Mitigating Bias in Facial Analysis through MST-based Skin Tone Classification on a Brazilian Electoral Dataset

1st Bruno Moreira
Faculdade de Computação (FACOM)
Universidade Federal de Uberlândia
Uberlândia, Brazil
brunomelo@ufu.br

***Abstract*—Facial analysis algorithms frequently exhibit significant bias, with disparate error rates across different demographic groups, particularly concerning skin tone. This issue is often rooted in unrepresentative training datasets and outdated classification scales that fail to capture the full spectrum of human skin tones. This paper proposes and evaluates an end-to-end pipeline to address these challenges in the context of the population of Minas Gerais. Our methodology includes the construction of a novel facial image dataset from a public Brazilian data source, the automated annotation of this dataset using the 10-point Monk Skin Tone (MST) scale, and the training of a MaxVit Vision Transformer for the classification task. The data collection process revealed a severe class imbalance, which necessitated adapting the problem to a four-class classification on the most representative tones. On this task, the model demonstrated high accuracy and robustness, validating the effectiveness of the proposed architecture for distinguishing between the selected skin tones. The results validate our pipeline as a viable approach for more equitable skin tone analysis but also empirically underscore that data imbalance remains a critical bottleneck, even when using modern classification scales.**

***Index Terms*—Facial Analysis, Algorithmic Bias, Skin Tone Classification, Vision Transformer, Monk Skin Tone Scale.**

I. Introduction

Dentre as partes do corpo humano, o rosto é a que provê o maior número de informações semânticas para identificar uma pessoa. A partir da face é possível estimar características como sexo, idade e expressões da mesma [1].

Algoritmos de aprendizado profundo são frequentemente aplicados para o reconhecimento de padrões entre um grande conjunto de dados, inclusive para imagens digitais de modo a reconhecer, segmentar ou processar objetos importantes contidos na imagem e realizar inferências a partir delas.

Segundo [2], a aplicação de técnicas de reconhecimento facial se destaca por ser não invasiva que funciona à distância, utilizando camêras de amplo acesso. Dessa forma, elas podem ser utilizadas para autenticação de identidade, vigilância ou controle de acesso de pessoas em ambientes monitorados.

No brasil essa tecnologia já está sendo utilizada pelo poder público, sendo levantados em [3], o uso dessa tecnologia em nove cidades: São Paulo, Rio de Janeiro, Brasília, Manaus, Belém, Porto Alegre, Campinas, São Luís e São Gonçalo, nas áreas de: Segurança Pública, Transporte Público, Saúde Pública, dentre outros.

Contudo, como demonstra [4], algoritmos de análise facial são frequentemente treinados em bases de dados desbalanceadas, o que gera taxas de erro drasticamente diferentes entre grupos sociais. Seu trabalho constata que esses algoritmos podem apresentar taxas de erro de 0.8% para homens brancos, enquanto que para mulheres pretas, de 34,7%.

Parte desse problema ocorre pela falta de baixa disponibilidade de dados com anotações étinico-raciais como apontado em [5] e pela baixa diversidade de base de dados populares na literatura, como afirmado em [6] e [7].

Em [8] é desenvolvido um método para detectar e reduzir o viés na classificação de tons de pele em grandes bases de dados de imagens faciais. O autor demonstra que modelos de reconhecimento facial, como a Facenet com ResNet50, têm pior desempenho ao distinguir pessoas de pele escura. Dessa forma, um método de rotulagem automática, combinando diferentes técnicas de processamento de imagem, é desenvolvido. Seus resultados demonstram que essa combinação é mais eficaz para reduzir o viés geral do que usar uma única abordagem.

Em [2], são desenvolvidos dois métodos que atuam na fase de extração de características, baseados na técnica Eigenface para realizar o reconhecimento de imagens de baixa resolução. Os métodos propostos trabalham redimensionando os próprios vetores de características (eigenvectors) para se adequarem à dimensão da imagem de baixa resolução. Os experimentos comparam essas novas abordagens com métodos tradicionais, avaliando o desempenho em métricas como sensibilidade, especificidade e acurácia em diferentes resoluções.

Este trabalho tem como objetivo avaliar um pipeline ponta-a-ponta para a classificação de tons de pele em um contexto de diversidade demográfica brasileira. O processo abrange a criação de um novo dataset a partir de fontes de dados públicas, e anotações automáticas com a escala Monk Skin Tone (MST) e, subsequentemente, o treinamento e a validação de um modelo Vision Transformer (MaxVit). Além disso, o artigo investiga os desafios práticos do desbalanceamento de dados e avalia a eficácia do modelo em um subconjunto representativo das classes.

