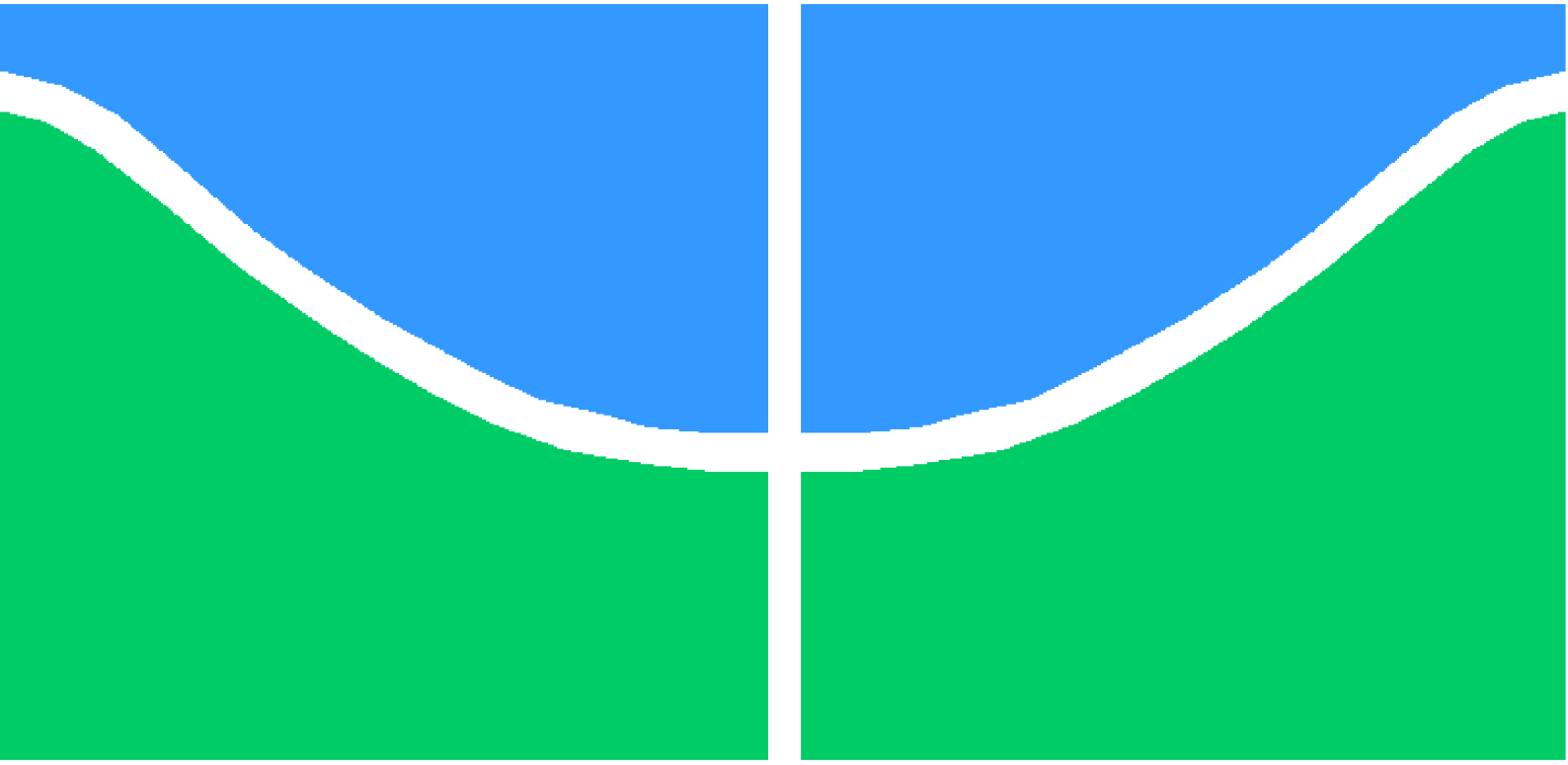

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Engenharia de Software

Racismo e Inteligência Artificial: Um Mapeamento Sistemático

Autor: Mar Akin Dantas Silva

Professor Orientador: Profa. Dra. Carla Rocha

Brasília, DF

2023

Mar Akin Dantas Silva
Racismo e Inteligência Artificial: Um Mapeamento Sistemático/ Mar Akin Dantas Silva. – Brasília, DF, 2023-
42 p. : il. (algumas color.) ; 30 cm.

Orientador: Profa. Dra. Carla Rocha

Trabalho de Conclusão de Curso – Universidade de Brasília – UnB
Faculdade UnB Gama – FGA , 2023.

1. . 2. Engenharia de Software. I. Profa. Dra. Carla Rocha. II. Universidade de Brasília. III. Faculdade UnB Gama. IV. Racismo e Inteligência Artificial: Um Mapeamento Sistemático

Mar Akin Dantas Silva

Racismo e Inteligência Artificial: Um Mapeamento Sistemático

Monografia submetida ao curso de graduação em Engenharia de Software da Universidade de Brasília, como requisito parcial para obtenção do Título de Bacharel em Engenharia de Software.

Universidade de Brasília – UnB

Faculdade UnB Gama – FGA

Orientador: Profa. Dra. Carla Rocha

Brasília, DF

2023

Agradecimentos

Dedico este trabalho à minha psicóloga, Brenda, cujo apoio foi fundamental durante os desafios ao longo da minha graduação e da elaboração desta pesquisa. Expresso minha gratidão aos amigos que estiveram sempre ao meu lado, com destaque para Renato Augusto, Rafaela Rosa, Lorrany Azevedo e Lucas Brilhante, pelos conselhos e apoio nos momentos mais desafiadores. Agradeço a todos os outros amigos que proporcionaram suporte emocional ao longo de todo o processo. E, por fim, dedico este trabalho à minha família e a mim mesmo, pelos esforços dedicados ao longo desses anos de graduação.

Lista de ilustrações

Figura 1 – Regra Perceptron 20
Figura 2 – Participação em grandes conferências de IA ao longo dos anos 20
Figura 3 – Neurônio Artificial 21
Figura 4 – Diagrama de blocos de uma CNN (Rede Neural Convolucional) 22
Figura 5 – Processo de Mapeamento Sistemático 26

Lista de tabelas

Tabela 1 – Strings de Buscas 27
Tabela 2 – Dados dos Artigos Coletados 28
Tabela 3 – Dados dos Artigos Selecionados para leitura 28
Tabela 4 – Dados dos Artigos Escolhidos 28

Lista de abreviaturas e siglas

TCC	Trabalho de Conclusão de Curso
IA	Inteligência Artificial
RNA	Redes Neurais Artificiais
MBN	Rede Meta Balanceada
GAC	Classificador Adaptativo de Grupo
MDP	Processo de Decisão de Markov
FPR	Taxa de Falso Positivo
MvCoM	Margem cosseno de múltiplas variações no nível da amostra.

