Investigando o Viés em Deep Learning para Classificação de Gênero e Raça: Um Estudo Comparativo com ConvNeXt em Datasets Balanceados e Enviesados

Gregory G. Ozaki Coelho[1], Taíza P. de Oliveira Lima[1], Leano Guerreiro Baba[1]

[1]Instituto de Ciências Exatas e Tecnologia (ICET) – Universidade Federal do Amazonas (UFAM) – Itacoatiara – AM – Brasil

{gregory.coelho, taiza.lima, leano.baba}@ufam.edu.br

Abstract. *This work investigates the impact of data bias on race and gender rediction through a comparative experimental approach. The balanced FairFace and unbalanced CelebA datasets were used, with attribute preprocessing and harmonization. A multitask ConvNeXt-Tiny model was trained in both intra- and cross-domain scenarios. The evaluation included traditional metrics and intersectional analysis to identify performance disparities. Results show that biased data lead to poorer performance and greater bias, while balanced data promote higher accuracy, robustness, and fairness. It is concluded that data balancing is essential for the effectiveness and fairness of deep learning systems.*

Resumo. *Este trabalho investiga o impacto do viés de dados na predição de raça e gênero por meio de uma abordagem experimental comparativa. Foram utilizados os datasets FairFace, balanceado, e CelebA, desbalanceado, com pré-processamento e harmonização dos atributos. Um modelo ConvNeXt-Tiny multitarefa foi treinado em cenários intra e cross-domain. A avaliação incluiu métricas tradicionais e análise interseccional para identificar desigualdades na performance. Os resultados mostram que dados enviesados levam a pior desempenho e maior viés, enquanto dados balanceados promovem maior precisão, robustez e equidade. Conclui-se que o balanceamento dos dados é essencial para a eficácia e justiça dos sistemas de aprendizado profundo.*

1. Introdução

A inteligência artificial (IA) tem sido amplamente empregada em tarefas de classificação de imagens, como a identificação de gênero e raça em sistemas de reconhecimento facial [Sarker, 2021]. No entanto, essas aplicações levantam preocupações éticas significativas, principalmente por sua tendência a reproduzir preconceitos históricos, uma vez que modelos de aprendizado profundo, ao serem treinados com dados desbalanceados, podem apresentar comportamentos enviesados com impactos críticos em áreas sensíveis como segurança, saúde e recursos humanos.

Estudos mostram que algoritmos de classificação demográfica tendem a ter maior acurácia para homens brancos, enquanto grupos minoritários – como mulheres negras – são frequentemente mal classificados [Buolamwini e Gebru, 2018; Pagano et al., 2023]. Esse viés algorítmico decorre, em grande parte, da composição desigual dos dados de treinamento, que refletem desigualdades sociais [Mehrabi et al., 2021].

Diante disso, este estudo investiga o impacto do desbalanceamento de dados no desempenho e na equidade da arquitetura ConvNeXt, uma rede neural convolucional de última geração [Liu et al., 2022]. Para isso, foi conduzido uma análise comparativa entre os conjuntos de dados CelebA, conhecido por seus vieses de representação, e FairFace, projetado para maior diversidade demográfica. A avaliação é realizada não apenas em cenários intra-domínio (treino e teste no mesmo dataset), mas também em cenários de domínio cruzado (*cross-domain*), expondo a robustez do modelo a distribuições de dados diferentes daquelas vistas no treinamento.

A hipótese central é que o treinamento com dados desbalanceados resulta em menor desempenho e maior viés interseccional, especialmente quando avaliado em outro domínio. Ao quantificar essas disparidades, o trabalho busca evidenciar a importância da curadoria e do balanceamento de dados para modelos de IA mais justos e confiáveis.

2. Objetivos

2.1 Objetivo Geral

Investigar o impacto do viés em dados de treinamento no desempenho e na equidade de um modelo de Deep Learning (ConvNeXt) para as tarefas de classificação de gênero e raça, por meio de uma análise experimental comparativa.

2.2 Objetivos Específicos

Para alcançar o objetivo geral, foram definidos os seguintes objetivos específicos:

- Realizar o pré-processamento e a harmonização dos datasets CelebA (enviesado) e FairFace (balanceado) para garantir a comparabilidade dos experimentos.
- Implementar e treinar um modelo de aprendizado multitarefa baseado na arquitetura ConvNeXt-Tiny para a classificação simultânea de gênero e raça.
- Conduzir uma análise comparativa em quatro cenários experimentais: intra-domínio e domínio cruzado (treino em um dataset e teste no outro).
- Avaliar e comparar o desempenho preditivo dos modelos utilizando métricas de classificação tradicionais, como acurácia, perda, recall e F1-Score.
- Analisar a equidade (*fairness*) dos modelos por meio de uma avaliação desagregada e interseccional, a fim de identificar e quantificar disparidades de desempenho entre diferentes grupos demográficos.

3. Fundamentação Teórica

3.1 Inteligência Artificial e Aprendizado Profundo

A Inteligência Artificial é a área da ciência da computação que busca simular a cognição humana, primariamente por meio do Aprendizado de Máquina (do inglês, Machine Learning – ML), que dota os sistemas da capacidade de aprender com dados. O Aprendizado Profundo (Deep Learning – DL), subcampo do ML, utiliza Redes Neurais com múltiplas camadas para modelar padrões complexos em dados não estruturados, como imagens, áudio e texto [LeCun et al., 2015; Sarker, 2021].

A arquitetura das Redes Neurais é uma abstração de neurônios que processam sinais de entrada e geram saídas por meio de funções de ativação não-lineares. Modelos como as Redes Neurais Convolucionais (CNNs) são proeminentes para o reconhecimento

de imagens e compõem a base dos sistemas de visão computacional, área que permite aos computadores interpretarem e extrair informações de dados visuais, enquanto as Redes Neurais Recorrentes (RNNs) se especializam em dados sequenciais [LeCun et al., 2015; Erickson et al., 2017; Barros et al., 2024]. As múltiplas camadas das redes permitem aprender representações cada vez mais abstratas, melhorando o desempenho em tarefas como classificação e detecção [Najafabadi et al., 2015; Sarker, 2021].

