DISSERTAÇÃO DE MESTRADO PROFISSIONAL

INFLUÊNCIA DO VIÉS RACIAL NO USO DE RECONHECIMENTO FACIAL APLICADO PARA CONTROLE DE ACESSO

ALEXANDRE MUNDIM DE OLIVEIRA

Programa de Pós-Graduação Profissional em Engenharia Elétrica

DEPARTAMENTO DE ENGENHARIA ELÉTRICA
FACULDADE DE TECNOLOGIA
UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA
Faculdade de Tecnologia

DISSERTAÇÃO DE MESTRADO PROFISSIONAL

INFLUÊNCIA DO VIÉS RACIAL NO USO DE RECONHECIMENTO FACIAL APLICADO PARA CONTROLE DE ACESSO

ALEXANDRE MUNDIM DE OLIVEIRA

Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia Elétrica como requisito parcial para obtenção do grau de Mestre em Engenharia Elétrica

Banca Examinadora

Luiz Antônio Ribeiro Júnior, Pós-doutor (IF/UnB) _____
Orientador

Rafael Rabelo Nunes, Doutor (FACE/ADM/UnB) _____
Examinador Interno

Alexandre de Barros Barreto, Doutor (GMU/EUA) _____
Examinador externo

William Ferreira Giozza, Doutor (ENE/UnB) _____
Examinador Suplente

FICHA CATALOGRÁFICA

DE OLIVEIRA, ALEXANDRE MUNDIM
INFLUÊNCIA DO VIÉS RACIAL NO USO DE RECONHECIMENTO FACIAL APLICADO PARA CONTROLE DE ACESSO [Distrito Federal] 2025.
xvi, 48 p., 210 x 297 mm (ENE/FT/UnB, Mestre, Engenharia Elétrica, 2025).
Dissertação de Mestrado Profissional - Universidade de Brasília, Faculdade de Tecnologia.
Departamento de Engenharia Elétrica

1. Reconhecimento Facial	2. Aprendizado de Máquina
3. Controle de Acesso	4. Viés Racial
I. ENE/FT/UnB	II. Título (série)

REFERÊNCIA BIBLIOGRÁFICA

DE OLIVEIRA, A.M. (2025). *INFLUÊNCIA DO VIÉS RACIAL NO USO DE RECONHECIMENTO FACIAL APLICADO PARA CONTROLE DE ACESSO*. Dissertação de Mestrado Profissional, Departamento de Engenharia Elétrica, Universidade de Brasília, Brasília, DF, 48 p.

CESSÃO DE DIREITOS

ALEXANDRE MUNDIM DE OLIVEIRA
Depto. de Engenharia Elétrica (ENE) - FT
Universidade de Brasília (UnB)
Campus Darcy Ribeiro
CEP 70919-970 - Brasília - DF - Brasil

AGRADECIMENTOS

A realização deste trabalho só foi possível graças ao apoio de muitas pessoas, às quais sou profundamente grato. Em primeiro lugar, agradeço à minha mãe, Maria Moisalina Mundim de Oliveira, aos meus filhos, Hugo Mundim de Abreu e Maria Helena Mundim Trancoso Amorim, e à minha esposa, Ana Carolina Trancoso Lopo de Amorim, especialmente a ela, por seu apoio constante e por ter sido minha principal incentivadora ao longo de todo o processo.

Sou imensamente grato aos professores do Programa de Pós-Graduação em Engenharia Elétrica da Universidade de Brasília: Rafael Rabelo Nunes, William Ferreira Giozza, Rafael Timóteo de Sousa Júnior, Demétrio Antônio da Silva e Alexandre Solon Nery. Em especial, agradeço ao professor Luiz Antônio Ribeiro Júnior, pela orientação dedicada durante o desenvolvimento deste trabalho e pela disponibilização dos recursos tecnológicos do laboratório do Instituto de Física da UnB, essenciais para a realização de testes e simulações.

Registro também minha gratidão aos colegas de turma: Fabrício Freire, pelas valiosas trocas de conhecimento em sala de aula, e a Tayná Gabriela, pelo eficiente suporte administrativo. Agradeço ainda ao colega Hugo Xavier Rodrigues, pelo apoio técnico com Python e pela contribuição na análise dos dados. E, por fim, aos amigos e professores da UnB, Alexandre Nascimento de Almeida e Flávio de Barros Vidal, pelos *insights* enriquecedores ao longo da jornada no programa.

RESUMO

O viés racial em tecnologias de reconhecimento facial, especialmente em aplicações de controle de acesso, é uma questão persistente. Este estudo analisa a adoção dessas tecnologias no contexto do aprendizado de máquina e sua integração com segurança da informação, cibersegurança e privacidade de dados. Essas soluções, apesar de amplamente utilizadas, frequentemente refletem desigualdades estruturais devido a conjuntos de dados e algoritmos enviesados, afetando desproporcionalmente grupos raciais sub-representados. A pesquisa combina revisão bibliográfica sistemática com simulações laboratoriais para avaliar como a composição dos conjuntos de dados afeta o desempenho de modelos de aprendizado de máquina em tarefas de reconhecimento facial. Os modelos foram treinados com bases de dados balanceadas e desbalanceadas, e os resultados confirmam que o uso de dados mais representativos pode reduzir o viés significativamente. Com base nessas evidências, foram formuladas 14 recomendações para mitigar o viés racial em sistemas de reconhecimento facial, incluindo a diversificação dos conjuntos de dados, o aumento da transparência dos algoritmos e a participação de equipes multidisciplinares nas decisões técnicas e éticas. Essas propostas estão alinhadas com os princípios do *NIST AI Risk Management Framework* (NIST AI 600), que orienta o desenvolvimento e uso responsável de sistemas de inteligência artificial, com foco em confiabilidade, equidade, explicabilidade e rastreabilidade. A adoção dessas medidas pode aprimorar a precisão e a equidade dos sistemas, promovendo maior confiança pública, especialmente em contextos sensíveis como o controle de acesso, e fornecer um roteiro para avançar a justiça algorítmica e a governança ética em IA.

Palavras-chave: Reconhecimento Facial; Aprendizado de Máquina; Inteligência Artificial; Controle de Acesso; Cibersegurança; Viés Racial; Racismo Algorítmico.

ABSTRACT

Racial bias in facial recognition technologies, especially in access control applications, is a persistent issue. This study analyzes the adoption of these technologies in the context of machine learning and their integration with information security, cybersecurity, and data privacy. These solutions, despite being widely used, often reflect structural inequalities due to biased datasets and algorithms, disproportionately affecting underrepresented racial groups. The research combines a systematic literature review with laboratory simulations to assess how the composition of datasets affects the performance of machine learning models in facial recognition tasks. The models were trained with balanced and unbalanced databases, and the results confirm that the use of more representative data can significantly reduce bias. Based on this evidence, 14 recommendations were formulated to mitigate racial bias in facial recognition systems, including diversifying datasets, increasing the transparency of algorithms, and the participation of multidisciplinary teams in technical and ethical decisions. These proposals align with the principles of the NIST AI Risk Management Framework (NIST AI 600), which guides the responsible development and use of artificial intelligence systems, with a focus on reliability, equity, explainability, and traceability. The adoption of these measures can improve the accuracy and fairness of the systems, promoting greater public trust, especially in sensitive contexts such as access control, and provide a roadmap to advance algorithmic justice and ethical governance in AI.

Keywords: Face Recognition; Machine Learning; Artificial Intelligence; Access Control; Cybersecurity; Racial Bias; Algorithmic Racism.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AI	Inteligência Artificial / Artificial Intelligence
CNN	Convolutional Neural Network (Rede Neural Convolucional)
FNMR	False Non-Match Rate (Taxa de Falsos Negativos)
FMR	False Match Rate (Taxa de Falsos Positivos)
GPU	Graphics Processing Unit (Unidade de Processamento Gráfico)
IAS	Inteligência Artificial Sensível
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
LGPD	Lei Geral de Proteção de Dados
MIT	Massachusetts Institute of Technology
NAC	Network Access Control (Controle de Acesso à Rede)
NIST AI 600	NIST AI Risk Management Framework
UTKFace	Conjunto de dados de imagens faciais
VGGFace2	Conjunto de dados de imagens faciais
Fairface	Conjunto de dados de imagens faciais
VGG16	Rede Neural Convolucional para classificar imagens
MobileNet	Rede Neural Convolucional para classificar imagens
ResNet50	Rede Neural Convolucional para classificar imagens

1 INTRODUÇÃO

Nesta seção são apresentadas a contextualização do tema investigado, a definição do problema de pesquisa, o objetivo geral, os objetivos específicos, a justificativa e a estrutura do trabalho.

1.1 CONTEXTUALIZAÇÃO DO TEMA

O reconhecimento facial se tornou uma tecnologia pervasiva na sociedade atual, sendo utilizada para diversas aplicações, como controle de acesso, segurança pública e autenticação de identidade (6). Essa afirmação se baseia nos conceitos discutidos por Harari sobre o impacto das tecnologias emergentes e sua crescente presença em diversas esferas da vida cotidiana. No entanto, pesquisas recentes têm demonstrado que sistemas de reconhecimento facial podem apresentar vieses raciais (1), o que significa que eles podem ter um desempenho inferior ao identificar pessoas de determinadas raças, especialmente de pele escura (3).

Esses vieses podem ter consequências graves, como discriminação e injustiças, além de reforçar estereótipos raciais e perpetuar desigualdades sociais (7). Na área de controle de acesso, por exemplo, o viés racial em sistemas de reconhecimento facial pode levar à negação indevida de acesso a locais e serviços, dificultar a locomoção de pessoas, negar autenticação e autorização em ambientes virtuais e até mesmo contribuir para a criminalização errônea de indivíduos (4).

1.2 PROBLEMA

O viés racial em sistemas de reconhecimento facial é um problema complexo que envolve diversos fatores, incluindo:

- Falta de diversidade nos dados de treinamento: Conjuntos de dados utilizados para treinar sistemas de reconhecimento facial compostos por imagens de pessoas brancas, pode levar os algoritmos a aprender características faciais associadas à branquitude e a ter um desempenho inferior ao identificar pessoas de outras raças (1).
- Algoritmos de reconhecimento facial que operam sem transparência em seus códigos dificultam a compreensão de como chegam a determinadas conclusões e, consequentemente, tornam complexa a identificação e mitigação de vieses intrínsecos (8).
- Falta de *accountability*: As empresas e instituições que desenvolvem e utilizam sistemas de reconhecimento facial nem sempre são responsabilizadas pelos vieses presentes em seus sistemas (9).

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo geral deste estudo é analisar a influência do viés racial no uso de reconhecimento facial aplicado para controle de acesso.

1.3.2 Objetivos Específicos

Para alcançar o objetivo geral, este estudo se propõe a:

1. Realizar revisão bibliográfica abrangente sobre o tema do viés racial em reconhecimento facial, incluindo suas causas, impactos e medidas de mitigação;
2. Desenvolver e implementar experimentos em laboratório para avaliar o desempenho de diferentes modelos de aprendizado de máquina em tarefas de reconhecimento facial;
3. Realizar o retreinamento (fine-tuning) de modelos de aprendizado de máquina previamente treinados, com o objetivo de adaptá-los à tarefa de classificação por etnia;
4. Analisar os resultados dos experimentos para identificar os modelos mais e menos propensos ao viés racial;
5. Com base nos resultados da revisão bibliográfica e dos experimentos, formular recomendações para mitigar o viés racial no uso de reconhecimento facial para controle de acesso. As recomendações formuladas nesta pesquisa estão alinhadas ao *NIST AI Risk Management Framework* (NIST AI 600), promovendo boas práticas na governança ética e responsável de sistemas de inteligência artificial.

1.4 JUSTIFICATIVA

Diante desse problema, é fundamental investigar as causas e os impactos do viés racial no reconhecimento facial e desenvolver medidas para mitigá-lo. Este estudo se propõe a contribuir para essa discussão por meio da revisão bibliográfica sobre o tema e da realização de experimentos em laboratório com diferentes modelos e dados de aprendizado de máquina. Além disso, as recomendações formuladas nesta pesquisa estão alinhadas ao *NIST AI Risk Management Framework* (NIST AI 600) (5), promovendo boas práticas na governança ética e responsável de sistemas de inteligência artificial.

1.5 PUBLICAÇÕES RESULTANTES DESTA PESQUISA

Fruto dessa pesquisa, foi publicado em 10/02/2025 na revista científica ***Research, Society and Development*****,** o artigo intitulado ***"Influence of racial bias in the use of facial recognition applied to access***

control: A critical analysis" (10), em que os autores evidenciaram, por meio de revisão bibliográfica, o viés racial no uso de sistemas de reconhecimento facial para controle de acesso.

1.6 ESTRUTURA

Este trabalho está organizado em seis capítulos, detalhando diversos aspectos da pesquisa sobre "Influência do viés racial no uso do reconhecimento facial aplicado para controle de acesso". O primeiro capítulo introduz o tema, problema, objetivos e justificativa da pesquisa. O segundo explora conceitos fundamentais sobre o viés racial, reconhecimento facial, controle de acesso, uso do reconhecimento facial para controle de acesso e o viés racial no uso do reconhecimento facial, estabelecendo uma base teórica. O terceiro capítulo descreve a metodologia de pesquisa utilizada, incluindo o fluxograma adotado e a descrição de cada etapa. O quarto capítulo descreve os resultados e discussões da investigação do viés realizada na literatura e nas simulações em laboratório dos modelos de aprendizado de máquina. O quinto capítulo apresenta as conclusões, as recomendações para mitigação do viés racial e sugestões para pesquisas futuras e, por último, o sexto capítulo descreve as considerações finais.

2 REVISÃO BIBLIOGRÁFICA

2.1 VIÉS RACIAL

O viés racial refere-se a tendências ou preconceitos inconscientes que afetam o tratamento e as decisões em relação a diferentes grupos raciais. Esses vieses podem manifestar-se de várias formas, incluindo estereótipos negativos, discriminação e tratamento desigual com base na raça (1).
As consequências do viés racial são devastadoras e afetam diversos aspectos da vida em sociedade, por exemplo, indivíduos de pele escura são frequentemente vítimas de discriminação, enfrentam barreiras no acesso a oportunidades e são desproporcionalmente submetidos à violência (11). Essa realidade reforça as desigualdades sociais existentes e perpetua um ciclo de exclusão e marginalização (12).

"o racismo é sempre estrutural, ou seja, [...] ele é um elemento que integra a organização econômica e política da sociedade. [...] é a manifestação normal de uma sociedade, e não é um fenômeno patológico ou que expressa algum tipo de anormalidade" (13).

2.2 RECONHECIMENTO FACIAL E O APRENDIZADO DE MÁQUINA

O reconhecimento facial utiliza algoritmos de aprendizado de máquina treinados em grandes conjuntos de dados de imagens rotuladas com a identidade de cada indivíduo (14). Através dessa análise, os modelos aprendem a identificar padrões e características faciais relevantes para o reconhecimento, e têm sido aplicados em várias áreas para analisar faces humanas por meio de câmeras, como segurança, monitoramento e autenticação biométrica (4). O reconhecimento facial baseia-se na extração de características faciais, como formato do rosto, posição dos olhos, nariz e boca, para criar representações matemáticas chamadas de "descritores faciais"(15), demonstrados na Figura 2.1.

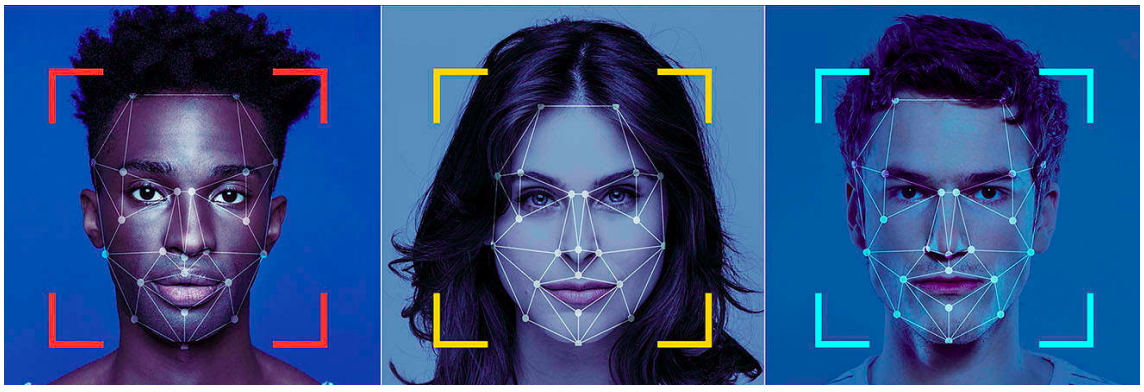

Figura 2.1: Extração da características faciais de imagem gerada por IA *Deepfake*.

Esses descritores são então comparados com os existentes em um banco de dados para identificar indivíduos ou verificar sua autenticidade. O modelo de processo da tecnologia de reconhecimento facial

envolve várias etapas para identificar e autenticar indivíduos com base em características faciais únicas, embora os detalhes específicos possam variar de acordo com as implementações e algoritmos utilizados.

O reconhecimento facial é um processo complexo que envolve várias etapas, desde a aquisição da imagem até a tomada de decisão final. Esse processo pode ser dividido em diversas fases que garantem a precisão e eficácia na identificação de indivíduos (16). As etapas incluem a aquisição de imagem, detecção facial, pré-processamento, extração de características, comparação com modelos existentes, correspondência e, por fim, a tomada de decisão, como ilustrado na Figura 2.2.

