COMPARAÇÃO ENTRE REDES NEURAIS CONVOLUCIONAIS NA CLASSIFICAÇÃO MULTICLASSE DE LESÕES DE PELE COM ANÁLISE DE ESTRATÉGIAS DE BALANCEAMENTO

Rubens A. de S. Ferreira[1], Luis F. P. de P. Pordeus[1], Rosana C. B. Rego[2]

[1] *Bacharelado em Engenharia de Software, Universidade Federal Rural do Semi-Árido (UFERSA), Rio Grande do Norte, Brasil (rubens.ferreira@alunos.ufersa.edu.br, luis.pordeus@alunos.ufersa.edu.br)*

[2]*Departamento de Engenharias e Tecnologia, Universidade Federal Rural do Semi-Árido (UFERSA), Rio Grande do Norte, Brasil*

Resumo: Este trabalho avalia estratégias de balanceamento, segmentação e transferência de aprendizado na classificação multiclasse de lesões de pele no HAM10000. Foram testados pesos por classe, SMOTE, ROS, máscaras, imagens segmentadas e ResNet50 com *fine-tuning*. O melhor resultado foi obtido com ResNet50 + *Class Weight* + *Fine-tuning*, alcançando F1-macro de 0,6016, indicando maior equilíbrio entre as classes. A análise com Grad-CAM complementou a avaliação quantitativa.

Palavras-chave: Câncer de pele; HAM10000; Redes neurais convolucionais; Balanceamento de dados; Transferência de aprendizado.

INTRODUÇÃO

O diagnóstico precoce do câncer de pele é determinante para a redução de complicações associadas à evolução da doença e para a ampliação das possibilidades de tratamento. Nesse contexto, técnicas de inteligência artificial têm sido investigadas como ferramentas de apoio à análise de imagens médicas, especialmente em tarefas de classificação automática de lesões cutâneas.

Entre os conjuntos de dados utilizados nesse tipo de estudo, destaca-se o HAM10000, composto por imagens dermatoscópicas distribuídas em sete classes diagnósticas (TSCHANDL; ROSENDAHL; KITTLER, 2018). Esse conjunto apresenta desafios para o treinamento de modelos de aprendizado profundo, como a semelhança visual entre lesões, a variação de cor, textura e formato, além do desbalanceamento entre as classes.

Em estudo anterior (FERREIRA; REGO, 2026), foi estabelecida uma linha de base para esse problema utilizando uma rede neural Perceptron Multicamadas (MLP). O modelo alcançou acurácia de $70,84\%$ e F1-score ponderado de $70,32\%$ em sua configuração com SMOTE, taxa de aprendizado $\alpha = 0,0001$ e arquitetura $128 - 64$. Os resultados evidenciaram limitações estruturais da abordagem, uma vez que a MLP exige o achatamento das imagens em vetores unidimensionais, com consequente perda de relações espaciais entre os pixels, dificultando a extração de padrões visuais para a classificação das lesões.

Diante dessas limitações, redes neurais convolucionais constituem alternativas para problemas de visão computacional, pois preservam relações espaciais e extraem características locais das imagens, como bordas, texturas e formas (GOODFELLOW et al., 2016).

Além disso, estratégias como aumento de dados, atribuição de pesos por classe, sobreamostragem com SMOTE e *Random Over Sampler*, segmentação e transferência de aprendizado com arquiteturas pré-treinadas, como a ResNet50, podem ser utilizadas para reduzir os efeitos do desbalanceamento e melhorar a generalização dos modelos (HE et al., 2016).

No entanto, a aplicação dessas técnicas não garante melhoria uniforme em todas as métricas de avaliação. Em problemas multiclasse e desbalanceados, a acurácia pode não representar adequadamente o desempenho do modelo, uma vez que resultados globais podem estar associados ao acerto nas classes majoritárias. Por isso, torna-se necessário avaliar o desempenho também por meio de métricas como precisão, *recall* e F1-score, considerando tanto médias macro quanto ponderadas.