Entre as arquiteturas modernas de redes neurais convolucionais, destaca-se o ConvNeXt, desenvolvido para otimizar o desempenho em tarefas de visão computacional. Essa arquitetura combina os princípios clássicos das CNNs com inovações recentes, como mecanismos de atenção e técnicas de regularização, oferecendo maior precisão e eficiência em comparação com modelos anteriores [Liu et al., 2022; Alsharif et al., 2023]. Sua estrutura baseada em múltiplas camadas convolucionais permite extrair características visuais relevantes das imagens, sendo especialmente eficaz em aplicações que demandam alto desempenho em reconhecimento facial e análise de atributos visuais. Porém, a alta capacidade de extração de características de modelos como o ConvNeXt também significa que, se treinados com dados enviesados, eles podem aprender e amplificar preconceitos sociais, tornando a análise de viés nessas arquiteturas um campo de estudo crucial.

Desse modo, o avanço do DL tem sido impulsionado pelo aumento do poder computacional e pela abundância de dados [Yuan, 2023]. No entanto, desafios permanecem, como a falta de interpretabilidade dos modelos [Kaushik, 2023] e a dificuldade de lidar com dados desbalanceados – problemas que demandam técnicas como *data augmentation* e regularização [Upadhyay, 2017]. Assim, o desenvolvimento do DL caminha junto à necessidade de enfrentar questões éticas e de transparência decorrentes do impacto social de suas aplicações [Najafabadi et al., 2015].

Nesse contexto, para lidar com tarefas complexas como a classificação simultânea de gênero e raça, o aprendizado multitarefa (*multi-task learning* – MTL) surge como uma abordagem relevante. Essa técnica tipicamente utiliza uma arquitetura com um extrator de características compartilhado (*backbone*), forçando o modelo a aprender representações mais ricas e generalizáveis. Ao fazer isso, o MTL pode não apenas melhorar a eficiência, mas também mitigar o viés, uma vez que o aprendizado conjunto desencoraja o modelo de se apoiar em correlações espúrias associadas a apenas uma tarefa [Barros et al., 2024].

3.2 Viés e Discriminação Algorítmica

O viés algorítmico refere-se à tendência de sistemas de inteligência artificial e aprendizado de máquina produzirem resultados injustos ou prejudiciais, refletindo preconceitos presentes nos dados de treinamento ou nas decisões de design dos próprios algoritmos. Esses vieses podem surgir tanto da sub-representação de determinados grupos nos dados quanto de escolhas técnicas feitas durante o desenvolvimento, o que resulta em um algoritmo enviesado – ou seja, em decisões automatizadas que tratam indivíduos de forma desigual com base em características como raça, gênero ou classe social [Rossetti e Angeluci, 2021; Henriques e Sampaio, 2022; Elias et al., 2023].

A discriminação algorítmica raramente ocorre em uma única dimensão, intensificando-se na interseção de identidades como raça e gênero. Uma análise que avalia o viés de forma isolada para cada atributo pode ser insuficiente e enganosa, pois oculta

disparidades de desempenho que afetam desproporcionalmente subgrupos específicos, a exemplo de mulheres negras [Buolamwini e Gebru, 2018]. Portanto, a adoção de uma abordagem interseccional é essencial para uma auditoria de justiça rigorosa, pois ao desagregar os resultados por grupos combinados, permite revelar vieses ocultos que uma análise unidimensional não capturaria [Lett et al., 2024].

O principal fator por trás desses vieses está na natureza dos dados. Conjuntos de dados com baixa representatividade de certos grupos – como negros, mulheres, idosos ou populações LGBTQIA+ – tendem a produzir modelos discriminatórios. Por exemplo, algoritmos de reconhecimento facial falham mais com pessoas negras devido à sua sub-representação nos dados de treinamento [Aragão et al., 2022].

O conjunto de dados CelebA, por exemplo, é amplamente utilizado em pesquisas relacionadas à visão computacional, especialmente em tarefas de reconhecimento facial. No entanto, é conhecido por ter uma super-representação de indivíduos de etnia branca e por retratar celebridades sob padrões de beleza eurocêntricos, o que pode levar modelos treinados nele a falhar na identificação ou a gerar atribuições estereotipadas para grupos sub-representados [Fiorati et al., 2018; Ferreira e Salles, 2022].

Por outro lado, o FairFace foi projetado com o objetivo de mitigar esses vieses, proporcionando uma representação mais equilibrada entre diferentes grupos demográficos. Embora tenha sido desenvolvido com a intenção de promover a equidade, ainda pode apresentar imperfeições e potenciais vieses [Ribeiro e Borges, 2022; Júnior et al., 2024]. A eficácia em combater desigualdades sociais por meio de algoritmos depende, em grande parte, da curadoria e dos métodos de seleção de dados utilizados no treinamento.

A performance de um modelo de aprendizado profundo pode degradar significativamente quando exposto a um conjunto de dados com distribuição diferente daquela utilizada em seu treinamento, um fenômeno conhecido como *domain shift*. Essa falha de generalização ocorre frequentemente quando o modelo, treinado em um domínio enviesado, aprende correlações espúrias – associações não causais, como vincular um cenário a um grupo demográfico – em vez de características semanticamente robustas. Por essa razão, a avaliação de domínio cruzado (*cross-domain*), que consiste em treinar um modelo em um domínio (e.g., enviesado) e testá-lo em outro (e.g., balanceado), estabeleceu-se como uma metodologia eficaz para expor e quantificar o viés algorítmico, tornando visível a dependência do modelo em atalhos aprendidos a partir de dados não representativos [Kim et al., 2019; Chen et al., 2020].

3.3 Casos de Discriminação Algorítmica

Casos de discriminação algorítmica têm surgido em diversas aplicações, refletindo as falhas inerentes a sistemas baseados em aprendizado de máquina que não levam em consideração a equidade nas suas operações. Vários estudos e análises destacam essas disparidades e suas consequências.

Um exemplo emblemático de viés algorítmico é o uso de sistemas de previsão de crimes, que têm revelado forte tendência ao enviesamento racial. Em diversas localidades, algoritmos foram empregados para identificar áreas com maior probabilidade de atividade criminosa. Contudo, esses sistemas frequentemente superestimam a criminalidade em regiões habitadas majoritariamente por minorias étnicas, reforçando estigmas sociais históricos [Kassam e Marino, 2022; Pagano et al., 2023]. Essa distorção ocorre, em

grande parte, pela chamada "*proxy discrimination*": atributos aparentemente neutros, como endereço ou bairro, funcionam como substitutos para variáveis sensíveis, como raça. Mesmo sem utilizar diretamente a raça, o algoritmo acaba reproduzindo seus efeitos ao se basear em dados correlacionados, como o CEP, marcada por uma segregação geográfica histórica [Kassam e Marino, 2022].

