

Análise de MLP e CNN em Datasets Balanceados e Não Balanceados

José Cunha¹ e Marta Antunes²

¹ uc2021223719@student.uc.pt

² uc2021216180@student.uc.pt

Resumo Este trabalho investiga o desempenho de Perceptrons Multi-Camadas (MLPs) e Redes Neurais Convolucionais (CNNs) na classificação de imagens dermatológicas, em contextos com dados balanceados e desbalanceados. Os experimentos analisam o impacto do balanceamento, da função de perda e do otimizador no desempenho dos modelos. Observou-se que o desbalanceamento favorece classes majoritárias, enquanto o balanceamento melhora o equilíbrio. A combinação CrossEntropy com RMSprop foi a mais eficaz para CNNs, e CrossEntropy com Adam, para MLPs. No geral, as CNNs superaram os MLPs em desempenho.

Keywords: Perceptrons Multi-Camadas · Redes Neurais Convolucionais · Balanceamento de Dados · Classificação de Imagens · Aprendizado Profundo

1 Introdução

As Redes Neurais Convolucionais (CNNs) e os Perceptrons Multi-Camadas (MLPs) são arquiteturas fundamentais no processamento de imagens, com abordagens distintas. Enquanto as CNNs exploram a estrutura espacial através de convoluções e pooling, capturando características hierárquicas como bordas, texturas e padrões complexos diretamente da disposição dos pixels, os MLPs processam imagens convertidas em vetores unidimensionais, perdendo a informação topológica mas mantendo a capacidade de aprender relações não-lineares entre os dados, servindo assim como base para comparação de desempenho.

Neste projeto, comparamos o desempenho desses modelos na classificação de imagens dermatológicas, analisando o impacto do balanceamento de dados, as funções de perda e os otimizadores. Além da precisão, avaliamos a equidade entre classes, essencial em cenários clínicos com distribuições desbalanceadas. O objetivo do nosso estudo é então esclarecer os benefícios e as restrições de cada arquitetura ao serem aplicadas a tarefas na área médica.

2 Fundamentação Teórica

2.1 Análise do Dataset

O conjunto de dados DermaMNIST apresenta um desbalanceamento significativo, com a classe "melanocytic nevi" dominando a maioria das amostras (aproximadamente 4700 no conjunto de treino). Em contraste, classes como "dermatofibroma" possuem apenas cerca de 100 amostras distribuídas entre os conjuntos de treino, validação e teste.

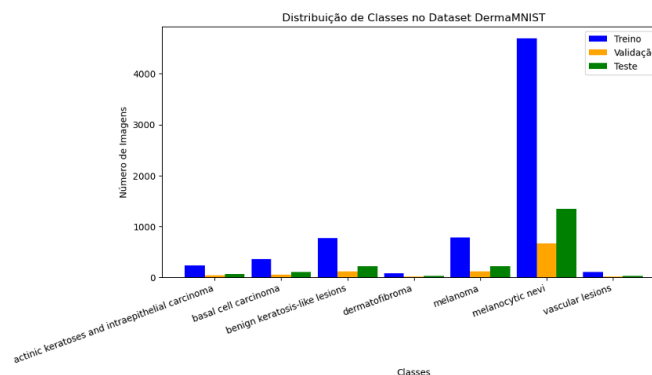


Figura 1. Distribuição das classes no conjunto de dados

Como mostra a Figura 1, esse desequilíbrio pode afetar negativamente o desempenho do modelo, gerando um viés em favor da classe majoritária e comprometendo a capacidade de classificação das classes com menos representatividade.

Em tarefas de classificação, esse tipo de desbalanceamento tende a direcionar as predições para as classes mais frequentes. Para reduzir esse problema, estratégias como o aumento de dados (data augmentation) podem ser empregadas com o objetivo de equilibrar a quantidade de amostras entre as classes.

2.2 Otimizadores e Funções de Perda

- **Otimizadores:**
 - **Adam:** Adapta as taxas de aprendizagem para cada parâmetro usando estimativas dos primeiros e segundos momentos dos gradientes. Combina as vantagens de AdaGrad e RMSProp.
 - **RMSprop:** Mantém uma média móvel do quadrado dos gradientes para normalizar a atualização de pesos, sendo eficaz para problemas com dados esparsos.
- **Funções de perda:**
 - **CrossEntropy:** Ideal para classificação multiclasse, mede a diferença entre distribuições de probabilidade preditas e reais.
 - **MultiMargin:** Versão generalizada da margem máxima, útil quando se deseja enfatizar a separação entre classes.