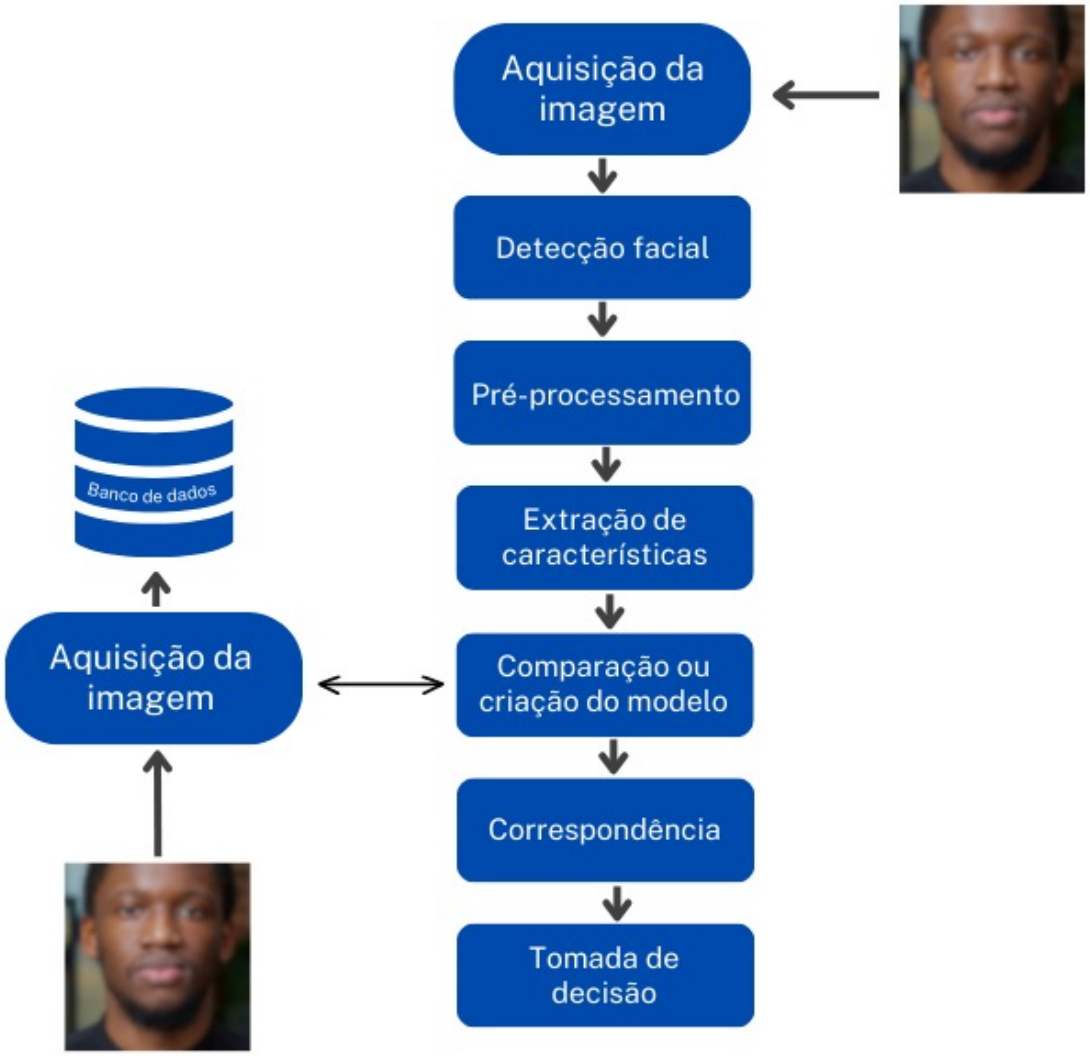

Figura 2.2: Modelo de processo padrão do reconhecimento facial.

Cada uma dessas etapas desempenha um papel crucial na formação de um sistema robusto de reconhecimento facial, que é amplamente utilizado em diversas aplicações, como segurança pública e controle de acesso. Nas seções que se seguem, o detalhamento de cada fase desse processo.

2.2.1 Aquisição de imagem:

O processo começa com a aquisição de uma imagem ou vídeo do rosto de uma pessoa. Isso pode ser feito usando câmeras de vigilância, câmeras em dispositivos móveis ou qualquer outra fonte de imagem.

2.2.2 Detecção facial:

A próxima etapa é a detecção facial, em que o sistema identifica a presença de um rosto na imagem ou vídeo. Isso envolve o uso de algoritmos de detecção de rosto que procuram características específicas, como olhos, nariz, boca e contornos faciais.

2.2.3 Pré-processamento:

Após a detecção facial, a imagem é pré-processada para melhorar a qualidade e a padronização. Isso pode incluir correções de iluminação, redimensionamento, alinhamento e normalização para garantir que as imagens estejam em um formato adequado para análise posterior.

2.2.4 Extração de características:

Nesta etapa, características faciais distintivas são extraídas da imagem, a fim de criar uma representação numérica única do rosto. Essas características podem incluir pontos de referência, como a posição dos olhos, nariz e boca, bem como padrões de textura e outras características discriminantes, como demonstrado na Figura 2.1.

2.2.5 Comparação ou criação de modelo:

Com base nas características extraídas, o sistema compara a imagem do rosto com uma base de dados existente contendo imagens e informações de identificação associadas a indivíduos conhecidos. Essa base de dados pode ser uma lista de funcionários, uma lista de suspeitos de crime, entre outras possibilidades.

2.2.6 Correspondência:

Nesta etapa, o sistema compara as características extraídas da imagem com as características armazenadas na base de dados e determina se existe uma correspondência significativa. Dependendo do cenário de uso, essa correspondência pode ser binária (sim / não) ou classificada em uma escala de confiança.

2.2.7 Tomada de decisão:

Com base na correspondência e na confiabilidade do sistema, uma decisão é tomada. Isso pode incluir a identificação do indivíduo, verificação de identidade ou acionamento de alarmes em caso de uma correspondência suspeita ou indesejada.

2.3 CONTROLE DE ACESSO

O controle de acesso se refere a um conjunto de medidas e procedimentos implementados para regular e gerenciar quem tem permissão para acessar determinados recursos, como locais físicos, sistemas de informação, dados confidenciais e outros ativos valiosos. O objetivo principal é proteger esses recursos contra acessos não autorizados, evitando roubos, sabotagens, espionagem e outras ameaças, é um dos pilares da segurança da informação e envolve três funções essenciais: **autenticação**, **autorização** e **auditoria** (17).

A **autenticação** consiste na verificação da identidade do usuário, geralmente realizada por meio de credenciais como senhas, cartões de acesso ou dados biométricos. O reconhecimento facial, nesse contexto, configura-se como um método de autenticação biométrica, ao utilizar características únicas do rosto para validar a identidade do indivíduo. Após a autenticação, ocorre a **autorização**, que define os recursos e serviços aos quais o usuário pode ter acesso, com base em regras e políticas previamente estabelecidas pelo sistema. A **auditoria**, por sua vez, corresponde ao registro e à análise das ações realizadas pelos usuários, sendo fundamental para fins de rastreabilidade, detecção de comportamentos anômalos e conformidade com normas e requisitos legais. Essas três funções atuam de forma integrada para assegurar que o acesso aos ativos da organização seja controlado, monitorado e compatível com os níveis de privilégio adequados.

2.3.1 Tipos de Controle de Acesso:

Existem diferentes tipos de controle de acesso, cada um com suas características e aplicações específicas, os mais comuns incluem:

1. **Controle de acesso físico:** Controla o acesso físico a locais, como edifícios, salas restritas e áreas protegidas. Isso pode ser feito através de portas com fechaduras, sistemas de biometria, cartões de acesso, seguranças e outros métodos (18);
2. **Controle de acesso lógico:** Controla o acesso a sistemas de informação, dados e recursos digitais. Isso pode ser feito através de senhas, autenticação multifator, *firewalls, VPNs* e outras medidas de segurança (17);
3. **Controle de acesso baseado em função:** Concede acesso a recursos com base nas funções ou cargos dos usuários. Isso significa que cada usuário tem acesso apenas aos recursos que são necessários para realizar seu trabalho (19);
4. **Controle de acesso baseado em atributos:** Concede acesso a recursos com base em atributos específicos dos usuários, como idade, localização, nível de segurança ou outros critérios (20).

2.4 USO DO RECONHECIMENTO FACIAL PARA CONTROLE DE ACESSO

O reconhecimento facial, tecnologia que emprega algoritmos para analisar características faciais humanas por meio de câmeras (21), encontrou ampla aplicação em diversos domínios, tais como segurança (22),

monitoramento (23) e autenticação biométrica (24). Seu uso como camada de controle de acesso desempenha papel fundamental em garantir os três pilares fundamentais da Segurança da Informação: integridade, disponibilidade e autenticidade (25). Esses princípios são essenciais para a segurança em ambientes físicos e lógicos, abrangendo desde a proteção de fronteiras (26) até a cibersegurança (27) e privacidade de dados (28). No Brasil, a Lei Geral de Proteção de Dados (LGPD) [Lei nº 13.709/2018] estabelece diretrizes para o tratamento de dados pessoais, incluindo dados biométricos como os utilizados em sistemas de reconhecimento facial. Essa lei impõe requisitos como o consentimento explícito do titular para o tratamento de seus dados biométricos, a limitação do uso para finalidades específicas e legítimas e a implementação de medidas de segurança para proteger os dados contra acessos não autorizados (29).

No setor bancário, por exemplo, instituições como o Hong Kong and Shanghai Banking Corporation (HSBC) adotaram essa tecnologia em seus aplicativos móveis (30), proporcionando aos clientes uma experiência bancária mais segura e conveniente, ao mesmo tempo em que combatem ativamente a fraude. Além disso, o reconhecimento facial é utilizado na esfera da cibersegurança para proteger o acesso a dispositivos e redes (31). A Microsoft, por exemplo, introduziu o *Windows Hello* no ecossistema do Windows 10, integrando autenticação biométrica por reconhecimento facial ao processo de login do sistema (32).

2.5 VIÉS RACIAL NO USO DO RECONHECIMENTO FACIAL

O reconhecimento facial surgiu como uma tecnologia transformadora no cenário contemporâneo, impulsionada pelo rápido avanço da visão computacional e do aprendizado de máquina (33). Essa abordagem se destaca como uma das áreas mais estudadas em visão computacional e reconhecimento de padrões, sendo aplicada em diversas áreas e setores (34). Entretanto, um dos desafios reconhecidos nesse campo é a questão do viés racial, que tem sido objeto de estudos e análises aprofundadas (1, 9, 3, 35, 36, 4, 37, 38, 8, 39). Esta é uma área de estudo em rápido crescimento, com o número de publicações e citações aumentando a cada ano, conforme demonstrado nas Figuras 2.3 e 2.4.

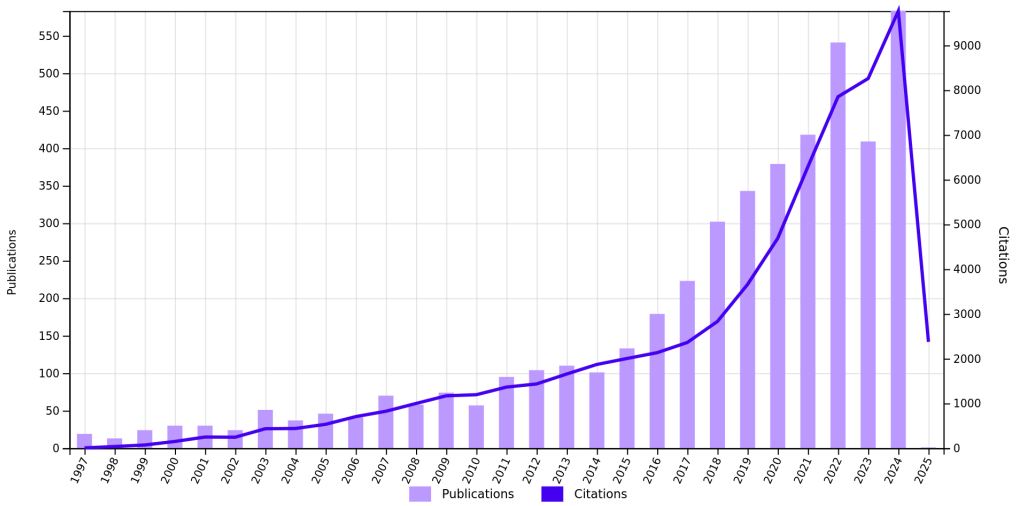

Figura 2.3: Evolução das publicações e citações, por ano, de 1997 a 2024, com títulos contendo o termo "*Facial Recognition*", extraídos do *Web of Science Report* em 9 de abril de 2025.

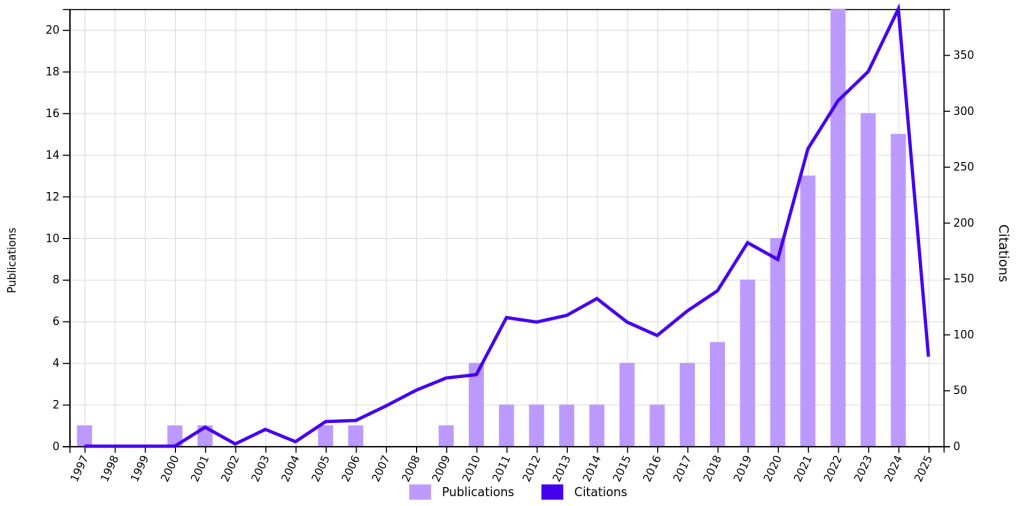

Figura 2.4: Evolução das publicações e citações, por ano, de 1997 a 2024, com títulos contendo os termos "*Facial Recognition*" e "*Racial Bias*", extraídos do Relatório da *Web of Science Report* em 9 de abril de 2025.

Em análise abrangente das disparidades de precisão em sistemas de classificação de gênero que utilizam tecnologia de reconhecimento facial (1). Seus resultados destacaram como grupos étnicos minoritários, especialmente mulheres com tons de pele mais escuros, eram frequentemente submetidos a classificações imprecisas, demonstradas na Figura 2.5.

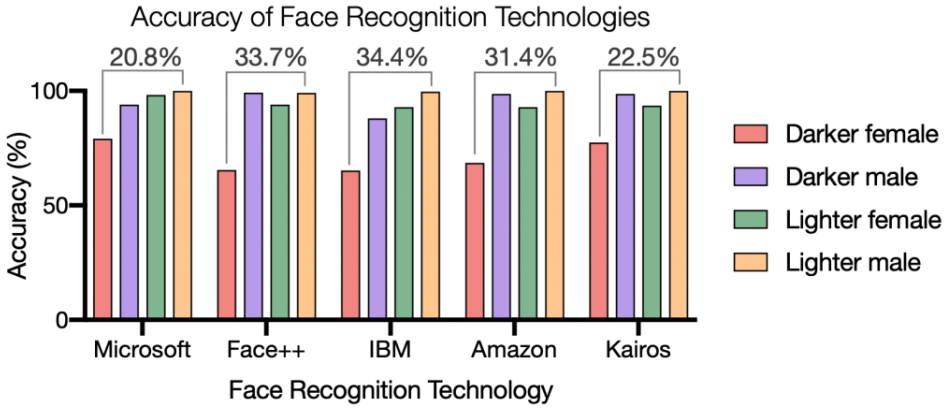

Figura 2.5: Em auditoria de cinco tecnologias de reconhecimento facial conduzida por (1), revelou discrepâncias na precisão de classificação dessas tecnologias para diferentes tons de pele e sexos. Consistentemente, os algoritmos demonstraram a menor precisão para mulheres com pele mais escura e maior precisão para homens com pele mais clara.

Este estudo iluminou os perigos inerentes de vieses e preconceitos algorítmicos, destacando a importância da justiça tecnológica. Além disso, reforçou a necessidade de adotar abordagens mais equitativas e inclusivas na pesquisa e implementação de sistemas de inteligência artificial.

Conforme descrito por (1), o viés racial se refere a tendências ou preconceitos inconscientes que influenciam o tratamento e as decisões em relação a diferentes grupos raciais. Esses vieses podem se manifestar de várias maneiras, incluindo estereótipos negativos, discriminação e tratamento desigual com base na raça. No contexto do reconhecimento facial, esses vieses podem ser incorporados aos sistemas devido a vários

fatores, como desequilíbrios nos conjuntos de dados de treinamento, algoritmos tendenciosos e falta de diversidade entre os desenvolvedores dessas tecnologias (1, 9, 3, 35, 36, 4, 15, 7, 37, 38, 8, 33, 39). Esse cenário levanta crescentes preocupações sobre a precisão e o potencial impacto discriminatório desses sistemas, bem como questões éticas e morais que devem ser consideradas no desenvolvimento da Inteligência Artificial, conforme preconizado pelo Instituto AI Now da Universidade de Nova York (9), para mitigar o "racismo algorítmico".

Vários estudos adicionais investigaram a influência do viés racial no reconhecimento facial, por exemplo (36), concluindo que existem disparidades significativas no desempenho dessas tecnologias com base na raça. Outro estudo (4), foi examinado a representação da raça nos conjuntos de dados usados para treinar algoritmos de reconhecimento facial e descobriu que a falta de diversidade nesses conjuntos pode contribuir para o viés. Conjuntos de dados públicos de imagens faciais existentes têm uma forte tendência a apresentar predominantemente rostos caucasianos, enquanto outras raças estão significativamente sub-representadas (3). Esse cenário leva a modelos treinados com classificação inconsistente, limitando a aplicabilidade dos sistemas de análise de reconhecimento facial a grupos raciais não caucasianos. Além disso, um estudo do MIT Media Lab (8) identificou que sistemas comerciais de reconhecimento facial exibiam taxas de erro mais altas ao identificar indivíduos com pele mais escura, e a falta de diversidade e representação nas equipes de desenvolvimento dessas tecnologias contribui para a perpetuação de injustiças (7).

Tecnologias modernas, incluindo algoritmos e inteligência artificial, podem perpetuar preconceitos raciais e desigualdades sociais, argumenta-se que a tecnologia não é neutra e muitas vezes reflete e amplifica preconceitos que já existem na sociedade (37). O estudo defende um movimento de "abolicionismo tecnológico" à medida que novas tecnologias emergem para desmantelar estruturas discriminatórias embutidas nessa tecnologia. Além disso, enfatiza a importância da inclusão e diversidade no desenvolvimento tecnológico, destacando que a falta de diversidade entre desenvolvedores e tomadores de decisão contribui para o viés racial e a discriminação em algoritmos e sistemas automatizados.

O algoritmo COMPAS, usado para tomar decisões judiciais nos Estados Unidos, exibia um viés racial substancial (38). O algoritmo previu desproporcionalmente que criminosos negros eram mais propensos a reincidir do que criminosos brancos, mesmo quando as características dos casos eram semelhantes. Essa tendência levanta preocupações sobre como esses algoritmos podem influenciar sentenças e tempos de prisão, potencialmente exacerbando disparidades raciais no sistema de justiça criminal. A análise também destaca a falta de transparência em relação ao funcionamento do COMPAS. Os autores encontraram dificuldades para obter informações detalhadas sobre como o algoritmo foi desenvolvido e como realizava suas previsões, o que levanta questões sobre transparência, responsabilidade e prestação de contas no uso desses sistemas para a sociedade.