Outro exemplo significativo ocorre na medicina, particularmente no diagnóstico de doenças. Pesquisas mostraram que algoritmos de aprendizado profundo usados em diagnósticos de imagens médicas podem refletir preconceitos existentes na formação dos dados. Por exemplo, um estudo que analisou modelos destinados a detectar câncer de pele revelou que esses sistemas tinham desempenho inferior em populações menos representadas nos conjuntos de dados de treinamento, resultando em diagnósticos imprecisos e na possibilidade de que certos grupos recebessem menos cuidados adequados [Lin et al., 2023]. Esses casos ressaltam a responsabilidade ética dos desenvolvedores ao criar modelos que impactam a vida e a saúde de indivíduos.

Além disso, sistemas de recrutamento e seleção frequentemente incorporam algoritmos que favorecem candidatos com base em características demográficas. Um relatório sobre discriminação algorítmica em sistemas de recrutamento revelou que esses algoritmos podem discriminar injustamente com base em gênero ou etnia, mesmo quando não há informações diretas sobre esses atributos no conjunto de dados [Kassam e Marino, 2022; Pagano et al., 2023]. Esse fenômeno é um exemplo claro de como a "discriminação algorítmica" pode se manifestar de forma insidiosa, levando a impactos negativos na vida profissional dos indivíduos afetados.

A exploração dessas questões tem gerado um aumento no interesse pela pesquisa em *fairness* no aprendizado de máquina, levando à criação de diretrizes e ferramentas para identificar e mitigar viés em sistemas algorítmicos [Kamiran et al., 2012; Mehrabi et al., 2021; Wang et al., 2023]. Por isso, a necessidade de uma abordagem consciente e responsável no desenvolvimento de algoritmos é mais evidente do que nunca, especialmente quando esses sistemas estão profundamente integrados em decisões que afetam diretamente indivíduos e comunidades.

4. Metodologia

Esta pesquisa quantitativa, experimental e comparativa investiga o impacto do viés de dados na predição de raça e gênero por redes neurais profundas. A partir de uma revisão bibliográfica, conduziu-se um experimento com os datasets FairFace [Kärkkäinen e Joo, 2021] e CelebA [Liu et al., 2015] usando o modelo ConvNeXt [Liu et al., 2022].

4.1 Conjunto de Dados

Para a condução dos experimentos, foram selecionados dois datasets públicos amplamente utilizados em pesquisas sobre atributos faciais:

- FairFace: Um dataset composto por 108.501 imagens, desenvolvido com o objetivo de promover o balanceamento entre classes raciais. Foi utilizado como baseline para um treinamento com menor viés de representação.
- CelebA (CelebFaces Attributes): Um dataset de larga escala contendo mais de 200.000 imagens de celebridades. É conhecido na literatura por possuir

desbalanceamentos demográficos significativos, o que o torna ideal para o estudo dos efeitos do viés de dados.

4.2. Preparação e Pré-processamento dos Dados

Ambos os datasets passaram por um rigoroso processo de preparação e padronização para garantir a comparabilidade. As seguintes etapas foram conduzidas:

- Padronização de dados: Padronização de atributos e formatos de arquivos.
- Criação de Atributo Racial para o CelebA (*Pseudo-labeling*): Dado que o CelebA não possui anotações de raça, empregou-se uma técnica de *pseudo-labeling*. Para esta tarefa, um modelo com arquitetura ResNet-34 (He et al., 2016), pré-treinado no dataset FairFace, foi utilizado para inferir e atribuir os rótulos de raça às imagens do CelebA.
- Limpeza de Dados e Seleção de Atributos: Descarte de colunas e atributos irrelevantes ao escopo da pesquisa (*feature selection*), focando nos dados pertinentes à classificação de raça e gênero.
- Downsampling e Divisão dos Dados (*Data Splitting*): Ambos os datasets foram processados para conter 95.000 imagens cada. Estas foram então divididas utilizando uma abordagem *hold-out*, destinando 90% (85.000) das imagens para o conjunto de treinamento e 10% (10.000) das imagens para o conjunto de teste.
- Gerenciamento de Rótulos (*Metadata-based Label Management*): Armazenamento dos rótulos e metadados em arquivos CSV, associando cada imagem ao seu respectivo conjunto e atributos para maior flexibilidade na análise. Esta abordagem foi preferida em detrimento da organização de imagens em subpastas por classe para maior flexibilidade na manipulação dos dados.

4.3 Arquitetura do Modelo e Treinamento

4.3.1. Arquitetura Multi-tarefa com ConvNeXt-Tiny

Foi implementado um modelo de aprendizado multi-tarefa (*multi-task learning*) com base na arquitetura ConvNeXt-Tiny, escolhida por seu desempenho estado da arte em tarefas de visão computacional. O modelo foi estruturado com um *backbone* ConvNeXt-Tiny, com pesos pré-treinados no dataset ImageNet [Deng et al., 2009], atuando como extrator de características. Sobre este *backbone*, duas cabeças de classificação (camadas lineares) foram adicionadas para as tarefas específicas de predição binária de gênero (Homem e Mulher) e predição de raça em sete categorias (Branco, Negro, Indiano, Asiático Oriental, Sudeste Asiático, Oriente Médio e Latino).

4.3.2. Configuração do Treinamento

O treinamento foi conduzido com os seguintes hiperparâmetros e configurações:

- Otimizador: AdamW [Loshchilov e Hutter, 2019], uma variante do Adam que aprimora a regularização por decaimento de peso (*weight decay*), favorecendo a generalização.
- Taxa de Aprendizagem: Fixada em 1e-4 (0,0001) para promover ajustes estáveis nos pesos do modelo.

