Viés Racial em Modelos de Inteligência Artificial para Classificação de Melanomas

José Alberto Souza Paulino[1]

Universidade Federal de Campina Grande (UFCG) – Campina Grande - PB, – Brasil.
Grupo Independente de Pesquisa em IA para Comunidade Afro - IA.fro

souzapaulino@gmail.com

Abstract. *The use of artificial intelligence (AI) for skin cancer detection has been the subject of much research and development in recent years. However, recent studies suggest that skin cancer classification algorithms may have racial bias, performing worse on patients with darker skin. In this article, we evaluated the performance of an AI model in classify melanomas across 10 different skin tones, according to the Monk Scale. As a result, it was observed that the models have worse performance in classifying melanomas on darker skin.*

Resumo. *O uso de inteligência artificial (IA) para a detecção de câncer de pele tem sido objeto de muita pesquisa e desenvolvimento nos últimos anos. No entanto, estudos recentes sugerem que alguns algoritmos de classificação de câncer de pele podem ter viés racial, com desempenho pior em pacientes com pele mais escura. Nesse artigo, avaliamos o desempenho de um modelo de IA ao classificar melanomas em 10 diferentes tons de pele, de acordo com a Escala Monk. Como resultado, foi observado que os modelos têm pior desempenho para classificar melanomas em peles mais escuras.*

1. Introdução

O câncer de pele é uma das principais causas de morte por câncer em todo o mundo, com mais de 150 mil novos casos registrados somente em 2020, de acordo com dados da [WCRF]. No entanto, acredita-se que a quantidade real seja ainda maior, uma vez que a sub-notificação é comum nesse tipo de câncer. Como resultado, o câncer de pele representa grande desafio para a saúde pública atualmente. Os tipos mais comuns de câncer de pele são o melanoma e o carcinoma, sendo o melanoma o mais agressivo. Geralmente, o melanoma é mais prevalente em pessoas de pele branca. De acordo com [Gupta et al. 2016], a cor da pele é determinada principalmente pela presença de melanina, uma fonte natural de proteção contra radiação solar. A pele escura tem mais melanina, o que resulta em maior proteção em camadas mais profundas da pele. Estima-se que o fator de proteção solar ou SPF (do inglês *Sun Protection Factor*) em pessoas com pele preta seja de 13,4, enquanto em pessoas com pele branca seja de aproximadamente 3,3. Essa variável está diretamente associada ao risco de câncer de pele decorrente da exposição ao sol, bem como a outros fatores ambientais e agentes químicos.

Apesar do melanoma ter uma incidência muito maior em pessoas brancas, a taxa de sobrevivência para pessoas pretas acometidas por esse tipo de câncer é substancialmente menor. De acordo com a pesquisa de [Gupta et al. 2016], a sobrevivência para pessoas pretas é de apenas 70%, enquanto que para pessoas brancas ultrapassa os 92%.

Um dos principais fatores para essa redução na taxa de sobrevivência é o diagnóstico tardio. O processo de diagnóstico do melanoma envolve a inspeção visual da pele e, conforme [Jicman et al. 2023], há dificuldade maior em identificar lesões malignas em tons mais escuros de pele. Além disso, há pouco conhecimento sobre quais partes do corpo são mais adequadas para avaliação em pessoas com pigmentação escura de pele. Ou seja, embora existam mais casos de câncer de pele em pessoas brancas, proporcionalmente, pessoas pretas morrem mais quando acometidas por essa doença.

Atualmente, a ocorrência desbalanceada de casos de câncer de pele tem influenciado diretamente as bases de dados utilizadas em pesquisas na área, gerando viés racial na produção acadêmica e comercial de técnicas eficientes para detectar melanomas. Nesse contexto, em que a inteligência artificial (IA) está cada vez mais presente na saúde, os métodos de aprendizado profundo são o estado da arte na classificação do câncer de pele, como descrito em [Bissoto et al. 2019]. No entanto, como esses métodos necessitam de grandes volumes de dados e as bases de dados validadas estão cada vez mais escassas, os modelos de IA desenvolvidos alimentam-se cada vez mais de bases de dados enviesadas. Conforme enfatizado em [Bissoto et al. 2019], se o viés está presente em grandes bases de dados como *ImageNet*, por exemplo, é ingênuo pensar que pequenas bases não sofrerão de viés. Os autores fazem duras críticas às bases de dados de câncer de pele mais adotadas na literatura, como a *Interactive Atlas of Dermoscopy* (Atlas), ISIC 2016 e ISIC 2018, por não refletirem necessariamente as características do mundo real.