Resumo

O reconhecimento facial inerentemente inclina-se a apresentar desigualdades na identificação e classificação de faces de indivíduos pertencentes a diferentes grupos demográficos. Essa inclinação é evidente quando o desempenho dos modelos favorece resultados mais precisos em rostos de um grupo específico, em detrimento de outros, sendo esse grupo, em sua maioria, composto por pessoas racializadas. Este estudo direciona seus esforços para buscar equidade e precisão no âmbito tecnológico, fornecendo um mapeamento sistemático de métodos destinados a mitigar o viés racial algorítmico. A pesquisa analisa minuciosamente estudos selecionados, oferecendo sínteses detalhadas das características e resultados obtidos. A abordagem abrangente busca compreender iniciativas que visam diminuir as disparidades observadas no reconhecimento facial, enfatizando a equidade, a adaptação de técnicas e a consideração de características demográficas durante o treinamento. Ao proporcionar uma visão ampla dessas abordagens, este trabalho contribui para uma compreensão crítica do panorama de pesquisa, impulsionando o desenvolvimento de sistemas mais equitativos no domínio do reconhecimento facial.

Palavras-chave: deep learning. reconhecimento facial. viés. equidade.

Abstract

Facial recognition inherently tends to manifest inequalities in the identification and classification of individuals' faces belonging to different demographic groups. This inclination becomes evident when model performance favors more accurate outcomes for faces from a specific group at the expense of others, with this group predominantly composed of racialized individuals. This study directs its efforts toward seeking equity and precision in the technological domain, offering a systematic mapping of methods aimed at mitigating algorithmic racial bias. The research meticulously examines selected studies, providing detailed syntheses of characteristics and results. The comprehensive approach seeks to comprehend initiatives aimed at reducing observed disparities in facial recognition, emphasizing equity, technique adaptation, and the consideration of demographic features during training. By offering a broad perspective on these approaches, this work contributes to a critical understanding of the research landscape, propelling the development of more equitable systems in the field of facial recognition.

Key-words: deep learning. facial recognition. bias. fairness

Sumário

1 INTRODUÇÃO 17
1.1 Contexto 17
1.2 Problema 18
1.3 Objetivos 18
1.3.1 Objetivo Geral 18
1.3.2 Objetivos Específicos 18
1.4 Organização do trabalho 18

2 REFERENCIAL TEÓRICO 19
2.1 Inteligência Artificial: Uma Breve Introdução 19
2.2 RNA - Redes Neurais Artificiais 21
2.2.1 Reconhecimento Facial 22
2.2.2 Vieses e Reconhecimento Facial 22
2.2.3 Impacto de Viéses 23

3 PROPOSTA 25
3.1 Metodologia de Pesquisa 25
3.1.1 Mapeamento Sistemático 25
3.1.2 Questão de pesquisa 26
3.1.3 Estratégia de Pesquisa 26
3.1.3.1 Bases de Dados 26
3.1.3.2 Strings de Buscas 26
3.1.3.3 Critérios de Inclusão e Exclusão 27
3.1.4 Resultados do Mapeamento Sistemático 27

4 RESULTADOS 29
4.1 Quais estratégias são propostas na literatura para mitigar o enviesamento racial em sistemas de inteligência artificial voltados para reconhecimento facial? 29
4.1.1 Sumário dos Achados na Literatura 30
4.1.1.1 Strategic Sampling Methods 30
4.1.1.2 Representation Disentanglement Methods 30
4.1.1.3 Domain Adaptation Methods 31
4.1.1.4 Domain Adaptation Methods e Domain Independent Training Method 31
4.1.1.5 Representation Disentanglement Methods e Domain Independent Training Method 32
4.1.1.6 Dataset Approach 32

4.1.2 Desvantagens e Limitações 32

5 CONCLUSÃO 35

REFERÊNCIAS 37

APÊNDICES 39

APÊNDICE A – MAPEAMENTO SISTEMÁTICO 41

1 Introdução

1.1 Contexto

Algoritmos podem carregar vieses e é inegável o potencial nocivo quando desenvolvidos longe de um olhar crítico aos problemas sociais. Isso não é diferente, e talvez até mais pronunciado, no contexto de algoritmos de inteligência artificial. Sistemas com inteligência artificial podem facilmente refletir vieses e preconceitos da sociedade em que estão inseridos.

O que o contexto atual parece revelar é que ao utilizar de soluções de *deep learning*, muitas vezes não existe uma preocupação em identificar possíveis vieses sociais. Surge então um questionamento ético sobre como a essa despreocupação afeta grupos marginalizados e como a tecnologia pode se tornar um perpetuador dos preconceitos já existentes.

Tomemos como exemplo o estudo realizado pela Rede de Observatório da Segurança, em cinco estados brasileiros que concluiu que 90% das 151 pessoas presas com base em câmeras de reconhecimento facial são negras [1]. Ou o caso de Joy Buolamwini, pesquisadora negra, que se deparou com essa questionamento ao averiguar que o modelo por ela utilizado para reconhecimento facial em sua pesquisa não era capaz de identificar seu rosto e de outras pessoas negras [2]. Podemos citar ainda o famoso caso do *Google Vision Cloud* que classificava um termômetro segurado por uma pessoa de pele escura como uma arma em contrapartida quando segurado por uma pessoa de pele clara esse mesmo termômetro era classificado como um dispositivo eletrônico [3]. Ou o mais recente caso do algoritmo de corte de imagens do *Twitter* que considerava como mais importante rostos de pessoas brancas em vez de rostos de pessoas negras [4]. E esses não são exemplos excepcionais, sendo comum circular nos meios de comunicação casos similares onde soluções de inteligência artificial reproduziram estereótipos preconceituosos contra pessoas negras ou de outros grupos marginalizados.

Para que as tecnologias de inteligência artificial não se tornem agentes perpetuadores do racismo e outras injustiças sociais, é preciso identificar onde podem surgir esses enviesamentos e criar formas de, não só identificá-los, como evitá-los ou tomar medidas compensatórias para minimizar os seus efeitos negativos.

---

1 http://observatorioseguranca.com.br/wordpress/wp-content/uploads/2019/11/1relatoriorede.pdf
2 https://www.netflix.com/br/title/81328723
3 https://algorithmwatch.org/en/google-vision-racism/
4 https://www.theguardian.com/technology/2020/sep/21/twitter-apologises-for-racist-image-cropping-algorithm

1.2 Problema

A partir do exposto na seção 1.1, busca-se então levantar quais métodos estão sendo utilizados e/ou discutido nos estudos de inteligência aritificial para mitigar viéses nocivos as grupos socialmente marginalizados, em particular, pessoas racializadas.

1.3 Objetivos

1.3.1 Objetivo Geral

Este trabalho tem o objetivo de mapear os principais métodos e estratégias que estão sendo discutido no campo de estudo para migitar racismo algoritimíco e enviesamento racial em soluções de inteligência artificial de reconhecimento facial.

1.3.2 Objetivos Específicos

- Identificar estudos relevantes no campo de estudo que sugerem formas de combate ao enviesamento de soluções de reconhecimento facial;
- Elaboração de um mapeamento compreendendo métodos e estratégias que propõem abordagens para resolver ou contribuir para a resolução da problemática central desta pesquisa.