Dessa forma, este trabalho tem como objetivo avaliar o impacto de diferentes estratégias de balanceamento, segmentação e transferência de aprendizado na classificação multiclasse de lesões de pele no dataset HAM10000. Para isso, foram testadas abordagens com pesos de classe, SMOTE, *Random Over Sampler*, imagens segmentadas, máscaras e modelos baseados em ResNet50 com *fine-tuning*, buscando analisar seus efeitos sobre o desempenho geral e sobre o equilíbrio da classificação entre as classes.

MATERIAL E MÉTODOS

Para este trabalho, foi utilizado o conjunto de dados HAM10000, composto por 10.015 imagens dermatoscópicas de lesões distribuídas em sete classes diagnósticas: Melanoma (MEL), Nevos Melanocíticos (NV), Carcinoma Basocelular (BCC), Queratose Benigna (BKL), Queratose Actínica (AKIEC), Dermatofibroma (DF) e Lesões Vasculares (VASC). A Ta-

bela 1 apresenta a distribuição das imagens por classe. Esse conjunto de dados foi escolhido por ser utilizado em pesquisas envolvendo classificação automática de lesões cutâneas (TSCHANDL; ROSENDAHL; KITTLER, 2018). Uma de suas características é o desbalanceamento entre as classes, uma vez que a classe Nevos Melanocíticos possui quantidade superior de amostras em relação às demais, representando cerca de 67% do total de imagens. Essa característica torna o problema desafiador, pois pode fazer com que o modelo aprenda os padrões da classe majoritária com maior facilidade e apresente desempenho inferior nas classes menos representadas.

Tabela 1: Distribuição dos dados por classe.

Classe	Quantidade	Porcentagem
MEL	1.113	$11,11\%$
NV	6.705	$66,95\%$
BCC	514	$5,13\%$
BKL	1.099	$10,97\%$
AKIEC	327	$3,27\%$
DF	115	$1,15\%$
VASC	142	$1,42\%$

As imagens foram redimensionadas para um tamanho fixo, a fim de padronizar a entrada dos modelos. Nos experimentos com redes convolucionais, foram utilizadas imagens com dimensões de $128x128$ pixels. Nos experimentos com a arquitetura ResNet50, as imagens foram redimensionadas para $160x160$ pixels. Os dados foram divididos em conjuntos de treino, validação e teste, utilizando a proporção de 70%, 15% e 15%, respectivamente. Também foram aplicadas técnicas de *data augmentation* no conjunto de treinamento, com o objetivo de aumentar a variabilidade das imagens e reduzir o risco de sobreajuste. As transformações utilizadas incluíram rotação, espelhamento horizontal, zoom e ajuste de contraste.

Estratégias de Balanceamento

O conjunto de dados HAM10000 apresenta um desbalanceamento significativo entre as classes, com a classe NV representando a maioria das amostras. Dessa forma, foram avaliadas estratégias para reduzir o impacto desse desbalanceamento e melhorar a capacidade dos modelos de aprender padrões relevantes para as classes minoritárias. Para isso, foram testadas as seguintes abordagens: atribuição de pesos por classe (*class weight*), SMOTE (*Synthetic Minority Oversampling Technique*) e ROS (*Random Over Sampler*).

A Figura 1 apresenta a comparação entre a quantidade de imagens por classe no conjunto de treino antes e após a aplicação das técnicas de sobreamostragem.

O SMOTE é uma técnica de sobreamostragem que cria amostras sintéticas para as classes minoritárias (CHAWLA et al., 2002). No contexto deste trabalho, essa técnica foi aplicada somente ao conjunto de treino, evitando que dados sintéticos fossem utilizados no conjunto de teste. Também foi avaliado o método ROS, que

