Investigação e aprimoramento de sistemas de diagnóstico auxiliado por computador na identificação de câncer de pele em tons de pele escura

Eduarda P. Magesk[1], **Pedro H. G. Bouzon**[1], **Luis A. de Souza Jr.**[1], **Andre G. C. Pacheco**[1]

[1]Departamento de Informática – Universidade Federal do Espírito Santo (UFES)
Caixa Postal 29075-910 – Vitória – ES – Brasil

{eduarda.magesk,pedro.bouzon}@edu.ufes.br

{la.souza,apacheco}@inf.ufes.br

Abstract. *The development of Computer-Aided Diagnosis (CAD) systems for skin cancer is a well-established research field. However, many proposed models exhibit significant racial bias, showing inferior performance when evaluating skin lesions in higher phototypes (darker skin). This paper assesses the presence of such bias in diagnostic models and proposes an adaptation of the Balanced Cross-Entropy loss function to mitigate performance disparity. Seven Deep Learning architectures were analyzed, including both Convolutional Neural Networks (CNN) and Vision Transformers (ViT). Results confirmed the existence of bias in conventional models; conversely, the proposed technique significantly reduced this inequality. In one of the evaluated architectures, the balanced accuracy for dark skin increased from* 0.58 ± 0.04 *to* 0.69 ± 0.06*, demonstrating the effectiveness of the approach in promoting performance equity.*

Resumo. *O desenvolvimento de sistemas de Diagnóstico Auxiliado por Computador (CAD) para o câncer de pele é uma área de pesquisa consolidada. Contudo, observa-se que muitos dos modelos propostos apresentam um viés racial significativo, com desempenho inferior na avaliação de lesões em peles com fototipos mais altos (escuras). Este artigo avalia a presença desse viés nos modelos de diagnóstico e propõe uma adaptação na função de perda Balanced Cross-Entropy para mitigar a disparidade de performance. Foram analisadas sete arquiteturas de Deep Learning, contemplando tanto Redes Neurais Convolucionais (CNN) quanto Vision Transformers (ViT). Os resultados confirmaram a existência de viés nos modelos convencionais; em contrapartida, a técnica proposta reduziu significativamente essa desigualdade. Em uma das arquiteturas, a acurácia balanceada para peles escuras saltou de* $0,58 \pm 0,04$ *para* $0,69 \pm 0,06$*, evidenciando a eficácia da abordagem na promoção de equidade de desempenho.*

1. Introdução

De acordo com o Instituto Nacional de Câncer (INCA), estima-se que cerca de 35% de todos os tumores malignos registrados no Brasil sejam câncer de pele [INCA 2026]. Apesar da estimativa de sobrevida em 5 anos ser mais de 96% para aqueles diagnosticados precocemente [SCF 2025a], a doença, quando não tratada, pode levar a deformidades físicas

e, em casos mais graves, à metástase, principalmente quando se trata do câncer de pele do tipo melanoma [MS 2025]. Portanto, o diagnóstico precoce é a chave para assegurar ao paciente o melhor prognóstico possível.

Como forma de auxiliar profissionais da área da saúde à identificarem o câncer de pele e outras dermatoses, têm-se destacado os Sistemas de Diagnóstico Auxiliado por Computador (CAD - do inglês: *Computer-Aided diagnosis*). Nos últimos anos, diversos modelos de machine learning foram propostos com essa finalidade [Pacheco and Krohling 2019, Vidya and Karki 2020, Souza et al. 2024, Bouzon et al. 2025]. Contudo, embora a área tenha alcançado resultados expressivos em termos de acurácia global, o desempenho de tais algoritmos raramente é equitativo para as populações parda e preta [Barros et al. 2023]. Esse viés racial decorre, fundamentalmente, da sub-representação de lesões em peles escuras nos conjuntos de dados públicos [Liu et al. 2023], o que compromete a generalização e a eficácia clínica desses sistemas nesse grupo. Tal desequilíbrio nas bases de dados de IA reflete disparidades históricas na própria educação médica, que serve como o referencial humano para a validação algorítmica. Auditorias em livros didáticos de residência e cirurgia revelam, por exemplo, que imagens de tons de pele escuros (fototipos V–VI na escala de Fitzpatrick) representam menos de 12% do material educacional disponível [Harp et al. 2022, Porras Fimbres et al. 2023].

Embora a ocorrência do câncer de pele seja habitualmente associada a indivíduos de pele clara [SCF 2025b], sua manifestação em pessoas de pele escura, apesar de menos frequente, apresenta taxas de mortalidade significativamente superiores [AIM 2024]. Diante dessa disparidade clínica, torna-se imperativo adotar estratégias que busquem mitigar a transferência desse viés racial para os modelos de machine learning, evitando que desigualdades históricas sejam automatizadas pelos algoritmos. Considerando que aproximadamente 55% da população brasileira se autodeclara parda ou preta [IBGE 2022], a viabilidade dos sistemas CAD como ferramentas de saúde pública no Brasil depende da implementação de mecanismos que assegurem a equidade diagnóstica, independentemente do fenótipo do paciente.