1.4 Organização do trabalho

Capítulo 1 - Introdução: apresenta o contexto, a justificativa, o problema de pesquisa e os objetivos.

Capítulo 2 - Referencial Teórico: apresenta os conceitos que irão direcionar o trabalho, como, Inteligência Artificial, Redes Neurais Artificiais (RNAs), Reconhecimento Facial e Viéses e seus imapctos no Reconhecimento Facial.

Capítulo 3 - Proposta: apresenta a metodologia e a proposta do trabalho a ser executada.

Capítulo 4 - Resultados: apresenta os resultados obtidos atraés da metodologia proposta.

Capítulo 5 - Conclusão: apresenta as conclusões derivadas do estudo proposto, oferecendo uma síntese analítica dos resultados obtidos.

2 Referencial teórico

2.1 Inteligência Artificial: Uma Breve Introdução

Diversos autores propuseram várias definições de inteligência artificial (IA) ao longo do tempo. Em seu livro "Artificial Intelligence: A Modern Approach", Russell (2010) realiza uma categorização dessas definições em duas dimensões: aquelas relacionadas aos processos de pensamento e raciocínio, e aquelas que abordam o comportamento. As definições na primeira dimensão estão preocupadas em avaliar a inteligência artificial com base em sua semelhança com o desempenho humano, enquanto as definições na segunda dimensão medem o sucesso em relação a um padrão ideal de desempenho chamado racionalidade. A abordagem na segunda dimensão envolve a incorporação de princípios matemáticos e técnicas de engenharia em seu desenvolvimento, dentro dessa abordagem, surge o conceito de "acting rationally", onde agentes racionais utilizam inferências e padrões racionais matematicamente bem definidos para alcançar o melhor resultado possível ou, em situações de incerteza, o melhor resultado esperado.

Uma outra forma de categorizar os estudos de IA foi proposta por Simon (1988), onde os estudos são categorizados quanto à sua abordagem, cognitiva ou conexionista. A abordagem cognitiva, também chamada de descendente, tem como objetivo compreender o comportamento inteligente a partir de aspectos psicológicos e processos algorítmicos. Por outro lado, a abordagem conexionista, também conhecida como ascendente, concentra seus esforços em entender o funcionamento do cérebro, dos neurônios e das conexões neurais. Os modelos mais amplamente conhecidos de inteligência artificial atualmente têm suas raízes na abordagem conexionista.

Em 1943 Warren McCulloch e Walter Pitts propuseram um modelo matemático de neurônios artificiais, definidos por um estado de ativação ou inativação, cuja transição para o estado ativado ocorre em resposta à estimulação proveniente de um número adequado de neurônios adjacentes. Posteriormente, Donald Hebb apresentou uma descoberta fundamental, na qual demonstrou que o incremento do peso das conexões entre neurônios artificiais, quando um estímulo de entrada influencia a produção de um estímulo de saída, possibilita que as redes de neurônios artificiais, conhecidas hoje como redes neurais, fossem capazes de aprender, essa regra ficou conhecida como regra de aprendizagem hebbiana. Outras regras bastante conhecidas que ajudaram no desenvolvimento da capacidade de apredizagem de redes neurais foram a regra de aprendizagem Perceptron (Rosenblatt, 1957) e Delta (Widrow e Hoff, 1960). A regra Percetron propõem que os pesos das conexões entre neuronios aritificais dependam de princípios probabilísticos em vez de determinísticos e que sua confiabilidade seja obtida através de propriedades das

medidas estatísticas obtidas de grandes populações de elementos. Em outras palavras, em vez de simplesmente aumentar os pesos quando a entrada influencia um estímulo de saída, é necessário calcular o efeito dessa entrada e ajustar os pesos para penalizar as conexões que resultam em saídas indesejadas. A regra Delta estabelece que a atualização dos pesos deve ser realizada com base no cálculo do erro entre a saída obtida e a saída desejada. Tanto a regra do Perceptron quanto a regra Delta foram desenvolvidas para capacitar os neurônios artificiais a classificar dois padrões com base em exemplos desses padrões.

$$w_i \leftarrow w_i + \alpha\,(y - h_{\mathbf{w}}(\mathbf{x})) \times x_i$$

weight — learning rate — target value — threshold function — input value

Figura 1 – Regra Perceptron

Fonte: RUSSEL e NORVIG (2010)

Apesar dos esforços para desenvolver algoritmos que melhorassem a capacidade de aprendizado das redes neurais, em 1969 Marvin Minsky e Seymour Papert demonstraram que, independentemente da regra de aprendizado utilizada, as redes neurais eram incapazes de resolver problemas de associação de padrões quando os conjuntos de pares de padrões não eram linearmente separáveis. Problemas como a porta lógica XOR, por exemplo, não poderiam ser solucionados. Somente na década de 80, com estudos como os de Hopfield (1982) e Soffer (1986), é que as redes neurais foram novamente consideradas como alternativas promissoras no desenvolvimento da inteligência artificial. A Figura 2 ilustra a evolução do tema ao longo dos anos, fundamentada no aumento e na emergência de conferências sobre IA.

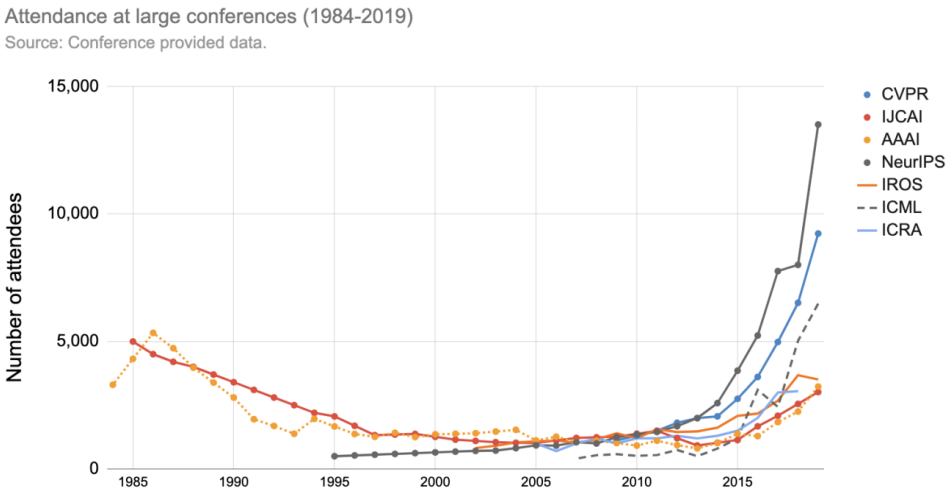

Figura 2 – Participação em grandes conferências de IA ao longo dos anos

Fonte: Artificial Intelligence Index Report 2019

2.2 RNA - Redes Neurais Artificiais

Ao longo do tempo, as redes neurais evoluíram, mantendo os princípios iniciais propostos por McCulloch e Walter Pitts. As redes neurais artificiais são compostas por unidades de processamento, chamadas neurônios, que são organizados em camadas e interconectados entre si. As conexões entre as camadas possuem pesos atribuídos a elas, que são utilizados para determinar a propagação de um sinal de entrada pela rede.