A segunda seção deste artigo detalha a metodologia empregada, desde a coleta de dados até o treinamento do modelo. A terceira seção apresenta os resultados da classificação e a análise das métricas de desempenho. A quarta seção discute as implicações desses resultados, seguida pela conclusão e direcionamentos para trabalhos futuros.

II. Methodology

A metodologia de pesquisa é divida nas etapas apresentadas em 1, sendo elas respectivamente: Coleta dos Dados, Detecção Facial, Anotação de tons de pele, Construção da base de dados, Treinamento do modelo e Análise dos resultados.

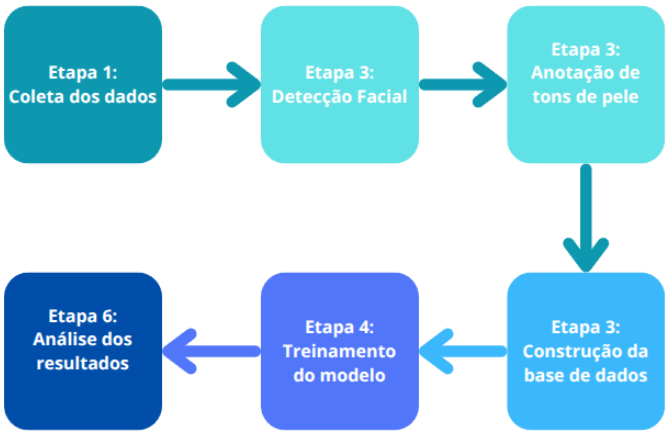

Fig. 1. Methodology Steps

As etapas de coleta de dados, detecção facial, anotação de tons de pele e análise dos resultados foram realizadas em um Desktop das seguintes configurações:

- RAM: 32GB DDR5
- CPU: AMD Ryzen 7 5700G
- GPU: NVIDIA RTX2060 Super
- OS: Linux Mint 21

O código foi desenvolvido em Python 3.8, utilizando as bibliotecas:

- Pytorch: Configuração do treinamento
- Timms: Importação de modelos
- Matplotlib: Plots
- Numpy: Processamento de vetores
- Webdataset: Construção do dataset
- PIL: Processamento de imagens
- Stone: Skin Tone Classification

Enquanto isso, os modelos foram treinados no ambiente compartilhado Google Colab, utilizando a versão Pro+ em uma máquina virtual do tipo GPU:A100. O código densenvolvido para a pesquisa está disponível em X.

A. Coleta de dados

Para compor a base de dados foram selecionadas imagens de acesso público do TSE (Tribunal Superior Eleitoral), reference aos candidatos às eleições de 2024 de Minas Gerais. A base é composta por 72939 imagens, cada uma anotada nas seguintes categorias: Brancos com 32859 entradas totais, Pretos com 10076, Pardos com 29638, Amarelos com 255, Indígenas com 111, com base na autodeclaração racial do candidato, apresentado pelo histograma em X. Dessa forma, o conjunto de dados apresenta-se diverso para as categorias: Pretos, Pardos e Brancos, contudo, não há amostragem suficiente de candidatos autodeclarados índigenas ou amarelos.

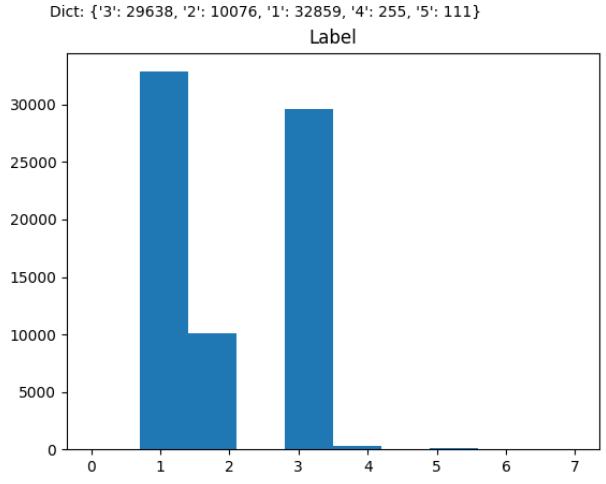

Fig. 2. Histograma de autodeclaração racial: Branco (1), Preto (2), Pardo (3), Amarelo (4) e Indígena (5)

As imagens apresentam a face dos candidatos, com variações de pose e iluminação, algumas apresentam características além da face, como ombros e torso do candidato, sendo necessário aplicar técnicas de detecção facial para extrair somente as regiões de interesse, as quais são: Contorno da face, olhos, boca, nariz e raiz do cabelo.

