Estudo de Modelos baseados em Redes Neurais Profundas para a Classificação de Tumores Melanocíticos Conjuntivais

Rafael B. dos Santos[1], **Matheus G. Pires**[1], **Fabiana C. Bertoni**[1]

[1]Departamento de Ciências Exatas – Universidade Estadual de Feira de Santana (UEFS)
Av. Transnordestina, s/n - Novo Horizonte – CEP 44036-900
Feira de Santana – BA – Brazil

rafa.brito295@gmail.com, {mgpires,fcbertoni}@uefs.br

Abstract. *Conjunctival melanoma is a malignant neoplasm which generally presents as a pigmented nodular conjunctival lesion. Variant cases with several atypical shapes can delay the diagnosis. To assist the doctor in early diagnosis, minimizing risks to the patient, this work conducted a comparative study of algorithms to classify conjunctival melanoma. For this purpose, models based on Convolutional Neural Networks were evaluated in binary and multiclass classification of tumors, based on VGG16, Xception and MobileNetV2 models, using the Transfer Learning technique to improve generalization. For final image classification, an approach based on an assembly of classifiers was performed, consisting of the PMC, SVM and KNN algorithms. The study used a dataset with 406 images, applying data balancing techniques, such as SMOTE and ADASYN. To find the best classification model, was used 5-folds cross validation technique. Considering all the tests carried out, Ensemble MobileNetV2 models was the one that obtained the best results.*

Resumo. *O melanoma conjuntival é uma neoplasia maligna, que geralmente se apresenta como uma lesão nodular pigmentada. Casos variantes com diversas formas atípicas podem atrasar a identificação. Com o intuito de auxiliar o médico no diagnóstico precoce, minimizando os riscos ao paciente, este trabalho tem como objetivo realizar um estudo comparativo de algoritmos para classificar os tumores melanocíticos conjuntivais. Para isso, foram avaliados modelos baseados em Redes Neurais Convolucionais de classificação binária e multiclasse dos tumores, a partir dos modelos VGG16, Xception e MobileNetV2, utilizando a técnica Transfer Learning para melhorar a generalização dos modelos. Para classificação final da imagem, foi realizada uma abordagem baseada em assembleia de classificadores, composta pelos algoritmos PMC, SVM e KNN. O estudo utilizou uma base de dados com 406 imagens, aplicando técnicas de balanceamento de dados, como SMOTE e ADASYN. Para encontrar o modelo de classificação com melhor desempenho, foi usada a abordagem de validação cruzada 5-folds. Considerando todos os testes realizados, o modelo Ensemble MobileNetV2 foi o que obteve os melhores resultados.*

1. Introdução

Os tumores melanocíticos conjuntivais, apesar de raros, representam a segunda lesão maligna conjuntival mais frequente. Devido a sua baixa incidência, oftalmologistas geralmente não estão familiarizados e preparados para identificar e conduzir os casos suspeitos.

Esse tipo de tumor representa um desafio para o oftalmologista, pois pode apresentar diversas formas e originar-se de lesões benignas como os nevos conjuntivais, ou até mesmo de lesões não pigmentadas como em alguns casos de melanose adquirida primária (MAP) [Novais and Karp 2012].

Para diagnosticar o tumor e determinar o risco ou não de malignidade, os pacientes são avaliados pelo médico oftalmologista por meio de um exame clínico típico de lâmpada de fenda para observação da superfície ocular. Quando o diagnóstico se torna inconclusivo, devido às similaridades presentes nas lesões, é necessário coletar algumas amostras para realização de biópsia [Jain et al. 2021]. O conhecimento das características do melanoma de conjuntiva, de suas lesões precursoras e dos sinais de transformação maligna são determinantes na identificação e na intervenção cirúrgica precoce, possibilitando a redução das taxas de recorrência, metástases e mortalidade.

Uma vasta área da computação, conhecida como *Machine Learning*, está focada no desenvolvimento de sistemas de apoio à decisão, que podem ser aplicados na análise de imagens médicas. Esses sistemas, chamados de CAD (Diagnóstico Assistido por Computador), ganharam popularidade como uma ferramenta útil para apoiar decisões clínicas para várias doenças, em especial a classificação de tumores oculares, com base em imagens de ressonância magnética ou tomografia [Allam et al. 2024].

Dada a dificuldade no diagnóstico dos tumores melanocíticos conjuntivais e a relevância em obtê-lo com brevidade, e considerando que foram encontrados poucos trabalhos na literatura de diagnóstico assistido por computador para este fim, este trabalho propõe o desenvolvimento de um modelo computacional para a classificação de tumores melanocíticos conjuntivais usando imagens da superfície ocular. Mais precisamente, modelos baseados em redes neurais profundas serão comparados na tarefa de classificação binária e multiclasse. Dentre essas redes, as Convolucionais (CNN) apresentam excelentes resultados na identificação de imagens médicas [Nazir et al. 2023], por isso, é motivador aplicar esse tipo de rede na classificação de imagens de tumores conjuntivais.