- Função de Perda: Para quantificar o erro do modelo em cada tarefa de classificação, utilizou-se a Entropia Cruzada (*Cross-Entropy Loss*). A perda total, que o otimizador busca minimizar, foi definida como a soma das perdas de gênero e raça, conforme as equações:

$$L_{CE} = -\sum_{c=1}^{C} y_c \log(\hat{y}_c)$$
$$L_{total} = L_{gênero} + L_{raça}$$

  Onde L_{CE} apresenta a perda para uma única tarefa, C é o número de classes, y_c é o rótulo verdadeiro (1 para a classe correta, 0 para as demais) e $\hat{y}_c$ é a probabilidade prevista pelo modelo para aquela classe.
- Épocas e Tamanho do Lote: O modelo foi treinado por 10 épocas, com um tamanho de lote (*batch size*) de 64 amostras, valores definidos para garantir a convergência do modelo dentro das limitações de hardware. O número reduzido de épocas visou evitar *overfitting*, já evidente com o dataset enviesado CelebA (Figura 1). Treinos mais longos agravariam a perda de generalização. Isso está de acordo com a literatura, que alerta que aumentar épocas sem balanceamento adequado eleva o risco de *overfitting* e compromete o desempenho [Ong et al., 2021; Tsai et al., 2021].
- Otimização de Desempenho: Foi utilizada a técnica de precisão mista (*mixed-precision training*) para acelerar o treinamento e reduzir o consumo de memória da GPU.

4.4 Delineamento Experimental

Para avaliar o impacto do viés nos dados, foram definidos quatro cenários experimentais:

- Intra-domínio (balanceado): Treino e teste no FairFace.
- Intra-domínio (desbalanceado): Treino e teste no CelebA.
- Domínio-cruzado: Treino no FairFace e teste no CelebA.
- Domínio-cruzado: Treino no CelebA e teste no FairFace.

Este delineamento investiga tanto a performance do modelo em seu domínio de origem quanto sua capacidade de transferência de aprendizado e robustez ao viés.

4.5 Métricas de Avaliação

A avaliação dos modelos foi conduzida em duas dimensões principais: desempenho preditivo global e análise de justiça.

4.5.1. Desempenho Geral

Para cada tarefa, foram calculadas as métricas de Acurácia, Perda, Recall e F1-Score (por classe), além da Matriz de Confusão.

4.5.2. Análise de Justiça Algorítmica

Para uma análise mais granular do viés, o desempenho foi desagregado e avaliado para diferentes subgrupos demográficos e interseccionais (ex: "mulheres negras", "homens

asiáticos"). Esta abordagem é uma prática recomendada em estudos de justiça algorítmica para detectar disparidades de desempenho que poderiam passar despercebidas em uma análise global [Buolamwini e Gebru, 2018].

4.6 Ferramentas e Ambiente Computacional

Os experimentos foram desenvolvidos em Python 3.10 no ambiente Google Colaboratory (GPU NVIDIA Tesla T4). As principais bibliotecas utilizadas foram PyTorch [Paszke et al., 2019], scikit-learn [Pedregosa et al., 2011], pandas [Mckinney, 2010], matplotlib e seaborn.

5. Resultados e Discussões

Esta seção apresenta e discute os resultados da análise comparativa, demonstrando como a composição do dataset de treinamento influencia diretamente o desempenho, a justiça e a robustez do modelo ConvNeXt. Os achados são organizados para primeiro estabelecer a causa fundamental das discrepâncias (a capacidade de generalização do modelo), depois quantificar suas consequências no desempenho e na justiça e, por fim, discutir as implicações destes resultados, incluindo a robustez em cenários de teste realistas. Os principais indicadores de desempenho e justiça para os quatro cenários experimentais estão consolidados na Tabela 1.

Tabela 1. Comparativo de desempenho e justiça dos quatro cenários experimentais

Cenário Experimental	Dataset de Treino	Dataset de Teste	Acurácia (Gênero)	Acurácia (Raça)	F1-Score Mín. (Raça)[1]	Recall Mín. (Raça)[2]
1. Intra-domain	CelebA	CelebA	51%	28%	0.04	0.6%
2. Cross-domain	CelebA	Fairface	51%	28%	0.04	0.6%
3. Intra-domain	FairFace	FairFace	93%	68%	0.53	8.0%
4. Cross-domain	FairFace	CelebA	92%	68%	0.54	8.0%

5.1. Causa Raiz: O Impacto do Dataset na Generalização do Modelo

A causa fundamental para a disparidade massiva nos resultados reside na capacidade de generalização do modelo, um comportamento diretamente influenciado pela qualidade e balanceamento do dataset de treinamento. A Figura 1 ilustra a dinâmica da perda (loss) ao longo das épocas para os quatro cenários experimentais, servindo como diagnóstico primário da saúde do treinamento.

Nos painéis superiores (treino em CelebA), a perda de validação (vermelho) diverge drasticamente da perda de treino (azul), indicando *overfitting* severo. Nos painéis inferiores (treino em FairFace), as curvas se mantêm muito mais próximas, demonstrando uma generalização eficaz.

Nos modelos treinados com o dataset enviesado CelebA (paineis superiores), observa-se um *overfitting* severo e imediato. A perda de validação (em vermelho) não apenas cresce, mas diverge exponencialmente da perda de treino (em azul). Isso indica que o modelo estava essencialmente memorizando os exemplos do conjunto de treino – em particular, as características do grupo racial majoritário – em vez de aprender padrões representativos e generalizáveis que pudessem ser aplicados a dados não vistos.

Em forte contraste, o treinamento no dataset balanceado FairFace (painel inferior esquerdo, intra-domain) resultou em um comportamento muito mais estável e desejável. As curvas de treino e validação permaneceram significativamente mais próximas, com a perda de validação subindo de forma muito mais contida. Este comportamento é a assinatura de um modelo que está aprendendo características distintivas e generalizáveis de todas as classes, sendo a base para um desempenho robusto e justo.