3 Parte Experimental

3.1 Experimentos sem Balanceamento dos dados de Treino

3.1.1 Número de Epochs Para determinar o número ideal de epochs, realizamos um treinamento inicial com hiperparâmetros estáticos.

No caso do MLP, utilizamos camadas fully connected com ativação ReLU, função de perda CrossEntropyLoss e otimizador Adam (com um learning rate inicial de 0.001 e weight decay de $1e-4$). Usámos um scheduler que reduzia a learning rate a cada 20 épocas. O modelo foi treinado por 150 epochs, e observou-se que a loss e a accuracy de treinamento se estabilizaram por volta da epoch 130, indicando que treinar além disso não traria ganhos significativos.

Para a CNN, adotou-se uma arquitetura com três camadas convolucionais (32, 64 e 128 filtros) com batch normalization, ativação ReLU e max pooling 2x2, seguidas por camadas fully connected com dropout de 0.5. O treino foi feito por 120 epochs com o otimizador Adam (learning rate de 0.001). A estabilização das métricas ocorreu por volta da epoch 100, sendo este o valor que adotamos como padrão para os experimentos seguintes.

3.1.2 Grid Search A grid search realizada teve como objetivo principal encontrar a combinação ideal de hiperparâmetros que maximizasse o desempenho dos modelos MLP e CNN no conjunto DermaMNIST, especialmente considerando o desafio do desbalanceamento de classes. A abordagem sistemática permitiu explorar metodicamente o espaço de parâmetros, mantendo uma estrutura flexível que se adapta às características específicas do dataset.

Tabela 1. Parâmetros testados na Grid Search do MLP

Parâmetro	Valores Testados
Número de neurónios por hidden layer	[512, 256, 128], [256, 128, 64], [512, 256, 128, 64], [512, 512, 256], [64, 64, 64, 64, 64]
Taxa de aprendizagem	[0.001, 0.0001, 0.0005]
Tamanho do batch	[64, 128, 256]
Taxa de dropout	[0.05, 0.1]

A Tabela 1 mostra que o espaço de busca teve 90 combinações possíveis ($5 \times 3 \times 3 \times 2$). Para cada combinação, o modelo foi instanciado com a arquitetura *Sequential*, treinado com **CrossEntropyLoss** e o otimizador **Adam**, e avaliado no conjunto de validação usando o F1-score ponderado como critério principal. Os resultados foram registados num arquivo CSV para rastreamento e comparação.

As características da melhor configuração para o MLP encontrada foram:

- **4 hidden layers ([512, 256, 128, 64]):** Camadas profundas permitem aprender características complexas, equilibrando a complexidade e a capacidade computacional.
- **Taxa de aprendizagem moderada (0.0005):** Proporciona uma convergência estável, evitando atualizações muito bruscas ou lentas.
- **Batch size grande (256):** Reduz a variância nas atualizações dos pesos e melhora a estabilidade do treinamento.
- **Dropout reduzido (0.05):** Regulariza o modelo sem perder muito da capacidade de aprendizado, ajudando a evitar overfitting.

Tabela 2. Parâmetros testados na Grid Search da CNN

Parâmetro	Valores Testados
Número de camadas convolucionais	[2, 3, 4]
Número de filtros iniciais	[32, 64, 128]
Taxa de aprendizagem	[0.005, 0.001, 0.0005]
Tamanho do batch	[64, 128, 256]

O espaço de busca totalizou 81 combinações possíveis ($3 \times 3 \times 3 \times 3$). Para cada combinação de parâmetros, o modelo foi instanciado com a arquitetura *FlexibleCNN* e treinado com o otimizador Adam. A avaliação foi realizada no conjunto de validação utilizando como métricas principais a accuracy e o F1-score ponderado, sendo este último o critério prioritário para seleção. Todos os resultados foram armazenados com um ID único para permitir o rastreamento e comparação entre as diferentes configurações testadas.