A questão do viés nas bases de dados utilizadas em pesquisas na área de melanoma é ainda mais complexa do que se pensava. Além do desbalanceamento natural da ocorrência de casos, [Bissoto et al. 2019] identificaram um novo tipo de viés nos algoritmos de aprendizado profundo utilizados para classificar a doença. Esses algoritmos não aprendem a classificar apenas a lesão, mas o contexto em que ela está inserida, ou seja, a pele do paciente. Isso significa que o modelo considera não apenas os aspectos clínicos da lesão, mas a pele do paciente, o que pode resultar em diagnósticos equivocados. De fato, uma análise estatística realizada pelos autores mostrou que mesmo cobrindo parte da lesão ou até mesmo toda ela, o modelo continuava classificando as imagens com uma preocupante taxa de acurácia acima de 70%.

Nesse contexto, somos intuitivamente levados a considerar que, quando os algoritmos estão usando também as características da pele para realizar a predição de uma lesão de câncer e são treinados em bases de dados com amostras de imagens predominantemente de pessoas brancas, terão dificuldade de realizar um diagnóstico efetivo em pessoas pretas. Apesar de ser impossível eliminar totalmente o viés das bases de dados, é importante identificar sua existência, entender suas origens e impactos com o objetivo de melhorar o processo de treinamento e utilização dos modelos num contexto de vida real. Desse modo, o objetivo dessa pesquisa reside em avaliar se o desempenho de um modelo de IA, capaz de classificar melanomas, sofre variações em diferentes tons de pele.

Para abordar a problemática descrita nesse capítulo introdutório, as sessões a seguir estão organizadas nesse formato: primeiro, na metodologia, é descrito como a pesquisa foi estruturada: são definias as bases de dados, os algoritmos utilizados, o processo de avaliação e como se dará a comparação das predições; na sessão seguinte são apresentados os resultados da pesquisa, e por fim é feita a discussão dos resultados e suas implicações para o uso da IA na sociedade na sessão de considerações finais.

2. Metodologia

A construção do processo de avaliação, objeto desse estudo, consiste em: inicialmente desenvolver um modelo de rede neural convolucional (RNC) capaz de classificar imagens de lesões de pele em benigna ou maligna (melanomas); em seguida, testar o modelo desenvolvido em uma segunda base de dados, segmentando as lesões dessa base e inserindo-as em 10 diferentes tons de pele de acordo com a Escala Monk[1] de Tons de Pele. Ou seja, usar o modelo desenvolvido para realizar a classificação das lesões do segundo conjunto de dados em diferentes tons de pele; por fim, avaliar e comparar os valores de AUC (*Area Under Curve*) resultantes das predições em cada tom de pele. Esse processo metodológico de validação externa em etapas está ilustrado na Figura1.