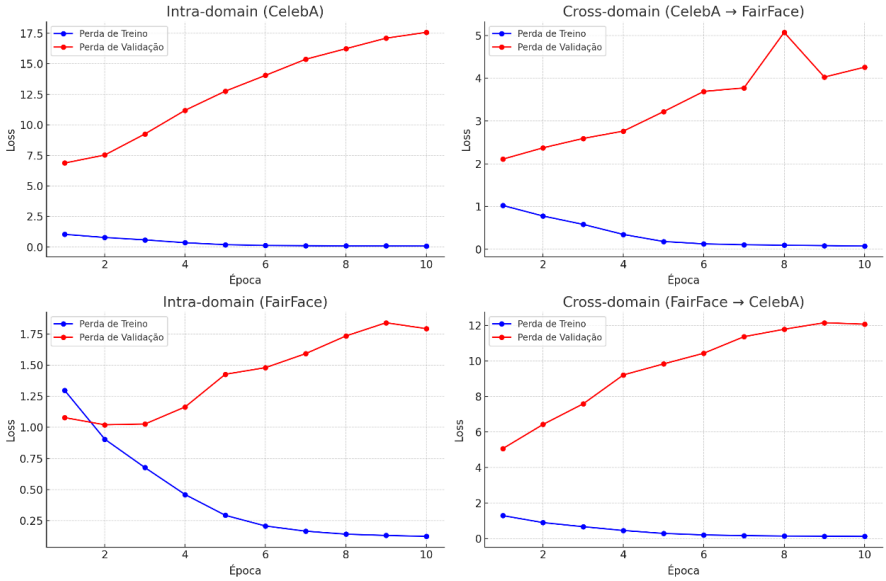

Figura 1. Overfitting — Desempenho dos 4 cenários

5.2. Consequência: Disparidade no Desempenho e na Justiça Racial

A consequência mais direta dessa falha de generalização é uma disparidade gritante no desempenho e na justiça do modelo, como quantificado na Tabela 1 e visualizado na Figura 2.

O gráfico expõe a diferença de performance entre o modelo treinado em dados enviesados (vermelho) e o modelo treinado em dados balanceados (verde escuro). O modelo balanceado não só eleva o desempenho de todos os grupos, mas o faz de forma muito mais equitativa.

O modelo treinado com dados enviesados (barras vermelhas) atingiu uma acurácia para raça de apenas 28%, um resultado que se manteve baixo mesmo quando testado em um dataset balanceado (cenário *cross-domain*). Em contrapartida, o modelo treinado em dados balanceados alcançou uma acurácia de 68%, um ganho de 40 pontos percentuais que se provou robusto em ambos os cenários de teste.

A Figura 2 expõe o cerne da injustiça algorítmica. Para o modelo enviesado, a capacidade de reconhecer corretamente (Recall/TPR) grupos minoritários é quase nula, com taxas de 0.64% para "Southeast Asian" e 0.65% para "Indian". Isso significa que a modelo falha em identificar corretamente mais de 99% dos indivíduos desses grupos. O modelo balanceado (barras verdes escuras) não só elevou o desempenho de todos os grupos, mas o fez de forma muito mais equitativa, com o recall para esses mesmos grupos saltando para 8.32% e 10.33%, respectivamente. O grupo "Black", que tinha um recall de

1.25% no modelo enviesado, alcançou 12.14% no modelo balanceado, o maior entre todos os grupos neste cenário.

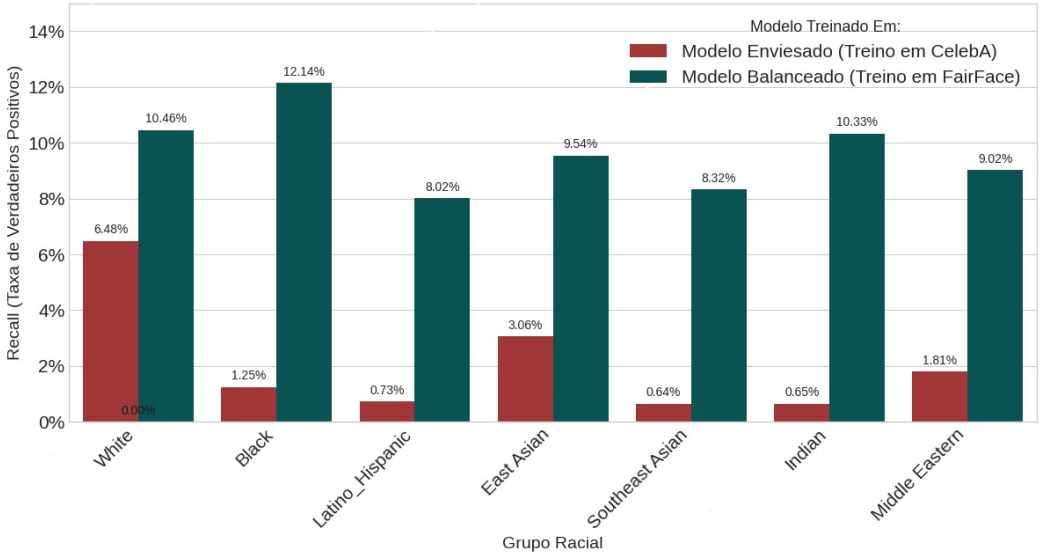

Figura 2. Comparativo de justiça - Recall (TPR) por raça

5.3. Aprofundando a Análise do Viés

O viés observado não é apenas geral, mas se manifesta de formas específicas e sistemáticas, o que pode ser observado através da análise dos padrões de erro e da interseccionalidade.

5.3.1. Padrões de Erro Sistemático: A Análise das Matrizes de Confusão

A Figura 3 compara as matrizes de confusão dos modelos no cenário *intra-domain* e ilustra como os modelos erram, revelando suas heurísticas internas.

