Construindo um *ensemble* de CNNs para classificação de imagens de lesões de pele

Pedro Lucas Silva Haga Torres
Departamento de Ciência da Computação, *Universidade de Brasília*
Brasília, Distrito Federal, Brasil
Email: pedrotorres@aluno.unb.br

***Resumo*—O câncer de pele é um dos cânceres de maior incidência no Brasil, podendo ser dividido em duas categorias: melanoma e não melanoma, sendo o primeiro a forma mais agressiva, dados seus altos índices de metástase. Há um grande interesse mundial no desenvolvimento de sistemas de diagnóstico auxiliado por computador para análise de imagens dermatoscópicas de lesões de pele, tanto para diagnóstico automático quanto para ajudar a tomada de decisão de dermatologistas. Neste trabalho, propõe-se os passos iniciais para a construção de um modelo de aprendizado profundo baseado em redes neurais convolucionais, especificamente, os resultados comparativos entre os otimizadores Adam e *stochastic gradient descent* com acelerador de Nesterov. Observou-se que esse último otimizador ofereceu os resultados mais robustos para a tarefa, pavimentando o caminho para a construção de um modelo baseado em *ensemble* para a tarefa de classificação de imagens dermatoscópicas de lesões de pele.**

***Index Terms*—Aprendizado de máquina, aprendizado profundo, redes neurais convolucionais, câncer de pele, dermatoscopia.**

I. Introdução

O câncer de pele é um dos cânceres de maior incidência no Brasil [1] e pode ser dividido em duas categorias: melanoma e não melanoma, o primeiro sendo a forma mais agressiva, dados seus altos índices de metástase. O diagnóstico precoce de melanoma é de suma importância, pois melhora significativamente o prognóstico do paciente [2]. Há um grande interesse mundial no desenvolvimento de sistemas de diagnóstico auxiliado por computador para processamento e análise de imagens dermatoscópicas de lesões de pele, objetivando-se auxiliar a tomada de decisões dos dermatologistas [3]–[5]. Imagens dermatoscópicas são adquiridas a partir de técnicas não invasivas, por meio de um dispositivo dermatoscópico que permite uma visualização mais detalhada dos padrões da lesão na superfície da pele, e têm sido amplamente usadas na literatura [6]. Apresentam-se alguns exemplos na figura 1.

Nesse âmbito, algoritmos de aprendizado de máquina (ML, do inglês, *machine learning*) se provam úteis para tarefas de classificação, podendo auxiliar o diagnóstico de lesões de pele. Uma tarefa de classificação em ML, consiste em atribuir um determinado rótulo a um conjunto de dados [7]. Um modelo de classificação aprende a rotular novos dados utilizando um conjunto de treinamento previamente rotulado, no que é chamado de aprendizado supervisionado. No contexto da dermoscopia, as principais tarefas de classificação consistem no diagnóstico de melanoma, classificando imagens de lesões de pele entre melanoma e não melanoma ou entre classes específicas dessas lesões, como nevo melanocítico, carcinoma basocelular, ceratose actínica, o próprio melanoma, entre outros.

O aprendizado profundo é uma área de ML que se sobressai em problemas simples e intuitivos para humanos, porém, difíceis de serem descritos por meio de formalizações matemáticas, como o reconhecimento da fala ou de rostos de pessoas [8]. Redes neurais convolucionais (CNN, do inglês *convolutional neural networks*) são modelos de aprendizado profundo amplamente utilizados para a tarefa de classificação de imagens [9]. CNNs têm em sua base filtros de convolução (ou mapas de ativação) que são utilizados para encontrar e extrair características relevantes dos dados. O treinamento de um modelo de classificação baseado em redes neurais convolucionais consiste em encontrar os filtros adequados para extrair as características relevantes de uma imagem para a classificar entre uma ou mais classes. Entretanto, esse tipo de abordagem exige grandes quantidades de dados para conseguir gerar uma solução adequada e generalizável. O International Skin Imaging Collaboration (ISIC) é um esforço internacional para a melhora do diagnóstico de melanoma por meio de imagens dermatoscópicas digitais, e são detentores do maior conjunto público de imagens dermatoscópicas de lesões de pele com controle de qualidade[1].

O uso de imagens dermatoscópicas é um tema amplamente estudado na literatura médica. Esse tipo de imagem permite uma maior acurácia no diagnóstico de lesões de pele por dermatologistas [10]. Entretanto, dermoscopia é difícil de se aprender [11], sua eficácia é limitada quando utilizada por profissionais mal treinados [12], [13] e, mesmo com treinamento adequado, a análise visual de uma imagem pelo dermatologista ainda é subjetiva [14]. Desse modo, estudos demonstram que modelos baseados em CNNs podem obter melhores resultados de classificação de imagens dermatoscópicas do que dermatologistas [15], [16]. Porém, como comentado anteriormente, esse tipo de modelo necessita de um grande volume de dados e mesmo as bases de imagens disponibilizadas pelo ISIC [17]–[19] sofrem com o desbalanceamento (vide a tabela I, por exemplo). Além do problema de desbalanceamento, não há uma padronização a ser seguida para a aquisição das imagens de lesões de pele, pois elas são adquiridas e compiladas a

