *internet*.

Para avaliar essas fragilidades, ataques de transferência (black-box) a modelos de classificação etária foram simulados, gerando perturbações adversariais a partir de um modelo substituto (white-box). Essa metodologia busca reproduzir cenários reais, em que o atacante não tem acesso direto aos parâmetros do modelo alvo, podendo atuar apenas através de consultas e da observação da saída do alvo [Papernot et al. 2016]. Os resultados apontam que a vulnerabilidade é real, causando distorções significativas na estimativa etária, com a gravidade variando a depender do gênero, etnia e arquitetura do modelo alvo.

Como contribuições, este trabalho (i) propõe um pipeline de auditoria de segurança de sistemas de verificação etária baseado em ataques adversariais de transferência, simulando um cenário realista de caixa-preta a partir de um modelo substituto *white-box*; (ii) define a métrica *Taxa de Inversão de Maioridade* (TIM), que quantifica o cruzamento do limiar de maioridade em ambas as direções; (iii) avalia empiricamente a transferibilidade do ataque em três modelos-alvo de paradigmas distintos (CNN, ViT e híbrido multi-entrada) sob diferentes orçamentos de perturbação e números de passos; e (iv) analisa o impacto demográfico do ataque, evidenciando variações por gênero e grupo étnico.

## 2. Trabalhos Relacionados

A biometria facial apoia-se em arquiteturas como o ArcFace [Deng et al. 2019] para a extração robusta de identidades, e em *Vision Transformers* (ViTs) [Dosovitskiy et al. 2021] para a modelagem semântica de atributos como a idade (*Soft Biometrics*). Contudo, a alta dimensionalidade dessas funções de decisão as tornam vulneráveis a perturbações imperceptíveis [Biggio and Roli 2018, Qiao et al. 2026]. No cenário *white-box*, em que o atacante tem acesso aos gradientes do modelo, o *Projected Gradient Descent* (PGD) [Madry et al. 2018] consolida-se como um ataque iterativo de primeira ordem eficiente [Cinà et al. 2025]. Em contrapartida, cenários sem acesso interno impulsionam ataques *black-box* baseados em *queries*, como o *Square Attack* [Andriushchenko et al. 2020], que utilizam buscas estocásticas sem computar derivadas.

Para a validação científica de ambos os cenários, a perturbação deve ser visualmente imperceptível e limitada pelo parâmetro $\epsilon$ sob as normas $L_p$ [Croce and Hein 2020]. A norma $L_\infty$ delimita o teto de alteração individual por pixel, sendo o padrão para o PGD, enquanto a norma $L_2$ restringe a distorção euclidiana total do ruído inserido. Na literatura de *Soft Biometrics*, tarefas de regressão densa como a estimativa etária — exemplificadas pelos modelos DEX [Rothe et al. 2018] e otimizações de *Mean-Variance Loss* [Pan et al. 2018] — operam em superfícies de

decisão complexas, facilitando a síntese de ruídos adversariais focados em manipulações semânticas em detrimento de alterações espaciais.

A relevância prática de ataques iterativos *white-box*, como o PGD, é amplificada pelo fenômeno da transferibilidade adversarial. Perturbações otimizadas em um modelo substituto local (*surrogate*) costumam preservar sua eficácia contra modelos alvo desconhecidos, explorando o alinhamento não-intencional de suas fronteiras de decisão [Goodfellow et al. 2015, Papernot et al. 2016]. Em cenários de ameaça real, isso significa que o atacante pode sintetizar o ruído PGD em uma arquitetura própria e transferir a imagem corrompida para o sistema alvo sem necessitar de acesso interno. A eficácia dessa transferência é governada pela complexidade do *surrogate* e pela vulnerabilidade intrínseca do alvo [Demontis et al. 2019]. Adicionalmente, abordagens como o *Adv-Attribute* [Ding et al. 2022] evidenciam que ataques focados na manipulação semântica de atributos maximizam essa transferibilidade.

## 3. Metodologia

A metodologia foi projetada para avaliar a robustez de sistemas biométricos diante de ataques de transferência e vulnerabilidades entre tarefas diferentes. A proposta se divide na seleção do ambiente de testes, na criação do vetor de ataque através de um modelo *surrogate* (*white-box*) usando o ataque PGD, na transferência do ataque para os alvos e por fim, na análise dos resultados, com a Figura 1 demonstrando o fluxo de geração do ataque.