Este trabalho está organizado da seguinte forma: na Seção 2, é realizado um levantamento de trabalhos diretamente relacionados ao problema abordado; na Seção 3, a metodologia aplicada no diagnóstico de melanoma conjuntival é descrita; na Seção 4 são descritos os experimentos realizados e os resultados obtidos; e por fim, na Seção 5 são apresentadas as conclusões.

2. Trabalhos relacionados

[Yoo et al. 2021] desenvolveram um modelo de *deep learning low-shot* para detectar melanoma conjuntival em imagens da superfície ocular. O aprendizado *low-shot* usa poucas amostras anotadas para treinamento do modelo. Várias CNN como a GoogleNet, a ResNet50 e a MobileNetV2 foram treinadas e testadas, sendo que a MobileNetV2 teve o melhor desempenho. O conjunto de treinamento foi aumentado usando Redes Adversárias Generativas (GAN), o que melhorou o desempenho de todos os modelos.

O trabalho desenvolvido por [Santos-Bustos et al. 2022], fornece uma abordagem exploratória utilizando redes neurais convolucionais para detectar anormalidades oculares, com um caso ilustrativo de melanoma uveal (MU), um tipo de câncer ocular. Em estudos anteriores foram empregadas diferentes técnicas computacionais com foco em

características discriminativas, usando sistemas fuzzy, redes neurais e sistemas neuro-fuzzy adaptativos. No entanto, dada a natureza hereditária do problema, decidiu-se por usar CNN com *Transfer Learning*, como uma alternativa promissora para melhorar a precisão dos resultados, os quais superaram os principais trabalhos de mesma finalidade, disponíveis na literatura.

Em [Li et al. 2022], foi desenvolvida uma rede neural convolucional baseada em região mais rápida (*Faster R-CNN*), associada à técnica de *Transfer Learning*, para localizar automaticamente tumores palpebrais e distingui-los entre malignos e benignos, em imagens fotográficas capturadas por câmeras digitais comuns. Sensibilidade, especificidade, precisão e AUC foram calculadas para avaliar o desempenho da *Faster R-CNN*. Comparando esses valores com análises feitas por oftalmologistas com diferentes níveis de experiência, não se observou diferença estatística entre os diagnósticos.

O trabalho de [Nazir et al. 2023] fornece uma revisão sistemática sobre o campo da Inteligência Artificial Explicável para diagnóstico de imagens biomédicas, discutindo os desafios e apontando trabalhos futuros. Os autores concluem que as técnicas de *deep learning*, associadas à Inteligência Artificial Explicável, constituem-se em técnicas bastante promissoras para auxiliar os especialistas no diagnóstico por imagem. Os autores destacam a baixa quantidade de repositórios de dados públicos disponíveis para treinamento e teste de modelos computacionais, o que dificulta a comparação entre resultados.

Levando em consideração o grande potencial no uso de aprendizado de máquina para diagnóstico, prognóstico e tratamento de várias condições médicas em oftalmologia, o artigo de revisão publicado por [Chandrabhatla et al. 2023], sintetiza o estado da arte da aplicação dessas técnicas em oncologia ocular. Foram avaliadas 804 publicações, coletando métricas sobre o desempenho dos algoritmos de aprendizado de máquina. A pesquisa apontou que as CNN são um dos algoritmos mais comumente usados e que a maioria dos trabalhos se concentrou no melanoma uveal e retinoblastoma. A maioria dos modelos foram desenvolvidos para diagnóstico e prognóstico. Algoritmos para diagnóstico utilizaram principalmente imagens, enquanto aqueles para prognóstico aproveitaram combinações da expressão genética, características do tumor e dados demográficos dos pacientes. As conclusões indicam que as CNN têm grande potencial na identificação de cânceres intraoculares, mas ocasionalmente eram limitadas pela pequena quantidade de dados disponíveis para treinamento, validação e teste.

O trabalho desenvolvido em [Koseoglu et al. 2023], forneceu uma atualização sobre a aplicação das técnicas de *deep learning* e aprendizado de máquina clássico para a detecção e prognóstico de malignidades intraoculares e da superfície ocular. O estudo mostrou que a maioria dos trabalhos se dedica a detecção e classificação de tumores intraoculares, como o melanoma uveal. Grande parte dos artigos utilizam técnicas de *deep learning* associadas a técnicas clássicas de aprendizado de máquina, buscando melhorar o desempenho dos modelos, dada a relativa escassez de casos oncológicos oculares.