As características da melhor configuração da CNN encontrada foram:

- **Batch size grande (256):** Proporcionou maior estabilidade durante o treino, reduzindo a variância nas atualizações dos pesos.
- **4 camadas convolucionais:** Esta profundidade mostrou-se adequada para capturar características hierárquicas relevantes no conjunto de dados.
- **Taxa de aprendizagem moderada (0.0005):** Permitiu ajustes precisos nos pesos da rede, equilibrando velocidade de convergência e estabilidade do treinamento.
- **64 filtros iniciais:** Esta configuração alcançou um bom equilíbrio entre capacidade computacional e poder de representação, extraindo features significativas sem excessiva complexidade.

3.1.3 Desempenho da Melhor Configuração A melhor configuração no conjunto de validação do MLP obteve 76% de accuracy, com F1-score ponderado de 0.75. As classes majoritárias, como a classe 5 (F1=0.88), tiveram bom desempenho, enquanto as classes minoritárias, como a 3 (F1=0.24), apresentaram desempenho mais baixo. A MLP teve uma boa performance geral, mas o modelo ainda enfrenta desafios com classes minoritárias, evidenciados pela baixa performance de algumas categorias.



Figura 2. Matriz de confusão no conjunto de validação do MLP

Classe	Precision	Recall	F1-score
Actinic keratoses	0.45	0.45	0.45
Basal cell carcinoma	0.51	0.60	0.55
Benign keratosis	0.51	0.43	0.47
Dermatofibroma	0.40	0.17	0.24
Melanoma	0.50	0.50	0.50
Melanocytic nevi	0.87	0.90	0.88
Vascular lesions	0.86	0.43	0.57
Accuracy		0.76	
Macro avg	0.59	0.50	0.52
Weighted avg	0.75	0.76	0.75

Figura 3. Relatório de classificação no conjunto de validação do MLP

A melhor configuração no conjunto de validação da CNN obteve 77% de accuracy, com F1-score ponderado de 0.77. As classes majoritárias, como a classe 5 (F1=0.89), tiveram bom desempenho, enquanto as minoritárias, como a classe 3 (F1=0.40), melhoraram moderadamente. A matriz de confusão indicou redução de falsos negativos em algumas classes. O treinamento mostrou estabilidade nas curvas de loss e accuracy, sem sinais de overfitting.

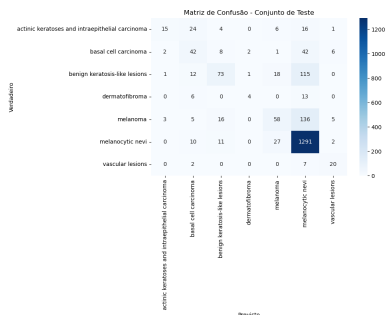


Figura 4. Matriz de confusão no conjunto de validação da CNN

Classe	Precision	Recall	F1-score
Actinic keratosis	0.44	0.42	0.43
Basal cell carcinoma	0.54	0.54	0.54
Benign keratosis	0.57	0.50	0.53
Dermatofibroma	0.58	0.53	0.40
Melanoma	0.51	0.53	0.52
Melanocytic nevi	0.88	0.90	0.89
Vascular lesions	0.82	0.64	0.72
Accuracy		0.77	
Macro avg	0.61	0.55	0.58
Weighted avg	0.77	0.77	0.77

Figura 5. Relatório de classificação no conjunto de validação da CNN

3.1.4 Testes com diferentes Funções de Perda e Otimizadores As combinações testadas consistem na combinação das funções de perda e otimizadores descritas na secção de fundamentação teórica.

Tabela 3. Resultados das combinações testadas (F1-score)

Combinação	MLP	CNN
CrossEntropy + Adam	0.7314	0.7536
CrossEntropy + RMSprop	0.7131	0.7546
MultiMargin + Adam	0.7225	0.7365
MultiMargin + RMSprop	0.7055	0.7278

A combinação que apresentou o melhor desempenho para o MLP foi a utilização de **CrossEntropy + Adam**, alcançando um F1-score de 0.7314, enquanto que para a CNN a combinação **CrossEntropy + RMSprop** apresentou o melhor desempenho geral (F1=0.7546).

3.1.5 Avaliação no Conjunto de Teste Submetemos o modelo com a melhor combinação de parâmetros da grid search e a melhor configuração de treino aos dados de teste não vistos durante o desenvolvimento.