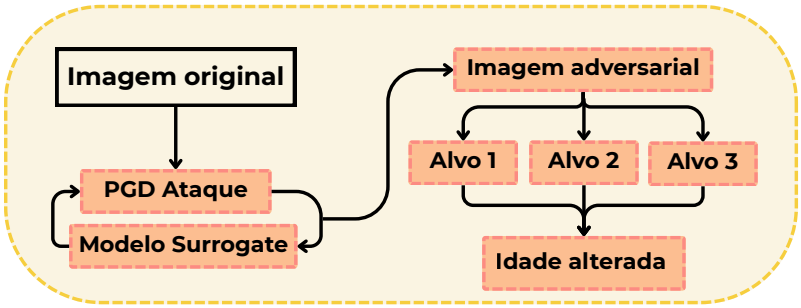



**Figura 1. Diagrama de funcionamento**

O ambiente utilizado para simular um ataque real foi estruturado para integrar diferentes frameworks de inferência. A geração dos vetores de ataque adversarial é centralizada no modelo substituto (*surrogate*) via PyTorch. Para a avaliação nos modelos alvos, o pipeline realiza a inferência direta utilizando as APIs nativas do *Hugging Face* para os modelos baseados em ViTs (ViT-UTK e MiVOLO v2), enquanto emprega o framework *deepface* [Serengil and Ozpinar 2020] especificamente como interface de execução para o modelo convolucional DEX. Essa arquitetura híbrida garante que cada modelo alvo seja avaliado sob suas condições originais e otimizadas de pré- processamento, garantindo que o ruído adversarial seja aplicado de forma correta. A Figura 2 exemplifica a influência do ataque com diferentes magnitudes de ruído na mesma imagem.

### 3.1. Modelos Alvos

Como alvos, foram selecionados três modelos de reconhecimento facial baseados em diferentes paradigmas e gerações de arquiteturas de aprendizado profundo. Os modelos, bem como os dados de treino, serão apresentados a seguir.




**Figura 2. Exemplo de imagem com diferentes perturbações.**

### 3.1.1. Deep EXpectation

O modelo Deep EXpectation (DEX) [Rothe et al. 2018] baseia-se na arquitetura convolucional profunda VGG-16, sendo um dos primeiros modelos da estimativa de idade. O modelo teve seus pesos ajustados sobre o dataset IMDB-WIKI, contendo mais de 500 mil imagens de rostos. Sob a perspectiva demográfica, a base de treinamento apresenta um viés herdado de fotos de celebridades da internet, sendo composta majoritariamente por indivíduos adultos e brancos [Kärkkäinen and Joo 2021], com baixa representatividade de crianças, idosos e de grupos étnicos minoritários. A distribuição por gênero do dataset correspondente (IMDB-Wiki) está detalhada na Tabela 1. No framework *deepface*, a predição adota uma classificação em 101 classes de idade seguida de um refinamento estatístico por valor esperado via *softmax*.

**Tabela 1. Distribuição por gênero do dataset IMDB-Wiki.**

| Gênero | Quantidade de Imagens | Proporção (%) |
|---|---|---|
| Masculino | 100.106 | 54,44 |
| Feminino | 83.781 | 45,56 |

### 3.1.2. MiVOLO v2

O MiVOLO v2 [Kuprashevich and Tolstykh 2023] baseia-se em aprendizado de representações *multi-input* com ViT para predição conjunta de idade e gênero. A versão adotada teve seus pesos otimizados de forma combinada sobre bases de dados de larga escala como IMDB-Clean, UTKFace e Open Images. Embora apresente maior resiliência a oclusões ao integrar características faciais e corporais (*dual-input*), os datasets de treinamento exibem assimetrias do ponto de vista demográfico, com predominância de adultos jovens, o que pode reduzir a acurácia e a robustez das fronteiras de predição nas faixas etárias extremas e em subgrupos étnicos minoritários [Kärkkäinen and Joo 2021]. A distribuição percentual de gênero e grupo étnico da base de fine-tuning UTKFace está detalhada na Tabela 2.