[1] https://challenge.isic-archive.com/

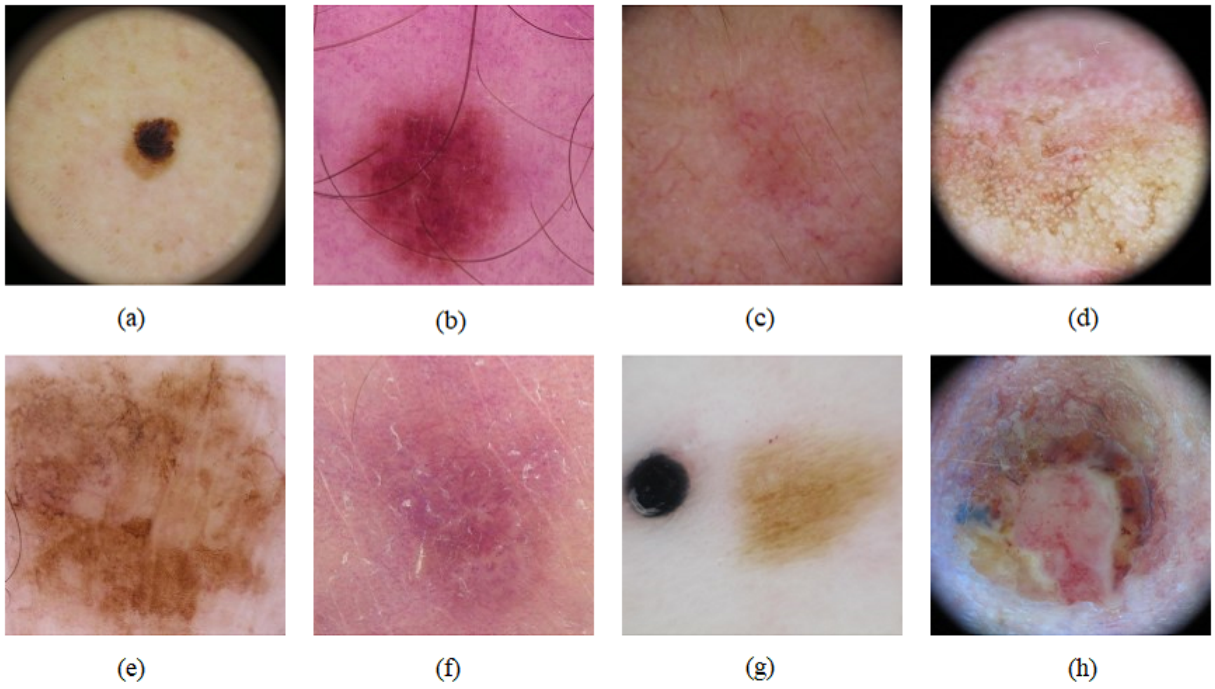

Figura 1. Exemplos de imagens dermatoscópicas de lesões de pele: (a) melanoma, (b) nevo melanocítico, (c) carcinoma basocelular, (d) ceratose actínica, (e) ceratose seborreica, (f) dermatofibroma, (g) lesão vascular e (h) carcinoma espinocelular.

partir de diferentes fontes, havendo variações de iluminação e, consequentemente, das cores detectadas pelo dispositivo de captura de imagens, o que leva a perdas na performance de classificação de algoritmos de ML [20].

O ISIC Challenge de 2019[2] consiste na classificação de imagens dermatoscópicas de lesões de pele entre 9 classes (tipos de lesões de pele) distintas: melanoma, nevo melanocítico, carcinoma basocelular, ceratose actínica, ceratose seborreica, dermatofibroma, lesão vascular e carcinoma espinocelular, com a 9ª classe representando imagens que não pertencem a nenhuma das classes anteriores (N.D.A.). Foram fornecidos dois conjuntos de imagens, um para treinamento e outro para teste com 25.331 e 8.238 instâncias, respectivamente. Exemplos das imagens do conjunto de treino e seus respectivos rótulos são apresentados na figura 1 e a distribuição das imagens em suas respectivas classes é apresentada na tabela I. É possível observar que não há uma padronização das imagens, com algumas delas contendo bordas pretas na região mais externa, assim como também não é possível garantir a consistência de cores entre uma imagem e a próxima, dado a falta de padronização e problemas inerentes aos sensores de captura de luz e a forma como eles registram as cores. Além disso, também é possível observar um severo desbalanceamento na base de dados, com a classe nevo melanocítico contendo mais de 50% das imagens, ao passo que as classes dermatofibroma e lesão vascular possuem aproximadamente 1% do total de imagens cada.

No presente trabalho, apresentam-se os passos iniciais para a construção de um modelo baseado no *ensemble* homogêneo de CNNs de arquitetura EfficientNet [21] para o problema apresentado no ISIC Challenge de 2019. *Ensemble learning* (do inglês, aprendizado por *ensemble*) é uma técnica de ML que consiste em combinar a saída, de diferentes classificadores para gerar uma predição mais robusta [22]. Como o *ensemble* proposto neste trabalho é composto por apenas uma única arquitetura de CNN, ele é descrito como homogêneo. Foram treinados dois modelos de EfficientNet-B2 para comparar o desempenho dos algoritmos de otimização Adam [23] e *stochastic gradient descent* com acelerador de Nesterov [24]. Para treinar esses modelos, utilizou-se as imagens disponibilizadas para treinamento pelo ISIC Challenge 2019, assim como imagens geradas seguindo a técnica de *data augmentation* [25], utilizando diferentes conjuntos de transformações para gerar mais imagens para o treinamento. Para garantir a consistência das cores entre as diferentes imagens, será utilizado o algoritmo Shades of Gray [20], [26].