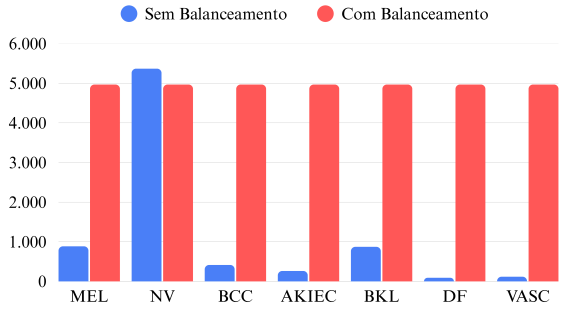


Figura 1: Comparação da quantidade de imagens por classe no conjunto de treino sem balanceamento e após a aplicação das técnicas SMOTE e ROS.

realiza a sobreamostragem das classes minoritárias por meio da duplicação de amostras existentes. Diferentemente do SMOTE, o ROS não gera novas amostras artificiais, mas aumenta a quantidade de exemplos das classes menos representadas a partir dos próprios dados disponíveis.

A atribuição de pesos por classe, por sua vez, foi aplicada durante o treinamento dos modelos, ajustando a função de perda para dar maior importância às classes minoritárias.

Redes Neurais Convolucionais

Foram avaliados dois modelos convolucionais distintos para a classificação das lesões de pele, uma redes convolucional simples e a arquitetura ResNet50 com transferência de aprendizado (HE et al., 2016).

A arquitetura convolucional base recebeu imagens RGB redimensionadas e iniciou com uma camada de normalização por *Rescaling*. Em seguida, foram utilizados blocos convolucionais com filtros 3×3, ativação ReLU, normalização em lote e *MaxPooling*, aumentando progressivamente o número de filtros de 32 para 64 e 128. Após a extração de características, foi aplicada uma camada GlobalAveragePooling2D, seguida por uma camada densa com 64 neurônios, regularização L2 de $0,001$, *Dropout* de $0,5$ e uma camada final com sete neurônios, correspondente às classes diagnósticas avaliadas.

Na arquitetura ResNet50, foram adicionadas camadas finais para adaptação ao problema de classificação multiclasse das lesões de pele. A Figura 2 apresenta a arquitetura final baseada em ResNet50 utilizada nos experimentos com *fine-tuning*.

Como forma de reduzir a quantidade de parâmetros antes das camadas densas finais, foi utilizado o GlobalAveragePooling2D em substituição ao *Flatten*. Além disso, nos experimentos com ResNet50, foi adotado o congelamento inicial da rede base e, posteriormente, o *fine-tuning* das últimas camadas. Esse procedimento permitiu que o modelo ajustasse parte dos pesos pré-treinados ao domínio das imagens dermatoscópicas.