### 3.1.3. ViT-UTK

O ViT-UTK [Dosovitskiy et al. 2021] fundamenta-se na arquitetura *Vision Transformer* (ViT) com a variante base (*google/vit-base-patch16-224*) ajustada especificamente no

**Tabela 2. Distribuição demográfica (%) por gênero e etnia do dataset de fine-tuning UTKFace.**

| Dataset | Gênero | Branca (%) | Parda (%) | Amarela (%) | Preta (%) | Outras (%) |
|---|---|---|---|---|---|---|
| UTKFace | Masculino | 23,10 | 9,54 | 6,64 | 9,78 | 3,21 |
| | Feminino | 19,41 | 7,23 | 7,84 | 9,31 | 3,93 |

dataset UTKFace para predição conjunta de idade e gênero. Embora o dataset UTKFace (cerca de 20 mil imagens) forneça uma rotulação mais diversa com informações demográficas detalhadas, ele ainda reflete disparidades étnicas e etárias inerentes à coleta de dados, Tabela 2, gerando instabilidade na estimativa de idade nas margens dessas distribuições étnico-etárias.

### 3.2. Modelo Substituto (Surrogate)

O modelo substituto (*surrogate model*) adotado como fonte de extração de gradientes sob condições de *white-box* foi um ViT (variante ViT-B/16) [Dosovitskiy et al. 2021]. Diferentemente dos modelos alvos baseados em CNNs, o ViT divide a imagem de entrada em *patches* de $16 \times 16$ pixels e utiliza blocos de autoatenção (*self-attention*) global para capturar correlações espaciais e semânticas complexas. O modelo específico empregado foi pré-treinado e passou por *fine-tuning* sobre o *dataset* FairFace [Kärkkäinen and Joo 2021] (banco de dados com 108 mil imagens balanceadas entre 7 grupos étnicos) para estimativa etária em 9 classes etárias discretas. O uso de um ViT como *surrogate* para atacar redes CNN testa a transferibilidade adversarial sob discrepância estrutural de arquitetura. Durante os testes, as perturbações adversariais geradas no *surrogate* são aplicadas às imagens e avaliadas nos modelos alvos, simulando um atacante em cenário de caixa-preta (*black-box*) que interage com o sistema de verificação em produção unicamente por meio de consultas e observação dos *scores* de distância retornados.

### 3.3. Base de Dados Face Age

Para a condução dos experimentos de auditoria de segurança, utilizou-se a base de dados pública Face Age. A amostragem do estudo focou em indivíduos com idade superior a 10 anos, totalizando um conjunto de 836 indivíduos únicos, onde cada indivíduo é representado por uma única imagem facial frontal. Para mitigar vieses de distribuição e assegurar uma representação homogênea ao longo do espectro etário, aplicou-se um teto amostral de até 10 indivíduos por idade. O dataset é caracterizado por diversidade demográfica distribuída em 4 grupos étnicos autodeclarados: branca (511 indivíduos), parda (153 indivíduos), amarela (141 indivíduos) e preta (31 indivíduos).

A Figura 3 apresenta a demografia detalhada da base utilizada. O gráfico (a) exibe o histograma de idade, evidenciando uma distribuição uniforme ao longo do espectro etário, consequência direta do teto amostral. O gráfico (b) apresenta boxplots da idade por gênero, demonstrando que a base é balanceada. O gráfico (c) detalha a distribuição de idade por etnia, indicando que os grupos branca, parda e amarela possuem perfis etários semelhantes e amplos, enquanto o grupo de pessoas pretas, embora menor em termos absolutos, também cobre faixas etárias diversas. Por fim, o gráfico (d) exibe a distribuição agrupada por etnia e gênero, confirmando que a homogeneidade etária e o equilíbrio entre os gêneros se mantêm consistentes mesmo com a segmentação étnica.