Este artigo está estruturado da seguinte maneira: a Seção II traz a revisão de literatura dividida entre as subseções II-A e II-B com o estado da arte e trabalhos relacionados, respectivamente. A Seção III aborda a metodologia adotada no trabalho para o comparativo. Por fim, a Seção V traz a conclusão e trabalhos futuros.

II. Revisão de Literatura

Nesta seção, serão apresentados o estado da arte e trabalhos relacionados à tarefa de classificação de imagens dermatoscópicas de lesão de pele.

[2]https://challenge.isic-archive.com/landing/2019

A. Estado da arte

Os cinco trabalhos melhor classificados no ISIC Challenge 2019 foram investigados no presente estudo. A metodologia de avaliação do desafio consiste na métrica de acurácia multiclasse normalizada, balanceada entre as categorias de lesão, descrita em (1). Nessa equação, n representa o número de classes e i é um identificador para cada uma das classes. Essa métrica é semanticamente equivalente a média aritmética da revocação de cada uma das classes. Em caso de empate, a decisão é tomada levando em consideração a área sob a curva característica do receptor (AUC, do inglês, *area under the receiver operating characteristic curve*).

$$BMAcc = \frac{1}{n} * \sum_{i=1}^{n} \frac{TP_i}{TP_i + FN_i} \tag{1}$$

Dessa forma, notou-se que a arquitetura EfficientNet foi utilizada nos dois trabalhos melhor classificados [27], [28]. Os demais trabalhos investigados utilizaram diferentes versões de arquiteturas como DenseNet, ResNet, Inception, entre outras [29]–[31], porém, foram classificados abaixo dos estudos citados anteriormente. A técnica de *ensemble* também foi utilizada por todos os trabalhos estudados, assim como o algoritmo de Shades of Gray. O trabalho de Pollastri et al. [29] fez uso de diferentes técnicas de *data augmentation* para treinar as redes neurais que compõem sua arquitetura. Chouhan [31] testou técnicas de *oversampling*, *undersampling* e o uso de pesos para as diferentes classes durante o cálculo da função de perda, constatando que essa última é a alternativa que produz os melhores resultados.

Iqbal et al. desenvolveram uma arquitetura de rede neural convolucional profunda especificamente para o problema de classificação de imagens dermatoscópicas de lesões de pele [5]. A CNN desenvolvida por eles é composta por camadas convolucionais, *max pooling* e *batch normalization*, com a camada de topo composta por *average pooling* e uma camada totalmente conectada para a saída. Os autores utilizaram as bases dos desafios ISIC de 2017, 2018 e 2019 porém, não submeteram sua solução. Com relação a esse último ano, eles dividiram a base de dados em 70% para treinamento, 10% para validação e 20% para teste e empregaram uma estratégia agressiva de *data augmentation* no conjunto de treinamento, equiparando o número de imagens de todas as classes a da classe de nevo melanocítico, entretanto, trataram apenas das 8 classes de lesões, desconsiderando a classe N.D.A. Obtiveram os seguintes resultados sobre seu conjunto de teste: revocação de $89,58\%$, AUC $0,991$, precisão $90,66\%$, especificidade $97,57\%$ e F1-score de $89,75\%$, com revocação, precisão e especificidade sendo a média aritmética entre as classes.

Ali et al. também desenvolveram uma arquitetura de CNN própria [4], porém, utilizaram a base de imagens dermatoscópicas HAM10000 [17], composta por 7 classes diferentes de lesões de pele (ceratose actínica, carcinoma basocelular, ceratose seborreica, dermatofibroma, melanoma, nevo melanocítico e lesão vascular). O conjunto de dados HAM10000 faz parte da base de imagens disponibilizadas para o desafio ISIC de 2019. Ali et al. desenvolveram uma arquitetura de CNN mais simples, utilizando apenas camadas convolucionais seguidas por *max pooling* ligadas a camadas totalmente conectadas por uma camada de *dropout* (com chance de 0,25) e uma camada de saída. Os autores não detalham a estratégia utilizada para o *data augmentation*, apenas citando as transformações que utilizaram. Os seguintes resultados foram obtidos sobre um conjunto de teste, com a divisão 80%, 10% e 10% para treino validação e teste, respectivamente: 96,57% de precisão, 93,66% de revocação, 95,09% de F1-score, 93,16% de acurácia no conjunto de treino e 91,43% de acurácia no conjunto de teste. Ali et al. também compararam seu modelo proposto com outras arquiteturas mais conhecidas como AlexNet, ResNet, VGG-16, DenseNet e MobileNet obtendo melhores resultados que todas na métrica F1-score.

B. Trabalhos relacionados

Como descrito anteriormente, todos os trabalhos investigados utilizaram a técnica de *data augmentation*, que consiste em aumentar o número de amostras de treino artificialmente por meio de transformações das imagens já disponíveis [25]. Perez et al. [32] testaram diferentes transformações das imagens originais para o problema de classificação de lesões de pele proposto no ISIC Challenge de 2017 (classificação binária de imagens dermatoscópicas de lesões de pele entre melanoma e nevo e entre melanoma e ceratose seborreica). Foram constatados ganhos na métrica AUC para todas as transformações propostas, como alterações nos parâmetros de cor da imagem (brilho, saturação, matiz, etc.), *random erasing*, inversões verticais e horizontais, rotações em até 90º, entre outras. Os autores propuseram 4 conjuntos de transformações:

1) **Conjunto básico:** cortes aleatórios, *affine*, inversões horizontais e/ou verticais e alterações nos parâmetros de saturação, contraste, brilho e matiz;
2) Conjunto básico e apagamento aleatório (*random erasing*);
3) Conjunto básico e transformação elástica; e
4) Conjunto básico e *mix* de lesões.