B. Detecção Facial

Para detecção facial, o modelo MTCNN [9] foi selecionado devido à sua robustez em condições do mundo real (variação de iluminação, oclusões, ângulos), superando alternativas como dlib-HOG [10] e FaceNet [11] conforme demonstrado em benchmarks independentes [12].

Após detecção, os landmarks faciais (Fig. 3a) são utilizados para extrair a região de interesse (ROI) centrada no nariz com padding de 40% da largura/altura facial. Esta abordagem segue protocolos de antropometria facial [13] para garantir cobertura consistente de regiões críticas. As ROIs são então redimensionadas para 300×300 pixels via interpolação bicúbica [14], otimizando o trade-off entre detalhamento e eficiência computacional [15].

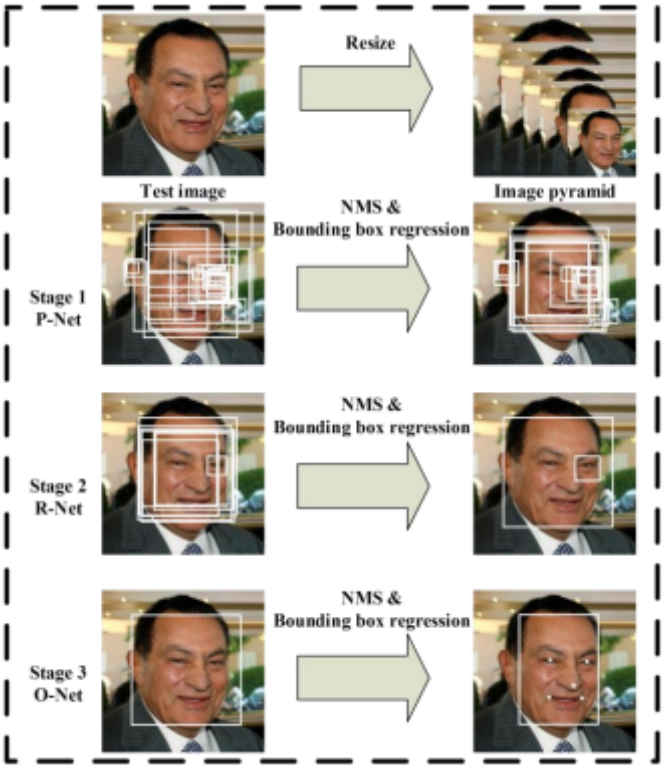

Fig. 3. MTCNN detection process and landmarks extraction [9]

C. Anotação de tons de pele

Diversas escalas de tons de pele foram consideradas para anotação do conjunto de dados, incluindo a escala de Fitzpatrick [16], desenvolvida para classificação de fototipos em dermatologia, e a escala Perla [**?**], baseada em percepções subjetivas de luminosidade. Entretanto, a escala Monk Skin Tone (MST) [17] foi selecionada por sua capacidade superior de representar a diversidade étnica da população brasileira.

Desenvolvida por [17], a MST emprega um espaço de cores perceptualmente uniforme (CIELAB) e agrupa tons de pele em 10 categorias equidistantes (Fig. **??**), evitando viés geográfico inerente a escalas tradicionais [18].

A anotação automática foi realizada utilizando a biblioteca stone (Skin TONE classification) [19], que implementa um pipeline de visão computacional otimizado para extração de tons de pele. O processo segue três etapas:

1) **Detecção facial**: Utiliza o detector MTCNN [9] para identificar regiões de interesse (ROIs).
2) **Extração de cor**: Calcula a média de pixels no espaço CIELAB na região da bochecha, minimizando efeitos de iluminação.
3) **Classificação MST**: Mapeia o valor médio de L^* (luminância) para o tom MST mais próximo via k-NN ($k = 5$) [20], gerando labels $\in \{1, 2, \ldots, 10\}$.