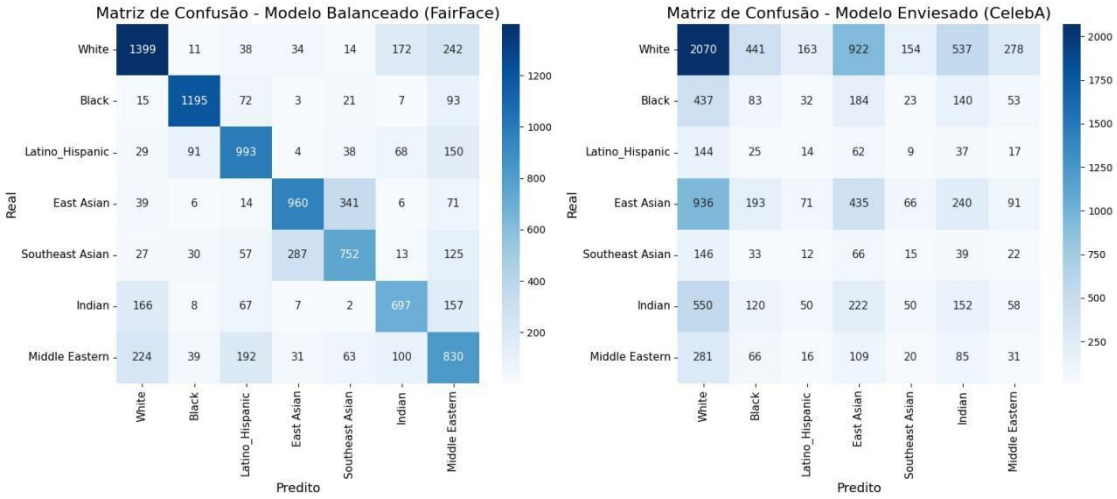

Figura 3. Matriz de Confusão - Comparativo entre Modelo Balanceado (esquerda) e Modelo Enviesado (direita)

A matriz da direita não possui uma diagonal forte, com erros massivamente concentrados na coluna "White". A matriz da esquerda exibe uma diagonal proeminente, sinal de um classificador competente.

A matriz da direita (modelo enviesado) não possui uma diagonal forte. Em vez disso, os erros se concentram massivamente na coluna "White", indicando que o modelo aprendeu uma heurística falha e perigosa: na dúvida, classificar como a classe majoritária do seu treinamento. Por exemplo, dos indivíduos do grupo "East Asian", 922 foram incorretamente classificados como "White", enquanto apenas 341 foram classificados corretamente.

Por outro lado, a matriz da esquerda (modelo balanceado) exibe uma diagonal proeminente e bem definida, um sinal claro de um modelo que aprendeu a distinguir as características de cada classe de forma muito mais competente. Os erros, embora existentes, são menores e mais distribuídos, não favorecendo uma única classe como padrão de falha.

5.3.2. Viés Interseccional: A Dupla Penalidade de Gênero e Raça

A Figura 4 revela o viés em seu nível mais granular, analisando o recall para grupos interseccionais (Gênero-Raça) no modelo enviesado. A figura demonstra que a injustiça não afeta todos os subgrupos igualmente, e que médias de desempenho em grupos raciais podem mascarar falhas críticas.

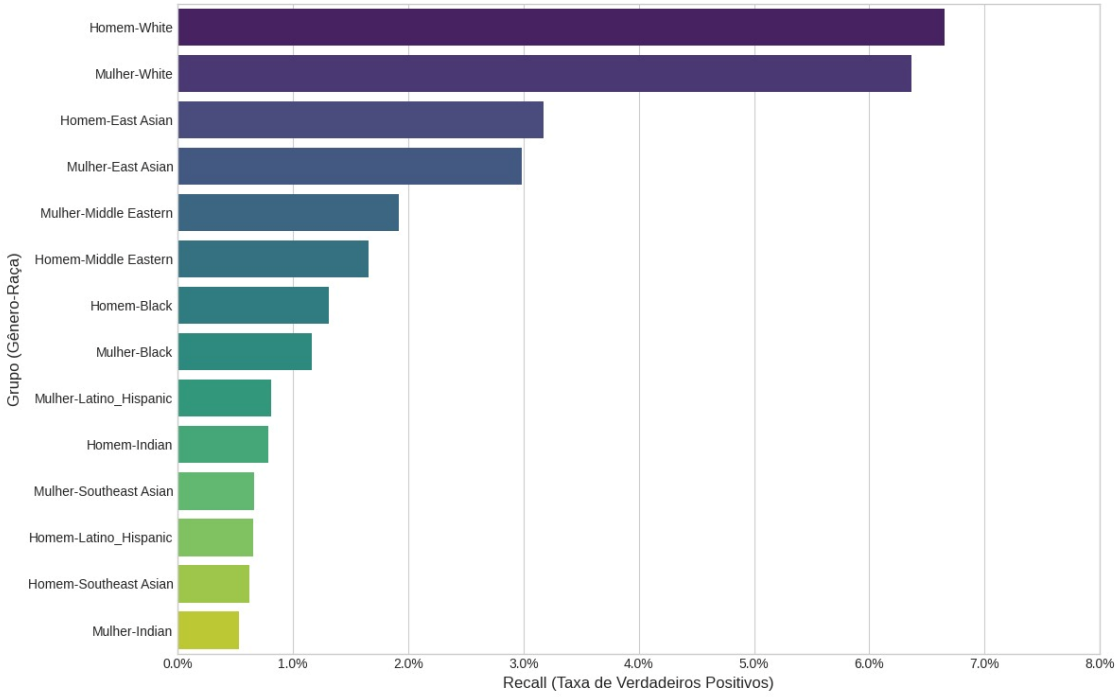

Figura 4. Justiça Interseccional - Recall (TPR) por Grupo (Modelo Enviesado)

A performance do modelo enviesado varia drasticamente entre subgrupos, prejudicando severamente combinações de gênero e raça de grupos minoritários. Enquanto "Homem-White" e "Mulher-White" apresentam o maior recall relativo (ainda que baixo, em torno de 6-7%), grupos como "Mulher-Indian" (recall de 0.53%) e "Homem-Southeast Asian" (recall de 0.62%) são os mais severamente

prejudicados. Isso comprova que a justiça de um modelo deve ser avaliada em múltiplas granularidades, pois médias gerais mascaram falhas críticas em subpopulações específicas, levando a uma "dupla penalidade" para indivíduos que pertencem a mais de um grupo minoritário.

5.4. A Prova de Robustez: O Teste Cross-Domain e a Resiliência do Modelo Justo

A análise mais reveladora vem dos testes em cenários *cross-domain*. Quando o modelo foi treinado no dataset enviesado CelebA e testado no FairFace, mesmo em um ambiente balanceado, ele manteve baixa acurácia (28%) e altos níveis de injustiça. Isso evidencia que a exposição a dados justos não corrige um modelo mal treinado – o viés aprendido permanece. Por outro lado, o modelo treinado com dados equilibrados do FairFace mostrou-se robusto ao ser testado no CelebA. Manteve acurácia elevada (68%) e melhor desempenho interseccional (Recall mínimo de 8,0%), como mostrado na Tabela 1.