2.6 GERENCIAMENTO DE RISCOS EM INTELIGÊNCIA ARTIFICIAL

A crescente adoção de tecnologias baseadas em inteligência artificial (IA), como o reconhecimento facial, tem despertado preocupações sobre confiabilidade, equidade, transparência e mitigação de riscos algorítmicos. Nesse contexto, o NIST *AI Risk Management Framework* (AI RMF 1.0), publicado oficial-

mente em janeiro de 2023 pelo *National Institute of Standards and Technology* (NIST) (5), constitui uma referência normativa essencial para o desenvolvimento e a implementação ética de sistemas de IA.

O *framework* foi estruturado para auxiliar organizações públicas e privadas no mapeamento, medição e mitigação dos riscos associados a sistemas de IA em todo o seu ciclo de vida, Tabela 2.1. Ele é composto por dois componentes principais: Core (Núcleo) e Profiles (Perfis), com ênfase nos quatro atributos de sistemas de IA confiáveis: Valiosos, Confiáveis, Justos e Responsáveis.

Componente	Descrição
Mapear	Identificar contextos organizacionais, riscos associados e partes interessadas envolvidas nos sistemas de IA.
Medir	Quantificar e qualificar impactos e incertezas relacionadas ao uso de IA, incluindo métricas de desempenho, segurança e viés.
Gerenciar	Implementar controles técnicos, administrativos e organizacionais para reduzir os riscos identificados.
Governar	Estabelecer políticas, responsabilidades, auditorias e prestação de contas para garantir governança eficaz e uso responsável da IA.

Tabela 2.1: Componentes do NIST *AI Risk Management Framework*.

A avaliação dos riscos relacionados ao viés racial nos sistemas de reconhecimento facial pode ser realizada também através de metodologias como a ISO/IEC 23894 (40), que fornece diretrizes para a gestão de riscos de tecnologias de inteligência artificial. Essa norma propõe um processo de gerenciamento de riscos que envolve a identificação, análise, avaliação e tratamento dos riscos (40).

No contexto do reconhecimento facial, esses riscos podem ser traduzidos em métricas como a taxa de falsos positivos (FMR) e a taxa de falsos negativos (FNMR) para diferentes grupos raciais, permitindo uma avaliação quantitativa do potencial de discriminação do sistema (4). Além disso, a ISO/IEC 23894 enfatiza a importância de considerar o contexto de uso do sistema de IA e as potenciais consequências negativas para os indivíduos afetados, o que é crucial na análise do impacto social e ético do viés racial.

2.7 LEI GERAL DE PROTEÇÃO DE DADOS (LGPD)

No Brasil, a Lei Geral de Proteção de Dados (LGPD) [Lei nº 13.709/2018] estabelece diretrizes para o tratamento de dados pessoais, incluindo dados biométricos como os utilizados em sistemas de reconhecimento facial. Esta dissertação se alinha aos princípios da LGPD ao propor medidas de mitigação do viés racial, que podem levar a um tratamento discriminatório e, portanto, inadequado dos dados dos indivíduos.

Nesse contexto, é essencial estudar o impacto do viés racial nos sistemas de reconhecimento facial para melhorar a eficiência dessas aplicações e eliminar esse viés prejudicial. Este estudo tem como objetivo examinar a ampla adoção dessas tecnologias na era do aprendizado de máquina, destacando sua integração nos *frameworks* de segurança da informação, cibersegurança e privacidade de dados. Por meio de uma investigação das abordagens de inteligência artificial voltadas para análise e engenharia de dados.

3 METODOLOGIA

3.1 INVESTIGAÇÃO DO VIÉS RACIAL NO USO DO RECONHECIMENTO FACIAL E ETAPAS DO PROCESSO.

Este trabalho envolveu uma análise criteriosa e profunda por meio da revisão de estudos científicos recentes que exploram a problemática do viés racial no uso do reconhecimento facial. Inicialmente, a busca por publicações foi realizada na base de dados do *Web of Science Report* utilizando os seguintes critérios de filtragem: período de publicação de 1º de janeiro de 1997 a 31 de dezembro de 2024, e as expressões "*Facial Recognition*" e "*Racial Bias*" em todos os campos. Essa busca resultou em 108 publicações, Figura 2.4. O refinamento da busca nos títulos e resumos dessas 108 publicações permitiu que a seleção fosse reduzida a 28 artigos. A seleção considerou publicações com o maior número de citações como indicador de relevância científica, e a leitura completa desses artigos resultou na seleção final de 12 trabalhos; a seleção final de artigos foi baseada na aderência ao objetivo da pesquisa: investigar o viés racial no uso do reconhecimento facial em sistemas de controle de acesso. Além disso, incluímos relatórios de organizações e instituições de destaque, como o MIT Media Lab (8), ACLU (41), Electronic Frontier Foundation (42) e NIST (4), que conduziram investigações sobre os impactos do viés nessa tecnologia. Consideramos também o trabalho que avaliou a precisão e o desempenho de sistemas de reconhecimento facial em diversos grupos raciais (3), o resultado desse processo é apresentado na Tabela 4.1. Posteriormente, foram executadas simulações em laboratório para analisar a acurácia de modelos de aprendizado de máquina e os dados coletados submetidos a uma análise abrangente, com foco na identificação de viés racial nos sistemas de reconhecimento facial. Avaliamos métricas de desempenho, como taxas de falsos positivos e falsos negativos, em diferentes grupos raciais, etapa descrita no Capítulo 3.3. Além disso, investigamos as causas subjacentes do viés, incluindo a falta de diversidade nos conjuntos de dados de treinamento e as abordagens de desenvolvimento de algoritmos, ou seja, na análise e engenharia dos dados. Por fim, foi possível propor 14 recomendações com a finalidade de obter impactos positivos ao mitigar o viés racial no uso de sistemas de reconhecimento facial, ver Tabela 4.5, e alinhá-las aos princípios do *NIST AI Risk Management Framework* (NIST AI 600) (5). Abaixo, demonstração do processo em fluxograma, Figura 3.1, e a descrição das etapas realizadas nesta pesquisa:

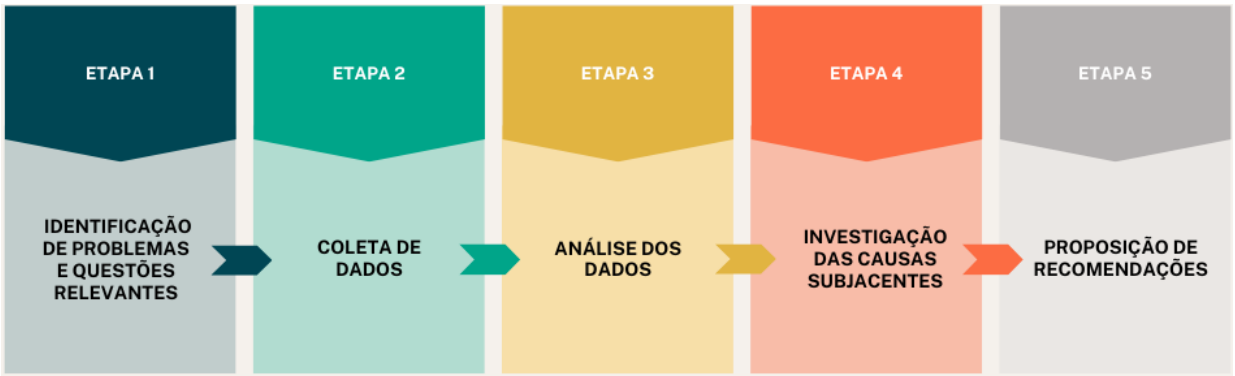

Figura 3.1: Fluxograma do processo com as etapas realizadas neste trabalho.

3.1.1 Etapa 1: Identificação de problemas e questões relevantes:

Inicialmente, foram identificadas as questões relacionadas ao reconhecimento facial e ao viés racial. Feito por meio da revisão da literatura científica, relatórios de organizações relevantes e pesquisas anteriores;

3.1.2 Etapa 2: Coleta de dados:

Foram coletados dados relevantes que abordam o viés racial em sistemas de reconhecimento facial. Isso incluiu estudos científicos, artigos, relatórios de organizações e dados de pesquisa disponíveis;

3.1.3 Etapa 3: Análise dos dados:

Os dados coletados foram analisados em profundidade. Isso envolveu a identificação de métricas de desempenho, como taxas de falsos positivos e falsos negativos, em diferentes grupos raciais;

3.1.4 Etapa 4: Investigação das causas subjacentes:

Além de identificar o viés, foi crucial investigar as suas causas subjacentes. Isso incluiu, por meio da revisão bibliográfica, Capítulo 2, a análise de como os algoritmos de reconhecimento facial são desenvolvidos, os conjuntos de dados de treinamento utilizados e as técnicas de engenharia de dados aplicadas; e,

3.1.5 Etapa 5: Proposição de recomendações:

Identificadas as causas subjacentes, foi possível propor recomendações específicas para mitigar o viés racial em sistemas de reconhecimento facial e correlacioná-las com os princípios do *NIST AI Risk Management Framework* (NIST AI 600) (5).

3.2 METODOLOGIA DA REVISÃO SISTEMÁTICA

A revisão da literatura foi conduzida com o objetivo de identificar e analisar criticamente os estudos existentes sobre o viés racial em sistemas de reconhecimento facial. No entanto, reconhece-se que a abordagem metodológica adotada apresenta algumas limitações.

Embora a seleção dos artigos tenha seguido critérios de relevância e qualidade, não foi estabelecido um protocolo formal e explícito para a revisão, como os preconizados por diretrizes como PRISMA (43) *[Preferred Reporting Items for Systematic Reviews and Meta-Analyses]* ou PICOC (44) *[Population, Intervention, Comparison, Outcome, Context]*. A ausência de um protocolo predefinido pode introduzir um grau de subjetividade no processo de seleção e análise dos estudos.

Além disso, a apresentação dos artigos revisados se deu de forma narrativa, sem a inclusão de uma tabela sistemática que sintetize informações cruciais como autores, ano de publicação, foco da pesquisa, metodologia empregada e principais resultados. A inclusão de uma tabela desse tipo poderia aprimorar a transparência e a reprodutibilidade da revisão.

Em futuras pesquisas, recomenda-se a adoção de protocolos formais de revisão sistemática e a apresentação dos estudos em formato de tabela, a fim de fortalecer a rigorosidade e a transparência da metodologia.

3.3 VIÉS RACIAL EM MÉTODOS DE APRENDIZADO DE MÁQUINA.

3.3.1 Seleção dos modelos de aprendizado de máquina

Para investigar a existência do viés racial no uso do reconhecimento facial, selecionamos três modelos de aprendizado de máquina: VGG16, MobileNet e ResNet50. Esses modelos são amplamente reconhecidos na literatura devido à sua arquitetura robusta e capacidade de generalização em tarefas de classificação de imagens, onde, através do uso do *active learning* é possível retreinar os modelos para novas tarefas como o pretendido neste trabalho.

Abaixo temos uma breve descrição de cada um dos modelos que foram utilizados para a classificação de imagens por etnia, a fim de constatar o viés racial dos modelos com respeito aos conjuntos de dados selecionados para o treinamento e aos ajustes de cada modelo.

3.3.2 VGG16

O VGG16 é conhecido por sua profundidade e capacidade de generalizar bem para uma grande gama de tarefas e conjuntos de dados, o que o torna ideal para uso com *active learning* na classificação de imagens (45). Originalmente treinado com imagens do ImageNet para participar do *ImageNet Large-Scale Visual Recognition Challenge* (ILSVRC) de 2014, onde obteve o primeiro e segundo lugares, o VGG16 consegue classificar imagens em 1.000 categorias distintas. Além disso, o VGG16 compreende uma rede neural com 16 camadas de peso, incluindo 13 camadas convolucionais e 3 camadas totalmente conectadas, com pequenos campos receptivos de 3x3, que permitem uma redução na quantidade de parâmetros e tornam a função de descrição mais discriminativa, facilitando a tarefa de classificação (45).

As imagens de entrada desse modelo devem estar em RGB e ter resolução de 224x224, não sendo necessário outro tipo de tratamento além da subtração da média dos valores RGB no tensor de entrada, o que agiliza o processo de preparação para o treinamento do modelo (45). A estrutura de rede do VGG é mostrada na Figura 3.2.

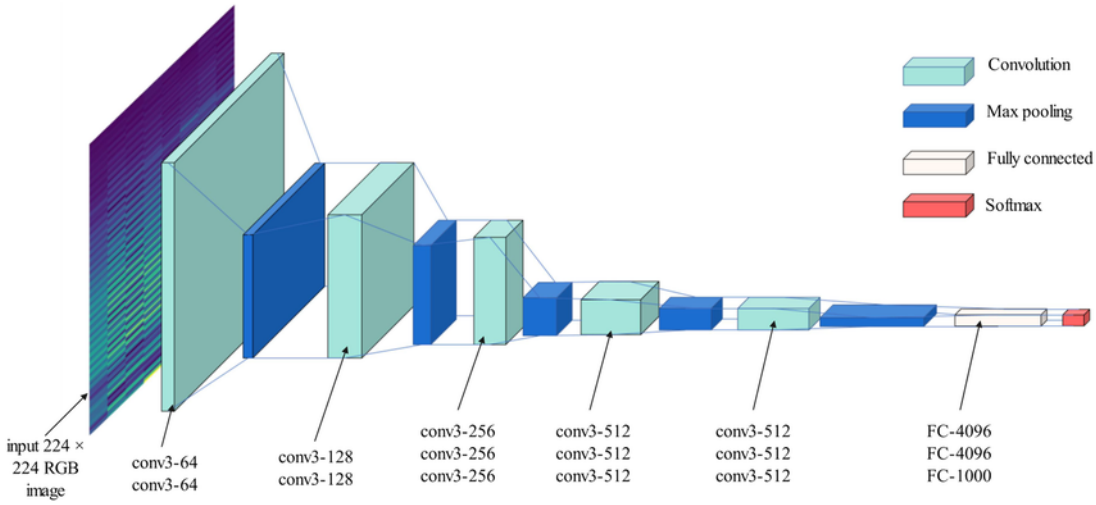

Figura 3.2: Estrutura da rede neural convolucional VGG (2).

Assim, o VGG16 se torna um modelo de fácil uso e eficiente para a tarefa de classificação, embora precise ser adaptado para a classificação de cada um dos conjuntos de dados analisados.

3.3.3 MobileNet

Os modelos MobileNet oferecem uma solução eficiente para dispositivos móveis e sistemas embarcados, permitindo previsões adaptadas a cenários do mundo real, onde os recursos computacionais são limitados (46). Também treinados com os dados do ImageNet, classificando imagens em 1.000 categorias, esses modelos utilizam camadas convolucionais separáveis em profundidade, com o objetivo de reduzir o custo computacional e o tamanho do modelo, mantendo uma acurácia próxima à de modelos como o VGG16 (46). Essas camadas separáveis são resultados da fatoração de uma camada convolucional completa em duas: uma convolução em profundidade e uma convolução ponto a ponto. A primeira aplica um único filtro a cada canal de entrada, enquanto a segunda, que sempre segue a primeira, combina as saídas da convolução em profundidade, realizando a mesma tarefa da camada convolucional completa em duas etapas (46).

Essa estrutura permite que o MobileNet reduza o custo computacional em cerca de 8 a 9 vezes em comparação com outros modelos. Além disso, o MobileNet possibilita o uso dos multiplicadores de largura, que reduzem o número de canais de entrada e saída de cada camada convolucional através de um fator α, e dos multiplicadores de resolução, que diminuem a resolução das imagens de entrada, originalmente 224x224 em RGB, assim como no VGG16 (46).

Essas reduções têm a capacidade de melhorar a eficiência do treinamento, permitindo o uso de lotes maiores sem redução drástica na performance. Para classificações, como a de faces, o MobileNet consegue manter uma precisão semelhante, mesmo com o uso das reduções aplicadas pelos multiplicadores (46).

Além disso, por ser um modelo com poucos parâmetros, o MobileNet não exige muitos regularizadores ou aumento de dados para evitar o sobreajuste (46).

Portanto, o MobileNet é uma alternativa eficiente para a classificação de imagens faciais, alcançando

boa performance com baixo custo computacional, podendo ser utilizado para a identificação de vieses raciais, desde que ajustado para a classificação conforme a base de dados fornecida.

3.3.4 ResNet50

O modelo ResNet50, cuja versão inicial venceu o ILSVRC de 2015, introduz as redes residuais, projetadas para mitigar o problema do gradiente em redes profundas, quando ocorre degradação na acurácia do treinamento devido a uma grande quantidade de camadas encadeadas (47).

Para isso, essas redes neurais utilizam resíduos das camadas anteriores no cálculo do gradiente nas camadas posteriores, usando "conexões de atalho"nas camadas convolucionais que permitem uma suavização na flutuação do gradiente, proporcionando desempenho superior em tarefas complexas, como reconhecimento facial. Essa abordagem permite um encadeamento maior de camadas, o que possibilita uma análise mais detalhada das características de uma imagem (47).

O ResNet50, por exemplo, apresenta 50 camadas convolucionais com aprendizado residual, aceitando imagens de 224x224 em RGB, como nos outros modelos mencionados. Embora seja naturalmente mais custoso computacionalmente que o MobileNet, o ResNet50 é significativamente mais rápido que o VGG16, pois utiliza uma quantidade menor de parâmetros e permite uma convergência mais rápida (47).

Além disso, o ResNet50, que também foi treinado com os dados do ImageNet, apresentou uma acurácia ligeiramente superior à do VGG16 na classificação dessas imagens, tornando-se uma ótima referência para comparação. Seu encadeamento de múltiplas camadas convolucionais pode ajudar na caracterização de imagens faciais, possibilitando a detecção de vieses raciais no treinamento com os conjuntos de dados selecionados.

3.4 SELEÇÃO DOS CONJUNTOS DE DADOS PARA SIMULAÇÕES

Além dos modelos, os conjuntos de dados selecionados para as simulações — VGGFace2, UTKFace e FairFace — são cruciais para uma análise abrangente do viés racial. Esses dados permitem verificar a relação entre a precisão do modelo na classificação e o percentual de imagens de cada etnia na base de dados, além de observar, por meio dos falsos positivos, como os modelos podem classificar erroneamente imagens de etnias distintas devido a características semelhantes.