Este trabalho propõe-se a investigar a presença de viés racial em um sistema CAD selecionado por seu estágio avançado de desenvolvimento e proximidade com a validação em ambientes clínicos reais. O classificador avaliado fundamenta-se na proposta de Castro et al. [2020], com aprimoramentos subsequentes introduzidos por Pacheco and Krohling **[?]**. Este modelo caracteriza-se por uma abordagem multimodal, integrando metadados clínicos a imagens de lesões cutâneas para o diagnóstico de dermatoses. A escolha por este sistema CAD justifica-se pela sua robustez e pela complexidade da fusão de dados, permitindo uma análise aprofundada de como o viés racial pode se manifestar em modelos que buscam mimetizar o raciocínio clínico dermatológico.

Inicialmente, será realizada uma análise de referência (*baseline*) na qual nenhuma técnica de mitigação de viés será aplicada durante o treinamento. O propósito desta etapa é quantificar a disparidade de desempenho entre amostras de peles claras e escuras, evidenciando a existência de viés algorítmico. Em seguida, visando reduzir essa assimetria, propõe-se uma adaptação da função de perda *Balanced Cross-Entropy Loss*. Diferente das abordagens convencionais de balanceamento de classes, a inovação deste trabalho reside na atribuição de pesos fundamentada na frequência conjunta entre a classe diagnóstica e o

fototipo da amostra. Essa estratégia define penalidades específicas para cada par (classe, fototipo), garantindo que grupos sub-representados — como lesões malignas em peles escuras — tenham uma contribuição equitativa no ajuste dos pesos do modelo. Para validar a robustez da proposta, serão avaliados sete diferentes backbones, contemplando arquiteturas baseadas em Redes Neurais Convolucionais (CNN) e Vision Transformers (ViT). Tal investigação é essencial para assegurar que sistemas CAD operem com equidade e segurança em populações diversas. As principais contribuições deste artigo são resumidas a seguir:

- Investigação sistemática do viés racial em sistemas de Diagnóstico Auxiliado por Computador (CAD) voltados à aplicação clínica, evidenciando limitações que podem comprometer a equidade no atendimento dermatológico.
- Proposição de uma função de perda customizada, baseada na frequência conjunta classe-fototipo, projetada para mitigar disparidades raciais no treinamento de modelos de deep learning sem degradar o desempenho global.
- Avaliação experimental exaustiva envolvendo sete arquiteturas de estado da arte (CNNs e ViTs), com análise comparativa sob as perspectivas de acurácia e equidade algorítmica.
- Disponibilização pública do código-fonte e dos protocolos experimentais, promovendo transparência, reprodutibilidade científica e incentivando avanços futuros na construção de sistemas CAD mais justos e inclusivos.

2. Trabalhos Relacionados

Outra problemática relacionada à origem do viés racial em modelos de machine learning reside na escassez de datasets que incluam informações explícitas sobre a tonalidade da pele [Groh et al. 2024]. Essa limitação compromete tanto a análise adequada da equidade dos modelos quanto o desenvolvimento de estratégias eficazes para mitigação de vieses. Neste contexto, Barros et al. **[?]** realizaram a avaliação de três datasets públicos e seus respectivos modelos de inteligência artificial em relação a disparidade de desempenho de acordo com a cor da pele. A escala de cor utilizada por eles foi a Fitzpatrick [Wolff et al. 2015], considerando como pele escura amostras com *Fitzpatrick skin type* (FST) 4 a 6. Os datasets avaliados foram DDI [Daneshjou et al. 2022], Fitzpatrick17k [Groh et al. 2021] e PAD-UFES-20 [Pacheco et al. 2020]. Nenhum dos modelos alcançou uma acurácia balanceada superior a 60% ao ser avaliado em amostras de pele escura, enquanto o desempenho padrão do modelo proposto para o PAD-UFES-20, por exemplo, alcançou 75% de acurácia balanceada [Pacheco and Krohling 2020]. Porém, apesar de comprovar a presença do viés racial nos datasets, o estudo não propôs nenhuma forma de mitigar esse problema.

Ademais, o emprego de modificações da *Balanced Cross-Entropy Loss* para otimizar a detecção de câncer de pele é um tópico recorrente na literatura [Hosseini and Baghshah 2024, Lyakhov et al. 2023, Lyakhova 2022]. Todavia, observa-se uma carência de investigações que adaptem essa função de perda especificamente para o enfrentamento do viés racial. Diante desse cenário, este trabalho propõe investigar o viés em modelos treinados a partir do conjunto de dados PAD-UFES-20+ [Bouzon et al. 2025] e implementar uma modificação na *Balanced Cross-Entropy Loss*, visando reduzir a disparidade de performance conforme o fototipo da amostra.

3. Materiais e Métodos

Nesta seção, são apresentados os principais elementos necessários para a realização dos experimentos, assim como a descrição das modificações feitas na *Cross Entropy Loss*.