Portanto, os resultados apresentados levam a uma afirmação inequívoca: a principal fonte de viés e baixo desempenho neste estudo não foi a arquitetura do modelo ConvNeXt, mas sim, o dataset de treinamento. Treinar com dados enviesados produz um modelo fundamentalmente falho, cujo desempenho ruim e injusto se mantêm. Em contrapartida, o treinamento com dados balanceados produziu um modelo não apenas mais preciso e justo, mas também mais resiliente, estabelecendo a curadoria de dados como o pilar fundamental para o desenvolvimento de uma IA ética e eficaz.

6. Considerações Finais

Este estudo evidenciou que a qualidade e o balanceamento dos dados de treinamento têm impacto decisivo sobre o desempenho e a justiça de modelos de classificação facial, superando inclusive a influência da arquitetura utilizada. A comparação entre os datasets FairFace (balanceado) e CelebA (enviesado) revelou que dados desbalanceados resultam em modelos tendenciosos e pouco confiáveis.

Embora melhorias técnicas, como o uso de *early stopping*, possam mitigar problemas como *overfitting*, limitações estruturais persistem – especialmente a classificação binária de gênero, que exclui identidades transgênero e não-binárias. Tal abordagem restrita compromete a equidade dos sistemas de IA e pode reforçar práticas discriminatórias em contextos sociais sensíveis como emprego, segurança e acesso a serviços. Para trabalhos futuros, propõem-se as seguintes questões:

- Analisar o impacto em grupos de gênero não-conformantes, desenvolvendo metodologias e datasets mais inclusivos.
- Aplicar técnicas ativas de mitigação de viés ao treinar com dados desbalanceados, comparando sua eficácia.

Conclui-se que a busca por sistemas de IA eficazes é indissociável de um compromisso rigoroso com a curadoria e o equilíbrio dos dados. A construção de um futuro digital justo exige que a equidade seja um pilar desde a concepção da tecnologia, garantindo que ela sirva a toda a sociedade, e não apenas a uma parcela dela.

Referências

ALSHARIF, B.; ALTAHER, A.; ALTAHER, A.; ILYAS, M.; ALALWANY, E. (2023). Deep learning technology to recognize american sign language alphabet. *Sensors*, 23(18), 7970. https://doi.org/10.3390/s23187970. Acesso em: 02 jun. 2025.

ARAGÃO, H.; SANTANA, J.; SILVA, G.; SANTANA, M.; SILVA, L.; OLIVEIRA, M.; MELO, C. (2022). Impactos da covid-19 à luz dos marcadores sociais de diferença: raça, gênero e classe social. *Saúde Em Debate*, 46(spe1), 338-347. https://doi.org/10.1590/0103-11042022e123. Acesso em: 02 jun. 2025.

BARROS, M.; LINS, R.; RODRIGUES, I. (2024). Avaliação de redes neurais deep learning para a classificação de câncer de mama. *Cuadernos De Educación Y Desarrollo*, 16(8), e5325. https://doi.org/10.55905/cuadv16n8-124. Acesso em: 02 jun. 2025.

BUOLAMWINI, J.; GEBRU, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Conference on Fairness, Accountability, and Transparency (FAT)*, Anais... https://proceedings.mlr.press/v81/buolamwini18a.html. Acesso em: 29 mai. 2025.

CHEN, Y.; WEI, C.; KUMAR, A.; MA, T. (2020). *Self-training avoids using spurious features under domain shift*. arXiv preprint. https://doi.org/10.48550/arxiv.2006.10032. Acesso em: 02 jun. 2025.

DENG, J. et al. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anais..., p. 248–255. https://image-net.org/static_files/papers/imagenet_cvpr09.pdf. Acesso em: 29 mai. 2025.

ELIAS, M.; FAVERSANI, L.; MOREIRA, J.; MASIERO, A.; CUNHA, N. (2023). Inteligência artificial em saúde e implicações bioéticas: uma revisão sistemática. *Revista Bioética*, 31. https://doi.org/10.1590/1983-803420233542en. Acesso em: 02 jun. 2025.

ERICKSON, B.; KORFIATIS, P.; AKKUS, Z.; KLINE, T.; PHILBRICK, K. (2017). Toolkits and libraries for deep learning. *Journal of Digital Imaging*, 30(4), 400-405. https://doi.org/10.1007/s10278-017-9965-6. Acesso em: 02 jun. 2025.

FERREIRA, C.; SALLES, A. (2022). Uma análise além da renda: o pioneirismo de gunnar myrdal na abordagem econômica sobre as desigualdades sociais. *Estudos Econômicos (São Paulo)*, 52(1), 155-183. https://doi.org/10.1590/1980-53575215ccas. Acesso em: 02 jun. 2025.

FIORATI, R.; CÂNDIDO, F.; SOUZA, L.; POPOLIN, M.; RAMOS, A.; ARCÊNCIO, R. (2018). Desigualdades sociais e os desafios à estratégia de eliminação da tuberculose no brasil. *Vittalle - Revista De Ciências Da Saúde*, 30(2), 59-72. https://doi.org/10.14295/vittalle.v30i2.7502. Acesso em: 02 jun. 2025.

HE, K.; ZHANG, X.; REN, S.; SUN, J. (2016). Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anais... https://arxiv.org/abs/1512.03385. Acesso em: 02 jun. 2025.

HENRIQUES, I.; SAMPAIO, I. (2022). Discriminação algorítmica e inclusão em sistemas de inteligência artificial - uma reflexão sob a ótica dos direitos da criança no

ambiente digital. *Direito Público*, 18(100). https://doi.org/10.11117/rdp.v18i100.5993. Acesso em: 02 jun. 2025.

JÚNIOR, J.; LIMA, P.; PASSOS, T.; MARTINS, P.; SILVA, M.; ROSADO, S.; HUBER, N. (2024). Educação na era dos algoritmos: como a hiperconectividade está moldando os processos de ensino e aprendizagem. *Contribuciones a Las Ciencias Sociales*, 17(5), e6486. https://doi.org/10.55905/revconv.17n.5-004. Acesso em: 02 jun. 2025.