Abaixo, encontra-se uma breve descrição de cada um desses conjuntos de dados, com o percentual de cada etnia e a forma como as imagens são categorizadas. Esses conjuntos foram escolhidos devido à disparidade no número de imagens de cada etnia, com exceção de um deles, permitindo a análise do impacto do viés racial na categorização de imagens faciais.

3.4.1 VGGFace2

O VGGFace2 fornece uma variedade de imagens de rostos de pessoas com diferentes idades (48), embora inicialmente não tivesse rótulos de etnia, com o VMER fornecendo essa categorização. Esse

conjunto de dados é composto por mais de 3 milhões de imagens de 9.129 indivíduos distintos, com 8.631 disponíveis para este estudo (48). Para garantir uma comparação razoável com os outros conjuntos de dados, as imagens do VGGFace2 foram filtradas, selecionando-se cinco imagens aleatórias de cada indivíduo, resultando em um conjunto final de 43.065 imagens.

Essas imagens foram categorizadas, através do rótulo do VMER, em quatro grupos étnicos e raciais: *African American*, *Asian Indian*, *Caucasian Latin*, e *East Asian*. A diversidade deste conjunto de dados, conforme definido pelo VMER, é detalhada na Tabela 3.1 (49).

Etnia	African American	Asian Indian	Caucasian Latin	East Asian
Porcentagem	8.25%	6.14%	79.43%	6.18%

Tabela 3.1: Percentagem racial das imagens VGGFace

A partir da Tabela 3.1, observa-se uma predominância significativa de imagens da categoria *Caucasian Latin*, que representam $79{,}43\%$ do total, contrastando com $8{,}25\%$ de *African American*, $6{,}14\%$ de *Asian Indian* e $6{,}18\%$ de *East Asian*. Essa distribuição evidencia um forte desequilíbrio na representatividade étnico-racial do conjunto de dados, com uma diferença percentual de mais de 71% entre o grupo majoritário e os demais. Tal desproporção não afeta apenas os indivíduos afro-americanos, mas também indica viés contra grupos asiáticos e indianos, que aparecem de forma igualmente sub-representada. Essas imagens, que variavam em resolução, foram ajustadas para atender aos padrões dos modelos e ficarem similares às dos outros conjuntos de dados, visando uma análise consistente.

3.4.2 UTKFace

O UTKFace inclui anotações de idade, gênero e raça, permitindo a avaliação do desempenho do modelo em diferentes grupos demográficos (50). Além disso, este conjunto de dados permite a categorização de imagens em cinco grupos étnicos e raciais: *Black*, *White*, *Asian*, *Indian* e *Others* (51). Assim, possui uma categoria a mais que o VGGFace2, com uma classificação um pouco distinta, mas onde é possível vincular as diferentes categorias: *White* pode ser associado a *Caucasian Latin*, *Black* a *African American*, *Indian* a *Asian Indian* e *Asian* a *East Asian*.

Além disso, o UTKFace apresenta um rótulo adicional, *Others*, que, embora represente uma minoria na base de dados, permite algum tipo de classificação para indivíduos que não se enquadram em nenhuma dessas categorias. O UTKFace, disponível em <https://susanqq.github.io/UTKFace/>, contém 24.125 imagens com resoluções variadas, que foram padronizadas para a categorização neste estudo. A diversidade étnica e racial do UTKFace é detalhada na Tabela 3.2 (52).

Etnia	Asian	Black	Indian	Others	White
Porcentagem	14.86%	18.89%	16.69%	7.09%	42.45%

Tabela 3.2: Percentagem racial das imagens UTKFace

Na Tabela 3.2, observamos que há uma distribuição mais equilibrada entre cada uma das categorias,

embora ainda haja uma grande discrepância entre a quantidade de imagens de pessoas brancas em relação às outras, tornando, portanto, esta base de dados enviesada, embora menos que o VGGFace2.

3.4.3 FairFace

O FairFace, por sua vez, foi especificamente criado para promover uma representação balanceada de atributos raciais e de gênero, oferecendo uma base sólida para testar e mitigar preconceitos em modelos de reconhecimento facial (3). Por conta disso, o FairFace traz uma maior margem de categorização, com o conjunto de dados possuindo sete grupos étnicos e raciais distintos, classificados como: *Black*, *White*, *Latino/Hispanic*, *Indian*, *East Asian*, *Southeast Asian* e *Middle Eastern*.

Este conjunto de dados possui 86.744 imagens, mais do que as filtradas nos outros dois conjuntos anteriores, todas na resolução de 224x224, ideal para uso nos modelos mencionados anteriormente (3). As porcentagens de diversidade étnica e racial neste conjunto de dados estão detalhadas na Tabela 3.3.

Etnia	Black	East Asian	Indian	Latino Hispanic
Porcentagem	14.10%	14.16%	14.20%	15.10%
	Middle Eastern	Southeast Asian	White	
	10.62%	12.44%	19.05%	

Tabela 3.3: Percentagem racial das imagens FairFace

Na Tabela 3.3, observa-se que há equidade entre as categorias, diferente dos conjuntos de dados anteriores, onde a maior discrepância percentual não ultrapassa os 9%. A partir disso, é possível verificar se os modelos, ao serem treinados com esta base, conseguem atingir uma performance melhor que os outros conjuntos de dados. A resposta a essa indagação levará à constatação do viés na identificação dos indivíduos por etnia e raça.

3.5 PARAMETRIZAÇÃO E PREPARAÇÃO DOS MODELOS

O *fine-tuning*, técnica de *transfer learning* que ajusta os pesos de um modelo pré-treinado em um novo conjunto de dados (53), é essencial para adaptar os modelos de reconhecimento facial a características específicas dos dados de controle de acesso. Esse ajuste permite refinar a capacidade do modelo de extrair e reconhecer características faciais relevantes para a identificação de indivíduos, ao mesmo tempo em que auxilia na mitigação do viés racial ao adaptar o modelo às variações étnicas presentes nos dados de treinamento (54). Além disso, o *fine-tuning* permite otimizar o modelo para a tarefa específica, reduzindo significativamente o custo computacional em comparação com o treinamento do zero, o que o torna uma alternativa viável para aplicações em larga escala (55).

Para a tarefa de classificação de imagens de rostos, foi realizado o procedimento de *active learning* com *fine-tuning*, no qual um modelo já treinado, utilizando um conjunto de dados diverso, é retreinado (podendo retreinar todas ou apenas algumas camadas) para um fim específico, que neste caso é a categorização das

imagens de faces por etnia.

Esse método foi escolhido uma vez que o uso combinado de *active learning* com *fine-tuning* é altamente justificado na tarefa de identificação facial, por aliar eficiência computacional, melhor uso de dados, capacidade de adaptação contínua e redução de viés. Em comparação com métodos convencionais, essa abordagem oferece um equilíbrio ideal entre custo, desempenho e robustez em cenários reais e em constante mudança (56, 57).

Neste trabalho, foi realizado um *fine-tuning* em cada um dos modelos já mencionados, removendo algumas camadas finais, a fim de adicionar outras relacionadas ao problema de classificação requerido, como a camada *Dense* com função de ativação *softmax*, que permite a categorização das imagens nas classes definidas no modelo.

Além disso, em alguns modelos, foram utilizadas camadas de regularização. Uma dessas camadas é a *L2 Regularization*, que aplica penalidades nos pesos das camadas para evitar que estes cresçam excessivamente, o que poderia ocasionar sobreajuste.

Também foi adicionada a camada *Dropout*, que, dado um certo valor, desativa uma porcentagem dos neurônios, fazendo com que os mesmos não transmitam valores para as próximas camadas, uma medida adicional para evitar o sobreajuste.

Outra estratégia utilizada em alguns dos modelos para auxiliar na mitigação do sobreajuste foram os *callbacks*, que visam reduzir a taxa de aprendizado quando a acurácia da perda aumenta ao longo das iterações, além da função *Early stopping*, que interrompe o treinamento caso seja detectado um sobreajuste utilizando o mesmo parâmetro.

Durante o treinamento de cada modelo, uma parte dos dados foi separada para validação, com 20% em todos os testes, permitindo verificar se há sobreajuste dos dados ao comparar a acurácia e perda neste conjunto em relação aos dados de treino. Após o treinamento dos modelos, o conjunto de dados de teste do FairFace, que não foi utilizado em nenhum treino, foi reservado para testar e validar a performance dos modelos treinados. Para isso, as imagens classificadas em sete categorias (para o FairFace) foram organizadas de acordo com suas equivalências, onde a categoria *Black* foi classificada como *African American* para os modelos treinados com VGGFace2; *East Asian* e *Southeast Asian* foram agrupadas como *East Asian* para o VGGFace2 e *Asian* para o UTKFace; *Indian* foi classificado como *Asian Indian* para o VGGFace2; *Latino/Hispanic* e *Middle Eastern* foram categorizados como *Caucasian Latin* (por não haver outra categoria similar e se enquadrarem razoavelmente nesta) no VGGFace2 e como *Others* no UTKFace; por fim, a categoria *White* foi classificada como *Caucasian Latin* no VGGFace.

Após essa etapa, foi realizado um tratamento de dados para remover imagens excedentes, de modo que, para o teste dos modelos treinados com VGGFace2 e UTKFace, cada categoria contava com 1.500 imagens de faces, enquanto as imagens de teste dos modelos treinados com o FairFace continham 1.200 imagens de faces cada.

A partir daí, através de uma *confusion matrix*, foi possível verificar os verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos, onde o modelo categoriza as imagens por grupos cujos valores verdadeiros são conhecidos, sendo a *confusion matrix* capaz de mostrar como está ocorrendo o viés de cada modelo treinado.

Por meio da *confusion matrix*, foram calculadas algumas métricas. Uma delas é o *Precision*, que verifica a fração de verdadeiros positivos em relação ao total de valores previstos para aquela categoria, incluindo os falsos positivos. Logo, essa métrica mede a precisão das previsões positivas do modelo.

Outra métrica é o *Recall*, que mede a fração dos verdadeiros positivos em relação à soma dos verdadeiros positivos e falsos negativos, avaliando, assim, a sensibilidade do modelo.

Por fim, temos o *F1-score*, que representa uma relação entre o *Precision* e o *Recall*, sendo uma média harmônica entre estes dois. Um *F1-score* igual a 1 indica que o modelo está prevendo corretamente todas as classificações.

Para todos esses cálculos, foi utilizada a API Keras do Python voltada para redes neurais, que contém os modelos requeridos e todas as funções necessárias para o treinamento, através do TensorFlow (biblioteca Python de Aprendizado de Máquina).

Além disso, o código utilizado, separado de acordo com os modelos e dividido em Parâmetros Individuais e Parâmetros Globais, onde o primeiro se refere aos parâmetros referentes a cada base de dados utilizada e o segundo às configurações do modelo e forma do treinamento, está disponível no Apêndice A. A seguir, apresentam-se os detalhes de cada parametrização utilizada nos modelos testados, onde cada um deles foi treinado com os três bancos de dados.

3.5.1 VGG16

Para o VGG16, foi utilizado um tamanho de lote de 64, ou seja, 64 imagens de treinamento foram empregadas por iteração, tanto nos dados de treino quanto nos de validação e teste. No VGG16, foram removidas as três últimas camadas para adicionar uma camada de achatamento, que organiza as entradas das camadas anteriores em uma dimensão. Em seguida, foram adicionadas duas camadas densas com 512 unidades, utilizando a função de ativação ReLU, que remove todos os valores menores ou iguais a zero, e um regularizador L2, finalizando com uma camada de *Dropout* de 0,5 e uma camada densa com N unidades, correspondentes às N categorias de cada banco de dados, com uma função de ativação *softmax*, que calcula a probabilidade final da imagem pertencer a uma das categorias pré-definidas.

Ademais, o modelo foi treinado a partir da 15ª camada para baixo, compilado com um otimizador Adam com taxa de aprendizado de 10^{-6}, e treinado por 30 épocas, com *callbacks* para redução da taxa de aprendizado caso a perda aumentasse de uma época para outra, além do *Early Stopping*, que interrompe o treinamento caso a perda aumente em cinco iterações.

3.5.2 MobileNet

No MobileNet, também foi utilizado um tamanho de lote de 64, mas apenas a última camada, a que faz a classificação, foi removida, substituindo-a por uma camada com a quantidade de unidades equivalente às categorias de cada conjunto de dados, não sendo necessário o uso de regularizadores, conforme já comentado na Seção 3.3.1.

Além disso, o MobileNet foi treinado durante 50 épocas com o mesmo otimizador e taxa de aprendi-

zado de 10^{-4} para o VGGFace2 e UTKFace e 10^{-6} para o FairFace, uma vez que o modelo treinado com FairFace parava o treinamento muito cedo, devido ao *early stopping*, com o mesmo parâmetro do VGG-Face2 e UTKFace, não alcançando a acurácia desejada. E embora a taxa de aprendizado fosse inicialmente alta, ela foi diminuída ao longo do treinamento a partir dos *callbacks*.

O treinamento foi feito em todas as camadas do modelo, uma vez que se trata de um modelo leve e foi observada uma melhor performance ao realizar esse tipo de treinamento. Foram utilizados *callbacks* com os mesmos métodos e parâmetros do VGG16.

3.5.3 ResNet50

Por fim, temos o ResNet50, que foi treinado utilizando um tamanho de lote de 15, com exceção do FairFace, onde somente a última camada foi removida, seguindo o mesmo procedimento do MobileNet, sem o uso de regularizadores, uma vez que as "conexões de atalho"ajudam o modelo a evitar o sobreajuste.

Todas as bases de dados foram treinadas durante 30 épocas com o ResNet50, tempo considerado suficiente para alcançar uma boa acurácia. Além disso, utilizou-se o mesmo otimizador e a mesma taxa de aprendizado dos outros modelos, também com os mesmos *callbacks*.

4 RESULTADOS

4.1 VIÉS RACIAL EM SISTEMAS DE RECONHECIMENTO FACIAL

Numerosos estudos encontraram evidências claras de viés racial em sistemas de reconhecimento facial, Tabela 4.1. Por exemplo, indivíduos com pele mais escura têm maior probabilidade de serem identificados incorretamente, o que pode levar a tratamento injusto, como a negação de acesso a serviços (3). A falta de diversidade nos dados de treinamento usados para treinar algoritmos de reconhecimento facial é uma causa significativa de viés racial nessa tecnologia. Quando os dados de treinamento não representam adequadamente a diversidade da população, os sistemas podem identificar erroneamente os rostos de grupos sub-representados (4).

Em pesquisa recente (3), o objetivo era criar um conjunto de dados de atributos faciais balanceado em termos de raça, gênero e idade. Os pesquisadores argumentam que os conjuntos de dados de atributos faciais existentes geralmente são tendenciosos, o que pode levar a preconceitos em sistemas de aprendizado de máquina treinados com esses dados. Os pesquisadores avaliaram o conjunto de dados FairFace usando vários métodos e descobriram que ele é significativamente mais balanceado do que os conjuntos de dados existentes. Além disso, eles descobriram que os sistemas de aprendizado de máquina treinados com o conjunto de dados FairFace têm menor probabilidade de serem tendenciosos. Os desequilíbrios em conjuntos de dados de atributos faciais comumente usados (LFWA+, CelebA, COCO, IMDB-WIKI, VGG2, DiF e UTK) e a natureza balanceada do conjunto de dados FairFace são demonstrados na Figura 4.1.

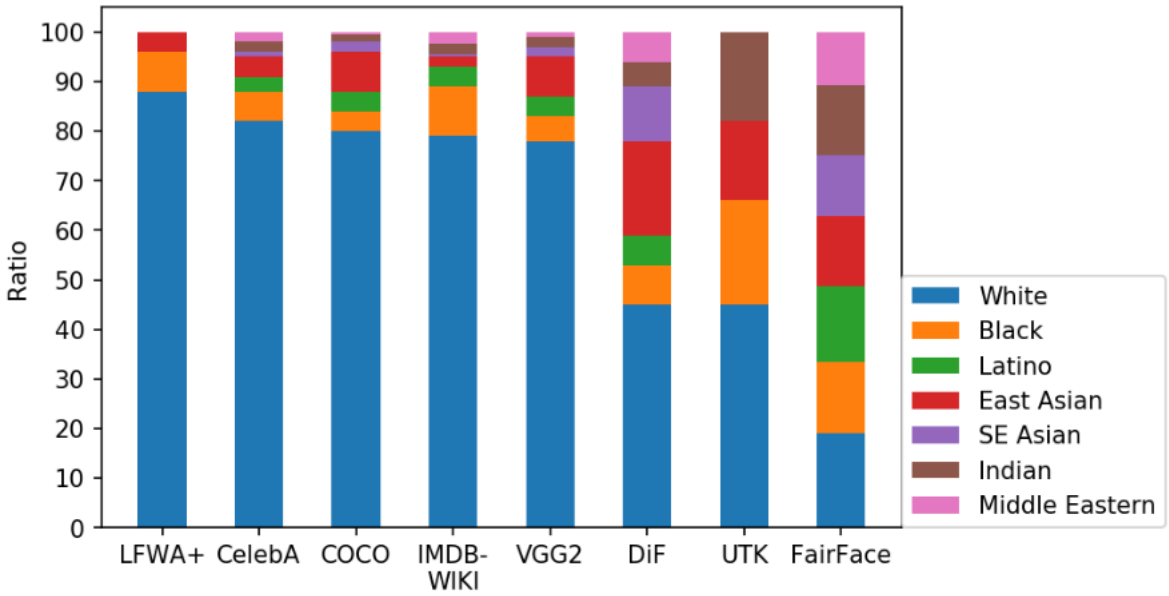

Figura 4.1: Composição racial em conjuntos de dados faciais(3).

O conjunto de dados LFWA+ (58), com seus atributos faciais rotulados, fornece uma base para a avaliação de sistemas de reconhecimento facial em cenários do mundo real. O CelebA (59), com sua extensa coleção de mais de 200.000 imagens de celebridades e 40 rótulos de atributos, oferece um rico

conjunto de dados para a análise de atributos faciais, contribuindo para as capacidades de generalização do modelo. O COCO (60), conhecido por suas tarefas de detecção e segmentação de objetos em larga escala, aumenta a robustez do modelo ao introduzir uma ampla gama de objetos e contextos, que são cruciais para tarefas de reconhecimento não facial. O conjunto de dados IMDB-WIKI (61), que foca em rótulos de idade e gênero em uma ampla gama de rostos de celebridades, auxilia no desenvolvimento de modelos abrangentes de reconhecimento facial. O VGG2 (62), com sua alta variabilidade em pose, idade, iluminação e etnia, é fundamental para melhorar a precisão do reconhecimento facial sob condições diversas. O conjunto de dados DiF (63), que enfatiza a diversidade, é crítico para reduzir viéses e aumentar a equidade no reconhecimento facial entre diferentes grupos demográficos. Por fim, o UTKFace (64), com sua representação equilibrada de idade, gênero e etnia, complementa os outros conjuntos de dados ao fornecer uma base robusta para a análise demográfica.