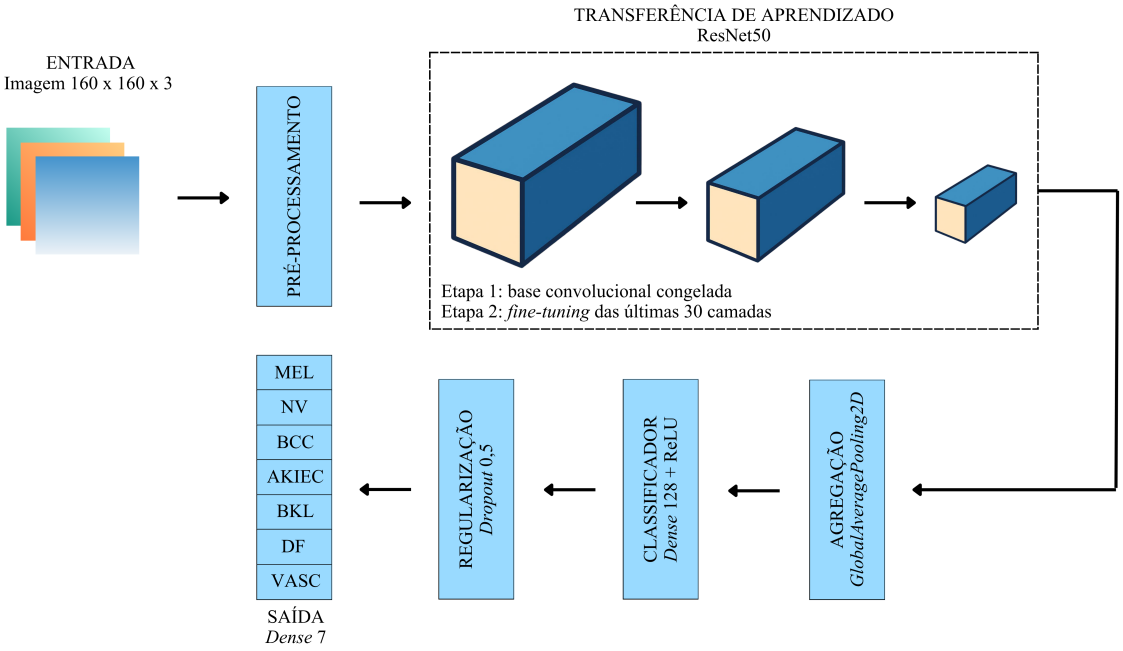


Figura 2: Arquitetura final baseada em ResNet50 utilizada nos experimentos com transferência de aprendizado.

Segmentação e Uso de Máscaras

Além das imagens originais, também foram realizados experimentos com imagens segmentadas e com o uso exclusivo de máscaras das lesões. Nos experimentos com imagens segmentadas, as máscaras foram aplicadas sobre as imagens dermatoscópicas originais. A região correspondente à lesão foi preservada, enquanto o restante da imagem foi removido.

A Figura 3 apresenta um exemplo desse processo, mostrando a imagem original, sua respectiva máscara e a imagem resultante após a segmentação.



Figura 3: Exemplo de imagem original da classe NV, máscara correspondente e imagem segmentada utilizada nos experimentos.

Nos experimentos que utilizaram apenas as máscaras, os modelos receberam como entrada imagens binárias (preto e branco) representando o formato e a área ocupada pela lesão. Nesse caso, informações visuais presentes na imagem original, como variação de cor e textura, não foram consideradas pelo modelo.

Treinamento e Avaliação

Durante o treinamento dos modelos, foram utilizados mecanismos para reduzir o risco de sobreajuste e controlar o processo de aprendizagem. Em todos os experimentos, foi aplicado o *Early Stopping*. Essa técnica monitorou a perda no conjunto de validação, interrompendo o treinamento quando não houvesse melhora por 10 épocas consecutivas.

Além disso, foi habilitada a restauração dos melhores pesos obtidos durante o treinamento, evitando que o modelo final correspondesse a uma época posterior com desempenho inferior de validação. Nos experimentos finais com ResNet50 e *fine-tuning*, também foram utilizados *ReduceLROnPlateau* e *Model Checkpoint*.

O *ReduceLROnPlateau* reduziu a taxa de aprendizado quando a perda de validação permanecia sem melhora.

Os modelos foram treinados utilizando o otimizador Adam (*Adaptive Moment Estimation*) e função perda *Sparse Categorical Crossentropy*. Durante essa etapa, o desempenho foi monitorado a partir da perda e da acurácia nos conjuntos de treino e validação, possibilitando acompanhar a evolução do aprendizado e identificar sinais de sobreajuste.

A avaliação final foi realizada sobre o conjunto de teste, considerando as métricas de acurácia, precisão, *recall* e F1-score. Além das métricas globais, também foram analisadas as médias macro e ponderada, bem como as métricas individuais de cada classe.

A média macro atribui o mesmo peso a todas as classes, enquanto a média ponderada considera a quantidade de amostras de cada classe. A análise individual por classe permite observar o desempenho do modelo em cada categoria diagnóstica, o que é importante em bases desbalanceadas.

Nesse contexto, a acurácia isolada pode favorecer modelos com desempenho nas classes majoritárias, sem representar adequadamente o comportamento nas classes

menos representadas.

Análise Visual com Grad-CAM

Como análise complementar, foi utilizada a técnica Grad-CAM para visualizar as regiões das imagens que mais contribuíram para as predições do modelo baseado em ResNet50 (SELVARAJU et al., 2017).