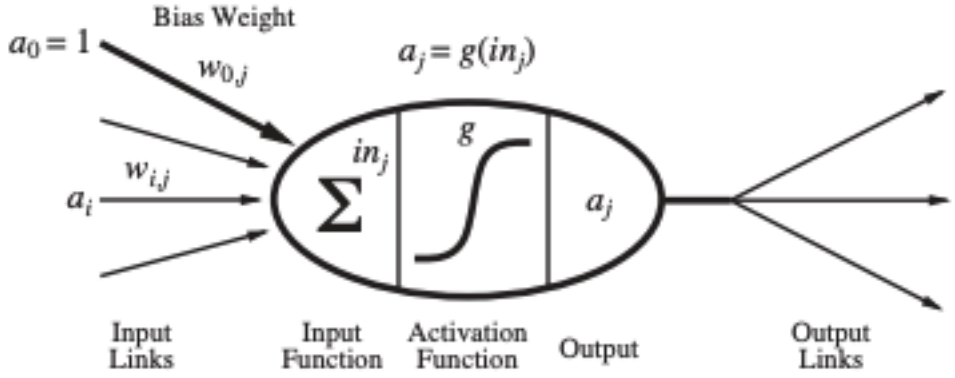

Figura 3 – Neurônio Artificial

Fonte: RUSSEL e NORVIG (2010)

Cada unidade de processamento em uma rede neuronal tem uma função de entrada, cujo papel é adaptar os dados de entrada para que se ajustem ao formato utilizado nas operações realizadas pelas unidades. Essa função de entrada permite que a unidade processe corretamente o sinal de entrada antes de transmiti-lo para as camadas seguintes da rede. Além disso, as unidades também possuem uma função de ativação, que determina se a unidade deve ser ativada ou não, utilizando um algoritmo de aprendizado que emprega uma regra de aprendizagem e produz uma saída.

De forma geral, RNAs são composta por três ou mais camadas, que podem ser classificadas como:

- Camada de Entrada: são responsáveis por receber os dados iniciais (sinais).
- Camada Intermediária ou Escondida: responsáveis por extrair características fazendo processamento do sinais através das conexões ponderadas.
- Camada de Saída: nessa camada o resultado do processamento da rede neural é concluído e apresentado.

Podemos categorizar redes neurais com base na direção em que o sinal é propagado através delas. Em determinados casos, os sinais fluem em uma única direção, partindo da

camada de entrada, atravessando a camada intermediária e chegando à camada de saída. RNAs com esse tipo de fluxo são conhecidas como *feedforward*. Enquanto as RNAs onde os sinais de entrada podem fluir em um ou mais loops rertoalimentando entradas de unidades de processamento anteorioes e por isso são conhecidas como *feedback*.

2.2.1 Reconhecimento Facial

O reconhecimento facial é dividido em duas modalidades principais: verificação facial, que busca determinar se duas imagens correspondem à mesma pessoa, e identificação facial, que tem como objetivo reconhecer a pessoa retratada em uma imagem específica.

O processo de reconhecimento facial ocorre em três etapas fundamentais: detecção facial, onde são localizadas as faces nas imagens; extração ou alinhamento de recursos, que envolve a identificação e representação de características faciais relevantes; e, por fim, a etapa de verificação ou identificação facial, onde ocorre a comparação ou correspondência das características extraídas.

No contexto do reconhecimento facial, as Redes Neurais Convolucionais (CNNs) são amplamente utilizadas para a extração de características faciais. Essas redes são projetadas para reconhecer padrões em imagens e são eficazes na identificação de características específicas do rosto. Por meio de treinamento em conjuntos de dados extensos, as CNNs aprendem a representar e distinguir características faciais, contribuindo significativamente para o sucesso de sistemas de reconhecimento facial.

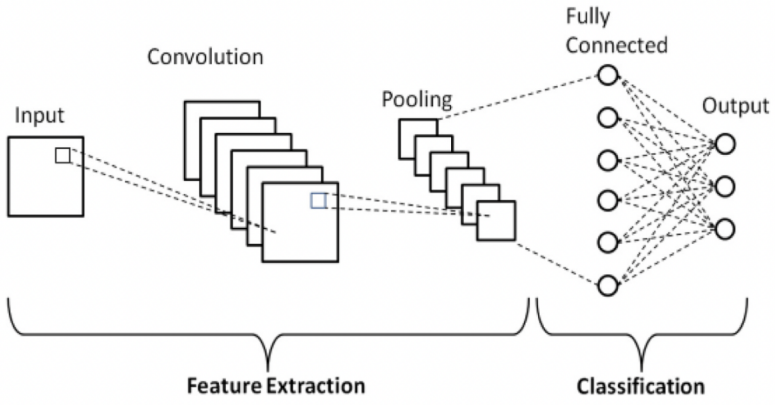

Figura 4 – Diagrama de blocos de uma CNN (Rede Neural Convolucional)

Fonte: (PHUNG; RHEE, 2018)

2.2.2 Vieses e Reconhecimento Facial

No âmbito do reconhecimento facial e *deep learning*, o conceito de "viés" refere-se à propensão dessas soluções para resultados específicos quando expostas a cenários do mundo real. O impacto palpável desses vieses torna-se evidente ao considerarmos a

eficácia dos sistemas de reconhecimento facial em situações específicas. Um exemplo notável é a observação recorrente de menor precisão na identificação de rostos pertencentes a indivíduos racializados. Esse cenário pode variar desde a incapacidade do sistema em reconhecer de forma precisa um rosto de uma pessoa racializada até situações mais preocupantes, como a identificação equivocada da identidade de um indivíduo (WANG et al., 2019). Ou, de maneira equivocada, avaliar incorretamente o acréscimo do risco de uma pessoa cometer outro crime com base em sua aparência.[1]

No contexto da literatura, é amplamente reconhecido que o desafio associado ao viés está frequentemente ligado aos dados utilizados. Essa constatação sugere que os vieses intrínsecos presentes na amostra terão impacto em qualquer modelo derivado desses dados (Caton; Haas, 2020). Contudo, é crucial salientar que essa não é a única fonte de vieses. Evidências de viés também foram identificadas em algoritmos analisados (MEHRABI et al., 2019), enfatizando a importância de uma avaliação crítica das possíveis fontes de parcialidade no contexto do reconhecimento facial que utilizam IA.

Essa análise evidencia que o desafio do viés não pode ser abordado de maneira isolada; requer uma compreensão profunda tanto das características dos conjuntos de dados quanto do funcionamento interno dos algoritmos. A interseção entre esses dois elementos é crucial para mitigar possíveis distorções e garantir que os sistemas de aprendizado de máquina forneçam resultados justos e equitativos em diversas situações.