A Figura 4.2 compara a taxa de falsos positivos entre mulheres de países da Europa Oriental e Leste Asiático.

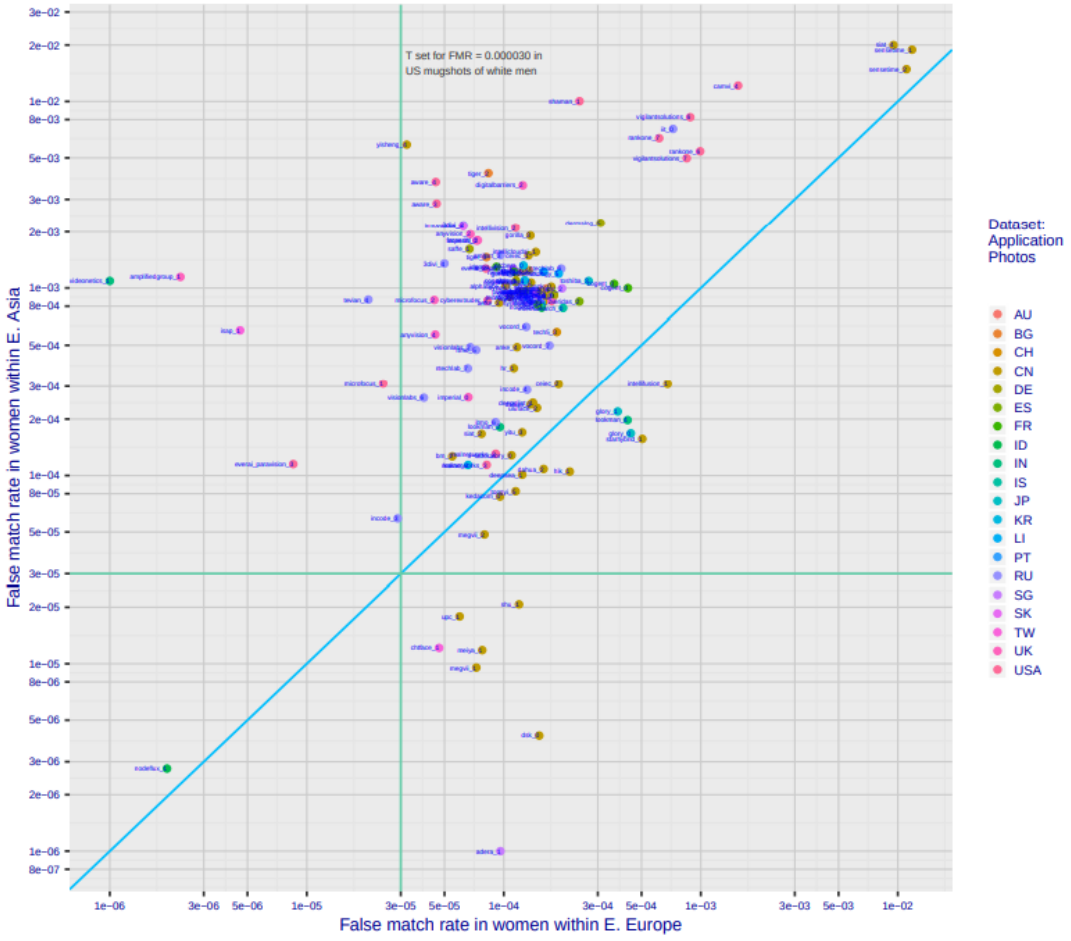

Figura 4.2: Representação esquemática para comparação da taxa de falsos positivos (FMR) entre mulheres da mesma idade. Os países considerados são Polônia, Rússia, Ucrânia, China, Japão, Coreia, Filipinas, Tailândia e Vietnã. As linhas verdes representam o limite para cada algoritmo em relação aos dados de FMR coletados para homens brancos no banco de dados de fotos de identificação dos EUA. A linha azul (y = x) indica paridade. O código de cores denota o domicílio do desenvolvedor, pois os dados de pesquisa e treinamento podem vir de locais diferentes. Esta Figura foi adaptada da referência (4).

No campo da segurança da informação e cibernética, o reconhecimento facial impreciso pode ser uma porta de entrada para acesso não autorizado a dados e sistemas sensíveis (35). Cibercriminosos podem explorar algoritmos tendenciosos em sistemas de segurança, potencialmente contornando verificações ou se passando por usuários autorizados por meio de *deepfakes*, que também podem herdar vieses se forem treinados em conjuntos de dados desbalanceados, como os discutidos neste trabalho.

4.2 SEGURANÇA CIBERNÉTICA E RECONHECIMENTO FACIAL

No campo da segurança da informação e cibernética, o reconhecimento facial impreciso pode ser uma porta de entrada para acesso não autorizado a dados e sistemas sensíveis (65). Cibercriminosos podem explorar algoritmos tendenciosos em sistemas de segurança, potencialmente contornando verificações ou se passando por usuários autorizados por meio de deepfakes, que também podem herdar vieses se forem treinados em conjuntos de dados desbalanceados, como os discutidos neste trabalho. Sistemas de reconhecimento facial enviesados podem representar uma vulnerabilidade de segurança, permitindo ataques de bypass de autenticação. Por exemplo, a Tática de Força Bruta (T1110) do MITRE ATT&CK pode ser utilizada para explorar inconsistências na autenticação de diferentes grupos raciais, onde um invasor pode se passar por um indivíduo de um grupo menos representado no treinamento do sistema (66). Podemos construir defesas robustas contra esses ataques priorizando a precisão e protegendo informações confidenciais e infraestrutura crítica.

4.3 PRIVACIDADE E CONTROLE DE ACESSO

As preocupações com a privacidade também ganham destaque. A identificação incorreta em sistemas tendenciosos pode levar a vigilância indevida, vazamentos de dados e até tratamento discriminatório. Garantir uma identificação precisa protege os indivíduos de violações e mantém a conformidade com regulamentos importantes de privacidade de dados (67). O reconhecimento facial preciso é igualmente essencial para um controle de acesso eficaz. Sistemas imprecisos correm o risco de conceder acesso a indivíduos não autorizados, comprometendo ativos físicos, propriedade intelectual e informações confidenciais. Em contraste, uma identificação precisa fortalece as medidas de segurança, protege áreas restritas e agiliza processos de acesso sem comprometer a segurança.

Neste trabalho, nos concentramos em entender as causas raízes do viés racial identificado em cada referência, buscando compreender os fatores que contribuem para o viés em sistemas de reconhecimento facial. O resultado é apresentado na Tabela 4.1 onde foi possível correlacionar as causas subjacentes com os problemas e questões relevantes.

Causas subjacentes	**Referências**	**Problemas e questões relevantes**
1. Desequilíbrio em conjuntos de dados de treinamento	(1), (4), (3), (8), (39)	Ressalta a importância de conjuntos de dados de treinamento diversificados. A falta de representatividade de diferentes grupos raciais contribui para algoritmos tendenciosos.
2. Desenvolvimento de algoritmos sem avaliações sistemáticas	(37), (36), (33)	Identificou problemas com algoritmos sem avaliações contínuas e que podem perpetuar vieses ao longo do tempo.
3. Falta de monitoramento	(36), (33)	Enfatiza a ausência de sistemas de monitoramento contínuo. Essa lacuna permite que vieses persistam sem intervenção oportuna.
4. Configuração inadequada de parâmetros do sistema	(37), (1)	Aborda o impacto da configuração inadequada de parâmetros do sistema no desempenho tendencioso entre grupos étnicos.
5. Ausência de equipe multidisciplinar	(8), (7)	Ressalta a importância de equipe diversificada com especialistas em ética, diversidade, aprendizado de máquina e direitos civis para tomada de decisão abrangente.
6. Falta de transparência nas decisões do sistema e ausência de logs de atividade	(4),(8)	Enfatiza a importância de sistemas transparentes com logs de atividade para a detecção eficaz de vieses.
7. Ausência de testes e auditorias externas por organizações independentes	(4), (8)	Esclarece a necessidade de testes e auditorias externas para uma avaliação imparcial de viés racial.
8. Falta de *feedback* constante do usuário	(8)	Enfatiza a importância do *feedback* contínuo, especialmente de usuários suscetíveis a vieses raciais.
9. Depender exclusivamente do reconhecimento facial em decisões críticas	(38), (36)	Discutidos os riscos associados à dependência exclusiva do reconhecimento facial em decisões críticas, pois isso pode contribuir para resultados discriminatórios.
10. Falta de atualizações regulares	(37), (33)	Enfatiza a importância de manter os sistemas de reconhecimento facial atualizados para alinhá-los com as melhores abordagens de mitigação de vieses.
11. Ausência de políticas e diretrizes claras	(35), (8)	Ressalta a falta de políticas claras para o uso ético do reconhecimento facial, levando em consideração o viés.

Tabela 4.1: Causas subjacentes identificadas em revisão de literatura.

4.4 RESULTADO DOS TREINAMENTOS

Para avaliar a eficácia das medidas propostas na Revisão Bibliográfica 2, especificamente no que se refere ao ajuste de parâmetros e ao aumento da diversidade nos dados de treinamento, etapas cruciais para mitigar o viés racial em modelos de reconhecimento facial impulsionados por IA, foi realizado o treinamento desses modelos.

Os modelos VGG16, MobileNet e ResNet50 foram treinados com ajuste fino nesses conjuntos de dados, até que cada modelo atingisse mais de 80% de acurácia nos dados de treinamento, apresentando boa precisão e sem sinais de sobreajuste grande o suficiente para comprometer a análise através do conjunto de dados de validação. Abaixo, fornecemos uma descrição detalhada do desempenho alcançado por cada modelo.

4.4.1 Desempenho com VGG16

Após o treinamento do modelo VGG16, utilizando a parametrização descrita na Seção 3.5, cada base de dados foi analisada separadamente. Para os conjuntos de dados UTKFace e VGGFace2, o modelo conseguiu ser treinado nas 30 épocas desejadas, sem necessidade de redução na taxa de aprendizado. Em contraste, para o conjunto de dados FairFace, o modelo atingiu a precisão alvo em 20 épocas, iniciando um sobreajuste relevante após esse ponto. Nesse caso, a parada antecipada foi utilizada, e a taxa de aprendizado, que começou com um valor inicial de 10^{-6}, foi reduzida para 10^{-8} ao final do treinamento.

A partir disso, as seguintes matrizes de confusão foram geradas para os dados de teste:

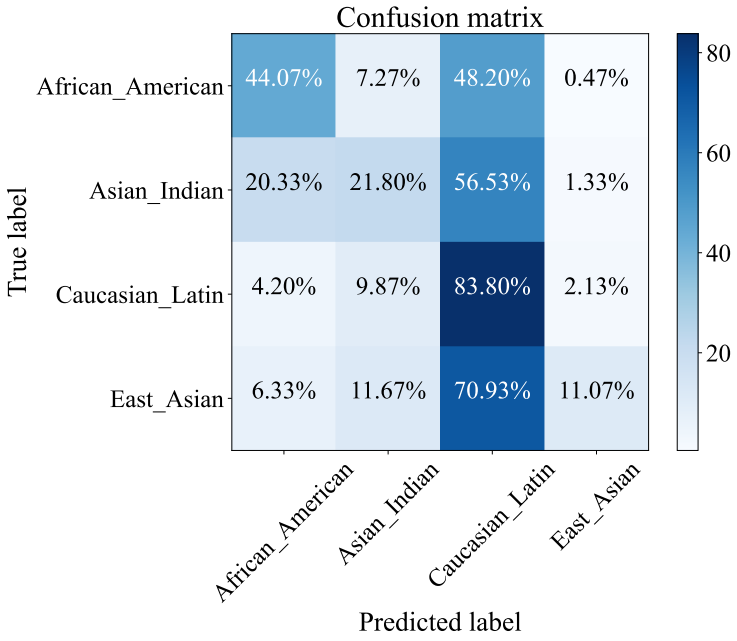

Figura 4.3: Matriz de confusão do VGG16 treinado com o conjunto de dados VGGFace2.

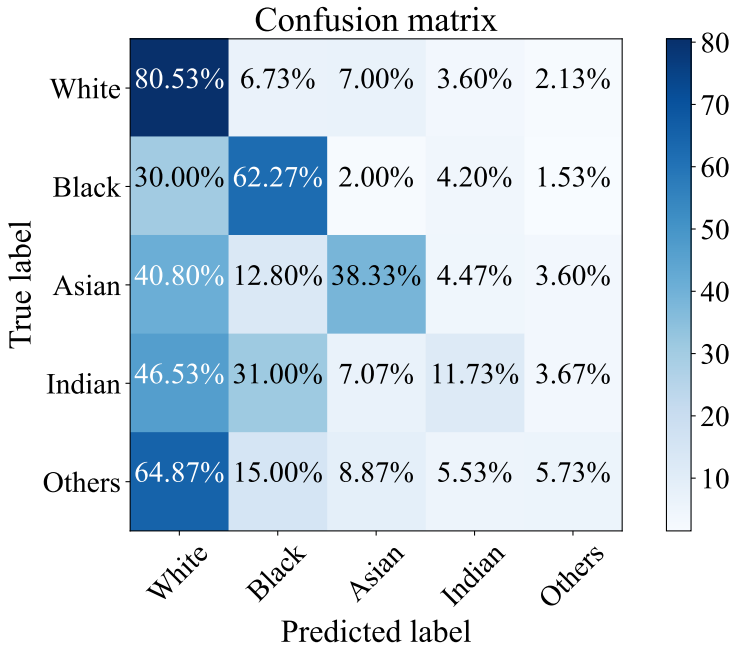

Figura 4.4: Matriz de confusão do VGG16 treinado com o conjunto de dados UTKFace.

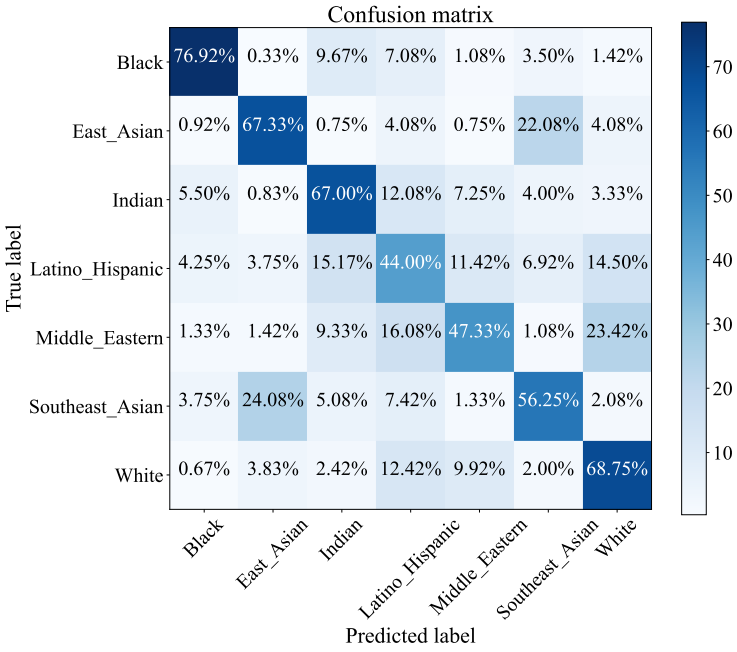

Figura 4.5: Matriz de confusão do VGG16 treinado com o conjunto de dados FairFace.

Nas imagens normalizadas, é evidente que o modelo treinado com o conjunto de dados VGGFace2 classifica predominantemente a maioria das imagens como *Caucasian Latin*, com 83,80% dos indivíduos sendo classificados nessa categoria. Essa tendência pode ser atribuída ao fato de que *Caucasian Latin* é o grupo majoritário no conjunto de dados, conforme detalhado na Tabela 4.3.

No conjunto de dados UTKFace, o modelo alcançou maior precisão para indivíduos *White* e *Black*, os grupos mais prevalentes neste conjunto, conforme mostrado na Tabela 4.4. O modelo também demons-

trou melhor precisão na classificação de outras etnias em comparação com o VGGFace2, com indivíduos asiáticos sendo classificados aproximadamente três vezes mais precisamente neste conjunto. No entanto, as imagens classificadas como *Others* (que incluem indivíduos *Middle Eastern* e *Latino Hispanic* no FairFace) são predominantemente classificadas como *White*, o que demonstra o viés do treinamento para esses grupos étnicos, uma vez que *Others* é uma categoria minoritária.

Os resultados para o modelo treinado no FairFace foram notavelmente diferentes. Ele exibiu alta precisão na classificação correta de imagens de acordo com a etnia, refletindo a maior diversidade do conjunto de dados, conforme detalhado na Tabela 4.5. Notavelmente, mais de 20% dos indivíduos *East Asian* foram incorretamente classificados como *Southeast Asian*, e há mais de 20% de chance de indivíduos *Middle Eastern* serem identificados incorretamente como *White*.

4.4.2 Desempenho com MobileNet

Com relação ao MobileNet, o modelo utilizando o UTKFace foi treinado por 44 épocas, alcançando a precisão desejada com uma taxa de aprendizado inicial de 10^{-4}. A parada antecipada foi implementada para evitar sobreajuste, e a taxa de aprendizado foi reduzida para 10^{-42} ao final do treinamento.

Para o VMER, o modelo foi treinado por 25 épocas, atingindo a precisão-alvo com uma taxa de aprendizado inicial de 10^{-4}. A parada antecipada foi utilizada, e a taxa de aprendizado foi reduzida para 10^{-11}.