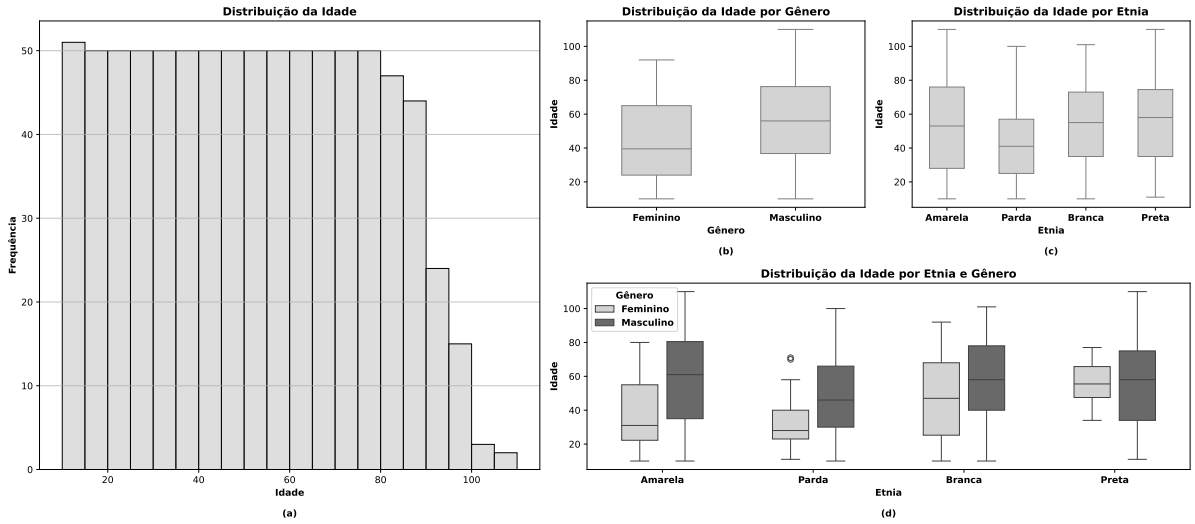



**Figura 3. Distribuição de idade, gênero e raça utilizados no ataque.**

### 3.4. Configuração do Ataque PGD

O ataque adversarial foi formulado utilizando o algoritmo *Projected Gradient Descent* (PGD), parametrizado sob a norma $L_\infty$ para assegurar a imperceptibilidade visual das perturbações ao olho humano. Os testes foram executados variando a perturbação máxima admissível ($\epsilon \in \{4/255, 8/255, 16/255, 24/255\}$) e o número de passos iterativos de otimização (steps $\in \{40, 100\}$). Ao aplicar essas variações sobre as imagens dos 836 indivíduos da base, obteve-se um total de 6.688 imagens adversariais. Estas imagens foram transferidas e avaliadas contra os três alvos, totalizando 20.064 simulações independentes para o mapeamento da robustez sistêmica.

## 4. Resultados e Discussão

Os experimentos revelaram que modelos de estimativa etária são vulneráveis à perturbação adversarial, fraudando a idade dos alvos com ruídos de diferentes magnitudes. Os desvios observados indicam que o ataque é capaz de comprometer gravemente a confiabilidade das predições etárias.

A Tabela 3 apresenta os desvios médios ($\Delta$) e absolutos ($|\Delta|$) de idade obtidos ao submeter as imagens perturbadas ao alvo em três experimentos independentes.

Para cenários em que o objetivo do sistema é mitigar o contato de menores de idade com maiores, os experimentos demonstraram que é possível tanto alterar a idade de um indivíduo do grupo de crianças para pertencer ao grupo de adultos quanto o contrário. Um comportamento observado é que o ataque nos alvos mostrou uma tendência a deslocar mais a idade de pessoas que estão em alguma extremidade etária, em um dos exemplos foi possível observar um $\Delta$ de 50 anos com a perturbação adversarial de menor magnitude do conjunto de testes, o que significa que um indivíduo poderia ficar significativamente mais novo ou mais velho.