2.2.3 Impacto de Viéses

> A tecnologia não é nem boa nem má; tampouco é neutra... A interação da tecnologia com a ecologia social é tal que os desenvolvimentos técnicos frequentemente têm consequências ambientais, sociais e humanas que vão muito além dos propósitos imediatos dos próprios dispositivos e práticas técnicas. (KRANZBERG, 1986)

Ao explorar a temática do viés na tecnologia, é imperativo ressaltar inicialmente que a tecnologia em si não carrega intrinsecamente aspectos problemáticos. Incapaz de gerar conteúdo a partir do vazio, a tecnologia constrói-se sobre preexistências, refletindo os vieses da sociedade que a concebeu e, na melhor das hipóteses, amplificando essas preexistências. Uma ilusão comum reside na suposição de neutralidade associada à tecnologia, presumindo que, por ser uma máquina responsável por tomar decisões, essas decisões serão automaticamente imparciais.

Um entendimento frequentemente subestimado no desenvolvimento de soluções de inteligência artificial é que a IA só adquire relevância em nossas vidas porque possui

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

dados suficientes para aprender a tomar decisões por conta própria. Contrariamente ao que pode parecer evidente, é crucial compreender que os dados, por si só, são enviesados, e os algoritmos, por sua vez, também carregam vieses, sendo desenvolvidos com base em conjuntos de dados que refletem os preconceitos da sociedade em que foram coletados.

Os impactos desses vieses são ubíquos, evidenciados por casos notáveis, como a ferramenta Amazon Rekognition, que erroneamente identificou 28 congressistas como criminosos, com um erro 39% maior para congressistas não brancos[2]. Outro exemplo é o Beauty.AI, utilizado para julgar um concurso de beleza, resultando na predominância de vencedoras brancas e/ou de pele clara[3]. O sistema COMPAS, utilizado por juízes para decisões de soltura com base na probabilidade de reincidência, também demonstrou viés, com mais falsos positivos para pessoas negras do que brancas[4]. No Brasil, o Observatório de Segurança revelou que a maioria das pessoas presas pelo sistema de monitoramento da polícia era negra. A lista de exemplos é extensa e tende a crescer sem uma atenção assertiva da sociedade e dos desenvolvedores para a justiça social no contexto de machine learning, especialmente no âmbito do deep learning usado para criar IAs de reconhecimento facial.

Entretanto, o aumento significativo de artigos sobre o tema, incluindo este documento, é uma evidência de que há indivíduos dedicados à reflexão sobre essas questões e que soluções estão sendo buscadas e/ou criadas para lidar com esse problema.

[2] https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28
[3] https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people
[4] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

3 Proposta

Com o intuito de fomentar a equidade e a imparcialidade no âmbito tecnológico do reconhecimento facial, a proposta do presente trabalho consiste na identificação e análise de abordagens recomendadas para abordar tais questões. Serão minuciosamente examinadas publicações científicas relevantes a fim de identificar estratégias propostas, avaliando suas características, aplicabilidade e eficácia. Os resultados advindos desse mapeamento sistemático oferecerão uma visão abrangente das abordagens existentes, contribuindo para a compreensão e o desenvolvimento de soluções mais equitativas e não discriminatórias no campo da inteligência artificial. A seguir, delineia-se o processo de pesquisa adotado neste estudo.

3.1 Metodologia de Pesquisa

3.1.1 Mapeamento Sistemático

Os estudos de mapeamento sistemático têm se mostrado valiosos no campo da engenharia de software. Esses estudos têm como objetivo identificar, avaliar e organizar a literatura de pesquisa existente em uma área específica. Ao fornecer uma visão geral do panorama de pesquisas, eles ajudam a identificar lacunas de pesquisa e direcionam futuras investigações, ao construir de uma base sólida de conhecimento para a área, fornecendo um ponto de partida para pesquisadores interessados em explorar novos temas e direções. (PETERSEN et al., 2008)

No processo de nosso estudo de mapeamento sistemático, são identificadas etapas essenciais que compõem o fluxo de trabalho. A primeira etapa consiste na definição das questões de pesquisa, que orientam a busca de artigos relevantes para a análise. Na segunda etapa, é realizada a busca de artigos por meio de bases de dados e outras fontes pertinentes, utilizando critérios de seleção específicos. Em seguida, ocorre a triagem dos artigos obtidos, por meio da avaliação dos títulos, resumos e, em alguns casos, textos completos, para determinar a relevância em relação às questões de pesquisa estabelecidas.

A quarta etapa envolve a definição de palavras-chave a serem extraídas dos resumos dos artigos selecionados. Essas palavras-chave são utilizadas para categorizar os estudos de acordo com os temas e aspectos relevantes identificados. Por fim, ocorre a extração e o mapeamento dos dados relevantes contidos nos artigos, que são sintetizados e organizados em um formato adequado.

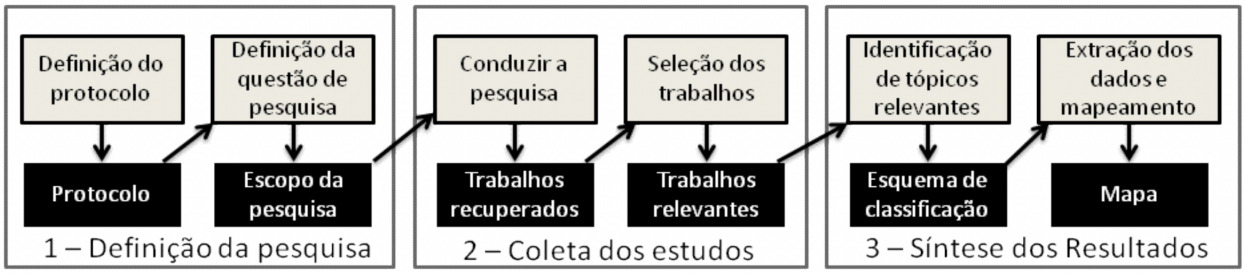

Figura 5 – Processo de Mapeamento Sistemático

Fonte: (JR.; OLIVEIRA; MEIRA, 2013)

Cada etapa do processo contribui para a construção gradual do mapa sistemático, que representa uma visualização estruturada e abrangente das informações obtidas ao longo do estudo. O resultado final desse processo é a criação do mapa sistemático, que oferece uma síntese dos estudos identificados, suas características e contribuições relevantes para a área de pesquisa em questão.

3.1.2 Questão de pesquisa

Conforme mencionado previamente, a primeira etapa do mapeamento sistemático consiste na formulação da questão de pesquisa que norteará o presente estudo. A seguir, é apresentado a questão de pesquisa definida.

- Quais estratégias são propostas na literatura para mitigar o enviesamento racial em sistemas de inteligência artificial voltados para reconhecimento facial?

3.1.3 Estratégia de Pesquisa

3.1.3.1 Bases de Dados

Foram realizadas buscas de artigos em duas bibliotecas digitais renomadas, a IEEExplore e a ACM Digital Library.

3.1.3.2 Strings de Buscas

Uma etapa crucial na estratégia de pesquisa é a definição das strings de busca, que desempenham um papel essencial na triagem precisa de estudos com temas relevantes. Para construí-las, foram selecionadas palavras-chave relacionadas aos tópicos de interesse.