3.1. Escala Fitzpatrick

Embora existam diferentes escalas propostas para a categorização de fototipos de pele, como a escala Monk e a escala Rihanna Fenty Beauty, a mais amplamente adotada na área da saúde e na pesquisa científica é a escala *Fitzpatrick Skin Type* (FST) [ARPANSA 2025]. Essa classificação baseia-se principalmente na suscetibilidade do indivíduo a queimaduras solares e em sua capacidade de bronzeamento após exposição à radiação ultravioleta. A Figura 1 apresenta uma representação visual da escala, acompanhada de uma breve descrição de cada um dos seus tipos.

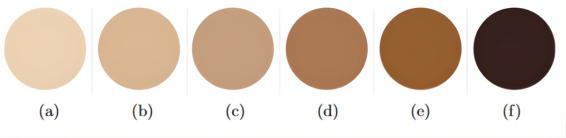

Figura 1. A escala de tons de pele de Fitzpatrick. (a) Tipo 1 (claro): pele pálida. (b) Tipo 2 (branco): pele clara. (c) Tipo 3 (médio): pele de branca a oliva. (d) Tipo 4 (oliva): pele marrom moderada. (e) Tipo 5 (marrom): pele marrom escura. (f) Tipo 6 (negro): pele de marrom muito escuro a preto. Fonte: [Alipour et al. 2024].

Em virtude da reduzida representação de amostras classificadas como FST IV, V e VI em datasets da área, optou-se, para fins de análise estatística e avaliação experimental, pelo agrupamento dos fototipos FST I a III na categoria "pele clara" e dos fototipos FST IV a VI na categoria "pele escura". Essa estratégia de agregação segue a abordagem metodológica adotada por [Barros et al. 2023], visando aumentar a robustez das estimativas e reduzir a variância associada a subgrupos com baixa frequência amostral.

3.2. Base de Dados

Os modelos de classificação diagnóstica foram treinados e avaliados utilizando o PAD-UFES-20+, uma extensão do PAD-UFES-20 [Bouzon et al. 2025, Pacheco et al. 2020] que compreende 17.578 imagens clínicas de lesões de pele, coletadas de 7.236 pacientes entre 2018 e 2025. A Tabela 1 apresenta a distribuição das imagens por classe e grupo FST. A população de pacientes é caracterizada por uma média de idade de $62,0 \pm 16,3$ anos, com uma distribuição de gênero de 53,75% feminino e 37,72% masculino. Todas as imagens foram adquiridas por meio de smartphones e tablets. O conjunto de dados inclui uma gama diversificada de tipos de lesões cutâneas, abrangendo condições benignas e malignas. Seis diferentes tipos de lesões estão mapeados no PAD-UFES-20+: melanoma (MEL), carcinoma basocelular (CBC), carcinoma espinocelular (CEC), ceratose actínica (ACT), ceratose seborreica (SEB) e nevo melanocítico (NEV). A Figura 2 apresenta um exemplo de cada classe. Além disso, o dataset inclui metadados abrangentes, como idade, gênero, histórico pessoal e familiar de câncer de pele, além de sintomas específicos da lesão, incluindo prurido (coceira), dor, crescimento recente e alterações morfológicas.

Classe	FST I–III	FST IV–VI	Total
MEL	329	25	354
CBC	3.475	82	3.557
CEC	2.409	47	2.456
ACT	9,750	285	10.035
SEB	664	92	756
NEV	381	39	420
Total	17.008	570	17.578

Tabela 1. Distribuição das amostras do classificador de diagnóstico em diferentes grupos de tipos de pele de Fitzpatrick.

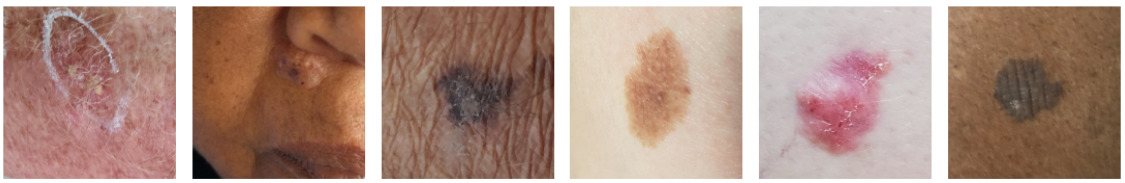

Figura 2. Um exemplo de cada classe e fototipo presente no PAD-UFES-20+.

3.3. Funções de Perda

A função de perda base deste trabalho é a *Balanced Cross Entropy Loss*. O uso de pesos para mitigar o desbalanceamento de classes constitui a estratégia central deste estudo. A seguir, detalham-se as duas implementações propostas para investigar o viés racial.

3.4. *Balanced Cross-Entropy Loss*

A primeira função de perda emprega a formulação padrão da Balanced *Cross-Entropy* (BCE) *Loss*. Nesta abordagem, os pesos das classes w_c são definidos como inversamente proporcionais à frequência da classe em relação ao conjunto de dados total:

$$w_c = \frac{N_{total}}{N_c}$$

Sob esta configuração, o processo de otimização é penalizado de acordo com a raridade da condição clínica, mas permanece agnóstico ao subgrupo racial da amostra. Consequentemente, a função de perda BCE serve como a linha de base (*baseline*) experimental para este estudo, destinada a identificar o viés racial inerente presente nos modelos.