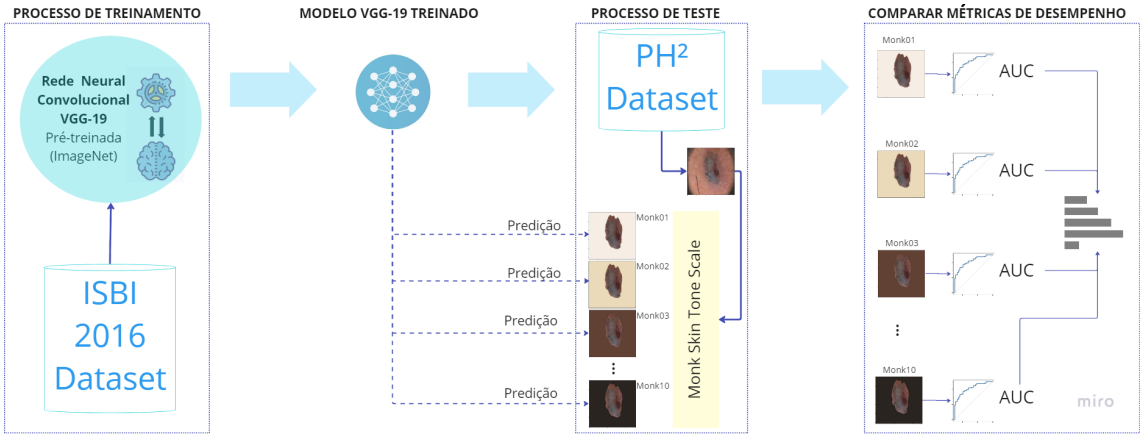

Figura 1. Etapas do processo de avaliação

2.1. Arquitetura de Rede Neural Convolucional

Para construção do modelo de RNC, foi utilizada a *Visual Geometry Group* (VGG), proposta por [Simonyan and Zisserman 2014], reconhecidamente uma arquitetura eficiente no campo de reconhecimento de imagens e possui um desenho de fácil construção. A VGG possui quatro versões: VGG-11, VGG-13, VGG-16 e VGG-19, sendo as duas últimas as mais difundidas na literatura. Nessa pesquisa, foi adotada a arquitetura VGG-19, uma arquitetura que faz uso de filtros convolucionais do tamanho 3x3 e são excelentes extratores de características. Ela considera como entrada uma imagem RGB com dimensões 224x224x3, possui 19 camadas profundas e totalmente conectadas, das quais 16 delas são convolucionais e 3 são camadas de classificação. Ela utiliza camadas *Max Pooling* no bloco de camadas convolucionais para reduzir dimensionalidade.

Para aprimorar o desempenho do modelo, adotou-se a técnica de Transferência de Aprendizagem. Essa técnica consiste em aproveitar o conhecimento prévio de modelos foram treinados com grande volume de dados em diferentes problemas [Menegola et al. 2017]. Dessa forma, foi construída a arquitetura do modelo usando apenas as camadas convolucionais da VGG-19 e fazendo uso da transferência de aprendizado a partir da base de dados do *ImageNet*, amplamente utilizada na literatura desde 2009 [Deng et al. 2009]. Em seguida, as camadas convolucionais foram ”congeladas”e novas camadas (de classificação) foram adicionadas, conforme descrito na Tabela 1.

[1]A Escala Monk foi desenvolvida em parceria com o Google para treinar modelos de inteligência artificial [Rezk et al. 2022].

Tabela 1. Camadas de classificação adicionadas manualmente

Descrição	Tipo
Camada convolucional de ajuste	GlobalAveragePooling2D
Camada de normalização 1	BatchNormalization
Camada de classificação 1	DenseLayer (128) Relu
Camada de normalização 2	BatchNormalization
Camada de regularização 1	Dropout com (0.5)
Camada de classificação 2 / saída	DenseLayer (2) Softmax

Os parâmetros utilizados no processo de treinamento incluem: uso do otimizador Adam com taxa de aprendizagem 10^{-4}; função de perda *Categorical Crossentropy*; para minimizar o platô do treinamento foi usada a função *ReduceLROnPlateau* do Keras com fator 0.2 se houver estabilidade por 20 épocas; foram utilizadas 120 épocas para o treinamento, com *bathsize* (lote de imagens) igual a 60; Por fim, haja visto que as bases de dados são extremamente desbalanceadas, foram adotados pesos diferentes para cada uma das classes, por meio do parâmetro *Weight* (peso), para a classe majoritária (benigna) foi dado peso 0.5 na classificação e para a classe minoritária (maligna) foi dado peso 0.95.