No conjunto básico, as transformações são feitas sequencialmente, ao passo que, nos demais, o passo acrescentado é feito entre uma das transformações. Foram observados ganhos significativos ao utilizar esses conjuntos de transformações.

Dada a forma como algoritmos de aprendizado profundo são treinados, i.e., os pesos são instanciados aleatoriamente e alterados de acordo com uma função de perda por meio do procedimento de retro propagação (*backpropagation*), classes com poucas instâncias têm seus gradientes significativamente reduzidos, o que leva o modelo a classificar todas as instâncias como uma das classes majoritárias [33]. Uma das formas de se combater esse fenômeno é pelo uso de pesos para as classes, de modo a penalizar acertos simples (em classes com muitas instâncias) e valorizar acertos nas classes menos representadas [34]. Essa técnica foi empregada neste trabalho por meio dos pesos calculados para as classes utilizando a Equação 3.

O algoritmo Shades of Gray [26], visa garantir a consistência de cores entre duas imagens digitais obtidas de

formas diferentes. A cor de um objeto no mundo real depende da distribuição espectral da luz que o ilumina assim como a quantidade de luz que esse objeto reflete. O sistema visual humano é capaz de deduzir um descritor estável da cor de um objeto independentemente da iluminação. Entretantos, sistemas digitais não conseguem reproduzir essa característica da visão humana de forma adequada. Para combater isso, os autores propõem que a média das cores de uma imagem é um tom de cinza (*shade of gray*) e utilizam essa característica para analisar e normalizar as cores de um grande conjunto de imagens utilizando a métrica de Minkowski, de modo que um tom de vermelho, por exemplo, seja representado da mesma forma em todas as imagens. O Shades of Gray produz os melhores resultados em tarefas de classificação de imagens quando comparado a outros algoritmos de consistência de cores [20].

A arquitetura EfficientNet [21] foi desenvolvida a partir de um estudo sistemático da escala dos modelos de CNNs em termos de profundidade, largura e resolução das imagens na camada de entrada da rede, sendo identificado que um balanço no crescimento das dimensões citadas anteriormente leva a uma melhora de performance. Esse estudo trata essas dimensões da rede como variáveis a serem otimizadas e propõe um método de *compound scaling* para encontrar a proporção ideal entre as variáveis. Isso pode ser utilizado para otimizar arquiteturas já existentes, como também para construir uma arquitetura base e ampliar ela de acordo com os parâmetros de otimização encontrados. A EfficientNet possui 8 iterações distintas (B0 a B7), partindo da arquitetura base (B0), que cresce de acordo com os parâmetros citados anteriormente até a arquitetura B7. Os resultados expostos neste trabalho foram obtidos utilizando-se a B2 (devido a limitação de hardware disponível).

III. Metodologia

No presente trabalho, foram treinadas duas instâncias distintas de EfficientNet-B2: uma utilizando o otimizador *stochastic gradient descent* (SGD) com gradiente acelerado de Nesterov [24] e a outra o otimizador Adam [23]. Esses otimizadores foram escolhidos por serem complementares, o Adam converge rapidamente e com qualidade satisfatória, porém pode ter problemas em alguns conjuntos de dados, problemas que não ocorrem com o SGD associado ao gradiente de Nesterov [35], entretanto, ele demora mais a convergir para a melhor solução. Ambos foram treinados sobre o mesmo conjunto de dados, i.e., com as mesmas imagens originais e as mesmas imagens criadas durante o processo de *data augmentation*, de modo a garantir a consistência dos comparativos.

A base de dados utilizada para treinar os modelos foi uma subdivisão da base de treino disponibilizada pelo ISIC Challenge 2019. Das 25.331 imagens, 70% foram destinadas para o treinamento, 20% para validação e 10% para teste. Essa divisão foi feita de forma estratificada, i.e., mantendo a proporção de lesões constante entre os diferentes conjuntos. Foi aplicado o conjunto básico de *data augmentation* descrito por Perez et al. [32] sobre o conjunto de treinamento, dobrando o número de imagens de todas as classes exceto nevo melanocítico - o mesmo não foi feito nos conjuntos de validação e teste. Ambos modelos foram treinados para classificar apenas os 8 tipos de lesões que possuem imagem na base, ainda estuda-se como fazer a classificação da 9ª classe (N.D.A.).

As métricas escolhidas para avaliação são: acurácia, perda e F1-score (média aritmética entre as classes) durante o treinamento; e revocação, sensibilidade, precisão e F1-score (todas descrevendo a média aritmética entre os resultados das classes) para avaliar os conjuntos de validação e teste. As métricas escolhidas para apresentar os resultados de treinamento, estão disponíveis para uso junto a API do TensorFlow[3], por isso sua escolha. A escolha das métricas para os conjuntos de validação e teste se dá pela conformidade com as métricas escolhidas para avaliação pelo ISIC Challenge 2019, no caso de revocação, e, as demais, para oferecer métricas que descrevem o comportamento dos classificadores com o rigor científico necessário. Naturalmente, espera-se que o comportamento do classificador seja marginalmente inferior no conjunto de teste com relação ao conjunto de validação, o que descreve uma solução bem generalizável.