TABLE I
MONK SKIN TONE

Key	Hex Value
Monk 01	#f6ede4
Monk 02	#f3e7db
Monk 03	#f7ead0
Monk 04	#eadaba
Monk 05	#d7bd96
Monk 06	#a07e56
Monk 07	#825c43
Monk 08	#604134
Monk 09	#3a312a
Monk 10	#292420

D. Construção da base de dados

O conjunto de dados foi dividida entre conjunto de treinamento e de validação de forma aleatória, através de um sorteador, resultando em um split de aproximadamente 80% e 20%, respectivamente. Para a construção da base de dados foi utilizada a biblioteca webdataset, de X, que é desenhada para lidar com grandes volumes de dados de forma otimizada.

E. Treinamento do modelo

Os modelos baseados em Vision Transformers, têm se tornado populares devido a seu mecanismo de autoatenção e capacidade de extrair relações entre uma sequência de dados. Dessa forma as imagens anotadas foram treinadas a partir da rede maxvit_nano_rw_256, disponível em [21], uma arquitetura baseada na MaxVit (Multi-Axis Vision Transformers) [22]. Para o treinamento foram utilizados os hiperparâmetros em II. Ainda, o treinamento foi feito em cima do modelo pré-treinado com o dataset ImageNet_V2 em [23].

1) Vision Transformer: O Vision Transformer (ViT), exemplificado em 4, adapta arquiteturas de transformers a imagens, a partir de sua divisão em patches linearizados, tratados como sequências [24]. Cada patch é mapeado em uma embedding de dimensão fixa, preservando relações espaciais via codificação posicional [25].

”O mecanismo de self-attention, exemplificado em 5, núcleo dos *transformers*, permite que todos os *patches* interajam globalmente, superando limitações de CNNs: enquanto *kernels* convolucionais capturam features locais com janelas fixas, a

TABLE II
HYPERPARAMETERS

Hyperparameter	Value
Learning Rate	1e-3
Batch Size	128
Scheduler Step Size	10
Scheduler	StepLR
Optimizer	AdamW
Loss Function	CrossEntropyLoss
Weight Decay	1e-2
Shuffle Size	57862
Gamma	0.95
Label smoothing	1e-1
Epochs	100

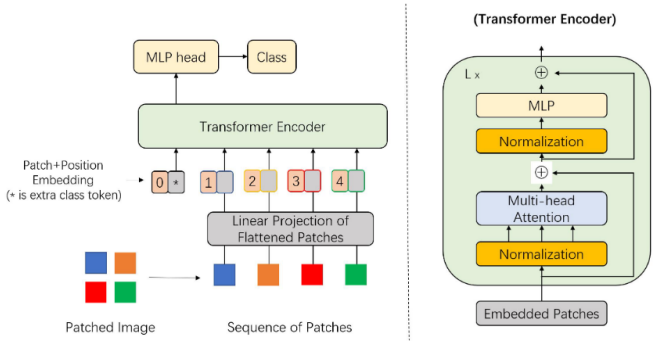

Fig. 4. Self-attention mechanism from X

self-attention modela dependências de longo alcance com pesos dinâmicos [25].”

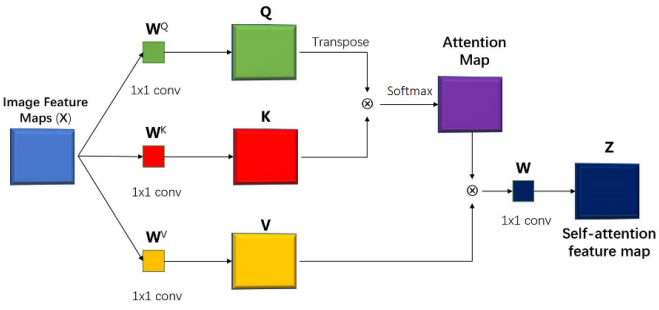

Fig. 5. Self-attention mechanism from X

As matrizes Query (Q), Key (K) e Value (V) são geradas por projeções lineares independentes. A atenção é calculada pela correlação entre Q e K, ponderando V conforme:

$$\mathbf{Z} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$$

onde d_k controla a escala do produto escalar [**?**]. A versão multi-head paraleliza esse processo, com múltiplos conjuntos de pesos extraindo hierarquias contextuais distintas.