Por fim, um trabalho mais recente desenvolvido por [Allam et al. 2024], propôs um sistema CAD para detectar várias formas de tumores orbitais, a partir de imagens de ressonância magnética, utilizando CNN. Pré-processamento e aumento de dados são aplicados às imagens para melhorar o desempenho do sistema. Os resultados mostraram que o sistema é capaz de detectar e classificar o tumor em cada tipo de imagem, reafirmando

a eficiência das CNN na identificação de tumores oculares.

Considerando o levantamento bibliográfico apresentado, este trabalho aplicou três modelos de CNN no diagnóstico de melanoma conjuntival, uma vez que a maioria dos trabalhos apontam que redes neurais profundas são promissoras na identificação de cânceres oculares, e que poucos trabalhos se dedicaram à identificação de tumores da superfície ocular. Na seção 3 descrevemos as etapas de desenvolvimento desta proposta.

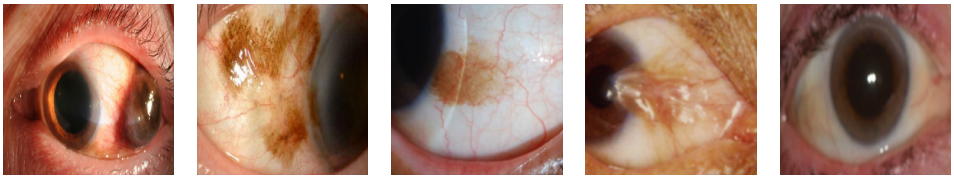
3. Metodologia

Os modelos de redes neurais foram implementados em *Python* 3.10 utilizando a biblioteca *Tensorflow*, versão 2.12.0 por meio da API (*Aplication Programming Interface*) *Keras*, *Scikit learn* versão 1.2.2. Para o processamento das imagens optou-se por utilizar a biblioteca OpenCV versão 4.5.4. Os modelos foram validados e testados utilizando duas GPUs (*Graphics Processing Units*) Nvidia Tesla T4 (versão do *driver* vídeo 470.161.03) de 15 GB (*Gigabyte*). O ambiente escolhido para a execução de todos os códigos e para o armazenamento de todos os dados foi a plataforma *Kaggle*.

3.1. Conjunto de dados

Os dados utilizados neste trabalho foram extraídos do estudo de [Yoo et al. 2021], que coletou as imagens da superfície ocular por meio do buscador de imagens do Google utilizando a estratégia de busca por imagens baseada na utilização das seguintes palavras-chave: "*conjunctiva*", "*pterygium*", "*conjunctival nevus*", "*conjunctival melanosis*", "*conjunctival melanoma*" e "*conjunctival malignant melanoma*". Os dados originais são compostos por imagens clínicas da superfície ocular de pacientes sem distúrbios na conjuntiva e com distúrbios na conjuntiva, os quais estão disponíveis publicamente. As imagens foram classificadas manualmente por dois oftalmologistas certificados, que de forma consensual selecionaram as imagens de modo a evitar ambiguidades e de estabelecer o domínio delas. Os dados foram disponibilizados por [Yoo et al. 2021] em um repositório público, *Mendeley Data*, disponível em `http://dx.doi.org/10.17632/t75wjsw6bw`.

O conjunto de dados original, obtido do repositório, possui 139 imagens da classe melanoma conjuntival, 10 imagens da classe MPA ou melanose, 86 imagens da classe nevo de conjuntiva, 75 imagens da classe pterígio e 96 imagens da classe conjuntiva normal, totalizando 406 imagens RGBs com resoluções variadas (maior = 1029 x 666, menor = 89 x 47, média = 310 x 210), no formato *Portable Network Graphics* (PNG). Na Figura 1 são ilustrados alguns exemplares das imagens que pertencem a este conjunto de dados.



(a) melanoma (b) mpa (c) nevo (d) pterígio (e) normal

Figura 1. Exemplares das classes pertencentes ao conjunto de dados.

Baseado no trabalho de [Yoo et al. 2021], adotamos a mesma estratégia utilizada por eles, a qual uniu as imagens das classes nevo de conjuntiva e MPA, formando uma

única classe. Vale destacar que ambas são consideradas lesões conjuntivais pigmentadas benignas. Sendo assim, o conjunto de dados ficou dividido em quatro classes: melanoma conjuntival (139 imagens), nevo & MPA (96), pterígio (75 imagens) e conjuntiva normal (96 imagens). Ainda de acordo com [Yoo et al. 2021], neste trabalho também realizamos a classificação dos tumores melanocíticos conjuntivais sob a perspectiva da classificação binária e multiclasse, ou seja, o conjunto de dados para a classificação binária foi dividido em duas classes, as lesões benignas e melanoma. Por outro lado, a classificação multiclasse considerou as quatro classes (melanoma, nevo & MPA, pterígio e normal). Para o aprendizado das redes neurais, em ambos os casos, os dados foram divididos em conjuntos de treinamento (70%), validação (10%) e teste (20%), utilizando validação cruzada de 5-*folds*.