Esse fenômeno é ilustrado de forma detalhada na Figura 4, que exibe a dispersão das idades reais (ordenadas de forma crescente em preto) em contraste com as idades hackeadas sob o ataque de maior intensidade ($\epsilon = 24/255$ em cinza) para os três modelos

**Tabela 3. Desvios médios e absolutos de idade sob diferentes limites de perturbação adversarial ($\epsilon$).**

| Alvo | Métrica | $\epsilon = 4/255$ | $\epsilon = 8/255$ | $\epsilon = 16/255$ | $\epsilon = 24/255$ |
|---|---|---|---|---|---|
| DEX | $\Delta$ Médio | -16,20 | -16,24 | -15,89 | -15,30 |
| | $\lvert\Delta\rvert$ Absoluto | 20,55 | 20,59 | 20,33 | 19,84 |
| MiVOLO v2 | $\Delta$ Médio | -1,47 | -1,13 | -0,52 | -0,19 |
| | $\lvert\Delta\rvert$ Absoluto | 4,26 | 4,41 | 4,78 | 5,11 |
| ViT-UTK | $\Delta$ Médio | -2,53 | -2,66 | -1,60 | -0,95 |
| | $\lvert\Delta\rvert$ Absoluto | 5,14 | 6,79 | 8,48 | 9,27 |

alvos. Nota-se que a perturbação, contra o modelo DEX, atua como uma espécie de "funil", forçando as predições de idades de quase toda a população a colapsarem em uma faixa central estreita (predominantemente entre 25 e 45 anos). Esse padrão visual explica por que jovens sofrem acréscimo etário significativo e indivíduos mais velhos sofrem reduções drásticas.

A Figura 4 também permite analisar os efeitos dos ataques em relação à maioridade. Observam-se casos em que indivíduos menores de 18 anos passam a ser classificados como adultos após a perturbação adversarial, bem como situações em que adultos são classificados como menores de idade. Ambos os cenários representam falhas críticas de segurança. Em aplicações reais, essas vulnerabilidades podem resultar, respectivamente, no acesso indevido de menores a conteúdos restritos ou na entrada de adultos em ambientes destinados ao público infantil.

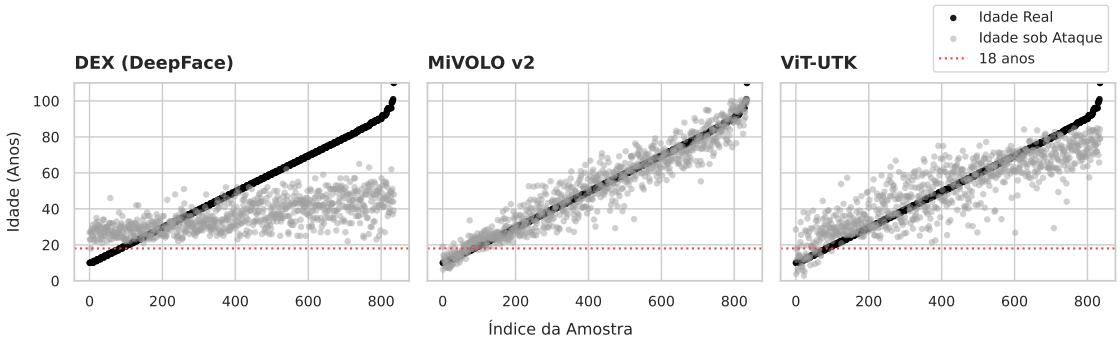



**Figura 4. Comparação entre as idades reais e as idades atacadas sob o ataque PGD com $\epsilon = 24/255$. A linha vermelha pontilhada representa a idade de 18 anos.**

Para quantificar os casos em que o ataque provoca uma mudança no status etário binário dos indivíduos, isto é, menores classificados como adultos e adultos classificados como menores, foi definida a métrica *Taxa de Inversão de Maioridade* (TIM), que mede a proporção de ocorrências em que a classificação do indivíduo cruza o limiar de maioridade em decorrência da perturbação adversarial. A Tabela 4 apresenta os valores de TIM, considerando especificamente os casos em que indivíduos menores de idade são classificados como adultos. De 81 indivíduos classificados como menores de idade, contra o modelo DEX o ataque fez com que 100% dos menores fossem classificados como

maiores, no ViT-UTK 55,56% e no MiVOLO v2 23,46%

**Tabela 4. TIM de maioridade por modelo alvo (menores de idade classificados como adultos).**

| Alvo | Total de Menores ($< 18$) | TIM (%) |
|---|---|---|
| DEX | 81 | 100,00(n=81) |
| ViT-UTK | 81 | 55,56(n=45) |
| MiVOLO v2 | 81 | 23,46(n=19) |