2) MaxVit: A MaxViT resolve limitações críticas de vision transformers puros (ViT), como a complexidade quadrática da autoatenção global [24] e a falta de viés indutivo para invariância espacial [26]. Seu mecanismo de atenção multi-eixo (Max-SA) decompõe operações globais em duas formas esparsas:

* Block Attention: Similar à abordagem de janelas do Swin Transformer [26], mas sem shifted windows, aplicando atenção local em blocos não sobrepostos.
* Grid Attention: Inspirado em técnicas de amostragem dilatada [27], permite interações globais com complexidade linear via grade adaptativa.

O destaque do modelo consiste na integração hierárquica de convolutions (MBConv) e atenção esparsa. Enquanto a CoAtNet [28] empilha blocos convolucionais e atencionais separadamente, a MaxViT unifica-os em um único bloco repetível, combinando eficiência de convoluções para padrões locais [29] e capacidade de transformers para dependências de longo alcance. Isso permite percepção global em altas resoluções, em contraste com a ViT original [24].

III. RESULTS

A partir da análise da base de dados construída a partir da anotação de tons de pele na escala MST das imagens coletadas, foi observado de que ela não possui quantidades relevantes de dados para as categorias: 1, 2, 3, 4, 9, 10 da escala MST, exemplificada em 6. Dessa forma, o modelo escolhido foi configurado para distinguir entre quatro classes, referentes às categorias: 5, 6, 7, 8 da escala MST. Para o treinamento entre as quatro classes, foram adotadas as labels: 0, 1, 2, 3, representando, respectivamente, as categorias acima.

A base de dados resultante dessa seleção de categorias, possui um tamanho total de 57862 para o conjunto de treinamento e 14364 para o conjunto de validação, representando respectivamente e aproximadamente 80% e 20% do total da base.

Para a avaliação do modelo, quatro métricas foram empregadas: Acurácia (Accuracy), que mede a proporção total de predições corretas em relação a todas as amostras do conjunto de teste; Loss (Perda), calculada pela função objetivo Cross-Entropy durante o treinamento e validação, indicando a convergência e qualidade da otimização; F1-Score, métrica harmônica que combina Precisão e Recall