3.2. Balanceamento do conjunto de dados

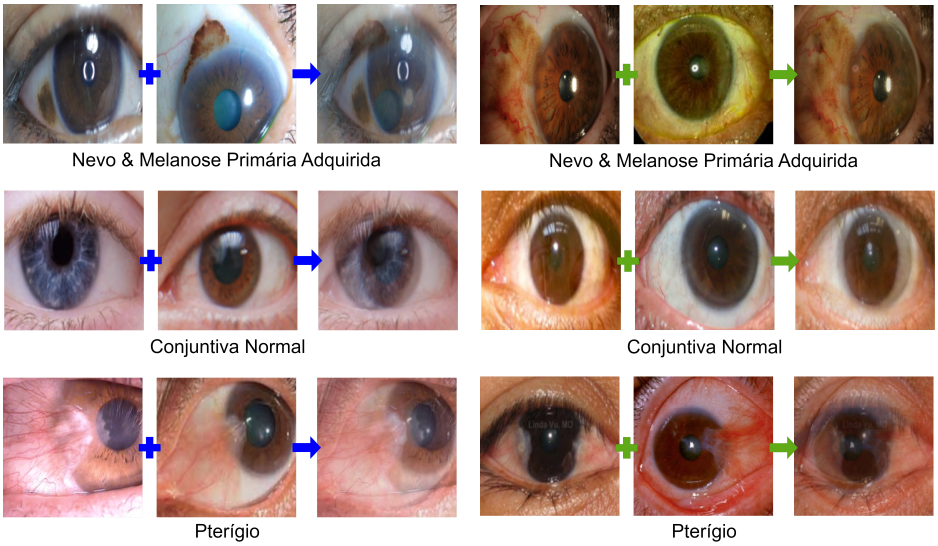
Com o propósito de evitar o enviesamento do aprendizado das redes neurais pela classe majoritária do conjunto de dados, aplicamos a técnica de *oversampling* para equilibrar os dados das classes minoritárias, no entanto, este balanceamento foi realizado somente no conjunto de treinamento. Uma particularidade observada na base de dados é que a maior quantidade de amostras pertence à classe de tumor maligno, e a menor quantidade de amostras pertence à classe de tumores benignos. Na maioria das situações, o que normalmente verifica-se em muitas bases de dados é o oposto, como em [Santos-Bustos et al. 2022].

O primeiro método a ser usado foi o SMOTE [Chawla et al. 2002], que realiza o equilíbrio dos dados por meio da superamostragem das classes consideradas minoritárias, ou seja, novas imagens sintéticas das classes minoritárias foram geradas para que houvesse o balanceamento de dados entre cada classe minoritária em relação a classe majoritária. O segundo método de *oversampling* aplicado foi o ADASYN [He et al. 2008], que assim como o SMOTE, realizou o equilíbrio do conjunto de dados produzindo novas imagens sintéticas, contudo, diferentemente do SMOTE, este método gerou mais dados para as classes minoritárias consideradas de difícil aprendizado do que para as classes minoritárias de fácil aprendizado, logo, como resultado final tem-se um conjunto de dados ampliado, porém, com um pequeno desequilíbrio entre as classes minoritárias.

Na Figura 2(a) e na Figura 2(b) são ilustrados alguns resultados das imagens sintéticas geradas pelos métodos SMOTE e ADASYN, respectivamente. Na Tabela 1 está descrito a composição dos conjuntos de treinamento, validação e teste, com e sem *oversampling*, para as duas abordagens de classificação, binária e multiclasse.

Tabela 1. Composição dos conjuntos de treinamento, validação e teste.

Conjunto de dados	Classificação binária		Classificação multiclasse			
	Lesões benignas	Melanoma	Melanoma	Nevo & MPA	Normal	Pterígio
Treinamento sem *oversampling*	188	97	97	67	68	53
Treinamento com SMOTE	188	187	97	97	98	93
Treinamento com ADASYN	188	187	97	100	96	101
Validação	26	14	14	10	9	7
Teste	53	28	28	19	19	15



(a) Imagens sintéticas: SMOTE. (b) Imagens sintéticas: ADASYN.

Figura 2. Exemplares de imagens sintéticas.