3.5. Função de Perda Balanceada por Classe e Cor

Para mitigar o viés racial durante a fase de otimização, propõe-se a Função de Perda Balanceada por Classe e Cor (BCC), que introduz sensibilidade aos subgrupos ao considerar a distribuição conjunta das classes diagnósticas e dos tons de pele. Diferente das abordagens padrão, esta estrutura dobra o número de pesos parametrizáveis, atribuindo coeficientes distintos a cada categoria diagnóstica com base em sua ocorrência em amostras de pele clara (FST I–III) ou escura (FST IV–VI). Essa estrutura permite a modelagem explícita da interação entre patologia e fototipo, viabilizando um mecanismo de balanceamento mais granular. Matematicamente, a função BCE base foi modificada para incorporar um parâmetro indicador de grupo (g_i), facilitando a ponderação dinâmica durante o

treinamento. A perda para uma única amostra i é definida como:

$$L_i = -W(c_i, g_i) \cdot \log(\hat{p}_{i,c_i})$$

no qual c_i é a classe de referência e $W(c_i, g_i)$ é o peso selecionado com base no par classe/fototipo. Consequentemente, a BCC impõe penalidades assimétricas para classificações incorretas, atribuindo custos mais elevados a erros em amostras sub-representadas de pele escura.

Contudo, a segmentação dos pesos pelo par (classe, fototipo) gerou valores com uma amplitude excessivamente elevada. No contexto de redes neurais profundas, a utilização de pesos de magnitude desproporcional na função de perda pode resultar na explosão do gradiente, desestabilizando a atualização dos parâmetros e impedindo a convergência do modelo. Além disso, valores extremos podem levar o otimizador a negligenciar classes com penalidades menores em favor de subgrupos com pesos massivos, prejudicando a generalização global. Para mitigar esses efeitos e promover um aprendizado mais estável e eficiente, aplicou-se a padronização sobre os pesos gerados. O reescalonamento é definido pela equação:

$$z = \frac{x - \mu}{\sigma}$$

no qual x representa o peso original, μ a média dos pesos e σ o desvio padrão dos pesos. Essa transformação assegura que a importância relativa entre os grupos seja preservada — mantendo o foco na mitigação do viés para peles escuras —, enquanto reajusta os valores para uma escala numérica que favorece a estabilidade do gradiente e a eficiência da convergência durante o treinamento.

4. Experimentos e Resultados

Esta seção apresenta a configuração experimental, as características do conjunto de dados e os protocolos de avaliação utilizados para validar a função de perda proposta, denominada BCC (*Balanced Cross-Entropy* com Balanceamento por Classe e Cor). Primeiramente, estabelece-se uma linha de base ao avaliar as disparidades de desempenho entre os grupos de pele clara e escura utilizando a *Balanced Cross-Entropy* (BCE) padrão. Subsequentemente, para demonstrar a eficácia da BCC na mitigação dessas disparidades raciais, comparamos seu desempenho em relação à *baseline* de BCE.

4.1. Configuração dos Experimentos

Avaliamos a função de perda proposta BCC em comparação com a *baseline* BCE utilizando um protocolo experimental controlado que visa quantificar seu impacto na redução das disparidades de desempenho entre populações de pele clara e escura. O viés racial foi avaliado em diversas redes neurais convolucionais (CNNs: EfficientNet-B4 [Tan and Le 2020], ResNet-50 [He et al. 2015], MobileNetV2 [Sandler et al. 2018]); e ViT e híbridos: (DaViT-tiny [Ding et al. 2022], CAFormer-S18 [Yu et al. 2024], MaxViT-tiny [Yurdakul et al. 2025], MViTv2-small [Li et al. 2022]).

Para a avaliação, selecionamos o conjunto de dados PAD-UFES-20+. Diferente de alternativas que dependem da estimativa de fototipos a partir de imagens (ex: Fitzpatrick17k [Groh et al. 2021]), o PAD-UFES-20+ fornece rótulos confiáveis de fototipos

de Fitzpatrick derivados de avaliações clínicas presenciais, evitando fatores de confusão causados por má qualidade de imagem ou iluminação. Os modelos foram implementados em PyTorch 2 [1], inicializados com pesos da ImageNet [Deng et al. 2009] e ajustados utilizando o otimizador Adam (taxa de aprendizado de 0,0001) por até 50 épocas, com critério de *early stopping* em 15 épocas. As imagens foram redimensionadas para 224×224 e submetidas a *data augmentation* via rotações, ajustes de saturação e ruído aleatório. Utilizou-se a arquitetura MetaBlock [Pacheco and Krohling 2020] para realizar a fusão das características das imagens com metadados clínicos abrangentes: demografia (idade, gênero), histórico clínico (histórico de câncer geral e câncer de pele), localização anatômica e sintomas relatados sobre a lesão (sangramento, prurido, crescimento, dor e alterações morfológicas). O desempenho foi avaliado por meio de validação cruzada de 5 folders, estratificados por lesão e pela frequência das classes.