Figura 6. Matriz de confusão no conjunto de teste do MLP

Classe	Precision	Recall	F1-score
Actinic keratosis	0.40	0.41	0.40
Basal cell carcinoma	0.52	0.52	0.52
Benign keratosis	0.48	0.41	0.45
Dermatofibroma	0.43	0.13	0.20
Melanoma	0.38	0.32	0.35
Melanocytic nevi	0.84	0.89	0.86
Vascular lesions	0.62	0.45	0.52
Accuracy		0.73	
Macro avg	0.52	0.44	0.47
Weighted avg	0.71	0.73	0.71

Figura 7. Relatório de classificação no conjunto de teste do MLP

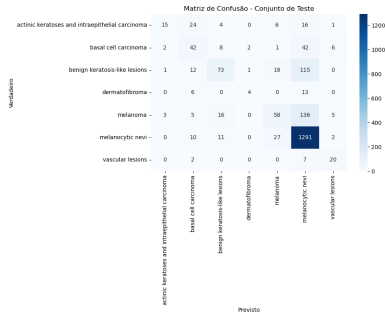


Figura 8. Matriz de confusão no conjunto de teste da CNN

Classe	Precision	Recall	F1-score
Actinic keratoses	0.71	0.23	0.34
Basal cell carcinoma	0.42	0.41	0.41
Benign keratosis	0.65	0.33	0.44
Dermatofibroma	0.57	0.17	0.27
Melanoma	0.53	0.26	0.35
Melanocytic nevi	0.80	0.96	0.87
Vascular lesions	0.59	0.69	0.63
Accuracy		0.75	
Macro avg	0.61	0.44	0.47
Weighted avg	0.72	0.75	0.71

Figura 9. Relatório de classificação no conjunto de teste da CNN

- **Accuracy:** Em ambos os modelos, os valores obtidos foram coerentes com os da validação, indicando um baixo nível de overfitting.
- **F1-score médio (weighted):** 0.71 (impacto do desbalanceamento nas classes minoritárias).

Problemas Identificados:

- **Viés para Classes Majoritárias:** O modelo prioriza "melanocytic nevi", negligenciando classes críticas como melanoma e actinic keratoses.
- **Confusão entre Lesões Benignas e Malignas:** Algumas lesões malignas foram classificadas como benignas, o que representa um erro crítico em aplicações clínicas.
- **Baixa Sensibilidade em Classes Raras:** Classes com menos de 100 amostras (ex.: dermatofibroma) tiveram recall abaixo de 20%.

3.1.6 Conclusões Nenhum dos modelos é confiável para detecção de cancro de pele (melanoma) e lesões pré-cancerosas (actinic keratoses) devido à alta taxa de falsos negativos.

3.2 Experimentos com Balanceamento dos dados de Treino

O dataset *DermaMNIST* foi balanceado para garantir uma distribuição equilibrada entre as classes, utilizando as seguintes estratégias:

- **Redução de Amostras:** Para classes com mais de **1000 amostras**, foi aplicada uma subamostragem aleatória, mantendo apenas 1000 exemplos por classe.
- **Aumento de Dados (Data Augmentation):** Para classes com menos de 1000 amostras, foram geradas novas imagens sintéticas através de transformações aleatórias, incluindo:
 - Rotação (até 15 graus)
 - Espelhamento horizontal (probabilidade de 50%)
 - Translação (até 10% da imagem)
 - Variação de brilho e contraste (20%)
 - Desfoque gaussiano (kernel de 3x3)
- **Distribuição Final:** Todas as classes foram balanceadas para conter exatamente **1000 amostras** cada, resultando em um dataset homogêneo para treinamento.

3.2.1 Número de epochs Assim como no dataset desbalanceado, o número de épocas foi definido como 130 para o MLP e 100 para a CNN, após uma análise cuidadosa da estabilização da train loss, utilizando a mesma metodologia adotada anteriormente. O monitoramento contínuo da convergência do modelo garantiu que o treino fosse realizado até que a perda se estabilizasse num nível mínimo, sem sinais de overfitting significativo.

3.2.2 Grid Search Semelhante ao que foi feito para o dataset desbalanceado, realizamos uma grid search com o objetivo de encontrar a combinação ideal de hiperparâmetros para maximizar o desempenho do MLP e da CNN. Utilizamos as mesmas arquiteturas definidas anteriormente, as mesmas combinações de teste e adotamos igualmente o F1-score ponderado como métrica de decisão.