3.3. Redes Neurais Convolucionais

Os modelos avaliados neste trabalho foram o VGG16 (*Visual Geometry Group*) [Simonyan and ZissermanK 2014], *Xception* [Chollet 2017] e *MobileNetV2* [Sandler et al. 2018]. Estes modelos foram pré-treinados no conjunto de dados *Imagenet*, que é um conjunto que abrange 1000 classes de objetos (contém 1,2 milhões de imagens de treinamento, 50000 de imagens de validação e 100000 de teste). Todos eles contaram com uma camada totalmente conectada pela rede *Perceptron* Multicamadas (PMC) que desempenhou o papel de classificador. O classificador PMC possui uma camada de entrada, uma camada de saída com função de ativação *softmax* e duas camadas ocultas com 4096 neurônios e mais duas camadas de *dropout* entre as camadas ocultas. As camadas de *dropout* foram usadas para descartar de forma aleatória parte dos neurônios das camadas ocultas, com um percentual de 0,5, para evitar o *overfitting*. Este procedimento seguiu os experimentos feitos por [Srivastava et al. 2014], afirmando que o *dropout* típico para as camadas ocultas devem estar entre 0,5 e 0,8.

Os modelos foram treinados usando o método de otimização baseado no algoritmo *Stochastic Gradient Descent* (SGD), com uma taxa de aprendizado de 0,01. O período de duração do treinamento dos modelos foi de no máximo 10 épocas usando um tamanho de lote total de oito. Para evitar o super ajuste dos modelos foi adotado o método de parada antecipada (*Early Stopping*).

Para que fosse possível a implementação dos modelos *Ensemble*, os modelos VGG16, *Xception* e *MobileNetV2* atuaram como extratores de recursos, os quais convertiam as imagens em vetores de recursos unidimensionais para alimentar um conjunto de classificadores. Neste trabalho, foram combinados três classificadores distintos: *Perceptron Multicamadas* (PMC), *Support Vector Machine* (SVM) e *K-nearest neighbors* (KNN). As matrizes de pesos pré-ajustados dos modelos pré-treinados VGG16, *Xception* e *MobileNetV2* foram consideradas como a matriz de pesos inicial para a classificação dos tumores da conjuntiva ocular. Em seguida, duas abordagens de *Transfer Learning*,

chamadas de ajuste fino e extração de recursos, foram aplicadas com o conjunto de dados descrito na Seção 3.1.

4. Experimentos e resultados

Os resultados que serão apresentados nesta seção são a média de cinco execuções durante a validação cruzada 5-*folds*. Em seguida, para cada métrica foram calculados os intervalos de confiança (IC) de 95%. Destacamos também que, na coluna **Modelos**, os modelos precedidos pela palavra **Ens.** significa que nestes casos foram usados uma assembleia de classificadores, composta pelos algoritmos PMC, SVM e KNN.

4.1. Avaliação dos modelos na classificação binária

Neste cenário de experimentos é considerada a abordagem binária, que classifica as lesões em benigna ou melanoma, e para avaliação dos modelos as seguintes métricas para classificação binária são calculadas: (i) Sensibilidade, que mede a taxa de acertos da classe positiva; (ii) Especificidade, que mede a taxa de acertos da classe negativa, e (iii) *G-Mean*, que avalia o equilíbrio entre a sensibilidade e a especificidade [Fawcett 2006]. Na Tabela 2 encontra-se o desempenho dos modelos durante o processo de aprendizado/validação, na classificação entre melanoma (classe positiva) e lesões benignas (classe negativa). Neste caso, os dados de treinamento não foram balanceados.

Tabela 2. Classificação binária sem *oversampling*: validação dos modelos.

Modelos	G-Mean (IC)	Sensibilidade (IC)	Especificidade (IC)
VGG16	70,0 (68,1-71,9)	71,5 (69,0-74,0)	70,0 (68,1-71,9)
Xception	78,4 (73,9-82,9)	77,5 (73,2-81,8)	78,4 (73,9-82,9)
MobileNetV2	67,5 (66,7-68,3)	69,0 (67,6-70,4)	67,5 (66,7-68,3)
Ens.VGG16	68,0 (67,2-68,8)	23,0 (18,5-27,6)	73,7 (72,1-75,2)
Ens.Xception	**86,5** (84,3-88,7)	85,4 (82,3-88,5)	94,6 (93,5-95,7)
Ens.MobileNetV2	84,0 (82,0-86,0)	**87,6** (84,7-90,5)	**95,2** (94,2-96,1)

De acordo com os resultados descritos na Tabela 2, o modelo Ens.MobileNetV2 obteve o melhor de desempenho nas métricas sensibilidade e especificidade, no entanto, o modelo Ens.Xception obteve melhor desempenho na métrica *G-mean*, que é uma métrica que melhor representa a qualidade de um classificador em um problema com classes desbalanceadas.