2.2. Bases de dados

Para o desenvolvimento dessa pesquisa, foi necessária a adoção de dois conjuntos de dados de imagens de câncer de pele, tal qual descrito anteriormente. São eles:

1. ISBI2016 Dataset: conforme [Gutman et al. 2016], é um conjunto de dados institucional com imagens dermatoscópicas em padrão de cores RGB e de alta qualidade (1022 x 1022), obtido no ISIC (Internacional Skin Imageing Collaboration). É dividida em dois subconjuntos: o primeiro subconjunto é rotulado como **TREINAMENTO** e possui 900 imagens, das quais 727 são de lesões benignas e 173 são de lesões malignas. O segundo subconjunto é denominado de **TESTE** e possui 379 imagens, das quais 273 são lesões benignas e 106 lesões malignas;
2. PH2 Dataset: conforme [Mendonça et al. 2013], é um conjunto de dados de imagens dermatoscópicas do Hospital Pedro Hispano, constituída de 200 imagens em padrão de cores RGB (768 × 560), das quais 160 são rotuladas como lesões benignas e 40 imagens rotuladas como lesões maligna (melanomas);

As lesões benignas do conjunto de dados PH2 se dividem em dois tipos, os **Nevos Comuns** e os **Nevos atípicos**. Uma vez que o objetivo da pesquisa é avaliar a influência do tom de pele na classificação, foi utilizado apenas o tipo **Nevos Comuns** para a classe de lesões benignas, construindo uma classificação binária, equivalente ao conjunto de dados ISBI2016: lesão benigna x lesão maligna.

No contexto do desenho apresentado na Figura 1, o primeiro conjunto de dados (ISBI2016) será utilizado no processo de treinamento do modelo e o conjunto de dados PH2, será utilizado para realização dos testes. Dessa forma, para cada lesão existente no PH2, serão criadas 10 novas imagens, cada uma delas corresponde um tom de pele de acordo com a Escala Monk, conforme mostrado na Figura 2. Ou seja, o conjunto de dados PH2 Dataset dará origem a 10 novos conjuntos de dados, um para cada tom de pele.