Essa técnica gera mapas de ativação sobrepostos às imagens originais, permitindo observar se o modelo concentrou sua atenção na região da lesão ou em áreas externas à região de interesse. A Figura 4 exemplifica como esse método foi utilizado.

Neste trabalho, o Grad-CAM foi aplicado ao melhor modelo salvo durante o treinamento, considerando exemplos corretamente classificados e exemplos classificados incorretamente.

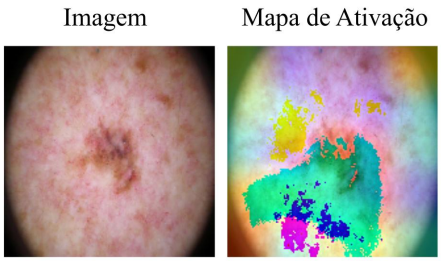


Figura 4: Imagem original da classe MEL e seu respectivo mapa Grad-CAM sobreposto.

RESULTADOS E DISCUSSÃO

Os modelos foram avaliados a partir de acurácia, precisão, *recall* e F1-score, considerando tanto os valores globais quanto as médias macro e ponderadas. Devido ao desbalanceamento do *dataset*, a acurácia isolada não é suficiente para indicar o desempenho em todas as classes. A média macro atribui o mesmo peso a todas as classes, enquanto a média ponderada considera a quantidade de amostras de cada classe, permitindo identificar se o modelo favorece classes majoritárias.

A Tabela 2 apresenta os resultados obtidos por todos os modelos avaliados neste trabalho, considerando diferentes estratégias de balanceamento, segmentação e transferência de aprendizado.

Tabela 2: Resultados dos modelos avaliados

Caso	Modelo	Acurácia	F1 p.	F1 m.
C1	CNN + SMOTE	0,76	0,75	0,56
C2	Seg. + SMOTE	**0,77**	0,74	0,55
C3	Máscaras	0,68	0,56	0,12
C4	ResNet50 + CW + FT	0,73	**0,76**	**0,60**
C5	ResNet50 + SMOTE + FT	0,75	**0,76**	**0,60**

CW: *Class Weight*; FT: *Fine-tuning*
F1 p.: F1 ponderado; F1 m.: F1 macro
Seg.: Imagens Segmentadas

Análise Comparativa das Estratégias

Entre os modelos baseados em CNNs simples, a configuração com SMOTE obteve desempenho competitivo, com acurácia de $0,7634$, F1-score ponderado de $0,7535$ e F1-score macro de $0,5641$. O uso de imagens segmentadas com SMOTE alcançou a maior acurácia ($0,7659$), porém apresentou F1-score macro inferior ($0,5457$), indicando que o desempenho global não se traduziu em equilíbrio entre as classes.

Esse resultado sugere que a remoção do fundo da imagem pode ter facilitado a classificação da classe majoritária, mas não contribuiu significativamente para o reconhecimento das classes minoritárias. A diferença entre F1-score ponderado e macro evidencia que o desempenho continuou mais favorável às classes com maior quantidade de amostras.

O uso exclusivo de máscaras apresentou F1-score macro de apenas $0,1173$, indicando que a forma da lesão, isoladamente, não foi suficiente para representar adequadamente as diferenças entre as sete classes diagnósticas. Informações como cor, textura, bordas e padrões internos da lesão mostraram-se fundamentais para a classificação, sendo parcialmente perdidas quando apenas a máscara é utilizada como entrada.

Transferência de Aprendizado com ResNet50

Os modelos baseados em ResNet50 com *fine-tuning* representam as configurações mais completas avaliadas, combinando transferência de aprendizado, controle do treinamento e ajuste de hiperparâmetros. A configuração com *class weight* obteve acurácia de $0,7292$, F1-score ponderado de $0,7599$ e F1-score macro de $0,6016$. Já a configuração com SMOTE alcançou acurácia de $0,7485$, F1-score ponderado de $0,7621$ e F1-score macro de $0,5991$.