Para lidar com o desequilíbrio das classes clínicas, o desempenho dos modelos foi avaliado utilizando a Acurácia Balanceada (BACC) e a Área Sob a Curva (AUC). Para quantificar a equidade algorítmica, calculou-se o *Unfairness Score* (US) [Sheng et al. 2024], que mede a dispersão do desempenho entre subgrupos utilizando a norma L_1:

$$US = \sum |A_{g_i} - A_{all}|$$

na qual A_{g_i} é a acurácia do subgrupo e A_{all} é a acurácia da população total. Por fim, a significância estatística das disparidades de desempenho entre os grupos de pele clara e escura foi avaliada por meio do teste de postos sinalizados de Wilcoxon ($\alpha = 0,05$).

4.2. Resultados

A primeira parte da nossa avaliação focou na quantificação da disparidade de desempenho entre os subgrupos de pele clara e escura utilizando a função de perda padrão *Balanced Cross-Entropy* (BCE). Como uma função de objetivo agnóstica ao grupo, a BCE opera sem considerar o fototipo da amostra. Os resultados, resumidos na seção BCE da Tabela 2, revelam uma degradação consistente de desempenho tanto em BACC quanto em AUC na transição de resultados de pele clara para pele escura. Nossa hipótese sobre a presença de viés racial inerente é reforçada pelo teste estatístico de Wilcoxon ($\alpha = 0,05$). A análise estatística indica que as diferenças de desempenho entre os grupos de tons de pele são significativas em quase todas as arquiteturas avaliadas, sugerindo que os modelos são de fato distintos em seu comportamento preditivo para diferentes fototipos. A única exceção foi o backbone DaViT-tiny, que apresentou um valor de p de 0.1602, falhando em rejeitar a hipótese nula de similaridade ao nível de significância de 5%.

Após essa avaliação de referência (*baseline*), a análise avaliou a eficácia da função de perda proposta, BCC (BCE com Balanceamento por Classe e Cor), na mitigação das disparidades de desempenho identificadas nos resultados da BCE. Os resultados, detalhados na seção BCC da Tabela 2, demonstram que o método proposto reduz efetivamente a disparidade de desempenho em todos os backbones. De acordo com o *Unfairness Score* (US) — que quantifica a magnitude da desigualdade de desempenho entre os dois grupos — todos os modelos exibiram uma redução substancial na disparidade, com melhorias que atingiram até cerca de 75% para um dos backbones. Além disso, o teste de Wilcoxon

[1] Disponível em: https://github.com/life-ufes/vies-racial-deteccao-cancer-de-pele.

confirma essa tendência: sob a perspectiva da função de perda BCC, todas as arquiteturas alcançaram equivalência estatística ao comparar os grupos FST I–III e FST IV–VI.