Na Tabela 3 estão os resultados dos modelos treinados com os dados de treinamento balanceados pela técnica SMOTE. Observa-se que o *Ens.MobileNetV2* foi o modelo que obteve o melhor valor de *G-mean* e de especificidade, enquanto que o modelo Ens.Xception obte melhor valor de sensibilidade. Comparando com os resultados da Tabela 2, o modelo Ens.MobileNetV2 melhorou os seus resultados em todas as métricas, evidenciando que o balanceamento dos dados contribuiu para melhora no desempenho do modelo. E isso aconteceu para todos os mobelos *Ensemble*, exceto para o valor de *G-mean* do modelo Ens.Xception, e também para o modelo MobileNetV2. Para o modelo Xception, houve melhora no valor de *G-mean* e especificidade, e empate no valor de sensibilidade. Por fim, para o modelo VGG16, ocorreu um comportamento inverso de todos os demais modelos, pois houve piora em todas as métricas.

Tabela 3. Classificação binária com SMOTE: validação dos modelos.

Modelos	G-Mean (IC)	Sensibilidade (IC)	Especificidade (IC)
VGG16	60,0 (54,8-65,2)	60,0 (54,6-65,4)	60,0 (54,8-65,2)
Xception	77,5 (69,5-85,5)	77,5 (70,2-84,8)	77,5 (69,5-85,5)
MobileNetV2	84,1 (80,6-87,5)	83,5 (80,3-86,7)	84,1 (80,6-87,5)
Ens.VGG16	78,0 (75,6-80,4)	94,0 (93,1-95,0)	92,6 (91,2-94,0)
Ens.Xception	85,0 (82,1-87,9)	**98,3** (97,8-98,8)	95,3 (94,7-95,9)
Ens.MobileNetV2	**85,5** (83,1-87,9)	96,2 (95,6-96,7)	**96,6** (96,3-96,9)

Na Tabela 4 estão os resultados dos modelos treinados com os dados de treinamento balanceados pela técnica ADASYN. Neste cenário, o melhor modelo em todas as métricas foi o Ens.MobileNetV2. Novamente comparando com os resultados da Tabela 2, todos os modelos obtiveram melhores resultados, com exceção do valor de *G-mean* do modelo Ens.Xception. Estes resultados evidenciam que a técnica ADASYN contribuiu para a melhora do desempenho dos modelos.

Tabela 4. Classificação binária com ADASYN: validação dos modelos.

Modelos	G-Mean (IC)	Sensibilidade (IC)	Especificidade (IC)
VGG16	75,0 (69,4-80,6)	73,5 (66,8-80,2)	75,0 (69,4-80,6)
Xception	79,7 (71,3-88,1)	78,0 (70,3-85,7)	79,7 (71,3-88,1)
MobileNetV2	77,2 (73,5-80,8)	77,0 (73,2-80,8)	77,2 (73,5-80,8)
Ens.VGG16	83,0 (81,2-84,8)	90,6 (89,9-91,3)	93,6 (93,2-93,9)
Ens.Xception	85,0 (81,6-88,4)	96,8 (96,6-97,0)	94,9 (94,5-95,4)
Ens.MobileNetV2	**86,0** (83,7-88,3)	**97,6** (97,2-98,1)	**97,1** (96,5-97,6)

Ao final do processo de aprendizado/validação dos modelos, escolhemos os modelos com melhor valor de *G-mean* para a fase de teste. Este critério se deve ao fato desta métrica melhor avaliar um classificador na classificação nas duas classes. Na Tabela 5 estão descritos os resultados obtidos pelos modelos, sendo que o Ens.MobileNetV2 treinado com dados balanceados pelo ADASYN foi o melhor modelo em todas as métricas.

Tabela 5. Classificação binária: resultados obtidos no teste.

Modelos	G-Mean	Sensibilidade	Especificidade
Ens.Xception (sem *oversampling*)	94,3	85,4	94,3
Ens.MobileNetV2 (SMOTE)	95,1	89,1	95,1
Ens.MobileNetV2 (ADASYN)	**95,8**	**90,5**	**95,8**

4.2. Avaliação dos modelos na classificação multiclasse

Neste cenário de experimentos é considerada a abordagem multiclasse (que classifica as lesões em melanoma, nevo & MPA, pterígio ou normal), e para avaliação dos modelos as seguintes métricas para classificação multiclasse são calculadas: (i) *G-Mean*, que avalia o equilíbrio entre as taxas de acerto de todas as classes; (ii) Acurácia, que mede a taxa

de acerto considerando todas as classes, e (iii) AUC-ROC, que representa o quão bem o classificador consegue distinguir as diferentes classes [Fawcett 2006]. Na Tabela 6 encontra-se o desempenho dos modelos, durante o processo de aprendizado/validação, na classificação entre melanoma, nevo & MPA, pterígio e normal. Neste caso, os dados de treinamento não foram balanceados.