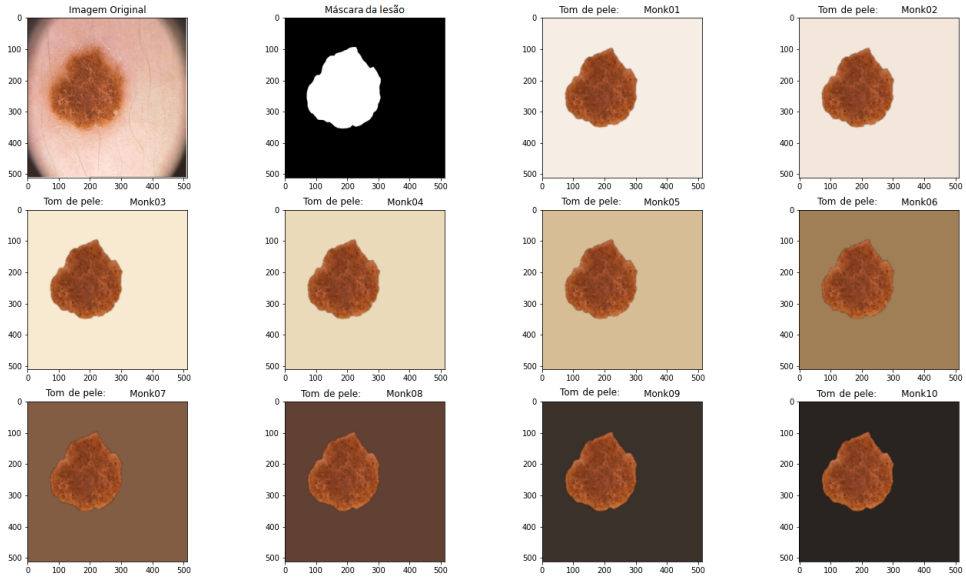

Figura 2. Representação de uma lesão do PH2 nos tons de pele da Escala Monk

2.3. Processo de avaliação

Com o modelo treinado e as previsões obtidas para cada subconjunto de imagens de tons de pele, seguindo o fluxo ilustrado na Figura 1, é possível avaliar o desempenho do modelo em cada um desses cenários. Para este propósito, optou-se por utilizar a métrica da área sob a curva (AUC - *Area Under Curve*) e, conforme explicado em [Bruce and Bruce 2019], AUC corresponde à área sob a curva ROC e é uma medida eficiente para avaliar o desempenho de classificadores. O valor da AUC varia entre 0 e 1, sendo que quanto mais próximo de 1, melhor é o desempenho do classificador. Embora a acurácia seja uma métrica comum adotada na literatura, ela pode levar a interpretações equivocadas em bases de dados muito desbalanceadas, como é o caso das bases utilizadas nesta pesquisa. Por sua vez, a curva ROC relaciona as métricas de sensibilidade e especificidade, diretamente relacionadas aos falsos positivos e negativos, possibilitando a apresentação de resultados mais robustos e confiáveis.

É importante ressaltar que o procedimento descrito na etapa dois do processo ilustrado na Figura 1 - que consiste na inserção das lesões da base de dados PH2 em diferentes tons de pele - foi realizado sem ajustes nas características originais das lesões, como: saturação, suavização ou normalização dos valores de pixels. Apenas foi realizada a sobreposição da lesão original em novas imagens cujo plano de fundo corresponde ao tom de pele da Escala Monk, conforme ilustrado na Figura 2. Por esse motivo, é esperado que os valores de AUC obtidos na etapa de avaliação dos novos conjuntos de dados sejam substancialmente inferiores aos valores obtidos na classificação das imagens originais. Entretanto, destaca-se que esse cenário não interfere diretamente no processo de avaliação, uma vez que o objetivo da pesquisa não é obter excelentes classificadores para os diferentes tons de pele, mas avaliar se o mesmo classificador apresenta variação de desempenho em diferentes contextos de pele.

3. Resultados

O modelo de rede neural convolucional treinado no conjunto de dados ISBI2016 na primeira etapa foi utilizado para classificar as imagens originais do conjunto de dados PH2, alcançando acurácia de 92,50%, *F1-Score* de 88,31% e AUC de 92,33%. vale destacar que esses esses resultados são compatíveis com outras pesquisas publicadas na literatura com essa mesma base de dados e fazendo uso de RNC, conforme observado em [Nambisan et al. 2023], [Arora et al. 2022], [Lynn and War 2019], [Brinker et al. 2019] e [Hekler et al. 2019]. A comparação entre esses estudos e o modelo ora desenvolvido foge ao escopo dessa pesquisa e os resultados supracitados são mencionados apenas para fins de compreensão sobre a eficiência do modelo treinado com a base ISBI2016 e testado com a base de dados PH2 com imagens originais. Reitera-se, entretanto, que a métrica de referência adotada nesta pesquisa será a AUC, como justificado na seção metodológica.

Os valores de AUC obtidos nesses testes para cada tom de pele da Escala Monk são apresentados no gráfico ilustrado na Figura 3:

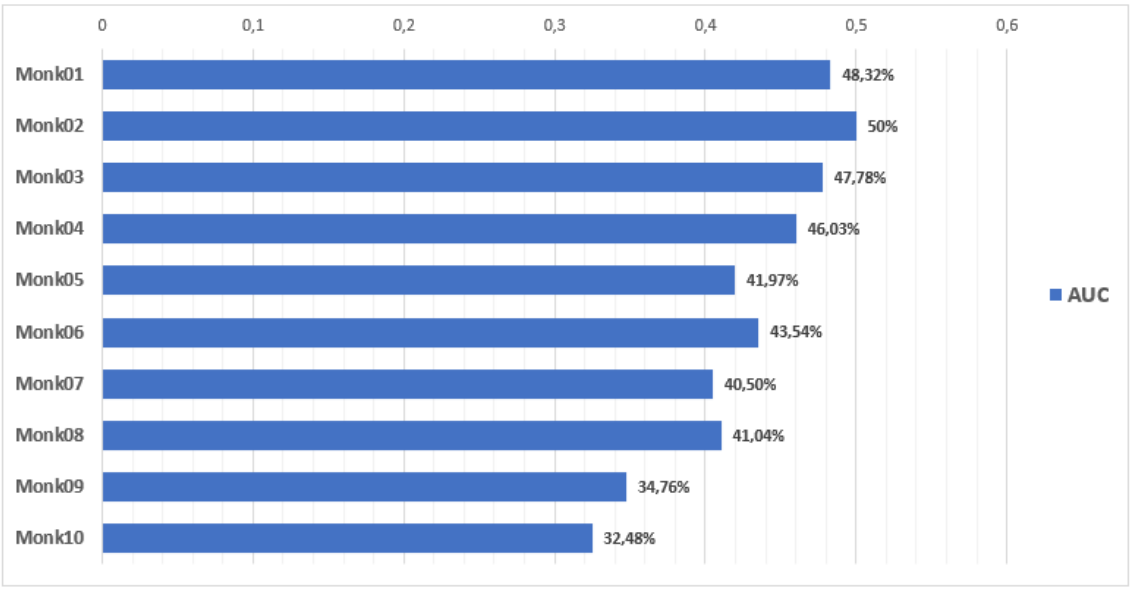

Figura 3. Valores de AUC para os tons de pele da Escala Monk

Os resultados mostram variação negativa da AUC a medida que o nível da Escala Monk aumenta, ou seja, a medida que o tom de pele é mais escuro. A maior diferença observada está entre o tom de pele Monk02 e o tom de pele Monk10, na qual os valores da AUC são 50% e 32,48%, respectivamente. Isso representa piora de 17,52 pontos percentuais no desempenho do classificador apenas com variação do tom de pele.

Essa variação ocorre porque o modelo passa a confundir o tom da pele com o padrão de cor da lesão, que normalmente tem pigmentação escura. Essa hipótese é reforçada ao avaliar a quantidade de amostras do grupo de verdadeiros negativos identificados pelo modelo. Os verdadeiros negativos, nesse contexto, são as amostras saudáveis que deveriam ser classificadas corretamente como saudáveis. Conforme ilustrado no gráfico da Figura 4, a quantidade de amostras classificadas corretamente como saudáveis diminui consideravelmente nos tons de pele mais escuros, Monk09 e Monk10.

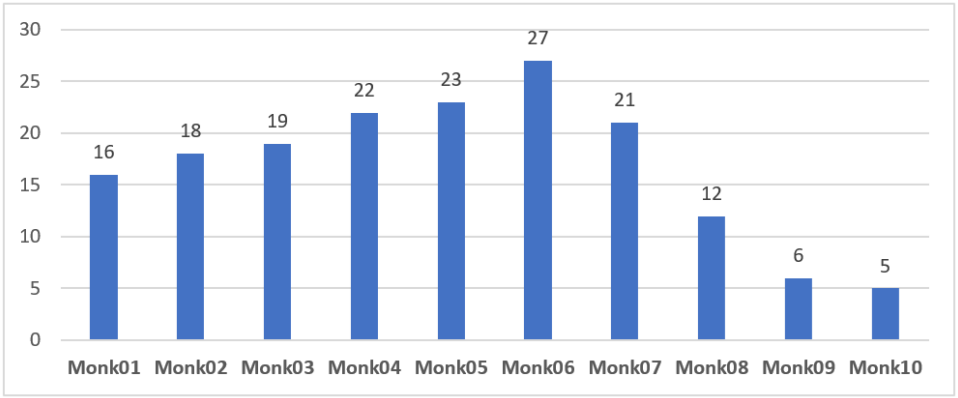

Figura 4. Quantidade de Verdadeiros negativos para tons de pele na Escala Monk

4. Considerações Finais

Os resultados dessa pesquisa corroboram a hipótese de que o desempenho do modelo apresenta variações ao classificar a mesma lesão em tons de pele diferentes e que o desempenho é pior ao classificar tons de pele mais escuros, cuja diferença é maior que 17% ao comparar tons mais claros e mais escuros. Esses achados contribuem para o entendimento acerca das limitações do uso de modelos de IA para classificação de melanomas, dado os vieses que essas aplicações podem incorporar e replicar, nesse caso, o viés racial.