Função de Perda		BCE				BCC			
Backbone	**FST**	**BACC**	**AUROC**	**US**	ρ_{value}	**BACC**	**AUROC**	**US**	ρ_{value}
ResNet-50	Todos	$0,71 \pm 0,02$	$0,93 \pm 0,01$	$0,06 \pm 0,03$	$0,0039$	$0,65 \pm 0,01$	$0,90 \pm 0,01$	$0,03 \pm 0,03$	$0,0645$
	I–III	$0,71 \pm 0,02$	$0,93 \pm 0,01$			$0,66 \pm 0,01$	$0,90 \pm 0,01$		
	VI–IV	$0,64 \pm 0,04$	$0,91 \pm 0,03$			$0,61 \pm 0,03$	$0,90 \pm 0,03$		
MobileNet-v2	Todos	$0,70 \pm 0,02$	$0,92 \pm 0,01$	$0,05 \pm 0,03$	$0,0098$	$0,65 \pm 0,02$	$0,90 \pm 0,01$	$0,04 \pm 0,01$	$0,0840$
	I–III	$0,70 \pm 0,02$	$0,92 \pm 0,01$			$0,65 \pm 0,02$	$0,90 \pm 0,01$		
	VI–IV	$0,64 \pm 0,05$	$0,90 \pm 0,03$			$0,70 \pm 0,06$	$0,90 \pm 0,03$		
EfficientNet-b4	Todos	$0,70 \pm 0,01$	$0,92 \pm 0,00$	$0,06 \pm 0,03$	$0,0059$	$0,63 \pm 0,02$	$0,88 \pm 0,02$	$0,02 \pm 0,02$	$0,5566$
	I–III	$0,70 \pm 0,01$	$0,92 \pm 0,00$			$0,63 \pm 0,02$	$0,88 \pm 0,02$		
	VI–IV	$0,60 \pm 0,05$	$0,91 \pm 0,02$			$0,61 \pm 0,08$	$0,88 \pm 0,04$		
Caformer-s18	Todos	$0,72 \pm 0,02$	$0,94 \pm 0,01$	$0,07 \pm 0,03$	$0,0039$	$0,69 \pm 0,02$	$0,92 \pm 0,01$	$0,04 \pm 0,03$	$0,6250$
	I–III	$0,72 \pm 0,02$	$0,94 \pm 0,00$			$0,69 \pm 0,02$	$0,92 \pm 0,01$		
	VI–IV	$0,63 \pm 0,05$	$0,91 \pm 0,02$			$0,67 \pm 0,11$	$0,92 \pm 0,03$		
Davit-tiny	Todos	$0,71 \pm 0,03$	$0,93 \pm 0,01$	$0,05 \pm 0,04$	$0,1602$	$0,69 \pm 0,01$	$0,92 \pm 0,00$	$0,02 \pm 0,02$	$0,4922$
	I–III	$0,71 \pm 0,03$	$0,93 \pm 0,01$			$0,69 \pm 0,01$	$0,92 \pm 0,00$		
	VI–IV	$0,66 \pm 0,09$	$0,91 \pm 0,03$			$0,67 \pm 0,06$	$0,92 \pm 0,02$		
Maxvit-tiny	Todos	$0,70 \pm 0,02$	$0,93 \pm 0,01$	$0,09 \pm 0,03$	$0,0039$	$0,66 \pm 0,02$	$0,91 \pm 0,01$	$0,02 \pm 0,02$	$1,0000$
	I–III	$0,70 \pm 0,02$	$0,93 \pm 0,01$			$0,66 \pm 0,02$	$0,91 \pm 0,01$		
	VI–IV	$0,58 \pm 0,05$	$0,91 \pm 0,02$			$0,65 \pm 0,09$	$0,92 \pm 0,02$		
Mvit2-small	Todos	$0,71 \pm 0,02$	$0,94 \pm 0,00$	$0,05 \pm 0,01$	$0,0020$	$0,64 \pm 0,05$	$0,90 \pm 0,02$	$0,03 \pm 0,02$	$0,6250$
	I–III	$0,71 \pm 0,01$	$0,94 \pm 0,00$			$0,64 \pm 0,05$	$0,90 \pm 0,02$		
	VI–IV	$0,63 \pm 0,05$	$0,93 \pm 0,01$			$0,64 \pm 0,11$	$0,91 \pm 0,02$		

Tabela 2. Desempenho classificador de diagnóstico utilizando as funções de perda BCE e BCC.

5. Discussão

O objetivo deste estudo foi elucidar as manifestações do viés racial em sistemas CAD baseados em *deep learning* e propor uma função de perda especializada para sua mitigação. Os resultados apresentados previamente indicam o alcance de ambos os objetivos. Ao implementar a função de perda BCC, diminuimos efetivamente as disparidades de desempenho observadas anteriormente na *baseline* (BCE). A eficácia dessa intervenção é confirmada estatisticamente; segundo o teste de Wilcoxon, a função de perda proposta induziu paridade estatística em todos os backbones avaliados, um indicativo de que o desempenho diagnóstico tornou-se mais consistente e equitativo para diferentes fototipos.

Contudo, apesar de sua capacidade de reduzir a disparidade de desempenho entre os subgrupos de tons de pele, é importante notar que este movimento em direção a um modelo mais igualitário resultou em uma redução no desempenho global, afetando os resultados para o grupo majoritário (FST I-III). Este fenômeno não é um achado isolado deste estudo; pelo contrário, alinha-se às observações de Sheng et al. **[?]** e Chiu et al. [2024], que documentaram um *trade-off* característico onde ganhos em equidade algorítmica frequentemente acarretam penalidades no desempenho global.

Apesar deste fato, alguns backbones demonstraram uma resiliência notável. A arquitetura DaViT-tiny não apenas aumentou sua BACC para pele escura, mas também manteve alta estabilidade, com uma redução negligenciável de aproximadamente 2% na BACC global. Notavelmente, este backbone exibiu similaridade estatística entre os grupos desde a fase de baseline, posicionando-se como a arquitetura inerentemente mais resistente ao viés racial neste estudo.

Por outro lado, a arquitetura MaxViT demonstrou a transformação mais significativa. Na fase de baseline, este modelo apresentou o menor desempenho para pele escura,

com BACC de aproximadamente 58% em comparação aos 70% para pele clara, resultando, consequentemente, no maior *Unfairness Score* (US) entre todos os backbones avaliados. No entanto, ao implementar a perda BCC, o MaxViT passou por uma compressão de disparidade notável, alcançando uma redução de mais de 75% no US registrado e um aumento de cerca de 7% na BACC para pele escura.

Ademais, os resultados indicam que as arquiteturas baseadas em *Vision Transformers* (ViTs) e modelos híbridos, como DaViT-tiny e CaFormer-s18, demonstraram maior resiliência à imposição da função de perda BCC em comparação às redes neurais convolucionais (CNNs). Enquanto CNNs tradicionais, como a ResNet-50 e a MobileNet-v2, apresentaram quedas mais acentuadas no desempenho global ao priorizar a equidade, os ViTs conseguiram diminuir o viés racial mantendo patamares de BACC superiores, com média em torno de 0,67. Esse comportamento sugere que os mecanismos de atenção global dos Transformers permitem que o modelo processe as penalidades agressivas da BCC de forma mais estável, preservando melhor o desempenho no grupo majoritário enquanto elevam a eficácia diagnóstica nas coortes de pele escura.