No caso do FairFace, o modelo alcançou a precisão alvo em 50 épocas com uma taxa de aprendizado inicial de 10^{-6}, sem que houvesse necessidade de redução da taxa de aprendizado durante o treinamento. As seguintes matrizes de confusão foram plotadas para os dados de teste:

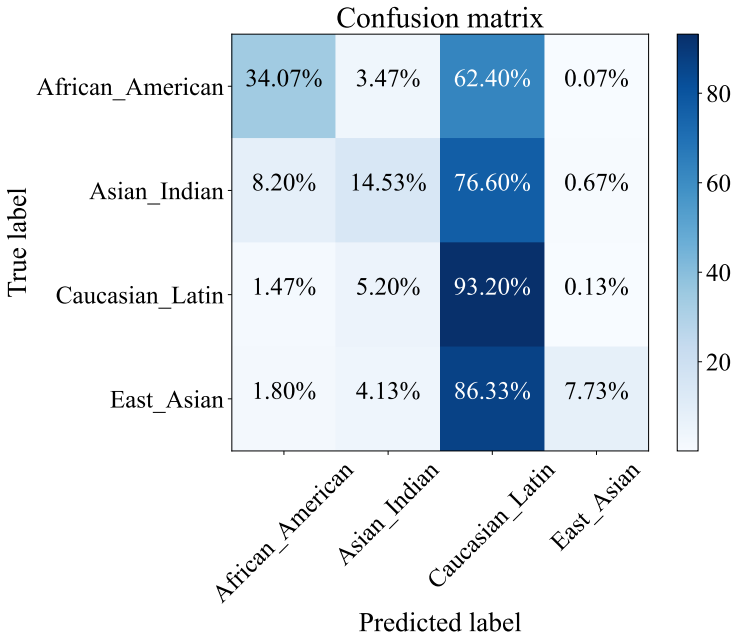

Figura 4.6: Matriz de confusão do MobileNet treinado com o conjunto de dados VGGFace2.

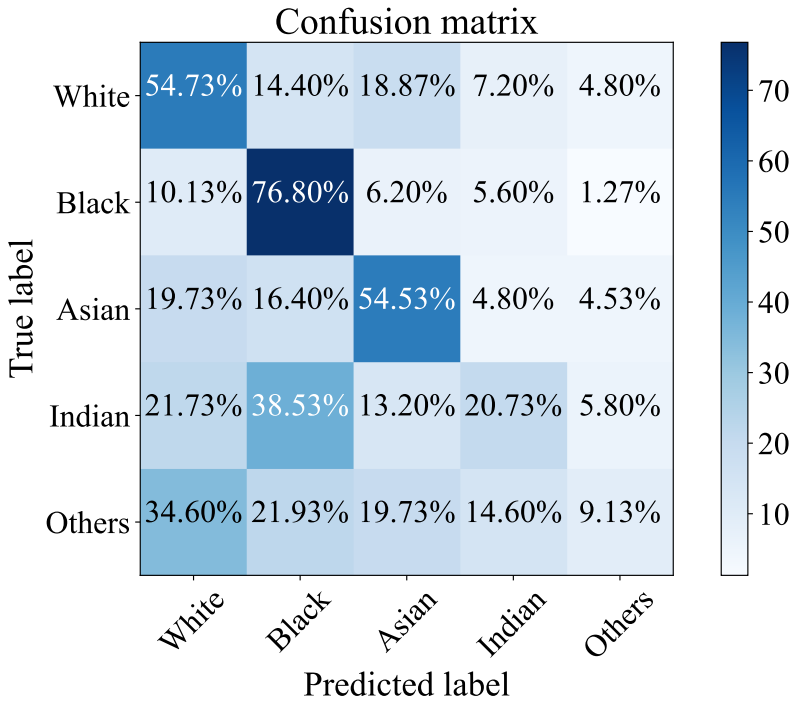

Figura 4.7: Matriz de confusão do MobileNet treinado com o conjunto de dados UTKFace.

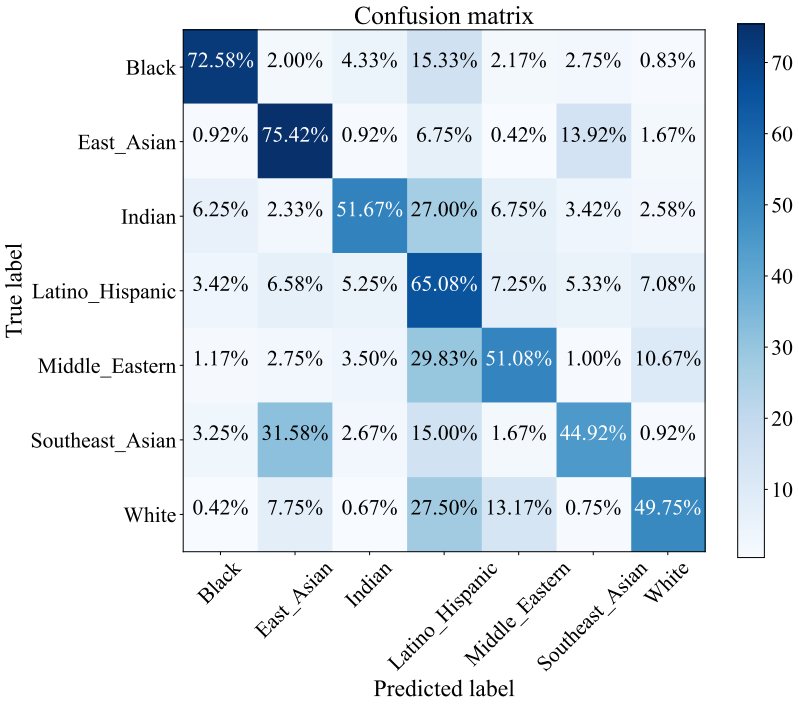

Figura 4.8: Matriz de confusão do MobileNet treinado com o conjunto de dados FairFace.

As matrizes de confusão revelam uma diferença notável na categorização em comparação com o VGG16, especialmente nos dados do VGGFace2. O MobileNet apresenta menor precisão na classificação de indivíduos *African American* e *East Asian* em relação ao VGG16. Há também uma tendência persistente de categorizar indivíduos de outros grupos étnicos como *Latino Caucasian*, sugerindo um pos-

sível viés no conjunto de dados. Para o UTKFace, o MobileNet reduz significativamente a má classificação de indivíduos de diferentes etnias como *White*, um problema observado com o VGG16. Isso sugere que tanto o conjunto de dados quanto a arquitetura do modelo influenciam esses erros de classificação. Quando treinado no FairFace, o MobileNet demonstra maior precisão entre os grupos étnicos em comparação com o VGG16. No entanto, o modelo tende a confundir indivíduos *East Asian* e *Southeast Asian* como pertencentes ao mesmo grupo e frequentemente rotula indivíduos *Middle Eastern* como *White*.

4.4.3 Desempenho com ResNet50

Com o ResNet50 e o conjunto de dados UTKFace, o modelo foi treinado por 25 épocas, alcançando a precisão esperada com uma taxa de aprendizado inicial de 10^{-6}. A parada antecipada foi empregada para evitar sobreajuste, e a taxa de aprendizado foi gradualmente reduzida para 10^{-20} ao final do treinamento. Para o conjunto de dados VGGFace2, o modelo foi treinado por 21 épocas, também alcançando a precisão desejada com uma taxa de aprendizado de 10^{-15}, utilizando a parada antecipada.

No caso do conjunto de dados FairFace, o modelo atingiu a precisão alvo em 22 épocas. A parada antecipada foi utilizada, iniciando com uma taxa de aprendizado inicial de 10^{-5}, que foi reduzida para 10^{-24} ao final do treinamento.

Para o FairFace, utilizou-se um tamanho de lote de 64, que se mostrou o mais eficaz neste estudo, em comparação com o tamanho de lote de 15 empregado para os outros conjuntos de dados. As previsões baseadas nos dados de teste do FairFace estão ilustradas nas matrizes de confusão abaixo:

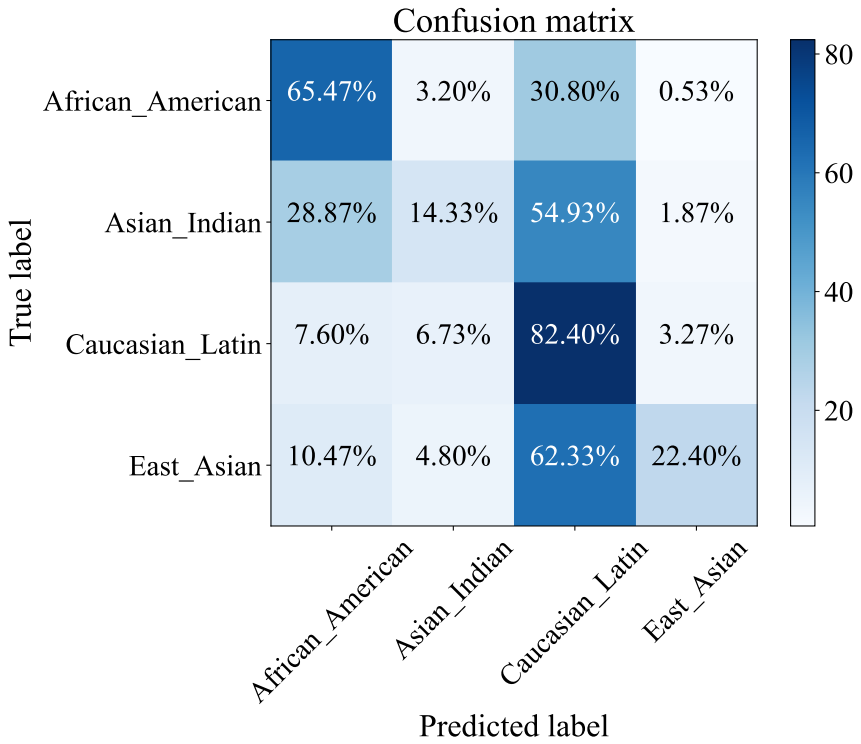

Figura 4.9: Matriz de confusão do ResNet50 treinado com o conjunto de dados VGGFace2.

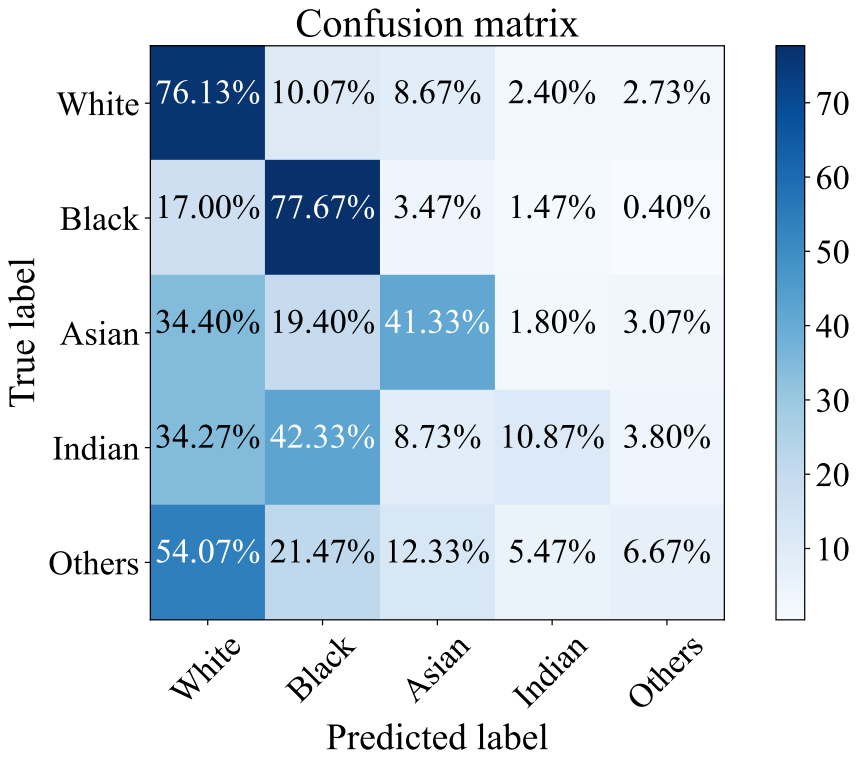

Figura 4.10: Matriz de confusão do ResNet50 treinado com o conjunto de dados UTKFace.

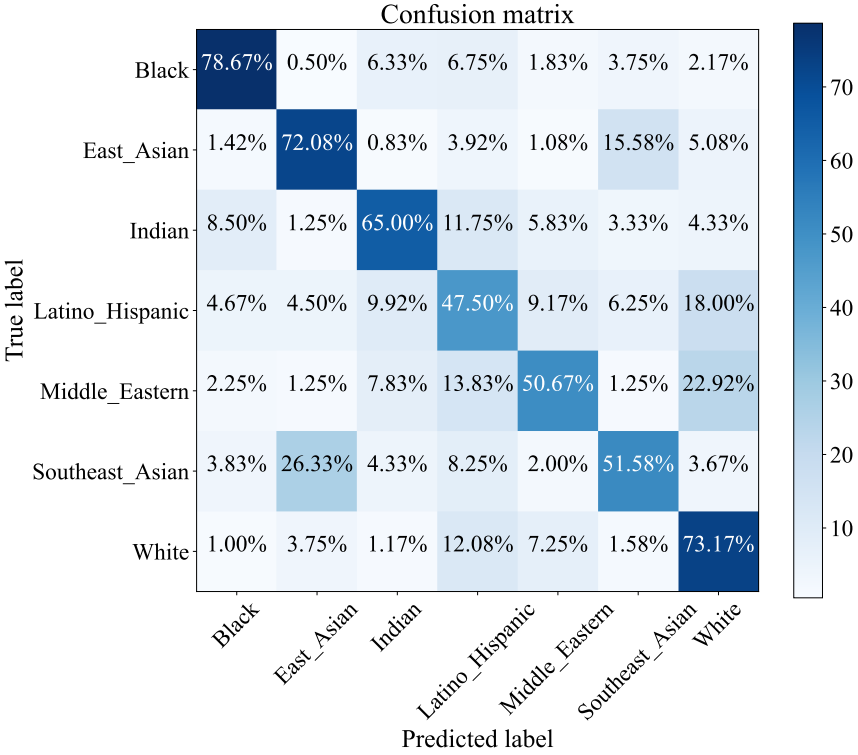

Figura 4.11: Matriz de confusão do ResNet50 treinado com o conjunto de dados FairFace.

Modelos treinados com o conjunto de dados VGGFace2 mostram uma tendência recorrente de clas-

sificar indivíduos de várias etnias como *Latino Caucasian*. No entanto, apresentam maior precisão na identificação de indivíduos *Black* em comparação com outros modelos. Com o conjunto de dados UTK-Face, o modelo teve mais dificuldade em identificar indivíduos *Indian*, frequentemente classificando-os incorretamente como *White* ou como *Black* em mais de 40% dos casos.

Em contraste, o modelo treinado com o FairFace alcançou a maior precisão entre os modelos testados. Para o FairFace, indivíduos *East Asian* e *Southeast Asian* foram confundidos entre si em mais de 15% das vezes. Além disso, indivíduos *Latino Hispanic* e *Middle Eastern* tiveram mais de 18% de chance de serem identificados incorretamente como *White*.

4.4.4 Resumo do desempenho e viés dos modelos

A Tabela 4.2 resume as principais métricas de *Precision*, *Recall* e *F1-score* de cada modelo nos três conjuntos de dados, além dos pontos de viés identificados, para facilitar a comparação:

Modelo	Dataset	Desempenho			Observações de Viés
		Train/Val Accuracy	Precision	F1-score	
VGG16	UTKFace	99.33%/73.17%	43%	35%	Viés para *White* e *Black*
	VGGFace2	99.60%/91.76%	52%	36%	Predomínio do grupo majoritário *Caucasian Latin*
	FairFace	93.07%/63.09%	61%	61%	Menor viés, com confusão entre os grupos *Asian*
MobileNet	UTKFace	86.76%/74.29%	42%	39%	Menor viés comparado ao VGG16
	VGGFace2	99.93%/93.04%	62%	32%	Viés para *Caucasian Latin*
	FairFace	86.34%/72.03%	63%	59%	Alta precisão, menor confusão entre grupos *Asian*, porém maior taxa de falsos negativos para *Latino Hispanic*
ResNet50	UTKFace	83.12%/73.19%	45%	36%	Maior precisão com menor viés para *White* e *Black*
	VGGFace2	94.91%/98.68%	56%	42%	Predomínio de *Caucasian Latin*, com viés similar ao VGG16
	FairFace	92.11%/62.91%	63%	63%	Maior precisão, com menor confusão entre grupos *Asian* e menos falsos negativos em comparação com o MobileNet

Tabela 4.2: Resumo do desempenho e viés para os modelos VGG16, MobileNet e ResNet50 nos conjuntos de dados UTKFace, VGGFace2 e FairFace.

A Tabela 4.2 evidencia que o conjunto FairFace, mais equilibrado em diversidade étnica, oferece um desempenho relativamente melhor e com menor viés em comparação aos outros conjuntos, mesmo que haja sobreajuste, como evidenciado pela comparação entre a acurácia (*Accuracy*) nos dados de treino (*Train*) e validação (*Val*). As características que diferenciam esses modelos e influenciam seu desempenho são sumarizadas na Tabela 4.3.

Modelo	Acurácia	Custo Computacional	Velocidade de Treinamento	Vantagens	Desvantagens
VGG16	Alta	Alto	Baixa	Alta acurácia em dados diversos	Alto custo computacional, treinamento lento
MobileNet	Média	Baixo	Alta	Baixo custo computacional, treinamento rápido	Menor acurácia comparado ao VGG16
ResNet50	Alta	Médio	Média	Boa acurácia, treinamento eficiente	Maior complexidade de implementação

Tabela 4.3: Comparação dos modelos de aprendizado de máquina.

4.5 IMPACTO DA DIVERSIDADE DO CONJUNTO DE DADOS NO DESEMPENHO DO MODELO E TRANSPARÊNCIA DO CÓDIGO

Este estudo destaca o impacto que a diversidade do conjunto de dados tem no desempenho do modelo de reconhecimento facial. Quando os conjuntos de dados carecem de diversidade, introduzem vieses significativos, o que pode resultar em classificações incorretas. O modelo pode ter dificuldades em distinguir entre indivíduos com características semelhantes, levando a uma categorização imprecisa devido aos vieses inerentes nos dados de treinamento.

As classificações dos modelos estudados revelam uma tendência de agrupar indivíduos com características étnicas próximas, porém distintas, em uma mesma categoria étnica. Por exemplo, indivíduos do Sudeste Asiático são frequentemente classificados como Asiáticos Orientais; Asiáticos e Asiáticos Orientais podem ser rotulados incorretamente como Brancos; indivíduos do Oriente Médio e latino-hispânicos são comumente classificados como brancos, independentemente do modelo utilizado. Esse viés reflete a distribuição das etnias nos conjuntos de dados de treinamento, destacando que a diversidade insuficiente nesses dados prejudica a capacidade de classificação precisa e justa dos modelos.

A análise dos resultados permite também uma reflexão sobre a relação entre acuidade e equidade nos modelos. Observa-se que, mesmo sem alcançar a maior acurácia no treinamento e validação com o conjunto de dados FairFace, o modelo obteve precisão consistentemente superior e mais equitativa na classificação de diferentes grupos étnicos. Isso sugere que uma maior diversidade no conjunto de dados favorece uma maior justiça no modelo, reduzindo o viés racial, mesmo que isso possa ocorrer em detrimento de uma acurácia bruta máxima. Assim, o uso de um conjunto de dados diverso se destaca como uma abordagem recomendada para equilibrar a precisão com a equidade, conforme apontado por este estudo.