As características da melhor configuração encontrada para o MLP foram:

- **4 hidden layers ([512, 256, 128, 64]):** Camadas profundas permitem aprender características complexas, equilibrando a complexidade e a capacidade computacional.
- **Taxa de aprendizagem moderada (0.0005):** Proporciona uma convergência estável, evitando atualizações muito bruscas ou lentas.
- **Batch size médio (128):** Um tamanho de batch médio permite um bom equilíbrio entre a utilização de memória e a estabilidade na atualização dos pesos, contribuindo para uma aprendizagem mais eficiente.
- **Dropout moderado (0.1):** O dropout ajuda a prevenir overfitting, desativando aleatoriamente 10% dos neurônios durante o treinamento, o que melhora a generalização do modelo.

As características da melhor configuração para a CNN encontrada foram as seguintes:

- **Batch size grande (256):** Proporcionou maior estabilidade durante o treino, reduzindo a variância nas atualizações dos pesos.
- **3 camadas convolucionais:** Esta profundidade mostrou-se adequada para capturar características hierárquicas relevantes no conjunto de dados.
- **Taxa de aprendizagem moderada (0.0005):** Permitiu ajustes precisos nos pesos da rede, equilibrando velocidade de convergência e estabilidade do treinamento.
- **64 filtros iniciais:** Esta configuração alcançou um bom equilíbrio entre capacidade computacional e poder de representação, extraindo features significativas sem excessiva complexidade.

3.2.3 Desempenho da Melhor Configuração O modelo MLP apresentou desempenho desigual após o balanceamento dos dados, com uma accuracy geral de 65% (inicialmente, com o dataset desbalanceado encontrava-se a 76%). Apesar do ajuste no número de amostras por classe, algumas categorias continuam a apresentar dificuldades significativas, enquanto outras mostraram evolução positiva.



Figura 10. Matriz de confusão no conjunto de validação do MLP

Classe	Precision	Recall	F1-score
Actinic keratoses	0.37	0.33	0.35
Basal cell carcinoma	0.46	0.63	0.53
Benign keratosis	0.37	0.51	0.43
Dermatofibroma	0.62	0.42	0.50
Melanoma	0.33	0.64	0.44
Melanocytic nevi	0.93	0.70	0.80
Vascular lesions	0.47	0.57	0.52
Accuracy		0.65	
Macro avg	0.51	0.54	0.51
Weighted avg	0.75	0.65	0.68

Figura 11. Relatório de classificação no conjunto de validação do MLP

Problemas Graves:

- **Classe 0** mantém um desempenho fraco, o que indica que o modelo ainda falha a detetar muitos casos.
- **Classe 4** apresenta uma precisão muito baixa (33%), o que sugere muitos falsos positivos.

Pontos Positivos:

- **Classe 1** ($F1 = 0.53$) apresenta um desempenho aceitável.
- Accuracy não se manteve estável após o balanceamento, tendo até diminuindo.

O modelo CNN apresenta desempenho desigual após balanceamento, com accuracy de 71%. A classe majoritária (5) domina com $F1 = 0.85$, enquanto classes minoritárias sofrem.

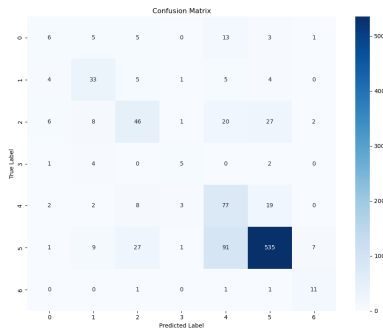


Figura 12. Matriz de confusão no conjunto de validação da CNN

Classe	Precision	Recall	F1-score
Actinic keratoses	0.30	0.18	0.23
Basal cell carcinoma	0.54	0.63	0.58
Benign keratosis	0.50	0.42	0.46
Dermatofibroma	0.45	0.42	0.43
Melanoma	0.37	0.69	0.48
Melanocytic nevi	0.91	0.80	0.85
Vascular lesions	0.52	0.79	0.63
Accuracy		0.71	
Macro avg	0.51	0.55	0.52
Weighted avg	0.75	0.71	0.72

Figura 13. Relatório de classificação no conjunto de validação da CNN

Problemas Graves:

- **Classe 0** tem recall catastrófico (18%), indicando que 82% dos casos não são detectados.
- **Classe 4** apresenta uma precisão muito baixa (37%), gerando muitos falsos positivos.