Coletivamente, esses achados corroboram o potencial da função de perda BCC para a diminuição de disparidades raciais em sistemas CAD dermatológicos. Contudo, a análise comparativa evidencia que a eficácia desta estratégia é intrinsecamente dependente da arquitetura do backbone selecionado. Portanto, a implementação de sistemas equitativos exige uma seleção criteriosa de modelos que possuam capacidade de representação suficiente para absorver as restrições de justiça sem comprometer a estabilidade do aprendizado. É imperativo buscar um ponto de equilíbrio no qual a paridade estatística seja alcançada sem induzir consequências negativas para a população majoritária, garantindo a segurança e a confiabilidade clínica do sistema para todo o espectro de fototipos.

6. Conclusão

Este estudo investigou as manifestações do viés racial em sistemas de Diagnóstico Auxiliado por Computador (CAD) e propôs a função de perda BCC (*Balanced Cross-Entropy* com Balanceamento por Classe e Cor) como uma estratégia de mitigação baseada na frequência conjunta de diagnósticos e fototipos. Os resultados experimentais demonstraram que a aplicação da BCC foi eficaz em diminuir as disparidades de desempenho observadas na baseline, alcançando paridade estatística ($p > 0,05$) em todas as arquiteturas avaliadas. Embora tenha sido observado um trade-off no desempenho global, a intervenção algorítmica reduziu significativamente o Unfairness Score, promovendo diagnósticos mais equitativos e seguros para pacientes com peles escuras (FST IV-VI).

Como trabalhos futuros, pretende-se investigar a aplicação da técnica de Número Efetivo de Amostras (*Effective Number of Samples* - ENS) [Cui et al. 2019] no cálculo dos pesos da função BCC. A técnica ENS oferece uma fundamentação teórica para suavizar as penalidades, baseando-se no volume de informação redundante em vez da contagem bruta de amostras. Essa abordagem visa mitigar a disparidade excessiva entre os pesos observados na BCC, reduzindo a instabilidade do gradiente e buscando um equilíbrio mais refinado entre a equidade algorítmica e a manutenção da acurácia no grupo majoritário.

7. Limitações

Apesar de contribuir para a compreensão e mitigação das disparidades de desempenho relacionadas à diferentes tonalidades de pele em sistemas de IA focados na classificação

de doenças dermatológicas, este trabalho apresenta algumas limitações, principalmente vinculadas ao uso da escala de fototipos de Fitzpatrick (FST) como um substituto para o tom de pele em si. Embora a FST seja amplamente adotada em conjuntos de dados dermatológicos e permita a comparação com *benchmarks* existentes, como o PAD-UFES-20, ela foi originalmente desenvolvida para avaliar a fotossensibilidade e não a cor visível da pele. Além disso, suas categorias discretas não capturam plenamente a diversidade de tons de pele, particularmente nos níveis mais elevados de FST. Como resultado, nosso agrupamento em 'Pele Clara' (FST I–III) e 'Pele Escura' (FST IV–VI) pode negligenciar disparidades mais sutis dentro desses grupos.

8. Agradecimentos

Os autores agradecem o apoio financeiro da Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES); o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); o Instituto Capixaba de Ensino, Pesquisa e Inovação em Saúde (ICEPi); o Ministério da Saúde (MS); e o Programa Nacional de Genômica e Saúde de Precisão (Genomas Brasil).

Referências

AIM (2024). Acral lentiginous melanoma. AIM at Melanoma Foundation. Disponível em: https://www.aimatmelanoma.org/melanoma-101/types-of-melanoma/cutaneous-melanoma/acral-lentiginous-melanoma/. Último acesso em: 04 de Fevereiro 2025.

Alipour, N., Burke, T., and Courtney, J. (2024). Skin Type Diversity in Skin Lesion Datasets: A Review. *Current Dermatology Reports*, 13(3):198–210.

ARPANSA (2025). Fitzpatrick skin phototype. Australian Radiation Protection and Nuclear Safety AGENCY. Disponível em: https://www.arpansa.gov.au/sites/default/files/legacy/pubs/RadiationProtection/FitzpatrickSkinType.pdf. Último acesso em: 10 de Setembro 2025.

Barros, L., Chaves, L., and Avila, S. (2023). Assessing the generalizability of deep neural networks-based models for black skin lesions. In *Iberoamerican Congress on Pattern Recognition*, pages 1–14. Springer.

Bouzon, P. H. G., Rocha, W. F. d., Souza, L. A., and Pacheco, A. G. C. (2025). Metablocks-se: A method to deal with missing metadata in multimodal skin cancer classification. *IEEE Journal of Biomedical and Health Informatics*, 29(12):8855–8862.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples.