Ao comparar os conjuntos de dados utilizados, como VGGFace2, UTKFace e FairFace, fica evidente que cada um possui limitações que impactam o desempenho dos modelos de maneira distinta. O VGGFace2, por exemplo, é amplamente utilizado, mas sua composição favorece grupos majoritários, contribuindo para um viés em favor de etnias como *Caucasian Latin*. O UTKFace também mostra uma distribui-

ção étnica limitada, embora mais equilibrada em comparação com o VGGFace2. Já o FairFace, projetado para promover a diversidade, possibilita uma redução significativa de vieses, como demonstrado pelos resultados, ainda que apresente menor acurácia no treinamento. Esses fatores ressaltam a importância da escolha de um conjunto de dados adequado para minimizar o viés, considerando o objetivo de aplicação do modelo.

Um aspecto relevante na escolha dos modelos e conjuntos de dados, como VGG16, MobileNet, ResNet50, VGGFace2, UTKFace e FairFace, é que todos estão disponíveis de forma fácil e gratuita. Essa característica promove maior transparência nos processos de treinamento e avaliação, facilitando a colaboração e a ética na comunidade científica. Com o uso da linguagem Python para as simulações e o código disponível no GitHub <https://github.com/HugoXR/Race-Bias>, pesquisadores e desenvolvedores podem revisar, reproduzir e aperfeiçoar os algoritmos utilizados, promovendo uma comunidade mais colaborativa e responsável. A transparência facilita a identificação e mitigação de vieses, permitindo que as limitações dos modelos sejam reconhecidas e abordadas de forma mais eficaz.

4.6 INTERPRETAÇÃO DAS MÉTRICAS DE DESEMPENHO NO CONTEXTO DO VIÉS

A análise das métricas de performance revelou que o modelo treinado com a base de dados VGGFace2 apresentou uma taxa de falsos negativos (*False Non-Match Rate*) significativamente maior para o grupo *African American* em comparação com o grupo *Caucasian Latin*. Essa disparidade implica um risco maior de negação de acesso para indivíduos afro-americanos, o que levanta sérias preocupações sobre a equidade e a justiça distributiva no uso desses sistemas.

4.7 IMPACTO NA SEGURANÇA CIBERNÉTICA E IMPLICAÇÕES DE ATAQUE

O viés racial em sistemas de reconhecimento facial apresenta implicações significativas para a segurança cibernética. Sistemas enviesados podem introduzir vulnerabilidades que facilitam ataques de *bypass* de autenticação, onde um invasor se faz passar por um usuário autorizado. Essa vulnerabilidade é particularmente crítica em sistemas de controle de acesso, onde a precisão na identificação é fundamental para a segurança física e lógica.

Cibercriminosos podem explorar essas falhas de segurança através de diversas táticas, técnicas e procedimentos (TTPs). Por exemplo, a técnica de força bruta (*Brute Force* - T1110) do *framework* MITRE ATTACK (68) pode ser adaptada para explorar variações no desempenho do reconhecimento facial entre diferentes grupos raciais (66). Um invasor pode realizar múltiplas tentativas de autenticação, direcionando seus esforços para perfis que se beneficiam do viés do sistema, como indivíduos de um grupo racial subrepresentado nos dados de treinamento.

Além disso, a capacidade de gerar *deepfakes*, que podem herdar vieses dos modelos de reconhecimento facial, representa uma ameaça adicional. *Deepfakes* podem ser usados para criar representações faciais que

contornam as verificações de autenticação, explorando as discrepâncias de precisão entre grupos raciais (69). Portanto, a mitigação do viés racial em sistemas de reconhecimento facial não é apenas uma questão de equidade e justiça social, mas também uma necessidade crítica para fortalecer a segurança cibernética e proteger contra ataques que exploram essas vulnerabilidades.

Neste capítulo, foram analisados de forma abrangente os fatores que embasam o viés racial em sistemas de reconhecimento facial. Com base nos problemas críticos que foram identificados, destacaram-se os principais impulsionadores, incluindo dados de treinamento desbalanceados, avaliação inadequada do algoritmo ao longo do tempo e falta de monitoramento contínuo. Essas revelações exigem medidas proativas para combater o viés em todo o desenvolvimento e implementação dessa tecnologia.

Na próxima Seção 4.8, são apresentadas soluções acionáveis, por meio de um conjunto cuidadosamente selecionado de 14 recomendações baseadas nas descobertas realizadas. Essas recomendações orientam desenvolvedores, formuladores de políticas e partes interessadas do setor, impulsionando o desenvolvimento de um futuro mais inclusivo, transparente e equitativo para a tecnologia de reconhecimento facial.

4.8 RECOMENDAÇÕES

Com base na identificação de 11 (onze) causas subjacentes realizadas por meio de uma revisão de literatura e à luz dos resultados e discussões, foi possível sugerir um conjunto abrangente de 14 (quatorze) recomendações para mitigar as causas, conforme detalhado na Tabela 4.4.

Para fortalecer a aplicabilidade das recomendações, considerou-se a aderência aos princípios do *NIST AI Risk Management Framework* (NIST AI 600) (5), que enfatiza práticas de equidade, explicabilidade, rastreabilidade e governança responsável no uso de sistemas de inteligência artificial. A Tabela 4.5 resume os impactos positivos esperados ao serem implementadas as recomendações sugeridas, e fornece uma visão geral dos benefícios que podem ser alcançados ao seguir as diretrizes propostas.

Ao implementar essas recomendações, espera-se melhorias na mitigação do viés racial em sistemas que utilizam reconhecimento facial para controle de acesso físico ou lógico. No geral, a tabela é útil para organizações que buscam melhorar suas operações e fortalecer sua postura de segurança no uso da tecnologia de reconhecimento facial. Importante destacar que os experimentos realizados com vários modelos de aprendizado de máquina treinados para classificar imagens faciais demonstraram que uma maior diversidade nos conjuntos de dados, aliada ao ajuste adequado de parâmetros, pode reduzir efetivamente o viés. Ao incorporar conjuntos de dados mais diversos, os modelos podem categorizar com maior precisão as imagens faciais por raça, levando a uma melhor identificação dessas imagens e de suas características únicas.

Recomendações	**Causas subjacentes**
1. Diversificar os dados de treinamento	1. Desequilíbrio em conjuntos de dados de treinamento
2. Avaliar regularmente a existência de vieses	2. Desenvolvimento de algoritmos sem avaliações sistemáticas
3. Monitorar continuamente	2. Desenvolvimento de algoritmos sem avaliações sistemáticas 3. Falta de monitoramento
4. Ajustar parâmetros do sistema	1. Desequilíbrio em conjuntos de dados de treinamento 3. Falta de monitoramento 4. Configuração inadequada dos parâmetros do sistema
5. Equipe de desenvolvimento multidisciplinar	5. Ausência de equipe multidisciplinar
6. Transparência	6. Falta de transparência nas decisões do sistema e ausência de logs de atividade
7. Execução de testes e auditorias externas	3. Falta de monitoramento 6. Falta de transparência nas decisões do sistema e ausência de logs de atividade 7. Ausência de testes e auditorias externas por organizações independentes
8. Coletar e analisar o *feedback* dos usuários	6. Falta de transparência nas decisões do sistema e ausência de logs de atividade 7. Ausência de testes e auditorias externas por organizações independentes 8. Falta de *feedback* constante do usuário
9. Limitar o uso em decisões críticas	4. Configuração inadequada dos parâmetros do sistema 9. Depender exclusivamente do reconhecimento facial em decisões críticas
10. Atualizar regularmente o sistema	5. Ausência de equipe multidisciplinar 10. Falta de atualizações regulares
11. Definir políticas e diretrizes	8. Falta de *feedback* constante do usuário 10. Falta de atualizações regulares 11. Ausência de políticas e diretrizes claras
12. Definir o âmbito de atuação	7. Ausência de testes e auditorias externas por organizações independentes 8. Falta de *feedback* constante do usuário
13. Solicitar o consentimento de uso	11. Ausência de políticas e diretrizes claras
14. Educar e conscientizar sobre o uso ético dos sistemas	9. Depender exclusivamente do reconhecimento facial em decisões críticas 10. Falta de atualizações regulares

Tabela 4.4: Correlação entre recomendações e causas subjacentes. As recomendações apresentadas foram estruturadas com base nas lacunas identificadas na literatura e nos experimentos realizados.

ID - Recomendações	Impactos positivos	Princípios do NIST AI 600
1. Diversificar os dados de treinamento	Garantir a diversidade do conjunto de dados de treinamento em termos de raça, etnia, gênero, idade e outras características relevantes. Aumentar a inclusão e precisão do sistema, evitar o domínio de um único grupo demográfico	*Fairness, Validity and Reliability*
2. Avaliar regularmente a existência de vieses	Realizar avaliações regulares do sistema para identificar qualquer viés racial que possa surgir ao longo do tempo	*Validity and Reliability, Risk Management*
3. Monitorar continuamente	Implementar sistemas de monitoramento contínuo para identificar e abordar possíveis problemas de viés à medida que surgirem. Garantir que o sistema permaneça justo ao longo do tempo e evite a amplificação de vieses existentes	*Continuous Monitoring, Trustworthiness*
4. Ajustar parâmetros do sistema	Configurar parâmetros do sistema de reconhecimento facial de maneira adequada para um desempenho equilibrado entre diferentes grupos étnicos, incluindo o ajuste de limites de decisão. Isso garante resultados justos e evita discriminação	*Safety, Fairness, Explainability*
5. Equipe de desenvolvimento multidisciplinar	Incluir na equipe especialistas em ética, diversidade, aprendizado de máquina e direitos civis irá garantir uma abordagem abrangente para lidar com questões de viés racial e tomar decisões assertivas	*Governance, Accountability*
6. Transparência	Garantir a transparência do sistema, torna suas decisões compreensíveis e registrar às atividades possibilita a detecção e correção eficaz de vieses	*Transparency, Explainability*
7. Execução de testes e auditorias externas	Submeter o sistema a testes e auditorias externas por organizações independentes para avaliar de forma imparcial a presença de viés racial garantirá resultados confiáveis e justos	*Accountability, Validity*
8. Coletar e analisar o *feedback* dos usuários	Solicitar *feedback* constante dos usuários do sistema, especialmente daqueles suscetíveis a vieses raciais irá auxiliar na identificação de problemas e na melhoria contínua do sistema	*Governance, User Engagement*
9. Limitar o uso em decisões críticas	Evite depender exclusivamente do reconhecimento facial em decisões críticas, como prisões ou contratações. Utilize-o como uma ferramenta complementar combinada a outras informações para uma tomada de decisão mais assertiva. Isso reduz o risco de discriminação e erros graves	*Risk Mitigation, Accountability*
10. Atualizar regularmente o sistema	Manter-se atualizado com as pesquisas e práticas recomendadas mais recentes para alinhar o sistema de reconhecimento facial às melhores abordagens de mitigação de viés. Melhoria contínua para mitigar viés por meio de correção de erros, atualizações de aplicativos, aprimoramentos de algoritmos e melhorias no conjunto de dados	*Continuous Improvement, Reliability*
11. Definir políticas e diretrizes	Estabelecer políticas e diretrizes claras para o uso ético do reconhecimento facial, considerando vieses. Promover o uso não discriminatório do reconhecimento facial	*Governance, Responsible Use*
12. Definir o âmbito de atuação	Limitar o uso do reconhecimento facial em contextos sensíveis, como aplicação da lei para evitar violações dos direitos humanos	*Use Case Limitation, Privacy*
13. Solicitar o consentimento de uso	Garantir o consentimento informado dos indivíduos em relação ao uso do reconhecimento facial, em ambientes públicos ou privados para respeitar a autonomia dos indivíduos quanto ao uso de suas imagens faciais	*Privacy-Enhanced, User Rights*
14. Educar e conscientizar sobre o uso ético dos sistemas	Promoção da educação e conscientização para combater o viés racial e garantir o uso ético do reconhecimento facial. Ampliar a compreensão dos riscos e benefícios dessas tecnologias	*Governance, Accountability, Fairness*

Tabela 4.5: Recomendações, impactos positivos e alinhamento com os princípios do NIST AI 600 (5).

5 CONSIDERAÇÕES FINAIS

Este estudo examinou criticamente o impacto do viés racial na aplicação do reconhecimento facial para controle de acesso. Os resultados destacaram a existência de viés, o que poderia levar a um tratamento desigual e injusto de grupos raciais específicos. Esta pesquisa contribui significativamente para a nossa compreensão do viés racial no reconhecimento facial. Ressalta a importância de abordar essa questão para garantir justiça e equidade na implementação dessa tecnologia.

É fundamental reconhecer as limitações deste estudo quanto à interpretação dos resultados, uma vez que a subjetividade pode influenciar a análise dos estudos selecionados. No entanto, essas limitações foram mitigadas pelo uso de uma abordagem rigorosa de seleção. Para fundamentar ainda mais o potencial das recomendações, é sugerida a incorporação de uma etapa adicional no processo de pesquisa 3.1, focada no impacto das 14 recomendações emitidas neste trabalho. Isso envolveria avaliar a eficácia das medidas propostas para mitigar o viés após sua implementação.

Os resultados desta pesquisa têm implicações éticas e sociais significativas; sistemas de reconhecimento facial racialmente tendenciosos podem levar a tratamento discriminatório, violações de privacidade e reforço de estereótipos nocivos. Esses problemas afetam negativamente comunidades sub-representadas e desfavorecidas, perpetuando sistemas de opressão e exacerbando as desigualdades raciais existentes na sociedade.

As contribuições deste trabalho vão além da identificação técnica do viés racial, ao propor recomendações práticas que se alinham a frameworks internacionais, como o *NIST AI Risk Management Framework* (5), fortalecendo o compromisso com a ética e a justiça algorítmica.

REFERÊNCIAS BIBLIOGRÁFICAS

1 BUOLAMWINI, J.; GEBRU, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, p. 81:1–15, 2018.

2 ZHU, H.; JI, Y.; WANG, B.; KANG, Y. Exercise fatigue diagnosis method based on short-time fourier transform and convolutional neural network. *Frontiers in Physiology*, v. 13, 08 2022.

3 KäRKKäINEN, K.; JOO, J. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, p. 1547–1557, 2021.

4 GROTHER, P.; NGAN, M.; HANAOKA, K. Face recognition vendor test part 3: Demographic effects. *NIST Interagency/Internal Report (NISTIR), National Institute of Standards and Technology, Gaithersburg, MD*, 2019.

5 National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. [S.l.], 2023. Available online. Disponível em: <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>.

6 HARARI, Y. N. *21 Lições para o Século 21*. [S.l.]: Companhia das Letras, 2018.

7 GORDON, F. Virginia eubanks (2018) automating inequality: How high-tech tools profile, police, and punish the poor. new york: Picador, st martin's press. *Law, Technology and Humans*, p. 162–164, 2019.

8 RAJI, I. D.; BUOLAMWINI, J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, p. 429–435, 2019.

9 WHITTAKER., M. Ai now report 2018. *AI Now Institute, New York University*, p. 1–44, 2018.

10 OLIVEIRA, A. M. d.; RODRIGUES, H. X.; NERY, A. S.; MENDONçA, F. L. L. d.; JUNIOR, L. A. R. Influence of racial bias in the use of facial recognition applied to access control: A critical analysis. *Research, Society and Development*, v. 14, n. 2, p. e3014248186, Feb. 2025. Disponível em: <https://rsdjournal.org/index.php/rsd/article/view/48186>.

11 BARBOSA, R. A. *Racismo e Desigualdade Social no Brasil*. [S.l.]: Editora XYZ, 2020.

12 SILVA, J. P. *Violência de Gênero e Raça: Desafios Contemporâneos*. [S.l.]: Editora ABC, 2019.

13 ALMEIDA, S. *Racismo Estrutural: Teoria e Prática*. [S.l.]: Editora GHI, 2019. 20 p.

14 ZHANG, Z. Deep learning for facial recognition: A comprehensive review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 42, n. 10, p. 2714–2731, 2020.

15 OLOYEDE, M.; HANCKE, G.; MYBURGH, H. A review on face recognition systems: recent approaches and challenges. *Multimed Tools and applications*, p. 27891–27922, 2020.

16 ZHANG, K.; ZHANG, Z.; CHEN, Y.; QIAO, Y. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Signal Processing Letters*, v. 26, n. 4, p. 657–661, 2020.

17 STALLINGS, W. *Computer Security: Principles and Practice*. [S.l.]: Pearson, 2015.

18 MCCUMBER, J. *Information Security: A Comprehensive Approach*. [S.l.: s.n.], 1991.

19 FERRAIOLO, D. F.; KUHN, D. R. Role-based access control. In: *1992 15th National Computer Security Conference*. [S.l.: s.n.], 1992.

20 HU, V. C.; FERRAIOLO, D. F.; KUHN, D. R. Attribute-based access control. In: *Computer Security: Principles and Practice*. [S.l.]: Pearson, 2015.

21 TURK, M.; PENTLAND, A. Eigenfaces for Recognition. *Journal of Cognitive Neuroscience*, v. 3, p. 71–86, 1991.

22 RATHA, N. K.; CONNELL, J. H.; BOLLE, R. M. Enhancing security and privacy in biometrics-based authentication systems. *IBM Systems Journal*, v. 40, p. 614–634, 2001.

23 LLAURADó, J.; PUJOL, F.; TOMáS, D. Study of image sensors for enhanced face recognition at a distance in the smart city context. *Scientific Reports*, p. 14713, 2023.

24 JAIN, A.; ROSS, A.; PRABHAKAR, S. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, v. 14, p. 4–20, 2004.

25 CABALLERO, A. Computer and information security handbook. In: . [S.l.: s.n.], 2017. p. 393–419.

26 CANTOR., J. R. *Privacy Impact Assessment for the Homeland Advanced Recognition Technology System (HART) Increment 1 PIA*. 2020. <https://www.dhs.gov/sites/default/files/publications/privacy-pia-obim004-hartincrement1-february2020_0.pdf>. [Online; accessed 27-August-2024].

27 CORNISH, P. The oxford handbook of cyber security. *Oxford University Press*, 2021.

28 SOLOVE, D. J. A taxonomy of privacy. *University of Pennsylvania Law Review*, v. 154, p. 477–564, 2006.

29 BERTOLINI, M.; OLIVEIRA, L. B. Reconhecimento facial e a lei geral de proteção de dados (lgpd): Análise crítica. *Revista de Direito da Tecnologia da Informação*, v. 10, n. 1, p. 45–62, 2020.