Em problemas de classificação multiclasse desbalanceada, o F1-score macro é a métrica mais adequada para avaliar o desempenho do modelo, pois atribui peso igual a todas as classes, independentemente de sua representação no conjunto de dados. A acurácia e o F1-score ponderado tendem a favorecer classes majoritárias, podendo mascarar o desempenho inferior em classes minoritárias. Portanto, considerando a métrica apropriada para este problema, ResNet50 com *class weight* apresentou o melhor desempenho, com F1-macro de $0,6016$, indicando maior capacidade de generalização equilibrada entre as sete classes diagnósticas.

Em bases desbalanceadas, o maior valor de acurácia nem sempre representa o modelo mais adequado. Um classificador pode obter desempenho global ao acertar majoritariamente a classe mais frequente, mas ainda apresentar desempenho inferior em classes menos representadas. Por isso, as médias macro e ponderadas foram utilizadas para interpretar os resultados obtidos.

A análise das matrizes de confusão reforçou esse comportamento. Conforme apresentado na Figura 5, em diferentes configurações observou-se que a classe NV permaneceu como a categoria com melhor desempenho, enquanto classes com menor quantidade de amostras,

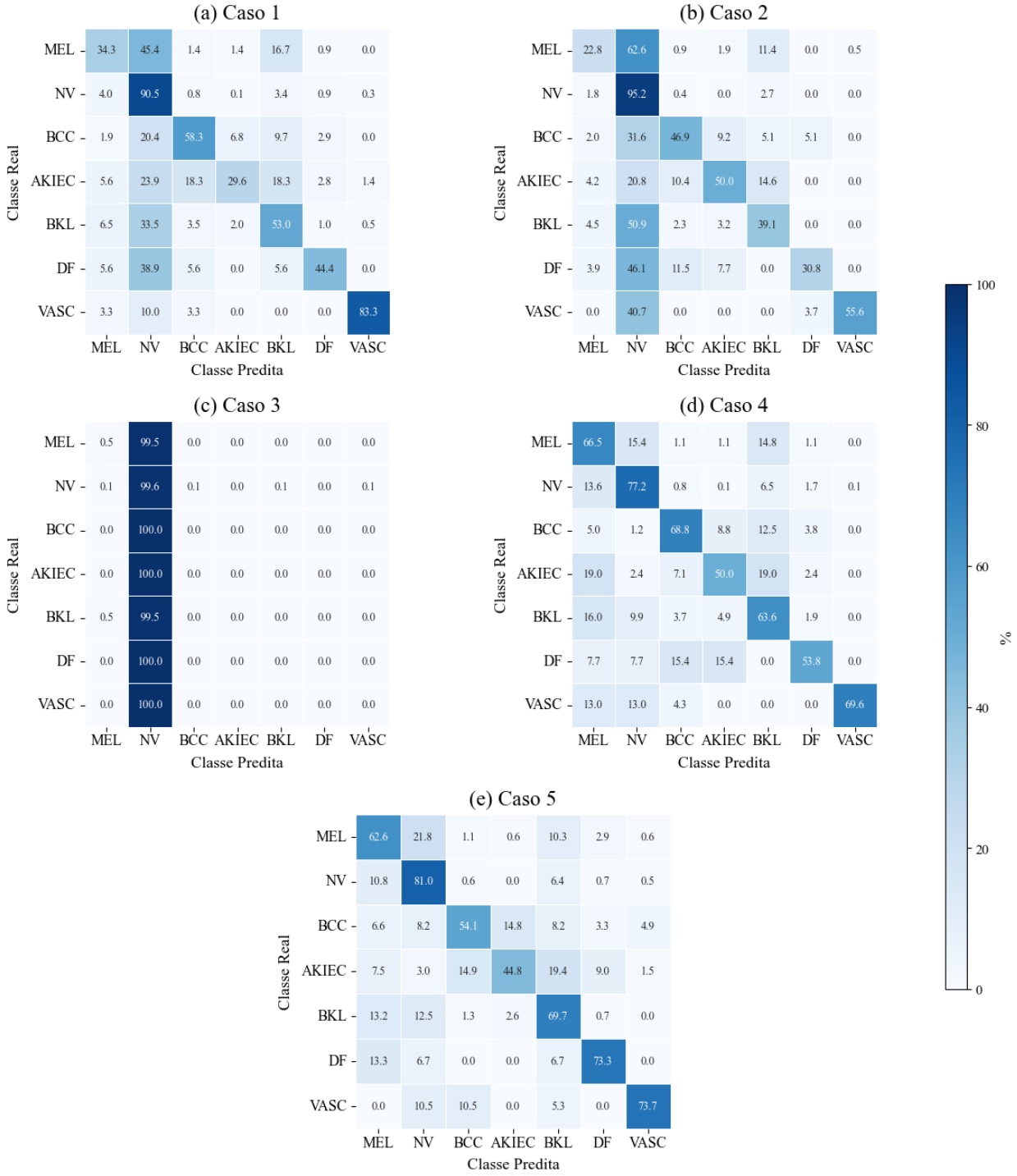


Figura 5: Matrizes de confusão dos modelos avaliados, com valores apresentados em porcentagem. Os casos C1, C2, C3, C4 e C5 correspondem às configurações descritas na Tabela 2

como DF, VASC e AKIEC, apresentaram maior instabilidade e taxas mais elevadas de confusão entre categorias. Esse padrão está diretamente relacionado ao desbalanceamento do conjunto HAM10000 e à elevada semelhança visual entre determinados tipos de lesões cutâneas.