Para os resultados expostos neste trabalho, as alterações feitas nas camadas do topo da rede neural consistem do aumento do *dropout* para $0,35$ seguido por uma camada densa de 8 neurônios com função de ativação softmax descrita em (2), para se calcular a probabilidade de uma imagem pertencer a uma classe, como requisitado pelo ISIC Challenge 2019. Utilizou-se *transfer learning* com os pesos da EfficientNet-B2 pré-treinados na base ImageNet [36]. Em ambos casos, as camadas do topo foram treinadas sozinhas por 60 épocas com taxa de aprendizado de 10^{-3}, dividindo-se a taxa de aprendizado por 10 a cada 20 épocas. Todas as camadas da rede foram treinadas em seguida por outras 80 épocas no caso da rede utilizando Adam e 50 no caso da rede utilizando SGD, observando o comportamento dos valores de perda de treinamento e validação ao longo do tempo. Para o treinamento de todas as camadas, a taxa de aprendizado inicial foi reduzida a 10^{-4}, novamente, realizando a divisão dessa por 10 a cada 20 épocas. O modelo treinando com o otimizador Adam fez uso de mais épocas, pois não estava se comportando como esperado (vide a Figura 3).

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{2}$$

Dado o desbalanceamento da base de dados, também utilizou-se pesos para as diferentes classes durante o cálculo da função de perda do modelo [34]. Apenas a estratégia de se usar *data augmentation* não é o suficiente para a melhora dos resultados e, em experimentos prévios, ao aplicar um segundo conjunto de transformações, triplicando o número de imagens das classes (excluindo nevo melanocítico), os ganhos não são tão expressivos e aumenta-se o risco de perda de generalização. Por isso, utiliza-se pesos para as classes durante o cálculo da função de perda do modelo. Esses pesos são computados a

[3]https://www.tensorflow.org/

partir da Equação (3), onde o peso de cada classe é dado pelo número total de imagens, dividido pelo número de classes vezes o número de amostras da respectiva classe, conforme implementação na biblioteca scikit-learn[4].

$$peso_{classe} = \frac{total_{imagens}}{num_{classes} * num_{amostras}} \quad (3)$$

IV. Resultados Experimentais

A base de dados para treinamento disponibilizada no ISIC Challenge 2019 [17]–[19] consiste de mais de 25 mil imagens divididas entre 8 classes: melanoma (MEL), nevo melanocítico (NV), carcinoma basocelular (BCC), ceratose actínica (AK), ceratose seborreica (BKL), dermatofibroma (DF), lesão vascular (VASC) e carcinoma espinocelular (SCC). Não foram disponibilizadas imagens para a 9ª classe (N.D.A.), porém, a organização do desafio informou que a avaliação dessa classe seria feita através de imagens dermatoscópicas de pele, sem nenhum tipo de lesão. Apesar de ser um volume grande de imagens, a base está severamente desbalanceada, com 50% das imagens pertencendo a classe de nevo melanocítico, ao passo que classes como dermatofibroma e lesão vascular possuem aproximadamente 1% do total de imagens cada. Pode-se observar o número de imagens para cada uma das classes, assim como seu percentual do total de imagens na Tabela I.

Tabela I
Distribuição das imagens disponibilizadas como conjunto de treinamento do ISIC Challenge 2019.

Diagnóstico	Número de amostras	Percentual do total
Melanoma	4522	17,85%
Nevo melanocítico	12875	50,83%
Carcinoma basocelular	3323	13,12%
Ceratose actínica	867	3,42%
Ceratose seborreica	2624	10,36%
Dermatofibroma	239	0,94%
Lesão vascular	253	1,00%
Carcinoma espinocelular	628	2,48%
N.D.A.	0	0,00%
Total	25331	100,00%

Os resultados do treinamento dos modelos utilizando os otimizadores SGD com acelerador de Nesterov e Adam podem ser encontrados nas Figuras 4, 5 e 8 a 10, no Apêndice. Os resultados obtidos nos conjuntos de validação e teste de ambos modelos estão expostos na Tabela II, na forma da média aritmética entre os resultados das diferentes classes. Para uma visão mais detalhada dos resultados de classificação, disponibiliza-se a matriz de confusão dos modelos nas Figuras 6, 7, 11 e 12. É possível observar que não há uma diferença significativa entre os resultados dos modelos, entretanto, o modelo com SGD obteve resultados marginalmente melhores com um número de épocas menor (47 contra 78 do Adam). Ao observar o comportamento da perda ao longo do tempo (Figuras 2 e 3, o modelo utilizando o SGD possui um resultado mais próximo do esperado, entretanto, dado o uso de *data augmentation* apenas no conjunto de treinamento e a camada de *dropout*, os resultados do modelo Adam são mais condizentes com esses fatores. Observando todos esses dados, acredita-se que o modelo utilizando SGD e Nesterov é o melhor para essa tarefa, uma vez que ele converge mais rapidamente (próximo da época 40, ao passo que o Adam ainda não regularizou nessa mesma época - característica que deveria fazer o Adam se sobressair) e seu comportamento está mais próximo do ideal, com uma perda no conjunto de treinamento superior ao do conjunto de validação.