Pontos Positivos:

- **Classe 1** ($F1 = 0.58$) e **Classe 6** ($F1 = 0.63$) apresentam desempenho aceitável.
- Accuracy mantém-se estável apesar do balanceamento.

O balanceamento aplicado não foi suficiente para resolver problemas críticos nas classes raras (0 e 3). A classe Actinic keratoses, em especial, manteve um desempenho insatisfatório, permanecendo inadequada para uso em contextos clínicos.

3.2.4 Testes com diferentes Funções de Perda e Otimizadores Semelhante ao que foi feito anteriormente, foram testadas as 4 combinações de funções de perda e otimizadores. Podemos observar estas combinações na tabela abaixo, acompanhado dos respectivos resultados.

Tabela 4. Resultados das combinações testadas (F1-score)

Combinação	MLP	CNN
CrossEntropy + Adam	0.6814	0.6968
CrossEntropy + RMSprop	0.6730	0.6881
MultiMargin + Adam	0.6623	0.6756
MultiMargin + RMSprop	0.6640	0.6879

Em ambos os modelos a combinação **CrossEntropy + Adam** obteve o melhor desempenho geral. A diferença entre CrossEntropy + Adam e CrossEntropy + RMSprop foi marginal, o que sugere que o otimizador teve menos impacto que a função de perda. O baixo desempenho das combinações MultiMargin indica que esta função não foi adequada para este problema específico.

3.2.5 Avaliação no Conjunto de Teste A MLP final utilizou a função de perda CrossEntropy, otimizador Adam ($lr=0.0005$), batch size de 128 e quatro camadas ocultas ([512, 256, 128, 64]) com dropout de 0.1, alcançando cerca de 68% no F1-score.

Apesar do treino com dados balanceados, o desempenho no teste ainda varia entre classes. A classe 5 (melanocytic nevi) domina os acertos ($F1=0.79$), contribuindo significativamente para a accuracy global (63%). Classes minoritárias, como “dermatofibroma” ($F1=0.24$), continuam com desempenho fraco.

A média macro de F1 (0.44) revela que o modelo ainda tem dificuldade em generalizar para classes menos representadas, indicando que estratégias adicionais podem ser necessárias.



Figura 14. Matriz de confusão no conjunto de teste do MLP

Classe	Precision	Recall	F1-score
Actinic keratoses	0.39	0.47	0.43
Basal cell carcinoma	0.43	0.56	0.49
Benign keratosis	0.36	0.49	0.42
Dermatofibroma	0.21	0.26	0.24
Melanoma	0.30	0.57	0.40
Melanocytic nevi	0.92	0.69	0.79
Vascular lesions	0.27	0.45	0.33
Accuracy		0.63	
Macro avg	0.41	0.50	0.44
Weighted avg	0.73	0.63	0.66

Figura 15. Relatório de classificação no conjunto de teste do MLP

A configuração final selecionada para a CNN após a otimização incluiu a função de perda CrossEntropy, o otimizador Adam com taxa de aprendizagem de 0.0005, tamanho do batch de 256, arquitetura de 3 camadas convolucionais e 64 filtros iniciais. Esta combinação demonstrou o melhor equilíbrio entre capacidade de generalização e eficiência computacional durante os experimentos.

O modelo, treinado com dados balanceados (1000 amostras por classe no treino), ainda apresenta disparidades significativas no conjunto de validação/teste (não balanceado). A classe 5 mantém bom desempenho (491 acertos, $F1=0.82$), mas classes minoritárias como 3 (0 acertos) e 6 ($F1=0.58$) continuam com dificuldades, indicando que o balanceamento não resolveu totalmente o problema.

O alto precision da classe 3 (0.86) com recall zero sugere que o modelo quase nunca a prevê, mas quando o faz, acerta. Já a classe 4 tem muitos falsos positivos (105), possivelmente por confusão com classes similares (ex: 2). A accuracy global (68%) é inflada pela classe 5 (majoritária), enquanto a média macro ($F1=0.56$) revela a fragilidade em classes menores.