A proposta desta pesquisa aborda uma questão fundamental na sociedade: o uso ético dos modelos de IA e as implicações negativas do uso indiscriminado desses modelos. A existência de viés racial na classificação de melanomas é apenas um dos vários exemplos desse problema, pois esse tipo viés está presente em muitos dos modelos utilizados atualmente. É alarmante constatar que as principais *APIs* de reconhecimento facial têm desempenho significativamente inferior em pessoas negras e mulheres, conforme [Buolamwini 2017]. O acesso às tecnologias deve ser democrático e inclusivo, alcançando diferentes classes, raças e gêneros, pois a exclusão desses aspectos só aumenta a segregação digital existente. Essa segregação digital agora ganha nova camada de complexidade com os modelos de IA, tornando ainda mais difícil combatê-la.

Como desdobramentos futuros, essa pesquisa tem como objetivo desenvolver estratégias para reduzir viés racial em modelos de IA que classificam melanomas. Dentre as possíveis vertentes está o uso de redes GAN para gerar amostras de pele na Escala Monk e uso da modelo de IA para colorir peles artificialmente em diferente bases de dados.

Referências

Arora, G., Dubey, A. K., Jaffery, Z. A., and Rocha, A. (2022). Bag of feature and support vector machine based early diagnosis of skin cancer. *Neural Computing and Applications*, pages 1–8.

Bissoto, A., Fornaciali, M., Valle, E., and Avila, S. (2019). (de) constructing bias on skin lesion datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0.

Brinker, T. J., Hekler, A., Enk, A. H., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., et al. (2019). Deep neural

networks are superior to dermatologists in melanoma image classification. *European Journal of Cancer*, 119:11–17.

Bruce, A. and Bruce, P. (2019). *Estatística prática para cientistas de dados*. Alta Books.

Buolamwini, J. A. (2017). *Gender shades: intersectional phenotypic and demographic evaluation of face datasets and gender classifiers*. PhD thesis, Massachusetts Institute of Technology.

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. Ieee.

Gupta, A. K., Bharadwaj, M., and Mehrotra, R. (2016). Skin cancer concerns in people of color: risk factors and prevention. *Asian Pacific journal of cancer prevention: APJCP*, 17(12):5257.

Gutman, D., Codella, N. C., Celebi, E., Helba, B., Marchetti, M., Mishra, N., and Halpern, A. (2016). Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*.

Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., Berking, C., Haferkamp, S., Klode, J., Schadendorf, D., et al. (2019). Superior skin cancer classification by the combination of human and artificial intelligence. *European Journal of Cancer*, 120:114–121.

Jicman, P. A., Smart, H., Ayello, E. A., and Sibbald, R. G. (2023). Early malignant melanoma detection, especially in persons with pigmented skin. *Advances in Skin & Wound Care*, 36(2):69–77.

Lynn, N. C. and War, N. (2019). Melanoma classification on dermoscopy skin images using bag tree ensemble classifier. In *2019 International Conference on Advanced Information Technologies (ICAIT)*, pages 120–125. IEEE.

Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., and Rozeira, J. (2013). Ph 2-a dermoscopic image database for research and benchmarking. In *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, pages 5437–5440. IEEE.

Menegola, A., Fornaciali, M., Pires, R., Bittencourt, F. V., Avila, S., and Valle, E. (2017). Knowledge transfer for melanoma screening with deep learning. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 297–300. IEEE.

Nambisan, A. K., Maurya, A., Lama, N., Phan, T., Patel, G., Miller, K., Lama, B., Hagerty, J., Stanley, R., and Stoecker, W. V. (2023). Improving automatic melanoma diagnosis using deep learning-based segmentation of irregular networks. *Cancers*, 15(4):1259.

Rezk, E., Eltorki, M., El-Dakhakhni, W., et al. (2022). Improving skin color diversity in cancer detection: Deep learning approach. *JMIR Dermatology*, 5(3):e39143.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

WCRF, W. C. R. F. I. Cancer trends: Skin cancer statistics.