[4]https://scikit-learn.org/stable/modules/generated/sklearn.utils.class_weight.compute_class_weight.html

Tabela II
Métricas obtidas sobre os conjuntos de validação e teste - média aritmética entre as classes.

Métrica	SGD + Nesterov		Adam	
	Validação	Teste	Validação	Teste
Revocação	12,66%	12,16%	12,39%	12,98%
Especificidade	87,47%	87,58%	87,45%	87,84%
Precisão	12,67%	12,62%	12,30%	12,03%
F1-Score	12,67%	12,12%	12,34%	12,49%
Acurárica	82,91%	82,91%	82,14%	82,15%

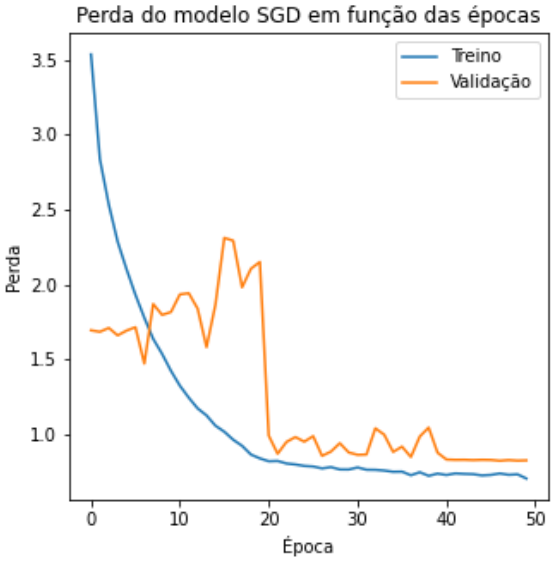

Figura 2. Gráfico da perda do modelo treinado com otimizador SGD e acelerador de Nesterov ao longo do treinamento.

V. Conclusão e Trabalhos Futuros

Neste trabalho apresentou-se os resultados de treinamento de dois modelos distintos, utilizando cada, um otimizador diferente. Ambos utilizaram o mesmo conjunto de treinamento e foram avaliados no mesmo conjunto de validação e teste, com todas as variáveis aleatórias minimizadas. Dessa forma, pode-se constatar que essa é uma tarefa em que o otimizador SGD com Nesterov se sobressai ao Adam, convergindo mais rapidamente e com um comportamento mais próximo do ideal. Entretanto, os resultados de classificação não são satisfatórios.

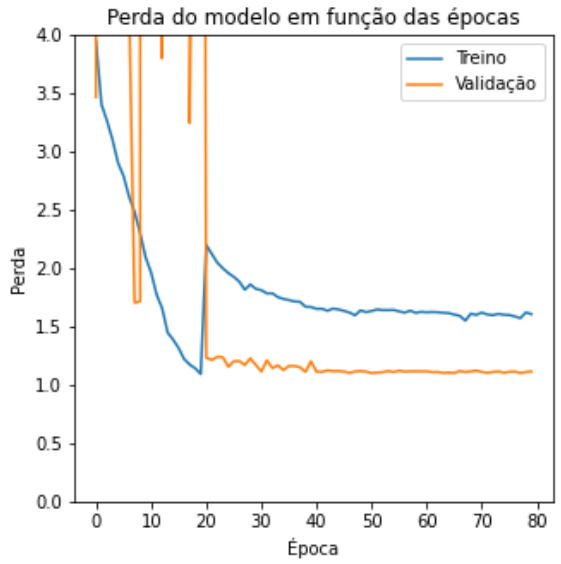

Figura 3. Gráfico da perda do modelo treinado com otimizador Adam ao longo do treinamento (cortado em $Y = 4$ para melhor visualização, o gráfico sem cortes está disponível na Figura 8, no Apêndice). Dado o valor de perda de 1420,41 durante a época 10 se fez necessário o corte para melhor visualização.

De fato, a acurácia de classificação é alta e os modelos não sofrem com *overfitting*, porém, observando a métrica principal de avaliação, ainda há muito a se melhorar. Mesmo com essas intempéries, considera-se o trabalho desenvolvido um sucesso: sabe-se que o otimizador SGD é melhor para a tarefa e há dados sobre o funcionamento da arquitetura ao longo de um grande número de épocas.

Para melhorar a performance de classificação do modelo, deseja-se, utilizar arquiteturas mais robustas da EfficientNet (B4 ou superior, a depender das limitações de hardware disponíveis). A arquitetura B2 claramente não é robusta o suficiente para essa tarefa. Além disso, também deseja-se construir *ensemble* homogêneo dessas EfficientNets treinadas com conjuntos de imagens que passaram por diferentes transformações de *data augmentation*. Espera-se que com isso seja possível alcançar melhores resultados de classificação no ISIC Challenge 2019 de acordo com as métricas do desafio. Estudar-se-á como lidar com a 9ª classe (N.D.A.).

Referências

[1] Instituto Nacional de Câncer José Alencar Gomes da Silva, *Estimativa 2020: incidência de câncer no Brasil*. INCA, 2019. [Online]. Available: https://www.inca.gov.br/publicacoes/livros/estimativa-2020-incidencia-de-cancer-no-brasil

[2] A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, "Early detection and treatment of skin cancer," *American Family Physician*, vol. 62, no. 2, pp. 357–368+375–376+381–382, 2000, cited By :266. [Online]. Available: www.scopus.com

[3] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Computerized Medical Imaging and Graphics*, vol. 71, pp. 19–29, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611118306050

[4] M. S. Ali, M. S. Miah, J. Haque, M. M. Rahman, and M. K. Islam, "An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models," *Machine Learning with Applications*, vol. 5, p. 100036, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666827021000177

[5] I. Iqbal, M. Younus, K. Walayat, M. U. Kakar, and J. Ma, "Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images," *Computerized Medical Imaging and Graphics*, vol. 88, p. 101843, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0895611120301385

[6] M. E. Celebi, T. Mendonca, and J. S. Marques, *Dermoscopy image analysis*. CRC Press, 2015, vol. 10.