Tabela 6. Classificação multiclasse sem *oversampling*: validação dos modelos.

Modelos	G-mean (IC)	Acurácia (IC)	AUC-ROC (IC)
VGG16	65,4 (60,3-70,5)	57,0 (50,5-63,5)	**81,6** (76,7-86,5)
Xception	12,4 (6,4-18,4)	39,5 (38,7-40,3)	66,4 (65,0-67,9)
MobileNetV2	66,8 (63,4-70,1)	65,5 (63,2-67,8)	87,4 (85,8-89,0)
Ens.VGG16	73,5 (71,4-75,6)	62,0 (59,2-64,8)	70,7 (69,0-72,5)
Ens.Xception	73,5 (72,0-75,1)	62,0 (59,8-64,2)	71,4 (69,7-73,1)
Ens.MobileNetV2	**80,0** (77,9-82,2)	**71,0** (68,0-74,0)	79,0 (76,7-81,3)

De acordo com os resultados da Tabela 6, o modelo Ens.MobileNetV2 obteve os melhores resultados nas métricas *G-Mean* e acurácia, enquanto que o modelo VGG16 obteve melhor valor para AUC-ROC.

Na Tabela 7 estão os resultados do processo de aprendizado/validação dos modelos, os quais foram treinados com dados balanceados pelo método SMOTE. Neste experimento, o modelo Ens.MobileNetV2 obteve o melhor desempenho em todas as métricas. Comparando com os resultados da Tabela 6, pode-se constatar que o balanceamento dos dados contribuiu com a melhora dos modelos, na maioria das métricas calculadas.

Tabela 7. Classificação multiclasse com SMOTE: validação dos modelos.

Modelos	G-mean (IC)	Acurácia (IC)	AUC-ROC (IC)
VGG16	69,9 (63,6-76,2)	60,0 (52,5-67,5)	79,9 (73,7-86,1)
Xception	19,2 (9,1-29,2)	41,0 (34,1-47,9)	63,6 (58,8-68,3)
MobileNetV2	56,0 (48,6-63,4)	53,5 (49,9-57,1)	77,2 (74,9-79,5)
Ens.VGG16	77,8 (75,2-80,5)	68,0 (64,4-71,6)	78,8 (76,2-81,4)
Ens.Xception	78,3 (77,5-79,1)	68,5 (67,4-69,6)	77,7 (77,0-78,4)
Ens.MobileNetV2	**82,6** (82,0-83,1)	**74,5** (73,7-75,3)	**82,9** (82,5-83,3)

Na Tabela 8 estão os resultados do processo de aprendizado/validação dos modelos, os quais foram treinados com dados balanceados pelo método ADASYN. Neste experimento, o modelo Ens.MobileNetV2 obteve o melhor desempenho nas métricas *G-Mean* e acurácia, enquanto que o modelo VGG16 obteve melhor valor para AUC-ROC. Comparando com os resultados da Tabela 6, é possível constatar que o balanceamento dos dados contribuiu com a melhora dos modelos, na maioria das métricas calculadas.

Ao final do processo de aprendizado/validação dos modelos, escolhemos os modelos com melhor valor de *G-mean* para a fase de teste, pois é uma métrica que melhor avaliar um classificador em um problema de classificação com dados desbalanceados. Na Tabela 9 estão descritos os resultados obtidos pelos modelos, sendo que o

Tabela 8. Classificação multiclasse com ADASYN: validação dos modelos.

Modelos	G-mean (IC)	Acurácia (IC)	AUC-ROC (IC)
VGG16	68,4 (60,3-76,5)	67,0 (61,2-72,8)	**85,2** (81,9-88,4)
Xception	9,8 (3,2-16,4)	33,0 (28,5-37,5)	60,2 (56,6-63,8)
MobileNetV2	62,0 (57,1-66,8)	53,0 (49,8-56,2)	80,1 (77,6-82,5)
Ens.VGG16	78,2 (75,8-80,7)	68,5 (65,1-71,9)	78,3 (75,4-81,2)
Ens.Xception	76,5 (75,3-77,6)	66,0 (64,4-67,6)	76,5 (75,7-77,3)
Ens.MobileNetV2	**80,4** (79,3-81,6)	**71,5** (69,9-73,1)	80,6 (79,4-81,7)

Ens.MobileNetV2 treinado com dados balanceados pelo ADASYN foi o melhor modelo em todas as métricas.