KAMIRAN, F.; KARIM, A.; ZHANG, X. (2012). Decision theory for discrimination-aware classification. In: *IEEE 12th International Conference on Data Mining*, Anais... https://doi.org/10.1109/icdm.2012.45. Acesso em: 02 jun. 2025.

KÄRKKÄINEN, K.; JOO, J. (2021). *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age*. arXiv preprint. https://arxiv.org/abs/1908.04913. Acesso em: 02 jun. 2025.

KASSAM, A.; MARINO, P. (2022). Algorithmic racial discrimination. *Feminist Philosophy Quarterly*, 8(3/4). https://doi.org/10.5206/fpq/2022.3/4.14275. Acesso em: 02 jun. 2025.

KAUSHIK, P. (2023). Deep learning and machinelearning to diagnose melanoma. *International Journal of Research in Science and Technology*, 13(01), 58-72. https://doi.org/10.37648/ijrst.v13i01.008. Acesso em: 02 jun. 2025.

KIM, B.; KIM, H.; KIM, K.; KIM, S.; KIM, J. (2019). Learning not to learn: training deep neural networks with biased data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Anais..., p. 9004-9012. https://doi.org/10.1109/cvpr.2019.00922. Acesso em: 02 jun. 2025.

LECUN, Y.; BENGIO, Y.; HINTON, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. https://doi.org/10.1038/nature14539. Acesso em: 02 jun. 2025.

LETT, E.; SHAHBANDEGAN, S.; BARAK–CORREN, Y.; FINE, A.; CAVA, W. (2024). *Intersectional consequences for marginal fairness in prediction models for emergency admissions*. medRxiv preprint. https://doi.org/10.1101/2024.11.05.24316769. Acesso em: 02 jun. 2025.

LIN, M.; LI, T.; YANG, Y.; HOLSTE, G.; DING, Y.; TASSEL, S.; PENG, Y. (2023). Improving model fairness in image-based computer-aided diagnosis. *Nature Communications*, 14(1). https://doi.org/10.1038/s41467-023-41974-4. Acesso em: 02 jun. 2025.

LIU, Z.; MAO, H.; WU, C.; FEICHTENHOFER, C.; DARRELL, T.; XIE, S. (2022). *A convnet for the 2020s*. arXiv preprint. https://doi.org/10.48550/arxiv.2201.03545. Acesso em: 02 jun. 2025.

LIU, Z. et al. (2015). Deep Learning Face Attributes in the Wild. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Anais... https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html. Acesso em: 29 mai. 2025.

LOSHCHILOV, I.; HUTTER, F. (2019). *Decoupled Weight Decay Regularization*. arXiv preprint. https://arxiv.org/abs/1711.05101. Acesso em: 02 jun. 2025.

MCKINNEY, W. (2010). Data Structures for Statistical Computing in Python. In: *Proceedings of the Python in Science Conference*, Anais...

https://conference.scipy.org/proceedings/scipy2010/mckinney.html. Acesso em: 29 mai. 2025.

MEHRABI, N.; MORSTATTER, F.; SAXENA, N.; LERMAN, K.; GALSTYAN, A. (2021). A survey on bias and fairness in machine learning. *Acm Computing Surveys*, 54(6), 1-35. https://doi.org/10.1145/3457607. Acesso em: 02 jun. 2025.

NAJAFABADI, M.; VILLANUSTRE, F.; KHOSHGOFTAAR, T.; SELIYA, N.; WALD, R.; MUHAREMAGIC, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1). https://doi.org/10.1186/s40537-014-0007-7. Acesso em: 02 jun. 2025.

ONG, J.-H. et al. Implementation of a deep learning model for automated classification of *Aedes aegypti* (Linnaeus) and *Aedes albopictus* (Skuse) in real time. *Scientific Reports*, v. 11, n. 1, p. 10431, Mai. 2021. Disponível em: https://doi.org/10.1038/s41598-021-89365-3. Acesso em: 11 jun. 2025.

PAGANO, T.; LOUREIRO, R.; LISBOA, F.; PEIXOTO, R.; GUIMARÃES, G.; CRUZ, G.; NASCIMENTO, E. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 15. https://doi.org/10.3390/bdcc7010015. Acesso em: 02 jun. 2025.

PASZKE, A. et al. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: *Advances in Neural Information Processing Systems (NeurIPS)*, Anais... https://arxiv.org/abs/1912.01703. Acesso em: 02 jun. 2025.

PEDREGOSA, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html. Acesso em: 29 mai. 2025.

RIBEIRO, E.; BORGES, D. (2022). Percepções de bem-estar nas favelas da maré. *Civitas - Revista De Ciências Sociais*, 22, e41764. https://doi.org/10.15448/1984-7289.2022.1.41764. Acesso em: 02 jun. 2025.

ROSSETTI, R.; ANGELUCI, A. (2021). Ética algorítmica: questões e desafios éticos do avanço tecnológico da sociedade da informação. *Galáxia (São Paulo)*, (46). https://doi.org/10.1590/1982-2553202150301. Acesso em: 02 jun. 2025.

SARKER, I. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *Sn Computer Science*, 2(6). https://doi.org/10.1007/s42979-021-00815-1. Acesso em: 02 jun. 2025.

TSAI, P.-Y. et al. Lumbar Disc Herniation Automatic Detection in Magnetic Resonance Imaging Based on Deep Learning. *Frontiers in Bioengineering and Biotechnology*, v. 9, p. 708137, Jul. 2021. Disponível em: https://doi.org/10.3389/fbioe.2021.708137. Acesso em: 11 jun. 2025.

UPADHYAY, A. (2017). Significant enhancements in machine translation by various deep learning approaches. *American Journal of Computer Science and Information Technology*, 05(02). https://doi.org/10.21767/2349-3917.100008. Acesso em: 02 jun. 2025.

WANG, P.; JIANG, Q.; LIU, B. (2023). Image processing of the special sensor microwave/imager based on passive microwave remote sensing. In: *Sixth*

International Conference on Computer Information Science and Application Technology, Anais... https://doi.org/10.1117/12.2684592. Acesso em: 02 jun. 2025.

YUAN, H. (2023). Current perspective on artificial intelligence, machine learning and deep learning. *Applied and Computational Engineering*, 19(1), 116-122. https://doi.org/10.54254/2755-2721/19/20231019. Acesso em: 02 jun. 2025.