De forma análoga, a Tabela 5 apresenta a TIM para os casos em que indivíduos adultos foram classificados como menores de idade. No modelo DEX, não foram observadas ocorrências desse tipo, enquanto os modelos ViT-UTK e MiVOLO v2 apresentaram uma e três ocorrências, respectivamente. Em conjunto com os resultados da Tabela 4, os dados indicam uma assimetria nas inversões de maioridade, com maior frequência de casos em que indivíduos menores são classificados como adultos. Esse comportamento sugere que, no cenário adversarial considerado, é mais provável que menores sejam classificados como adultos do que o inverso.

**Tabela 5. TIM de maioridade por modelo alvo (adultos classificados como menores de idade).**

| Modelo Alvo | Total de Maiores ($\geq 18$) | TIM (%) |
|---|---|---|
| DEX | 755 | 0,00(n=0) |
| ViT-UTK | 755 | 0,13(n=1) |
| MiVOLO v2 | 755 | 0,40(n=3) |

A Figura 5 apresenta uma análise complementar baseada em uma definição binária de sucesso, na qual um ataque é considerado bem-sucedido quando $\Delta > \tau$. Dessa forma, é possível avaliar a robustez dos modelos sob diferentes critérios de sucesso. Sob a perturbação de menor intensidade ($\epsilon = 4/255$), observa-se uma região de maior impacto para limiares mais baixos ($\tau \leq 15$ anos) no modelo DEX, cuja taxa de sucesso atinge 54,1%, evidenciando sua maior vulnerabilidade. Em contrapartida, os modelos baseados em ViTs (MiVOLO v2 e ViT-UTK) demonstram elevada robustez, mantendo taxas de sucesso inferiores a 5% para $\tau \geq 15$ anos. Para todos os modelos, a eficácia do ataque decai à medida que o critério de sucesso se torna mais rigoroso, com o DEX reduzindo sua taxa para 8,5% em desvios superiores a 45 anos, enquanto os demais convergem para 0%.

De acordo com o benchmark realizado por [Ren et al. 2026], uma taxa de erro prevista para os modelos alvo é de $\Delta < 5$, portanto definir $\tau = 5$ mede a taxa de certeza real do trabalho. Além disso, a Figura 5 mostra a tendência à convergência da curva a partir de $\tau \geq 5$, o que torna a escolha de $\tau = 5$ apropriada para avaliar o sucesso dos modelos. A Figura 6 demonstra a taxa de sucesso do ataque considerando os diferentes modelos, categorizando os resultados por gênero e grupo étnico. Foi observado que, em termos de distribuição étnico-demográfica, o ataque foi consistentemente mais eficaz no conjunto de indivíduos negros. No modelo DEX, o sucesso atingiu 89,6% no público feminino e 95,5% no masculino; no ViT-UTK, alcançou 68,8% (feminino) e 63,5% (masculino); e no MiVOLO v2, obteve 47,9% (feminino) e 41,5% (masculino).

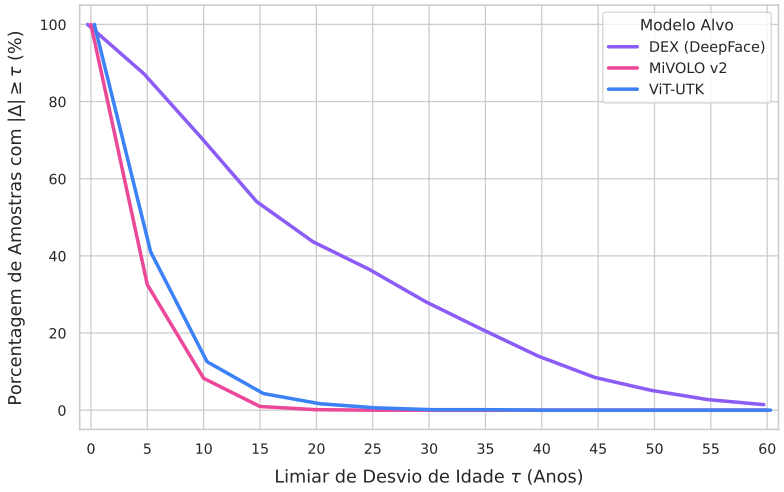



**Figura 5. Taxa de sucesso do ataque adversarial (%) em função do limiar de desvio de idade exigido ($\tau$ em anos) para o limite de perturbação $\epsilon = 4/255$.**