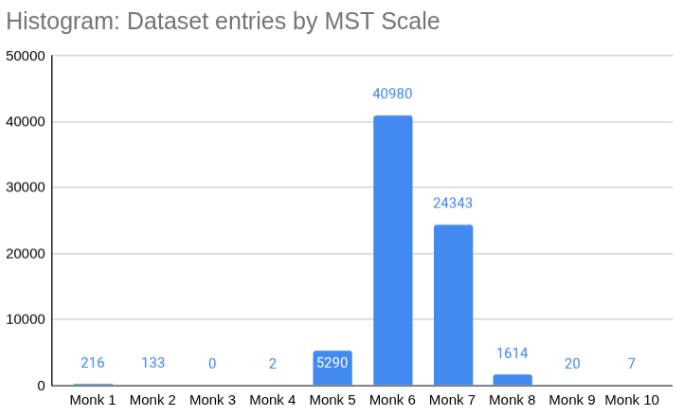

Fig. 6. Histogram of dataset labels by MST categories

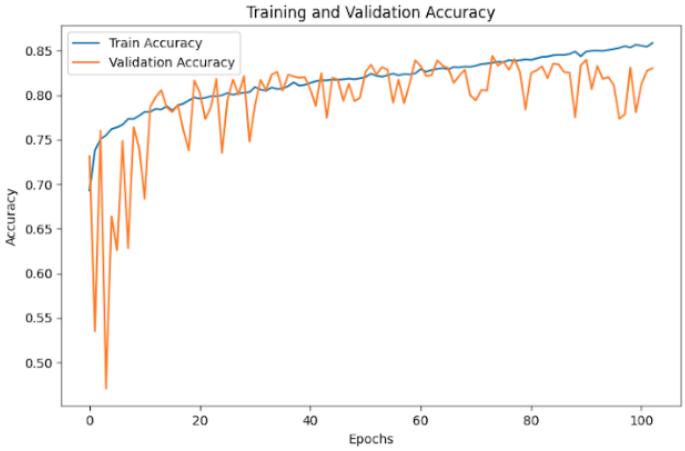

Fig. 7. Accuracy for Training and Validation along Epochs

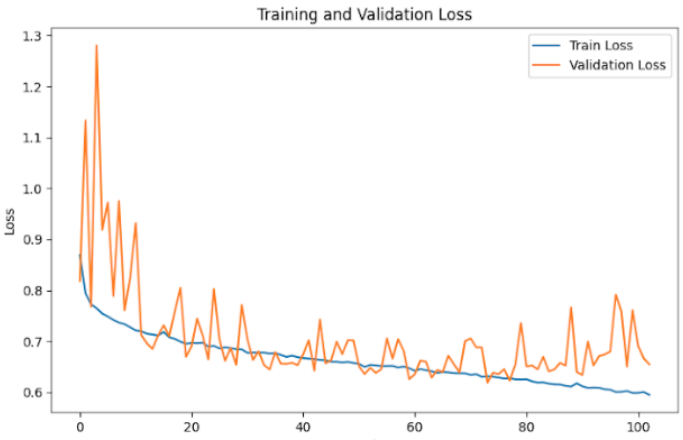

Fig. 8. Loss for Training and Validation along Epochs

(sensibilidade), crítica em cenários com classes desbalanceadas; e AUC (Area Under the ROC Curve), que avalia a capacidade de discriminação entre classes, independentemente da distribuição dos dados.

Conforme sumarizado na Tabela III, o modelo alcançou acurácia de 84,39% e AUC de 0.9626, com F1-Score de 0.84 em todas as classes. Este último demonstra que o desbalanceamento das classes afetou consideravelmente o modelo.

TABLE III
METRICS

Metric	Training	Validation
Accuracy	0.8363	0.8438
Loss	0.6314	0.6189
F1-Score	0.8336	0.8439
AUC	0.957	0.9626

De acordo com o gráfico 7 é possível observar que o conjunto de treinamento convergiu suavemente ao longo das épocas. O conjunto de validação também convegiu, porém apresentou elevada variação nas épocas iniciais e decresceu ao se aproximar do final do treinamento. Ainda, a acurácia do modelo não estibilizou, sinalizando que o modelo ainda pode melhorar sua capacidade de aprendizado se treinado para mais épocas

Ainda, considerando o gráfico 8, foi possível observar que o erro de treinamento e validação se manteram próximos a partir da época 20, porém a partir da época 80 o erro de validação começou a cair drásticamente, indicando a presença de um mínimo local no modelo e a presença de overfitting.

O desbalanceamento das classes afetou a capacidade preditiva do modelo, apresentando, segundo gráfico 9 um F1-Score que variou muito nas épocas iniciais, estabilizou por volta da época 60 e voltou a cair a partir da 80. Ainda, o crescente AUC, identificado em 10 indica que o modelo aprendeu a diferenciar cada classe suficientemente para predizê-la, porém esse valor começa a cair para o conjunto de validação a partir da época 80.

A análise da matriz de confusão revela um desempenho díspar do modelo para as quatro classes, o que está diretamente ligado ao desbalanceamento dos dados. O modelo apresenta alta proficiência na classificação da classe 1 (majoritária), com 7.134 acertos, e também um bom desempenho na classe 2, com 4.003 acertos. Em contrapartida, o desempenho mais fraco ocorre na classe minoritária 3, que obteve apenas 239 predições corretas, evidenciando a dificuldade do modelo em aprender com um número reduzido de amostras.

Além disso, matriz evidencia padrões de confusão significativos. Nota-se uma confusão mútua e expressiva entre as classes 1 e 2, onde 690 amostras da classe 2 foram incorretamente previstas como classe 1 e 767 da classe 1 foram previstas como 2. Enquanto