Nos Casos 1 e 2, apesar de existir melhor distribuição das predições entre as categorias, ainda são observadas elevadas taxas de confusão entre classes, especialmente entre MEL, BKL e NV. Já nos Casos 4 e 5, percebe-se uma melhora no equilíbrio geral da classificação, com aumento da taxa de acerto em diferentes categorias e redução parcial do viés para NV, embora ainda persistam dificuldades na distinção entre algumas lesões dermatologicamente similares.

No Caso 3, a matriz de confusão evidencia um comportamento crítico do modelo, caracterizado pela predominância quase absoluta da classe NV nas predições. Observa-se que praticamente todas as amostras, independentemente da classe real, foram classificadas como

NV.

A Figura 6 apresenta exemplos de imagens corretamente e incorretamente classificadas pelo modelo. Observa-se que erros ocorreram entre classes visualmente semelhantes, como MEL, NV e BKL, o que demonstra a dificuldade do problema.

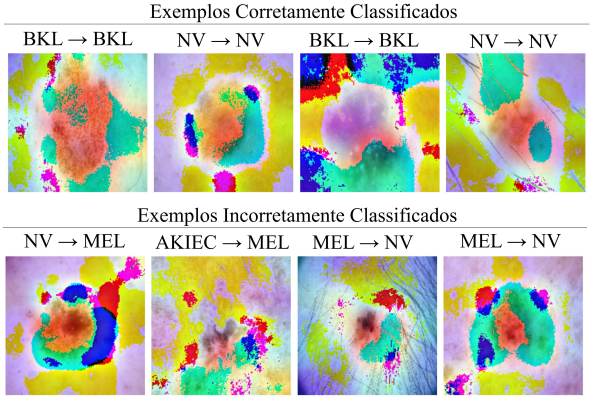


Figura 6: Exemplos de imagens corretamente e incorretamente classificadas pelo modelo baseado em ResNet50. Os rótulos indicam a classe real e a classe predita.

Em comparação com estudos disponíveis na literatura, os resultados obtidos ficaram abaixo de abordagens que utilizam *pipelines* mais elaborados para classificação de lesões de pele. O DODL Net, por exemplo, combina pré-processamento com filtro de mediana adaptativo, segmentação com U-Net, otimização híbrida baseada em BFO e PSO e classificação por *Deep* CNN, alcançando resultados superiores no HAM10000 (GOMATHI et al., 2023).

Outros estudos, como Deep Skin, também exploram balanceamento, segmentação e transferência de aprendizado com arquiteturas como DenseNet169 e ResNet50. Nesse trabalho, os autores observaram que o desempenho variou conforme a combinação entre arquitetura, técnica de balanceamento e divisão entre treino e teste, demonstrando que a escolha da estratégia experimental influencia os resultados obtidos (GURURAJ et al., 2023).