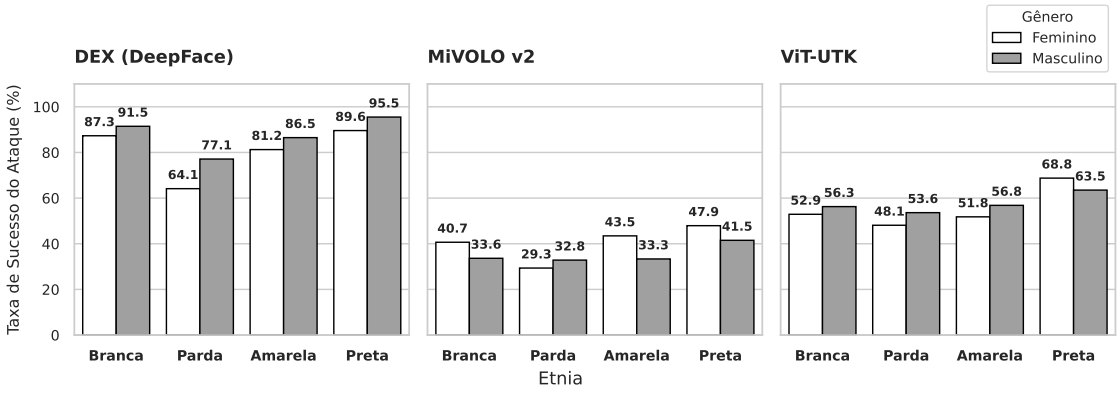



**Figura 6. Eficácia do ataque por gênero e grupo étnico assumindo sucesso para $|\Delta| \geq 5$ anos, segmentado para os modelos alvos DEX, MiVOLO v2 e ViT-UTK.**

Em contrapartida, as menores taxas de sucesso foram consistentemente registradas para o grupo de indivíduos pardos (Parda). Para o público feminino pardo, a taxa de sucesso foi de apenas 29,3% no MiVOLO v2, 48,1% no ViT-UTK e 64,1% no DEX. Entre os modelos, o DEX apresentou-se como o mais vulnerável em todas as categorias demográficas, apresentando taxas de sucesso sempre superiores a 64%. O modelo ViT-UTK exibiu uma vulnerabilidade intermediária (entre 48% e 69%), enquanto o MiVOLO v2 destacou-se como o mais robusto frente ao ataque de transferência, mantendo a taxa de sucesso do ataque abaixo de 48% em todas as segmentações.

A maior taxa de sucesso do ataque em grupos de pessoas negras pode ter relação com o processo de treinamento dos modelos utilizados como alvo. A literatura constatou que há um viés estrutural em modelos de reconhecimento facial, que são treinados majoritariamente em grupos de pessoas brancas e subtreinados em públicos de pessoas negras ou pardas [Grother et al. 2019], o que cria uma vulnerabilidade em sistemas de reconhecimento, que é transferida entre modelos [Papernot et al. 2016], demonstrando maior taxa de erro para indivíduos de pele escura [Buolamwini and Gebru 2018]. No entanto, ao analisar a influência demográfica dos conjuntos de *fine-tuning*, os resultados mostram-se inconclusivos. Para o modelo DEX, a base de dados utilizada não disponibiliza rótulos de raça, restringindo-se a gênero. Já para os demais modelos, as proporções das minorias étnicas (pretos, pardos e amarelos) apresentam um desvio padrão de apenas 2,31% entre si, o que não caracteriza uma diferença absoluta nos dados de ajuste capaz de justificar a variação na vulnerabilidade observada sob ataque.