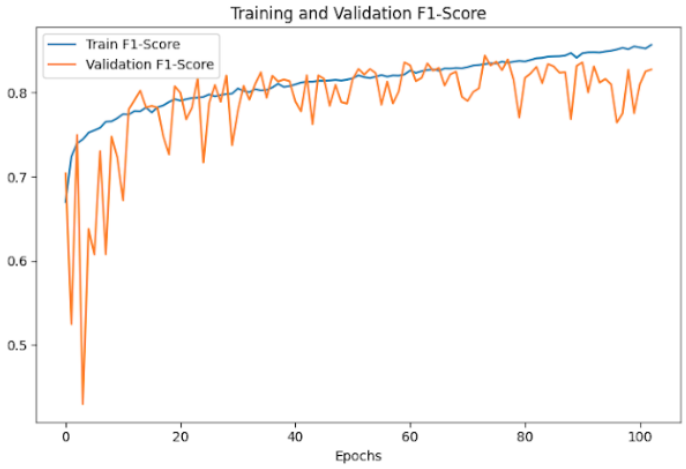

Fig. 9. F1-Score for Training and Validation along Epochs

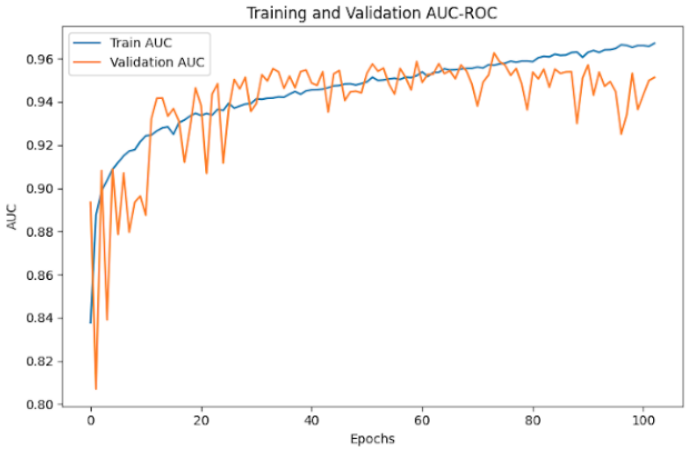

Fig. 10. AUC-ROC for Training and Validation along Epochs

isso, a classe 3 foi frequentemente confundida com a classe 2.

IV. Discussions

Os resultados demonstram que a arquitetura MaxVit, pré-treinada na ImageNet-V2, foi capaz de atingir um desempenho relevante na tarefa de classificação de tons de pele, com métricas crescentes, mesmo diante do desbalanceamento de classes no subconjunto de quatro categorias. Isso sugere que a combinação de convoluções locais e mecanismos de atenção esparsa da MaxVit é eficaz para extrair características discriminatórias relevantes para a identificação de tons de pele.

A análise das curvas de treinamento (Fig. 7, 8, 9 e 10) revela uma dinâmica de aprendizado rápida nas épocas iniciais, seguida por uma fase de estabilização. No entanto, a divergência entre as curvas de perda de treinamento e validação indica a presença overfitting. Embora esse problema não

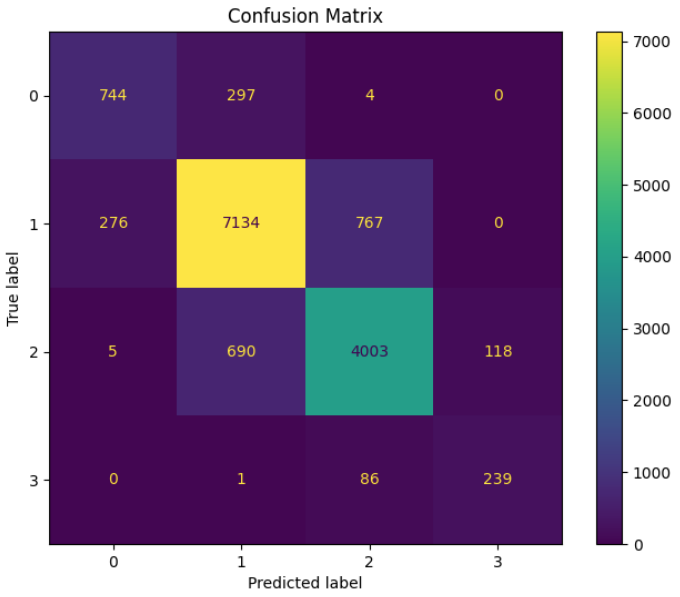

Fig. 11. Confusion Matrix for Validation Set

tenha impactado drasticamente as métricas finais de classificação, ele evidencia a necessidade de estratégias de regularização mais robustas ou de um volume de dados ainda maior para treinar modelos de alta capacidade como os Vision Transformers.

Este trabalho reforça as descobertas de [4], que apontaram o desbalanceamento de dados como uma causa primária para o viés em sistemas de reconhecimento facial. Ao construir um dataset a partir de uma fonte de dados brasileira, observamos empiricamente a dificuldade em obter uma representação uniforme entre os 10 tons da escala MST, um desafio também apontado por [5]. A concentração de amostras nos tons intermediários (5 a 8) reflete uma realidade demográfica que, se não tratada, levaria inevitavelmente a um modelo com baixo desempenho para tons de pele sub-representados.