Tabela 1 – Strings de Buscas

ID	**String de Busca**
SB1	(((Deep learning) AND face recognition) AND bias)
SB2	(((artificial intelligence) AND face recognition) AND bias)
SB3	((racial AND bias) AND artificial intelligence)
SB4	((racial OR bias) AND (artificial intelligence AND database))

3.1.3.3 Critérios de Inclusão e Exclusão

A aplicação de critérios de inclusão e exclusão desempenhou um papel essencial na condução deste estudo de pesquisa. Esses critérios determinaram as características necessárias para a inclusão de estudos na pesquisa, bem como as razões para a rejeição ou exclusão. Eles desempenharam um papel vital na busca, triagem e seleção de estudos, assegurando que estivessem alinhados com os objetivos da pesquisa, minimizando possíveis vieses e mantendo a integridade dos resultados. Nesse sentido, os seguintes critérios foram empregados:

- CI-1 O artigo deve estar no idioma inglês.
- CI-2 O artigo deve possuir versão virtual disponível de forma gratuita.
- CI-3 O artigo deve apresentar estudos relevantes ao tema proposto.
- CI-4 O artigo deve apresentar estudos no campo de reconhecimento facial.
- CI-5 O artigo deve propor uma solução para o problema tema do presente estudo.
- CI-6 O artigo deve ter sido publicado em periódicos ou conferências com revisão por pares.

3.1.4 Resultados do Mapeamento Sistemático

Utilizando as strings de busca delineadas na Tabela 1, foram reunidos ao todo 682 artigos. Com base nos critérios apresentados na seção 3.1.3.3, 95% (649) dos artigos foram rejeitados. Dentro desse percentual, 61,94% (402) foram excluídos devido ao critério CI4, que refere-se a artigos que não abordavam o tema do reconhecimento facial. Outros 31,21% (202) foram eliminados pelo critério CI3, indicando que não apresentavam estudos relevantes para o tema em questão, 5,24% (34) foram rejeitados pelo critério CI2, pois não dispunham de uma versão gratuita disponível. Adicionalmente, 2,16% (14) foram rejeitados pelo critério CI5, não apresentavam solução para o problema tema da pesquisa. Todos os artigos coletados atenderam aos critérios CI1 e CI6.

Tabela 2 – Dados dos Artigos Coletados

	IEEE	**ACM**	**Total**
Total de Artigos Analisados	260	22	282
Total de Artigos Rejeitados	589	60	649
Total de Artigos Coletados	622	60	682

Dos 682 artigos, 41,34% (282) foram examinados por meio da leitura de seus títulos e resumos. Dentro desse conjunto, 10,3% (29) foram selecionados para uma leitura integral.

Tabela 3 – Dados dos Artigos Selecionados para leitura

	IEEE	**ACM**	**Total**
Total de Artigos Selecionados	29	0	29

Dos 29 artigos selecionados para leitura integral 41,38% (12) foram escolhidos para integrar o mapeamento sistemático (ver apêndice A).

Tabela 4 – Dados dos Artigos Escolhidos

	IEEE	**ACM**	**Total**
Total de Artigos que propõem resolução	12	0	12
Total de Artigos que não propõem resolução	17	0	17

4 Resultados

4.1 Quais estratégias são propostas na literatura para mitigar o enviesamento racial em sistemas de inteligência artificial voltados para reconhecimento facial?

Adotando a classificação proposta por Wang (2020), os artigos mapeados foram categorizados como pertencentes a uma das seguintes classes:

- *Strategic Sampling Methods*: Refere-se aos métodos que empregam oversampling e reponderação para ajustar a seleção dos dados de treinamento, com o propósito de enfrentar desequilíbrios nas distribuições de classes.

- *Representation Disentanglement Methods*: São métodos que utilizam treinamento adversário para eliminar atributos demográficos específicos, visando maximizar a capacidade do classificador de prever a classe, ao mesmo tempo em que minimizam a habilidade do adversário em prever a variável protegida com base nas características aprendidas.

- *Domain Adaptation Methods*: Métodos nos quais se adquirem representações invariantes para grupos demográficos ao maximizar o desempenho no reconhecimento de identidade e ao minimizar a capacidade de prever atributos protegidos através de uma perda de discriminação.

- *Domain Independent Training Methods*: Os métodos dessa estratégia compreendem um tipo de treinamento no qual o modelo é instruído a adquirir um conjunto de classificadores individualizados para cada grupo demográfico, ao mesmo tempo que utiliza uma base compartilhada de características para realizar as respectivas classificações.

Devido à não conformidade do artigo A5 com qualquer uma das abordagens anteriormente mencionadas, foi estabelecida uma nova categoria denominada “Abordagem de Conjunto de Dados” (*Dataset Approach*). Essa abordagem tem como objetivo mitigar o viés em um modelo de reconhecimento facial por meio da criação de um conjunto de dados de treinamento que minimiza os efeitos de enviesamento do modelo. Além disso, foram identificados estudos que empregam a combinação de duas das estratégias aqui já mencionadas.

4.1.1 Sumário dos Achados na Literatura

4.1.1.1 Strategic Sampling Methods

O estudo [4] emprega a técnica de reponderação como abordagem para ajustar os pesos associados a grupos específicos. Grupos com representação reduzida no conjunto de dados de treinamento são designados com pesos superiores, buscando, assim, modular o processo de aprendizado do modelo de modo a equalizar a influência desses grupos em comparação aos grupos mais representados.

O estudo [6] adota a técnica de *oversampling*, a qual envolve o aumento do número de instâncias da classe menos frequente, com o objetivo de alcançar um equilíbrio entre as classes na base de dados de treinamento em relação às frequências de domínio-classe.

4.1.1.2 Representation Disentanglement Methods

O estudo [2] emprega a abordagem de “SensitiveNets” para mitigar a presença de informações sensíveis ou confidenciais durante o treinamento. "SensitiveNets"são redes que recebem como entrada a saída de uma rede pré-treinada, a tarefa que se deseja reforçar e a tarefa que se deseja prevenir. Essa incorporação no modelo visa a minimização da capacidade da rede em prever informações sensíveis, como raça e gênero.

O estudo [3] emprega rótulos de identidade e de subgrupo para desenvolver um mapeamento que preserva informações de identidade através de uma função de perda dedicada à identidade, enquanto simultaneamente elimina informações de subgrupo por meio de uma função de perda associada ao atributo (informações sensíveis que identificam um subgrupo).

O estudo [8] introduz o conceito de “sample-level multi-variation cosine margin (MvCoM)” como uma abordagem flexível para capturar as variações em nível de amostra e enfrentar desequilíbrios nas distribuições de dados, representando diversas fontes de variação ao nível da amostra. O MvCoM é ajustado dinamicamente por meio de uma abordagem de meta-aprendizado em três estágios. Este método considera explicitamente fatores de variação, tais como classe, etnia, pose da cabeça, desfoque e oclusão, e procede ao treinamento de um classificador para cada variação com o objetivo de quantificar o viés associado.