[7] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Pearson, 2010, vol. 3, ch. 18, p. 696.

[8] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, ch. 1, pp. 1–2, http://www.deeplearningbook.org.

[9] W. Wang, Y. Yang, X. Wang, W. Wang, and J. Li, "Development of convolutional neural network and its application in image classification: a survey," *Optical Engineering*, vol. 58, no. 4, pp. 1 – 19, 2019. [Online]. Available: https://doi.org/10.1117/1.OE.58.4.040901

[10] M. Vestergaard, P. Macaskill, P. Holt, and S. Menzies, "Dermoscopy compared with naked eye examination for the diagnosis of primary melanoma: a meta-analysis of studies performed in a clinical setting," *British Journal of Dermatology*, vol. 159, no. 3, pp. 669–676, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2133.2008.08713.x

[11] A. Forsea, P. Tschandl, V. del Marmol, I. Zalaudek, H. Soyer, E. W. Group, A. Geller, and G. Argenziano, "Factors driving the use of dermoscopy in europe: a pan-european survey," *British Journal of Dermatology*, vol. 175, no. 6, pp. 1329–1337, 2016. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/bjd.14895

[12] M. Binder, M. Schwarz, A. Winkler, A. Steiner, A. Kaider, K. Wolff, and H. Pehamberger, "Epiluminescence Microscopy: A Useful Tool for the Diagnosis of Pigmented Skin Lesions for Formally Trained Dermatologists," *Archives of Dermatology*, vol. 131, no. 3, pp. 286–291, 03 1995. [Online]. Available: https://doi.org/10.1001/archderm.1995.01690150050011

[13] H. Kittler, H. Pehamberger, K. Wolff, and M. Binder, "Diagnostic accuracy of dermoscopy," *The Lancet Oncology*, vol. 3, no. 3, pp. 159–165, 2002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470204502006794

[14] M.-L. Bafounta, A. Beauchet, P. Aegerter, and P. Saiag, "Is Dermoscopy (Epiluminescence Microscopy) Useful for the Diagnosis of Melanoma?: Results of a Meta-analysis Using Techniques Adapted to the Evaluation of Diagnostic Tests," *Archives of Dermatology*, vol. 137, no. 10, pp. 1343–1350, 10 2001. [Online]. Available: https://doi.org/10.1001/archderm.137.10.1343

[15] H. Haenssle, C. Fink, R. Schneiderbauer, F. Toberer, T. Buhl, A. Blum, A. Kalloo, A. B. H. Hassen, L. Thomas, A. Enk, L. Uhlmann, C. Alt, M. Arenbergerova, R. Bakos, A. Baltzer, I. Bertlich, A. Blum, T. Bokor-Billmann, J. Bowling, N. Braghiroli, R. Braun, K. Buder-Bakhaya, T. Buhl, H. Cabo, L. Cabrijan, N. Cevic, A. Classen, D. Deltgen, C. Fink, I. Georgieva, L.-E. Hakim-Meibodi, S. Hanner, F. Hartmann, J. Hartmann, G. Haus, E. Hoxha, R. Karls, H. Koga, J. Kreusch, A. Lallas, P. Majenka, A. Marghoob, C. Massone, L. Mekokishvili, D. Mestel, V. Meyer, A. Neuberger, K. Nielsen, M. Oliviero, R. Pampena, J. Paoli, E. Pawlik, B. Rao, A. Rendon, T. Russo, A. Sadek, K. Samhaber, R. Schneiderbauer, A. Schweizer, F. Toberer, L. Trennheuser, L. Vlahova, A. Wald, J. Winkler, P. Wölbing, and I. Zalaudek, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Annals of Oncology*, vol. 29, no. 8, pp. 1836–1842, 2018, immune-related pathologic response criteria. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0923753419341055

[16] C. Fink, A. Blum, T. Buhl, C. Mitteldorf, R. Hofmann-Wellenhof, T. Deinlein, W. Stolz, L. Trennheuser, C. Cussigh, D. Deltgen, J. Winkler, F. Toberer, A. Enk, A. Rosenberger, and H. Haenssle, "Diagnostic performance of a deep learning convolutional neural network in the differentiation of combined naevi and melanomas," *Journal of the European Academy of Dermatology and Venereology*, vol. 34, no. 6, pp. 1355–1361, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jdv.16165

[17] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[18] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[19] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[20] C. Barata, M. E. Celebi, and J. S. Marques, "Improving dermoscopy image classification using color constancy," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 3, pp. 1146–1152, 2015.