Diferentemente do trabalho de [8], que focou em métodos para detectar e reduzir o viés em datasets já existentes, nossa abordagem se concentrou na criação de um novo recurso de dados anotado com a escala MST. A performance obtida na tarefa de 4 classes sugere que, quando havendo dados suficientes para uma categoria, os modelos modernos de Vision Transformer podem, de fato, aprender a distinguir tons de pele com precisão.

Outra limitação reside na fonte de dados. Embora as imagens do TSE sejam públicas e de ampla cobertura nacional, elas representam um subconjunto específico da população (candidatos a eleições) de Minas Gerais e possuem um formato semi-padronizado (fotos 3x4). Isso pode não capturar a variabilidade de iluminação, pose e cenários encontrada ”no mundo

real".

Para trabalhos futuros, os esforços devem se concentrar em três frentes principais: Resolver o problema original de classificação de 10 classes, explorando técnicas avançadas para lidar com o desbalanceamento de dados; Expandir e diversificar a base de dados com imagens de outros estados para aprimorar a robustez e a capacidade de generalização do modelo em cenários do mundo real; Análise comparativa de outras arquiteturas de estado da arte, como Swin Transformers e EfficientNetV2.

REFERENCES

[1] M. Obayya, S. S. Alotaibi, S. Dhahb, R. Alabdan, M. Al Duhayyim, M. A. Hamza, M. Rizwanullah, and A. Motwakel, "Optimal deep transfer learning based ethnicity recognition on face images," *Image and Vision Computing*, vol. 128, p. 104584, Dec. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S026288562200213X

[2] J. I. S. da Silva, "Reconhecimento facial em imagens de baixa resolução," 2015.

[3] R. Brandão and J. L. Oliveira, "Reconhecimento facial e viés algorítmico em grandes municípios brasileiros," in *Workshop sobre as Implicações da Computação na Sociedade (WICS)*. SBC, 2021, pp. 122–127.

[4] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," 2018.

[5] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, "Diversity in faces," *arXiv preprint arXiv:1901.10436*, 2019.

[6] S. Yucer, S. Akcay, N. Al-Moubayed, and T. P. Breckon, "Exploring Racial Bias within Face Recognition via per-subject Adversarially-Enabled Data Augmentation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 83–92. [Online]. Available: https://ieeexplore.ieee.org/document/9150678/

[7] K. Kärkkäinen and J. Joo, "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age," Aug. 2019, arXiv:1908.04913 [cs]. [Online]. Available: http://arxiv.org/abs/1908.04913

[8] M. LAV, "Reduzindo viés em classificação de tons de pele em bases de dados de imagens [dissertação]. são carlos: Universidade de são paulo, instituto de ciências matemáticas e de computação; 2022 acesso em 22/10/2023."

[9] K. Zhang *et al.*, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[10] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009. [Online]. Available: http://www.jmlr.org/papers/v10/king09a.html

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823. [Online]. Available: https://ieeexplore.ieee.org/document/7298682

[12] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Style aggregated network for facial landmark detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 8453–8465, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9157541

[13] L. G. Farkas, *Anthropometry of the Head and Face*. Raven Press, 1994.

[14] R. G. Keys, "Cubic convolution interpolation for digital image processing," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 6, pp. 1153–1160, 1981. [Online]. Available: https://ieeexplore.ieee.org/document/1163711

[15] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *ICML*, 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[16] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types i through vi," *Archives of dermatology*, vol. 124, no. 6, pp. 869–871, 1988.

[17] C. Schumann, G. O. Olanubi, A. Wright, E. Monk Jr, C. Heldreth, and S. Ricco, "Consensus and subjectivity of skin tone annotation for ml fairness," *arXiv preprint arXiv:2305.09073*, 2023.

[18] J. Adams *et al.*, "Fairness in skin tone estimation: A comparative analysis," in *CVPR Workshops*, 2021, pp. 88–95.

[19] R. A. Rejón Pia and C. Ma, "Classification algorithm for skin color (casco): A new tool to measure skin color in social science research," *Social Science Quarterly*, vol. n/a, no. n/a. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ssqu.13242

[20] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[21] R. Wightman, "Pytorch image models," https://github.com/huggingface/pytorch-image-models, 2019.

[22] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "Maxvit: Multi-axis vision transformer," *ECCV*, 2022.

[23] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[25] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision transformers for image classification: A comparative survey," *Technologies*, vol. 13, no. 1, p. 32, 2025.

[26] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[28] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *arXiv preprint arXiv:2106.04803*, 2021.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.