O estudo [10] propõe a adoção de um algoritmo de aprendizado que desvincula a dependência do modelo em relação a atributos sensíveis, comumente denominados como covariáveis de viés. O algoritmo Filter-Drop realiza a remoção de filtros convolucionais responsáveis por codificar informações específicas associadas a atributos sensíveis, como etnia. Ao longo do treinamento, os filtros que encapsulam informações sobre a etnia são eliminados, tornando as previsões independentes desse atributo sensível. A técnica Filter-Drop é concebida para eliminar características vinculadas a atributos sensíveis, promo-

vendo, assim, previsões desprovidas de viés. O método envolve o aprendizado de quais filtros devem ser removidos durante o treinamento por meio de uma rede multi-tarefa, que executa tanto a tarefa principal (previsão de gênero) quanto uma tarefa secundária de previsão de covariáveis de viés (classificação de etnia). A remoção de filtros é realizada com base nas contribuições ponderadas para a previsão correta do atributo sensível.

4.1.1.3 Domain Adaptation Methods

O estudo [7] introduz um novo protocolo de avaliação para viés demográfico, fundamentado nas Taxas de Falsos Positivos (FPRs), evidenciando uma consideração explícita do domínio demográfico. Apresenta uma estratégia para mitigar o viés no reconhecimento facial, com foco em aumentar a consistência da Taxa de Falsos Positivos por instância. Desenvolve uma penalidade na função de perda softmax com o intuito de atenuar o viés e aprimorar a equidade de desempenho no reconhecimento facial. Propõe o conceito de FPR por instância como um caso extremo de FPR demográfico, indicando uma abordagem de adaptação de domínio para tornar as representações mais invariantes em relação aos grupos demográficos.

O estudo [12] apresenta a concepção de uma margem adaptativa para lidar com o desafio do equilíbrio racial. Nessa proposta, as margens permanecem inalteradas para indivíduos caucasianos, enquanto margens ótimas são dinamicamente selecionadas para cada grupo racial, com o intuito de minimizar a assimetria nos ângulos entre as raças e alcançar um desempenho equilibrado. A substituição da margem fixa m por um parâmetro relacionado à raça e à etapa de treinamento $m_j(t)$ sugere uma abordagem dinâmica, considerando as demandas variáveis das distintas raças ao longo do processo de treinamento. A formulação do problema para determinar $m_j(t)$ como um Processo de Decisão de Markov (MDP) e a aplicação de deep Q-learning para ajustar a margem em cada iteração indicam uma estratégia adaptativa ao longo do tempo.

4.1.1.4 Domain Adaptation Methods e Domain Independent Training Method

O estudo [1] emprega duas abordagens distintas, fazendo uso de kernels dinâmicos e mapas de atenção, com o propósito de aprender representações invariantes em relação ao grupo demográfico. Esta metodologia visa maximizar o desempenho no reconhecimento de identidade, enquanto minimiza a capacidade de predição de atributos protegidos, por meio da incorporação de uma função de perda de discriminação associada a um Classificador Adaptativo de Grupo (GAC). O GAC é composto por módulos adaptativos, os quais englobam camadas específicas destinadas a cada grupo demográfico, proporcionando, assim, uma abordagem adaptativa e discriminativa no processo de aprendizado.

O estudo [9] propõe uma estratégia para equilibrar as margens específicas de diferentes grupos populacionais, visando mitigar o viés algorítmico. Introduz o método

denominado “meta balanced network” (MBN), que emprega um conjunto de dados meta reduzido para aprender de maneira automática margens ótimas através de uma técnica de diferenciação automática de segunda ordem na meta-otimização. A utilização de meta-aprendizado sugere uma abordagem adaptativa de domínio, buscando atender a diferentes requisitos em grupos demográficos específicos, com o propósito de aprimorar a equidade no reconhecimento facial, independentemente das características distintivas do domínio em questão.

4.1.1.5 Representation Disentanglement Methods e Domain Independent Training Method

O estudo [11] apresenta uma proposta de perda de atributo consciente, que busca regularizar a proximidade das características em relação à proximidade de atributos, como gênero, etnia e idade. O objetivo primordial é otimizar a coerência das características aprendidas com base na similaridade dos atributos. A perda de atributo consciente é concebida para aproximar clusters de características associadas a atributos semelhantes, indicando uma tentativa de dissociar a representação de características de atributos específicos. A metodologia aborda atributos, como gênero, etnia e idade, considerados independentes do processo de imagem. Ao otimizar a proximidade das características com base nos atributos, a abordagem proposta sugere um método para a aprendizagem de representações independentes do domínio.

4.1.1.6 Dataset Approach

O estudo [5] emprega um conjunto de dados sintético, incorporando uma pequena proporção de imagens reais para treinamento do modelo. Essa abordagem singular permite que o modelo alcance maior precisão ao direcionar sua atenção para regiões específicas, tais como boca, nariz e olhos. Em contraste, modelos treinados com conjuntos de dados exclusivamente reais tendem a ter uma abrangência mais generalizada e uma focalização mais dispersa, incluindo até mesmo elementos do cenário circundante.

4.1.2 Desvantagens e Limitações

Embora todas as estratégias discutidas anteriormente tenham revelado resultados significativos na atenuação do viés em soluções de reconhecimento facial, é imperativo abordar algumas das desvantagens inerentes a cada abordagem.

Tomando, por exemplo, a metodologia de *Strategic Sampling Methods*, destaca-se o risco de sobreajuste, caracterizado pela adaptação excessiva a comportamentos específicos durante o treinamento. Esse fenômeno pode resultar em um modelo eficaz na previsão de dados já contemplados no treinamento, mas com desempenho deficiente quando confrontado com novos dados. Adicionalmente, essa abordagem enfrenta o desafio do aumento

do tempo de aprendizado, sem que haja uma correspondente elevação na quantidade de informação útil adquirida durante esse período.

Na abordagem *Representation Disentanglement Methods*, embora seja uma estratégia intuitiva, ela apresenta uma desvantagem significativa denominada "codificação redundante". Mesmo na ausência de um atributo protegido específico na representação de características de um classificador, a combinação de outros atributos pode ser utilizada como um substituto, resultando na introdução de um viés indesejado. Esta limitação destaca a complexidade associada à busca pela equidade em modelos, uma vez que há o risco inadvertido de excluir informações relevantes.

Já na abordagem *Domain Adaptation Methods*, uma possível limitação pode ser a dificuldade em lidar com variações complexas e não lineares nos dados demográficos, pois o método visa adquirir representações invariantes. Além disso, se a relação entre as características demográficas e as classes não for linear ou se houver sobreposição substancial entre diferentes grupos demográficos, isso pode representar um desafio para o desempenho eficaz do modelo.

As estratégias de *Domain Independent Training Methods* podem apresentar uma desvantagem ao demandar a criação de conjuntos de classificadores individualizados para cada grupo demográfico. Tal abordagem pode exigir uma quantidade considerável de dados específicos para cada grupo, o que pode se tornar impraticável em cenários nos quais alguns grupos demográficos têm uma representação limitada nos dados de treinamento. Adicionalmente, a gestão e a atualização eficientes de múltiplos classificadores destinados a diferentes grupos demográficos podem ser tarefas complexas, demandando recursos significativos. Portanto, a escalabilidade e a aplicabilidade generalizada dessa abordagem para uma variedade de grupos demográficos podem se apresentar como desafios relevantes.