Os resultados deste trabalho indicam que a aplicação de técnicas como SMOTE, segmentação ou transferência de aprendizado não garante melhoria uniforme no desempenho. Algumas estratégias favoreceram a acurácia, enquanto outras apresentaram maior equilíbrio nas métricas macro e ponderadas, evidenciando a necessidade de análise baseada em múltiplas métricas.

O desempenho obtido sugere que estratégias de balanceamento, isoladamente, podem não ser suficientes para melhorar a generalização dos modelos. Etapas de pré-processamento, como remoção de artefatos e preservação de informações visuais, podem exercer papel relevante no desempenho final. Também se destaca a necessidade de investigar o uso de imagens em maiores resoluções, uma vez que detalhes de cor, textura, bordas e padrões internos da lesão podem ser perdidos em imagens redimensionadas para 128×128 ou 160×160 pixels.

Em alguns experimentos, observou-se tendência ao sobreajuste, especialmente quando a acurácia de treinamento passou a se manter superior à acurácia de validação. No entanto, o uso da parada antecipada permitiu interromper o treinamento antes que esse comportamento se intensificasse, restaurando os melhores pesos obtidos durante a validação.

Ainda assim, a diferença entre treino e validação sugere limitações de generalização, possivelmente associadas à complexidade do problema, ao desbalanceamento entre as classes e à variabilidade visual das lesões.

CONCLUSÃO

Este trabalho avaliou estratégias de balanceamento, segmentação e transferência de aprendizado na classificação multiclasse de lesões de pele no HAM10000. Foram testadas abordagens com *class weight*, SMOTE, ROS, imagens segmentadas, máscaras e ResNet50 com *fine-tuning*.

Diferentes estratégias apresentaram desempenho superior em métricas distintas. Imagens segmentadas com SMOTE obtiveram maior acurácia $(0,77)$, enquanto ResNet50 com *class weight* e SMOTE alcançaram melhor F1-ponderado $(0,76)$ e F1-macro $(0,60)$. Considerando que o F1-macro é a métrica mais adequada para problemas de classificação multiclasse desbalanceada, o modelo ResNet50 com *class weight* e *fine-tuning* apresentou o melhor desempenho, demonstrando maior capacidade de generalização equilibrada entre as classes.

O uso exclusivo de máscaras apresentou F1-macro de $0,12$, indicando que características como cor, textura e padrões internos são necessárias para a classificação. A análise com Grad-CAM permitiu visualizar regiões que influenciaram as predições. Os resultados indicam que técnicas de balanceamento e transferência de aprendizado, quando aplicadas isoladamente, apresentam limitações para a classificação multiclasse de lesões de pele. A escolha da métrica de avaliação deve considerar o desbalanceamento do problema.

Como trabalhos futuros, propõe-se investigar pré-processamentos, imagens em maiores resoluções, outras arquiteturas e validação cruzada.

REFERÊNCIAS

CHAWLA, N. V. et al. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, v. 16, p. 321–357, 2002.

FERREIRA, R.; REGO, R. Rede neural perceptron multicamadas para classificação multiclasse de câncer de pele. *Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)*, v. 1, n. 9, 2026.

GOMATHI, E. et al. Skin cancer detection using dual optimization based deep learning network. *Biomedical Signal Processing and Control*, v. 84, p. 104968, 2023.

GOODFELLOW, I. et al. *Deep learning*. [S.l.]: MIT press Cambridge, 2016. v. 1.

GURURAJ, H. L. et al. Deep skin: A deep learning approach for skin cancer classification. *IEEE Access*, 2023.

HE, K. et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: [s.n.], 2016. p. 770–778.

SELVARAJU, R. R. et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice: [s.n.], 2017. p. 618–626.

TSCHANDL, P.; ROSENDAHL, C.; KITTLER, H. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, v. 5, p. 180161, 2018.