Tabela 9. Classificação multiclasse: resultados obtidos no teste.

Modelos	G-mean	Acurácia	AUC-ROC
Ens.MobileNetV2 (sem *oversampling*)	89,5	84,5	88,8
Ens.MobileNetV2 (SMOTE)	91,0	86,7	91,0
Ens.MobileNetV2 (ADASYN)	**92,0**	**88,2**	**92,3**

5. Conclusões

Neste trabalho foi realizado um estudo comparativo entre redes neurais profundas para a classificação do melanoma conjuntival, mais especificamente, foram comparados o VGG16, Xception e MobileNetV2, utilizando a técnica *Transfer Learning* para melhorar a generalização dos modelos. Além disso, dois algoritmos de balanceamento de dados, o SMOTE e ADASYN, também foram aplicados nos dados de treinamento com o objetivo de evitar o enviesamento no treinamento dos modelos pela classe majoritária. O resultado da classificação final destes modelos foram a partir de classificadores únicos e também por uma assembleia de classificadores, composta pelos algoritmos PMC, SVM e KNN. Os experimentos foram divididos em dois cenários, classificação binária e multiclasse.

Os resultados na fase de aprendizado/validação da classificação binária mostraram que o balanceamento dos dados de treinamento contribuíram para um melhor desempenho dos modelos, e o Ens.MobileNetV2 foi, na maioria dos casos, o melhor modelo nas métricas *G-mean*, sensibilidade e especificidade. Na fase de teste, o modelo Ens.MobileNetV2 treinado com os dados balanceados com o método ADASYN obteve o melhor desempenho em todas as métricas. Por outro lado, na classificação multiclasse, o balanceamento dos dados também foi positivo, aumentando o desempenho dos modelos. Na fase de aprendizado/validação, novamente o Ens.MobileNetV2 foi o melhor, e na fase de teste, Ens.MobileNetV2 com ADASYN obteve os melhores resultados para *G-mean*, acurácia e AUC-ROC.

Como trabalhos futuros pretendemos investigar outros métodos de balanceamento de dados, pois como foi demonstrado, o balanceamento dos dados melhorou o desempenho dos modelos na classificação. Além disso, pretendemos também aplicar o nosso estudo na identificação de cânceres intraoculares.

Referências

Allam, E., Salem, A., and Alfonse, M. (2024). Classification of orbital tumors using convolutional neural networks. *Neural Computing and Applications*, 36:6025–6035.

Chandrabhatla, A., Horgan, T., Cotton, C., Ambati, N., and Shildkrot, Y. (2023). Clinical applications of machine learning in the management of intraocular cancers: A narrative review. *Investigative ophthalmology visual science*, 64:29.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research*, 16:321–357.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874. ROC Analysis in Pattern Recognition.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Jain, P., Finger, P. T., Fili, M., Damato, B., Coupland, S. E., Heimann, H., Kenawy, N., Brouwer, N. J., Marinkovic, M., Duinen, S. G. V., Caujolle, J. P., Maschi, C., Seregard, S., Pelayes, D., Folgar, M., Yousef, Y. A., Krema, H., and Calle-Vasquez, B. G. A. (2021). Conjunctival melanoma treatment outcomes in 288 patients: a multicentre international data-sharing study. *British Journal of Ophthalmology*, 105(10):1358–1364.

Koseoglu, N., Corrêa, Z., and Liu, T. Y. (2023). Artificial intelligence for ocular oncology. *Current opinion in ophthalmology*, Publish Ahead of Print.

Li, Z., Qiang, W., Chen, H., Pei, M., Yu, X., Wang, L., Li, Z., Xie, W., Wu, X., Jiang, J., and Wu, G. (2022). Artificial intelligence to detect malignant eyelid tumors from photographic images. *npj Digital Medicine*, 5:23.

Nazir, S., Dickson, D. M., and Akram, M. U. (2023). Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine*, 156:106668.

Novais, G. A. and Karp, C. L. (2012). Redes neurais profundas e ensemble de classificadores: uma aplicação em imagens médicas. *Arquivos Brasileiros de Oftalmologia*, 75:289–295.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Santos-Bustos, D. F., Nguyen, B. M., and Espitia, H. E. (2022). Towards automated eye cancer classification via vgg and resnet networks using transfer learning. *Engineering Science and Technology, an International Journal*, 35:101214–101226.

Simonyan, K. and ZissermanK, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, pages 1–14.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Yoo, T. K., Choi, J. Y., Kim, H. K., Ryu, I. H., and Kim, J. K. (2021). Adopting low-shot deep learning for the detection of conjunctival melanoma using ocular surface images. *Computer Methods and Programs in Biomedicine*, 205:1–10.