## 5. Conclusão

Este trabalho apresentou uma análise crítica da confiabilidade de sistemas automáticos de estimativa etária facial, propondo um pipeline baseado em ataque adversarial por transferibilidade. A aplicação desse método em diferentes regimes de perturbação e passos iterativos constatou que ruídos imperceptíveis à visão humana são capazes de evadir sistemas de reconhecimento etário com desvio superior a 10 anos. Em cenários em que esses tipos de sistemas são responsáveis por proteger crianças segregando grupos etários, os resultados mostraram uma vulnerabilidade que pode ser explorada contra diferentes alvos, expondo os usuários desses sistemas a riscos.

A análise comparativa revelou que a vulnerabilidade é influenciada pelo paradigma arquitetural do modelo alvo. Arquiteturas baseadas em redes convolucionais profundas (DEX) mostraram-se mais vulneráveis ao ataque. Em contrapartida, modelos baseados em ViTs exibiram maior robustez. No MiVOLO v2, a integração de múltiplas modalidades de entrada apresentou-se como uma barreira de defesa eficaz, introduzindo redundâncias que reduzem a eficácia do ataque focado na face.

Como trabalhos futuros, propõe-se expandir a análise experimental por meio do teste de outras arquiteturas de modelos substitutos, da variação de parâmetros do ataque e da avaliação dos modelos sob cenários limpos para isolar o viés de predição nativo. Pretende-se também propor mecanismos de defesa e detecção específicos para a estimativa etária. A questão demográfica e o impacto racial devem ser abordados com maior profundidade, identificando contra-medidas que diminuam a vulnerabilidade contra públicos específicos. Dessa forma, buscaremos consolidar metodologias para a construção de modelos biométricos mais seguros em aplicações no mundo real.

## Referências


Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. (2020). Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer.

Biggio, B. and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Cinà, A. E., Rony, J., Pintor, M., Demetrio, L., Demontis, A., Biggio, B., Ayed, I. B., and Roli, F. (2025). AttackBench: Evaluating gradient-based attacks for adversarial examples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):2600–2608.

Croce, F. and Hein, M. (2020). Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., and Roli, F. (2019). Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 321–338.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. (2019). ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699.

Ding, S., Jia, S., Ma, C., Shen, C., Yang, X., Yao, T., and Yin, B. (2022). Adv-Attribute: Inconspicuous and transferable adversarial attack on face recognition. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, pages 34136–34147. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth $16\times16$ words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*.

Grother, P., Ngan, M., and Hanaoka, K. (2019). Face recognition vendor test (FRVT) part 3: Demographic effects. Technical Report NISTIR 8280, National Institute of Standards and Technology.

Kärkkäinen, K. and Joo, J. (2021). FairFace: Face image dataset for gender, race, and age balanced by race. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1548–1558.

Kuprashevich, M. and Tolstykh, I. (2023). MiVOLO: Multi-input transformer for age and gender estimation. In *Analysis of Images, Social Networks and Texts (AIST 2023)*, volume 14486 of *Lecture Notes in Computer Science*, pages 212–226. Springer.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*.

Pan, H., Han, H., Shan, S., and Chen, X. (2018). Mean-variance loss for deep age estimation from a face. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5285–5294.

Papernot, N., McDaniel, P., and Goodfellow, I. (2016). Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Qiao, Y., Sathyanarayana, N. B., Shi, C., He, Z., Wang, T., and Hou, T. (2026). A survey on adversarial machine learning: Attacks, defenses, real-world applications, and future research directions. *Neurocomputing*, 671:132670.

Ren, S., Shen, X., Raj, A., Dai, A., Zhang, C. M., Xu, Y., Chen, Z., Wu, S., Gong, C., and Zhang, Y. (2026). Out of the box age estimation through facial imagery: A comprehensive benchmark of vision-language models vs. out-of-the-box traditional architectures. arXiv preprint arXiv:2602.07815.

Rothe, R., Timofte, R., and Van Gool, L. (2018). Deep expectation of apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126:144–157.

Serengil, S. I. and Ozpinar, A. (2020). LightFace: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE.

Sousa, M., Viegas, E., Santin, A., and Geremias, J. (2025). Avaliação da Competência Parental na Gestão da Privacidade dos Filhos em Ambientes Digitais. In *Anais do XXV Simpósio Brasileiro de Cibersegurança*, pages 195–210, Porto Alegre, RS, Brasil. SBC.