Por fim, a estratégia aqui intitulada como *Dataset Approach* é frequentemente pouco explorada, pois não apresenta eficácia na redução dos vieses preexistentes nos algoritmos de reconhecimento facial. Importante observar que tais vieses não dependem exclusivamente do conjunto de treinamento para sua existência.

5 Conclusão

O presente estudo aborda de maneira crítica a problemática dos vieses incorporados em algoritmos de *deep learning* no contexto de reconhecimento facial. A pesquisa evidencia a necessidade de uma abordagem ética na identificação e mitigação desses vieses, destacando a urgência na criação de estratégias para evitar impactos negativos em grupos marginalizados.

A metodologia de mapeamento sistemático empregada neste estudo revelou-se eficaz para identificar e organizar a literatura existente na área de estudo, proporcionando uma visão abrangente das práticas e tendências atuais. As etapas definidas para o processo, desde a formulação de questões de pesquisa até a extração e mapeamento de dados relevantes, forneceram uma estrutura sólida para analisar e compreender a pesquisa existente.

Os resultados apresentados no apêndice A e no capítulo 4 oferecem uma visualização estruturada e abrangente das informações obtidas, destacando características e contribuições dos estudos identificados. Além disso, o mapeamento sistemático sintetiza esses resultados, proporcionando uma visão consolidada da pesquisa na área.

Como perspectiva para futuras investigações, sugerimos a identificação e análise das métricas e indicadores utilizados para medir enviesamentos em reconhecimento facial. Explorar como essas métricas e indicadores podem ser incorporadas nas estratégias para reduzir esses enviesamentos pode contribuir significativamente para o avanço da área de pesquisa.

Referências

Caton, S.; Haas, C. Fairness in Machine Learning: A Survey. 2020. Available online: <https://dl.acm.org/doi/10.1145/3616865>. Citado na página 23.

CHOLLET, F. Xception: Deep learning with depthwise separable convolutions. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, 2017. p. 1800–1807. ISSN 1063-6919. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.195>. Nenhuma citação no texto.

GONG, S.; LIU, X.; JAIN, A. K. Mitigating face recognition bias via group adaptive classifier. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2021. p. 3413–3423. Nenhuma citação no texto.

JR., J. J. D.; OLIVEIRA, J. A. de; MEIRA, S. Estudo empírico sobre adoção de soa: Um mapeamento sistemático da literatura. In: *Anais do XII Simpósio Brasileiro de Qualidade de Software*. Porto Alegre, RS, Brasil: SBC, 2013. p. 238–252. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbqs/article/view/15286>. Citado na página 26.

KRANZBERG, M. Technology and history: 'kranzberg's laws'. *Technology and Culture*, v. 27, n. 3, p. 544, 1986. Citado na página 23.

LI, J.; ABD-ALMAGEED, W. Information-theoretic bias assessment of learned representations of pretrained face recognition. In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. [S.l.: s.n.], 2021. p. 1–8. Nenhuma citação no texto.

LITTMAN, M. L. Reinforcement learning improves behaviour from evaluative feedback. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 445, 2015. Nenhuma citação no texto.

MEHRABI, N. et al. A survey on bias and fairness in machine learning. 2019. Disponível em: <https://doi.org/10.1145/3457607>. Citado na página 23.

PERRAULT, R. et al. *The AI Index 2019 Annual Report*. Stanford, CA, 2019. Nenhuma citação no texto.

PETERSEN, K. et al. Systematic mapping studies in software engineering. 2008. Available online: <https://www.researchgate.net/publication/228350426_Systematic_Mapping_Studies_in_Software_Engineering>. Citado na página 25.

PHUNG, V.; RHEE, E. A deep learning approach for classification of cloud image patches on small datasets. *J. Inf. Commun. Converg. Eng*, v. 16, n. 3, p. 173–178, 2018. Citado na página 22.

ROSENBLATT, F. The perceptron, a perceiving and recognizing automaton. *Project Para, Cornell Aeronautical Laboratory*, 1957. Nenhuma citação no texto.

RUSSELL, S. J.; NORVIG, P. *Artificial intelligence a modern approach*. London: Pearson Education, Inc, 2010. (Third Edition). ISBN 9780136042594. Nenhuma citação no texto.

SIMONS, G. T. Introdução a inteligência artificial. *Classe*, 1988. Nenhuma citação no texto.

WANG, M. et al. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. IEEE Computer Society, Seoul, Korea (South), 2019. Available online: <https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00078>. Citado na página 23.

WANG, Z. et al. Towards fairness in visual recognition: Effective strategies for bias mitigation. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.: s.n.], 2020. p. 8916–8925. Nenhuma citação no texto.

Apêndices

APÊNDICE A – Mapeamento Sistemático

ID	Título	Autor(es)	Abordagem
1	Mitigating Face Recognition Bias via Group Adaptive Classifier	Sixue Gong; Xiaoming Liu; Anil K. Jain	Domain Adaptation Method e Domain Independent Training Method
2	SensitiveNets: Learning Agnostic Representations with Application to Face Images	Aythami Morales; Julian Fierrez; Ruben Vera-Rodriguez; Ruben Tolosana	Representation Disentanglement Method
3	Balancing Biases and Preserving Privacy on Balanced Faces in the Wild	Joseph P. Robinson; Can Qin; Yann Henon; Samson Timoner; Yun Fu	Representation Disentanglement Method
4	Reducing Geographic Performance Differentials for Face Recognition	Martins Bruveris; Pouria Mortazavian; Jochem Gietema; Mohan Mahadevan	Strategic Sampling Method
5	A Dataless FaceSwap Detection Approach Using Synthetic Images	Anubhav Jain; Nasir Memon; Julian Togelius	Dataset Approach
6	Epistemic Uncertainty-Weighted Loss for Visual Bias Mitigation	Rebecca S Stone; Nishant Ravikumar; Andrew J Bulpitt; David C Hogg	Strategic Sampling Method
7	Consistent Instance False Positive Improves Fairness in Face Recognition	Xingkun Xu; Yuge Huang; Pengcheng Shen; Shaoxin Li; Jilin Li; Feiyue Huang; Yong Li; Zhen Cui	Domain Adaptation Method
8	Learning to Learn across Diverse Data Biases in Deep Face Recognition	Chang Liu; Xiang Yu; Yi-Hsuan Tsai; Masoud Faraki; Ramin Moslemi; Manmohan Chandraker; Yun Fu	Representation Disentanglement Method

9	Meta Balanced Network for Fair Face Recognition	Mei Wang; Yaobin Zhang; Weihong Deng	Domain Adaptation Method e Domain Independent Training Method
10	Attribute Aware Filter-Drop for Bias-Invariant Classification	Shruti Nagpal; Maneet Singh; Richa Singh; Mayank Vatsa	Representation Disentanglement Method
11	Robust RGB-D Face Recognition Using Attribute-Aware Loss	Luo Jiang; Juyong Zhang; Bailin Deng	Representation Disentanglement Method e Domain Independent Training Method
12	Mitigate Bias in Face Recognition using Skewness-Aware Reinforcement Learning	Mei Wang; Weihong Deng	Domain Adaptation Method