30 HSBC. *Touch ID and Face ID: A simple and secure alternative to your Digital Security Device Passcode*. <https://www.us.hsbc.com/mobile-banking/biometrics/>. [Online; accessed 27-August-2024].

31 ZHAO, W.; CHELLAPPA, R.; PHILLIPS, P. J.; ROSENFELD, A. Face recognition: A literature survey. *Association for Computing Machinery*, v. 35, p. 399–458, 2003.

32 MICROSOFT. *Windows Hello face authentication*. 2021. <https://learn.microsoft.com/en-us/windows-hardware/design/device-experiences/windows-hello-face-authentication>. [Online; accessed 27-August-2024].

33 HUA, G.; YANG, M.-H.; LEARNED-MILLER, E.; MA, Y.; TURK, M.; KRIEGMAN, D. J.; HUANG, T. S. Introduction to the special section on real-world face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 33, p. 1921–1924, 2011.

34 LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, v. 36, p. 86(11), 2278–2324, 1998.

35 WAYMAN, J. L. Biometrics in identity management systems. *IEEE Security and Privacy*, v. 6, p. 30–37, 2008.

36 KLARE, B. F.; BURGE, M. J.; KLONTZ, J. C.; BRUEGGE, R. W. V.; JAIN, A. K. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, v. 7, p. 1789–1801, 2012.

37 BENJAMIN, R. Review of “Race After Technology: Abolitionist Tools for the New Jim Code”. *Social Forces*, v. 98, p. 1–3, 2019.

38 LARSON, J.; MATTU, S.; KIRCHNER, L.; ANGWIN, J. How we analyzed the compas recidivism algorithm. *ProPublica*, 2016.

39 NAJIBI, A. *Racial Discrimination in Face Recognition Technology*. 2020. <https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/>. [Online; accessed 27-August-2024].

40 INTERNATIONAL ORGANIZATION FOR STANDARDIZATION/INTERNATIONAL ELECTROTECHNICAL COMMISSION. *Information technology – Artificial Intelligence – Guidance on risk management*. 2021.

41 STANLEY, J. The dawn of robot surveillance ai, video analytics, and privacy. *AMERICAN CIVIL LIBERTIES UNION*, p. 38–48, 2019.

42 LYNCH, J. Face off: Law enforcement use of face recognition technology. *Electronic Frontier Foundation (EFF)*, p. 10, 78–95, 2020.

43 PAGE, M. J.; MCKENZIE, J. E.; BOSSUYT, P. M.; BOUTRON, I.; HOFFMANN, T. C.; MULROW, C. D.; SHAMSEER, L.; TETZLAFF, J. M.; AKL, E. A.; BRENNAN, S. E. et al. The prisma 2020 statement for systematic reviews and meta-analyses: an updated guideline for reporting systematic reviews. *bmj*, British Medical Journal Publishing Group, v. 372, n. n71, p. n71, 2021.

44 PETTICREW, M.; ROBERTS, H. Systematic reviews in social sciences: A practical guide. *Wiley Online Library*, Blackwell Publishing Ltd, 2006.

45 SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: . [S.l.]: Computational and Biological Learning Society, 2015. p. 1–14.

46 HOWARD, A. G.; ZHU, M.; CHEN, B.; KALENICHENKO, D.; WANG, W.; WEYAND, T.; ANDREETTO, M.; ADAM, H. *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*. 2017.

47 HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2016.

48 CAO, Q.; WU, H.; SHEN, L.; AL. et. Vggface2: A dataset for recognising faces across pose and age. *arXiv preprint arXiv:1710.08092*, 2018.

49 GRECO, A.; PERCANNELLA, G.; VENTO, M.; VIGILANTE, V. Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*, v. 31, n. 7, p. 67, 7 2020. ISSN 1432-1769. Disponível em: <https://doi.org/10.1007/s00138-020-01123-z>.

50 NECH, M.; O’MAHONY, N.; AL. et. Utkface: A large scale face dataset with age, gender, and race labels. *arXiv preprint arXiv:1701.06912*, 2017.

51 ZHANG, Z.; SONG, Y.; QI, H. Age progression/regression by conditional adversarial autoencoder. In: IEEE. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.], 2017.

52 BELCAR, D.; GRD, P.; TOMIčIć, I. Automatic ethnicity classification from middle part of the face using convolutional neural networks. *Informatics*, v. 9, n. 1, 2022. ISSN 2227-9709. Disponível em: <https://www.mdpi.com/2227-9709/9/1/18>.

53 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>.

54 PARKHI, O. M.; VEDALDI, A.; ZISSERMAN, A. Deep face recognition. In: *British Machine Vision Conference (BMVC)*. [S.l.: s.n.], 2015.

55 YOSINSKI, J.; CLUNE, J.; BENGIO, Y.; LIPSON, H. How transferable are features in deep neural networks? In: *Advances in neural information processing systems*. [S.l.: s.n.], 2014. p. 3320–3328.

56 SENER, O.; SAVARESE, S. Active learning for convolutional neural networks: A core-set approach. In: *International Conference on Learning Representations (ICLR)*. [s.n.], 2018. Disponível em: <https://arxiv.org/abs/1708.00489>.

57 TAN, C.; SUN, F.; KONG, T.; ZHANG, W.; YANG, C.; LIU, C. A survey on deep transfer learning. In: *International Conference on Artificial Neural Networks (ICANN)*. [s.n.], 2018. Disponível em: <https://arxiv.org/abs/1808.01974>.

58 HUANG, G. B.; RAMESH, M.; BERG, T.; LEARNED-MILLER, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition*. [S.l.], 2007.

59 LIU, Z.; LUO, P.; WANG, X.; TANG, X. Deep learning face attributes in the wild. In: *Proceedings of the IEEE conference on computer vision (ICCV)*. [S.l.: s.n.], 2015. p. 3730–3738.

60 LIN, T.-Y.; MAIRE, M.; BELONGIE, S. J.; HAYS, J.; PERONA, P.; RAMANAN, D.; DOLLÁR, P.; ZITNICK, C. L. Microsoft coco: Common objects in context. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 740–755.

61 ROTHE, R.; TIMOFTE, R.; GOOL, L. V. *DEX: Deep EXpectation of apparent age from a single image*. 2015. Disponível em: <https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/>.

62 SIDDIQI, R. Effectiveness of transfer learning and fine tuning in automated fruit image classification. In: *Proceedings of the 2019 3rd International Conference on Deep Learning Technologies*. New York, NY, USA: Association for Computing Machinery, 2019. p. 91–100. ISBN 9781450371605. Disponível em: <https://doi.org/10.1145/3342999.3343002>.

63 WANG, Y.; XU, J.; ZHANG, S.; SHAN, S.; CHEN, X. Dif: Dataset for individual fairness in face detection. *arXiv preprint arXiv:2008.05800*, 2020.

64 ZHANG, Z.; SONG, Y.; QI, H. Age progression/regression by conditional adversarial autoencoder. p. 581–589, 2017.

65 WAYMAN, J. L. Biometrics in identity management systems. *IEEE Security & Privacy*, IEEE, v. 6, n. 2, p. 30–37, 2008.

66 MITRE CORPORATION. *MITRE ATT&CK®*. 2013–2024. Acessado em: 13 de maio de 2025. Disponível em: <https://attack.mitre.org/>.

67 SOLOVE, D. J. A taxonomy of privacy. *University of Pennsylvania Law Review*, v. 154, n. 3, p. 477–564, 2006.

68 MITRE CORPORATION. *T1110 - Brute Force*. 2013–2025. Acessado em: 13 de maio de 2025. Disponível em: <https://attack.mitre.org/techniques/T1110/>.

69 SILVA, J.; PEREIRA, M.; OLIVEIRA, C. Deepfakes como ameaça à autenticação biométrica. In: *Anais do Simpósio Brasileiro de Segurança da Informação e Sistemas Computacionais*. [S.l.: s.n.], 2024. p. 112–125.

A - CÓDIGO UTILIZADO

A.1 VGG16

A.1.1 Parâmetros Individuais

```
train_path_utkface = 'UTKFace/'
train_path_vggface2 = 'Dataset_VggFace2/'
train_path_fairface = 'Dataset_Fair_Face/'

classes_utkface= ['White', 'Black', 'Asian', 'Indian', 'Others']
classes_vggface2 = ['African_American', 'Asian_Indian', \
'Caucasian_Latin', 'East_Asian']
classes = ['Black', 'East_Asian', 'Indian', 'Latino_Hispanic',  \
'Middle_Eastern', 'Southeast_Asian', 'White']

epochs_utkface = 50
epochs_vggface = 50
epochs_fairface = 30


# Test Dataset
test_path_utkface = 'Dataset_Fair_Face_test/'
test_path_vggface = 'Dataset_Fair_Face_test_3/'
test_path_fairface = 'Dataset_Fair_Face_test_2/'

if not os.path.isdir("Dados"):
    os.makedirs("Dados")

if not os.path.isdir("Dados/Vgg16"):
    os.makedirs("Dados/Vgg16")

# Save confusion matrix to plot with another program
data_save_utkface = np.savetxt("Dados/Vgg16/UTKFace.dat", \
cm_n, fmt='%.2f')
data_save_vggface = np.savetxt("Dados/Vgg16/VGGFace2.dat", \
cm_n, fmt='%.2f')
data_save_fairface = np.savetxt("Dados/Vgg16/FairFace.dat", \
cm_n, fmt='%.2f')
```

A.1.2 Parâmetros Globais

```

datagen = ImageDataGenerator(preprocessing_function=tf.keras \
```

```
.applications.vgg16.preprocess_input, validation_split=0.2)

# Train Dataset
train_batches = datagen \
    .flow_from_directory(train_path, target_size=(224,224), \
    classes=classes, subset='training', batch_size=64)

# Validation Dataset
validation_batches = datagen \
    .flow_from_directory(train_path, target_size=(224,224), \
    classes=classes, subset='validation', batch_size=64)

vgg16_model = VGG16(include_top=False, weights="imagenet", \
input_shape=(224,224,3))

x = Flatten()(vgg16_model.output)
x = Dense(512, activation='relu', kernel_regularizer=l2(0.02))(x)
x = Dense(512, activation='relu', kernel_regularizer=l2(0.02))(x)
x = Dropout(0.5)(x)
output = Dense(units=4, activation='softmax')(x)

model = Model(inputs=vgg16_model.input, outputs=output)

for layer in model.layers[:-15]:
    layer.trainable = False

model.compile(optimizer=Adam(learning_rate=0.000001), \
loss='categorical_crossentropy', metrics=['accuracy'])

callbacks = [
    ReduceLROnPlateau(monitor='val_loss', factor=0.1, \
    patience=3, verbose=1),
    EarlyStopping(monitor='val_loss', patience=5, \
    verbose=1, restore_best_weights=True)
]

epochs = 30
history = model.fit(x=train_batches, epochs=epochs, verbose=2, \
validation_data=validation_batches, callbacks=callbacks)


test_batches = ImageDataGenerator(preprocessing_function=tf.keras \
.applications.vgg16.preprocess_input) \
    .flow_from_directory(directory=test_path, target_size=(224,224), \
    classes=classes, batch_size=64, shuffle=False)

predictions = model.predict(x=test_batches, verbose=0)
y_true = test_batches.classes
y_pred = np.argmax(predictions, axis=-1)

precision = precision_score(y_true, y_pred, average='weighted')
recall = recall_score(y_true, y_pred, average='weighted')
```

```
55 f1 = f1_score(y_true, y_pred, average='weighted')
56
57 cm = confusion_matrix(y_true=test_batches.classes, \
58 y_pred=np.argmax(predictions, axis=-1))
59
60 cm_n = (cm.astype('float') / cm.sum(axis=1)[:, np.newaxis])*100
```

A.2 MOBILENET

A.2.1 Parâmetros Individuais

```
1
2 train_path_utkface = 'UTKFace/'
3 train_path_vggface2 = 'Dataset_VggFace2/'
4 train_path_fairface = 'Dataset_Fair_Face/'
5
6 classes_utkface= ['White', 'Black', 'Asian', 'Indian', 'Others']
7 classes_vggface2 = ['African_American', 'Asian_Indian', \
8 'Caucasian_Latin', 'East_Asian']
9 classes = ['Black', 'East_Asian', 'Indian', 'Latino_Hispanic',  \
10 'Middle_Eastern', 'Southeast_Asian', 'White']
11
12 epochs_utkface = 50
13 epochs_vggface = 50
14 epochs_fairface = 50
15
16
17 # Test Dataset
18 test_path_utkface = 'Dataset_Fair_Face_test/'
19 test_path_vggface = 'Dataset_Fair_Face_test_3/'
20 test_path_fairface = 'Dataset_Fair_Face_test_2/'
21
22 if not os.path.isdir("Dados"):
23     os.makedirs("Dados")
24
25 if not os.path.isdir("Dados/MobileNet"):
26     os.makedirs("Dados/MobileNet")
27
28 # Save confusion matrix to plot with another program
29 data_save_utkface = np.savetxt("Dados/MobileNet/UTKFace.dat", \
30 cm_n, fmt='%.2f')
31 data_save_vggface = np.savetxt("Dados/MobileNet/VGGFace2.dat", \
32 cm_n, fmt='%.2f')
33 data_save_fairface = np.savetxt("Dados/MobileNet/FairFace.dat", \
34 cm_n, fmt='%.2f')
```

A.2.2 Parâmetros Globais

```
datagen = ImageDataGenerator(preprocessing_function=tf.keras \
    .applications.mobilenet.preprocess_input, validation_split=0.2)

train_batches = datagen \
    .flow_from_directory(train_path, target_size=(224,224), \
    classes=classes, subset='training', batch_size=64)
validation_batches = datagen \
    .flow_from_directory(train_path, target_size=(224,224), \
    classes=classes, subset='validation', batch_size=64)


mobile_model = tf.keras.applications.mobilenet \
.MobileNet(weights="imagenet", include_top=True)

x = mobile_model.layers[-2].output
output = Dense(units=5, activation='softmax')(x)

model = Model(inputs=mobile_model.input, outputs=output)

model.compile(optimizer=Adam(learning_rate=0.000001),  \
loss='categorical_crossentropy', metrics=['accuracy'])

callbacks = [
    ReduceLROnPlateau(monitor='val_loss', factor=0.1, \
    patience=1, verbose=1),
    EarlyStopping(monitor='val_loss', patience=5, \
    verbose=1, restore_best_weights=True)
]

history = model.fit(x=train_batches, epochs=epochs,  \
validation_data=validation_batches, callbacks=callbacks)
test_batches = ImageDataGenerator(preprocessing_function=tf.keras \
.applications.mobilenet.preprocess_input) \
    .flow_from_directory(directory=test_path, target_size=(224,224), \
    classes=classes, batch_size=64, shuffle=False)

predictions = model.predict(x=test_batches, verbose=0)

precision = precision_score(y_true, y_pred, average='weighted')
recall = recall_score(y_true, y_pred, average='weighted')
f1 = f1_score(y_true, y_pred, average='weighted')

cm = confusion_matrix(y_true=test_batches.classes, \
y_pred=np.argmax(predictions, axis=-1))

cm_n = (cm.astype('float') / cm.sum(axis=1)[:, np.newaxis])*100
```

A.3 RESNET50

A.3.1 Parâmetros Individuais

```
1
2  train_path_utkface = 'UTKFace/'
3  train_path_vggface2 = 'Dataset_VggFace2/'
4  train_path_fairface = 'Dataset_Fair_Face/'
5
6  classes_utkface= ['White', 'Black', 'Asian', 'Indian', 'Others']
7  classes_vggface2 = ['African_American', 'Asian_Indian', \
8  'Caucasian_Latin', 'East_Asian']
9  classes = ['Black', 'East_Asian', 'Indian', 'Latino_Hispanic',  \
10 'Middle_Eastern', 'Southeast_Asian', 'White']
11
12 epochs_utkface = 30
13 epochs_vggface = 30
14 epochs_fairface = 30
15
16
17 # Test Dataset
18 test_path_utkface = 'Dataset_Fair_Face_test/'
19 test_path_vggface = 'Dataset_Fair_Face_test_3/'
20 test_path_fairface = 'Dataset_Fair_Face_test_2/'
21
22 if not os.path.isdir("Dados"):
23     os.makedirs("Dados")
24
25 if not os.path.isdir("Dados/ResNet50"):
26     os.makedirs("Dados/ResNet50")
27
28 # Save confusion matrix to plot with another program
29 data_save_utkface = np.savetxt("Dados/ResNet50/UTKFace.dat", \
30 cm_n, fmt='%.2f')
31 data_save_vggface = np.savetxt("Dados/ResNet50/VGGFace2.dat", \
32 cm_n, fmt='%.2f')
33 data_save_fairface = np.savetxt("Dados/ResNet50/FairFace.dat", \
34 cm_n, fmt='%.2f')
```

A.3.2 Parâmetros Globais

```
1 datagen = ImageDataGenerator(preprocessing_function=tf.keras\
2 .applications.resnet50.preprocess_input, validation_split=0.2)
3
4 train_batches = datagen \
5     .flow_from_directory(train_path, target_size=(224,224), \
6     classes=classes, subset='training', batch_size=15)
```

```
validation_batches = datagen \
    .flow_from_directory(train_path, target_size=(224,224), \
    classes=classes, subset='validation', batch_size=15)

resnet_model = tf.keras.applications.resnet50.ResNet50()

x = resnet_model.layers[-2].output
output = Dense(units=5, activation='softmax')(x)

model = Model(inputs=resnet_model.input, outputs=output)

model.compile(optimizer=Adam(learning_rate=0.000001), \
loss='categorical_crossentropy', metrics=['accuracy'])

callbacks = [
    ReduceLROnPlateau(monitor='val_loss', factor=0.1, \
    patience=1, verbose=1),
    EarlyStopping(monitor='val_loss', patience=5, \
    verbose=1, restore_best_weights=True)
]

history = model.fit(x=train_batches, epochs=epochs, verbose=2,  \
validation_data=validation_batches, callbacks=callbacks)

test_batches = ImageDataGenerator(preprocessing_function=tf.keras \
.applications.resnet50.preprocess_input) \
    .flow_from_directory(directory=test_path, target_size=(224,224), \
    classes=classes, batch_size=15, shuffle=False)


predictions = model.predict(x=test_batches, verbose=0)

precision = precision_score(y_true, y_pred, average='weighted')
recall = recall_score(y_true, y_pred, average='weighted')
f1 = f1_score(y_true, y_pred, average='weighted')

cm = confusion_matrix(y_true=test_batches.classes, \
y_pred=np.argmax(predictions, axis=-1))

cm_n = (cm.astype('float') / cm.sum(axis=1)[:, np.newaxis])*100
```