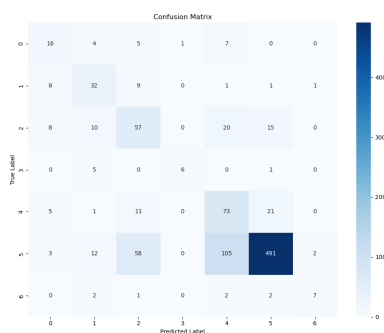


Figura 16. Matriz de confusão no conjunto de teste da CNN

Classe	Precision	Recall	F1-score
Actinic keratoses	0.40	0.48	0.44
Basal cell carcinoma	0.48	0.62	0.54
Benign keratosis	0.40	0.52	0.45
Dermatofibroma	0.86	0.50	0.63
Melanoma	0.35	0.66	0.46
Melanocytic nevi	0.92	0.73	0.82
Vascular lesions	0.70	0.50	0.58
Accuracy		0.68	
Macro avg	0.59	0.57	0.56
Weighted avg	0.76	0.68	0.71

Figura 17. Relatório de classificação no conjunto de teste da CNN

Embora o balanceamento artificial tenha equalizado a distribuição das classes durante o treino, a avaliação em dados reais (desbalanceados) revelou que o modelo mantém um viés significativo face a classes majoritárias. Isso sugere que estratégias como oversampling podem ser insuficientes para cenários clínicos, onde a detecção confiável de classes raras é crítica. A dificuldade em gerar imagens sintéticas realistas limita a eficácia do data augmentation, pois as variações criadas podem não capturar adequadamente a complexidade e diversidade dos cenários reais.

3.3 Outras Tentativas de Aprimoramento da CNN

Na busca por melhorar o desempenho do modelo, foram realizadas várias tentativas adicionais de otimização:

- **Variação no Balanceamento de Dados:** Testamos diferentes quantidades de amostras por classe (500, 1500 e 2000) durante o treinamento, mantendo o balanceamento artificial. No entanto, nenhuma dessas variações trouxe melhorias significativas nos resultados. Isso reforça que o simples aumento artificial de amostras não resolve o problema de representatividade quando as características intrínsecas das classes não são suficientemente distintas.
- **Aumento de Neurônios e Profundidade:** Adicionamos mais camadas e neurônios à arquitetura da CNN, visando capturar features mais complexas. Contudo, essa abordagem não melhorou o desempenho.

Estes resultados destacam dois desafios fundamentais:

1. A dificuldade inerente de trabalhar com datasets desbalanceados, onde o balanceamento artificial não garante melhoria automática no desempenho.
2. As limitações práticas na síntese de imagens, onde a criação de variações realistas e discriminativas se mostra complexa, especialmente para classes com poucos exemplos originais.

4 Conclusões e Trabalhos Futuros

Este estudo comparou o desempenho de MLPs e CNNs na classificação de imagens dermatológicas sob cenários balanceados e desbalanceados. Observou-se que o desbalanceamento favorece classes majoritárias, enquanto técnicas como subamostragem e aumento de dados ajudaram a mitigar esse viés — embora as classes minoritárias ainda apresentem desempenho inferior.

No geral, a combinação **CrossEntropy** + **Adam** foi a mais eficaz em ambos os classificadores. As CNNs superaram os MLPs, mostrando maior capacidade de extrair padrões espaciais relevantes.

Ainda assim, a classificação de classes raras permaneceu um desafio, indicando limitações do balanceamento artificial.

Trabalhos futuros incluem:

- Uso de GANs para aumento de dados realista;
- Teste de arquiteturas avançadas (e.g., ResNet, EfficientNet);
- Aplicação de aprendizado por transferência;
- Inclusão de métricas mais robustas (e.g., AUC-ROC);
- Análise de viés e justiça nas predições.

O trabalho reforça a importância de abordagens mais sofisticadas para lidar com desbalanceamento em contextos clínicos sensíveis.

Nota: Durante o desenvolvimento deste trabalho, foram utilizadas ferramentas baseadas em Modelos de Linguagem de Grande Escala (LLMs) tanto na redação do texto quanto na elaboração e depuração do código, com o objetivo de melhorar a clareza, eficiência e qualidade geral da implementação.