Daneshjou, R., Vodrahalli, K., Novoa, R. A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S. M., Bailey, E. E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J. A. C., Okata-Karigane, U., Zou, J., and Chiou, A. S. (2022). Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Ding, M., Xiao, B., Codella, N., Luo, P., Wang, J., and Yuan, L. (2022). Davit: Dual attention vision transformers.

Groh, M., Badri, O., Daneshjou, R., Koochek, A., Harris, C., Soenksen, L. R., Doraiswamy, P. M., and Picard, R. (2024). Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature Medicine*, 30(2):573–583.

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., and Badri, O. (2021). Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. arXiv:2104.09957 [cs].

Harp, T., Militello, M., McCarver, V., Johnson, C., Gray, T., Harrison, T., Presley, C., and Dellavalle, R. P. (2022). Further analysis of skin of color representation in dermatology textbooks used by residents. *Journal of the American Academy of Dermatology*, 87(1):e39–e41.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hosseini, S. M. and Baghshah, M. S. (2024). Dilated balanced cross entropy loss for medical image segmentation. *arXiv preprint arXiv:2412.06045*.

IBGE (2022). Censo demográfico 2022. Acessado em: 4 de Fevereiro 2026.

INCA (2026). Incidência do câncer no Brasil. Instituto Nacional do Câncer (INCA). Disponível em: https://ninho.inca.gov.br/jspui/handle/123456789/17914. Último acesso em: 06 de Março 2026.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., and Feichtenhofer, C. (2022). Mvitv2: Improved multiscale vision transformers for classification and detection.

Liu, Y., Primiero, C. A., Kulkarni, V., Soyer, H. P., and Betz-Stablein, B. (2023). Artificial intelligence for the classification of pigmented skin lesions in populations with skin of color: a systematic review. *Dermatology*, 239(4):499–513.

Lyakhov, P. A., Lyakhova, U. A., and Kalita, D. I. (2023). Multimodal analysis of unbalanced dermatological data for skin cancer recognition. *IEEE Access*, 11:131487–131507.

Lyakhova, U. A. (2022). Neural network skin cancer recognition with a modified cross-entropy loss function. In *International Conference on Actual Problems of Applied Mathematics and Computer Science*, pages 353–363. Springer.

MS (2025). Câncer de pele. Ministério da Saúde. Disponível em: https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-pele. Último acesso em: 19 de Julho 2025.

Pacheco, A. G. and Krohling, R. A. (2019). Recent advances in deep learning applied to skin cancer detection. In *Neural Information Processing Systems at Retrospectives workshop*, pages 1–8.

Pacheco, A. G. and Krohling, R. A. (2020). The impact of patient clinical information on automated skin cancer detection. *Computers in biology and medicine*, 116:103545.

Pacheco, A. G., Lima, G. R., Salomão, A. S., Krohling, B., Biral, I. P., de Angelo, G. G., Alves Jr, F. C., Esgario, J. G., Simora, A. C., Castro, P. B., et al. (2020). PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:1–10.

Porras Fimbres, D. C., Jacobs, J., Diamond, C., Rundle, C. W., Rodriguez-Homs, L., Presley, C., and Stamey, C. (2023). Representation of fitzpatrick skin phototype in dermatology surgical textbooks. *Archives of Dermatological Research*, 315(8):2463–2465.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4510–4520.

SCF (2025a). Skin cancer facts & statistics. Skin Cancer Foundation (SCF). Disponível em: https://www.skincancer.org/skin-cancer-information/skin-cancer-facts/. Último acesso em: 05 de Julho 2025.

SCF (2025b). Your skin type & skin cancer. Skin Cancer Foundation (SCF). Disponível em: https://www.skincancer.org/risk-factors/skin-type/. Último acesso em: 04 de Fevereiro 2025.

Sheng, Y., Yang, J., Li, J., Alaina, J., Xu, X., Shi, Y., Hu, J., Jiang, W., and Yang, L. (2024). Data-algorithm-architecture co-optimization for fair neural networks on skin lesion dataset. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 153–163. Springer.

Souza, L., Pacheco, A., Angelo, G., Oliveira-Santos, T., Palm, C., and Papa, J. (2024). Liwterm: A lightweight transformer-based model for dermatological multimodal lesion detection. In *Anais da XXXVII Conference on Graphics, Patterns and Images*, Porto Alegre, RS, Brasil. SBC.

Tan, M. and Le, Q. V. (2020). Efficientnet: Rethinking model scaling for convolutional neural networks.

Vidya, M. and Karki, M. V. (2020). Skin cancer detection using machine learning techniques. In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pages 1–5.

Wolff, K., Johnson, R., and Saavedra, A. (2015). *Dermatologia de Fitzpatrick - 7.ed.: Atlas e Texto*. AMGH Editora.

Yu, W., Si, C., Zhou, P., Luo, M., Zhou, Y., Feng, J., Yan, S., and Wang, X. (2024). Metaformer baselines for vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):896–912.

Yurdakul, M., Uyar, K., and Tasdemir, S. (2025). Maxglavit: A novel lightweight vision transformer-based approach for early diagnosis of glaucoma stages from fundus images.