[21] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946

[22] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*. Upper Saddle River, NJ, USA: Pearson, 2010, vol. 3, ch. 18, p. 748.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[24] Y. Nesterov, "A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$," *Doklady AN USSR*, vol. 269, pp. 543–547, 1983. [Online]. Available: https://ci.nii.ac.jp/naid/20001173129/en/

[25] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, 2019, cited By :516. [Online]. Available: https://doi.org/10.1186/s40537-019-0197-0

[26] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," *Color and Imaging Conference*, vol. 2004, no. 1, pp. 37–41, 2004. [Online]. Available: https://www.ingentaconnect.com/content/ist/cic/2004/00002004/00000001/art00008

[27] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution efficientnets with meta data," *MethodsX*, vol. 7, p. 100864, 2020.

[28] S. Zhou, Y. Zhuang, and R. Meng, "Multi-category skin lesion diagnosis using dermoscopy images and deep cnn ensembles," *ISIC 2019 Challenge Leaderboard*, 2020.

[29] F. Pollastri, J. Maroñas, M. Parreño, F. Bolelli, R. Paredes, C. Grana, and A. Albiol, "Aimagelab-prhlt at isic challenge 2019," *ISIC 2019 Challenge Leaderboard*, 2020.

[30] A. G. Pacheco, A.-R. Ali, and T. Trappenberg, "Skin cancer detection based on deep learning and entropy to detect outlier samples," *arXiv preprint arXiv:1909.04525*, 2019.

[31] V. Chouhan, "Skin lesion analysis towards melanoma detection with deep convolutional neural network," *ISIC 2019 Challenge Leaderboard*, 2020.

[32] F. Perez, C. Vasconcelos, S. Avila, and E. Valle, "Data augmentation for skin lesion analysis," in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*. Springer, 2018, pp. 303–311.

[33] R. Anand, K. Mehrotra, C. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Transactions on Neural Networks*, vol. 4, no. 6, pp. 962–969, 1993.

[34] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[35] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," 2018.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

APÊNDICE

Métricas do modelo treinado com otimizador SGD e acelerador de Nesterov

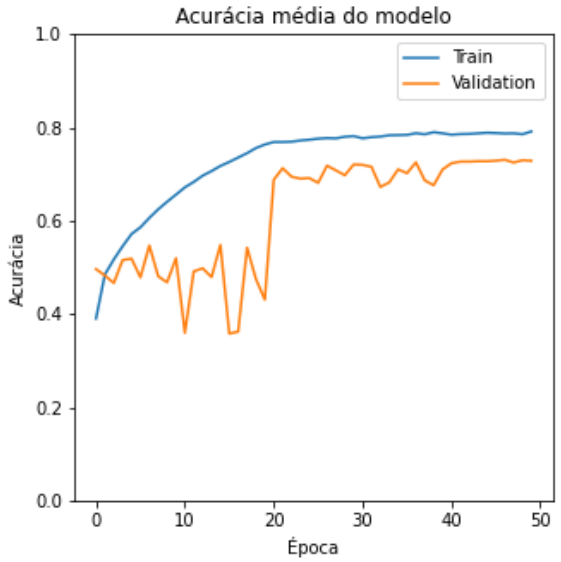

Figura 4. Gráfico da acurácia do modelo treinado com otimizador SGD e acelerador de Nesterov (média aritmética entre as classes) de treino e validação pelas épocas - durante o treinamento.

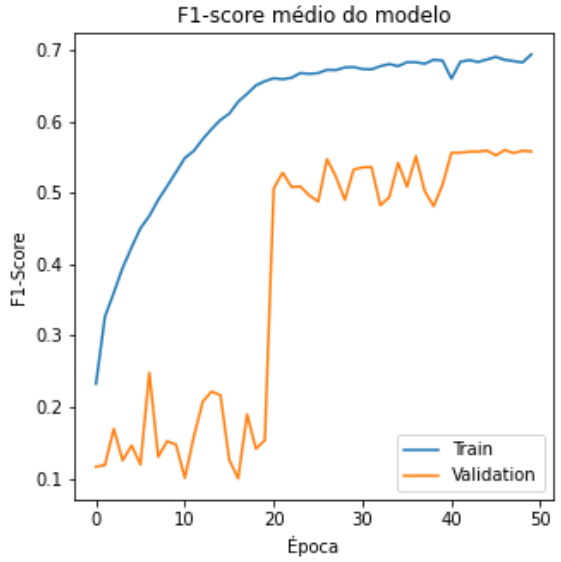

Figura 5. Gráfico do F1-score do modelo SGD com Nesterov (média aritmética entre as classes) de treino e validação pelas épocas - durante o treinamento.

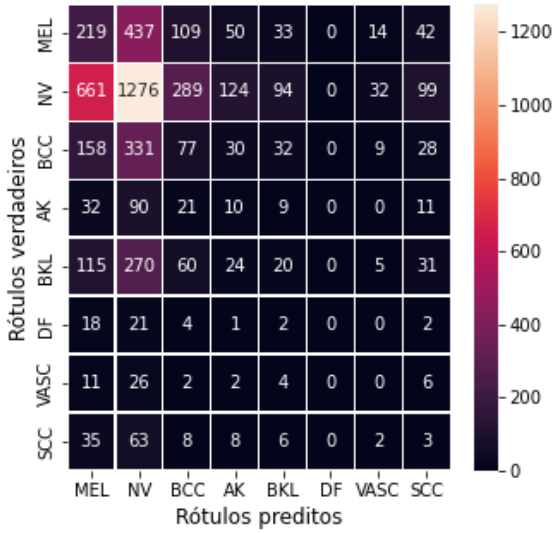

Figura 6. Gráfico da matriz de confusão do modelo SGD com Nesterov para o conjunto de validação.

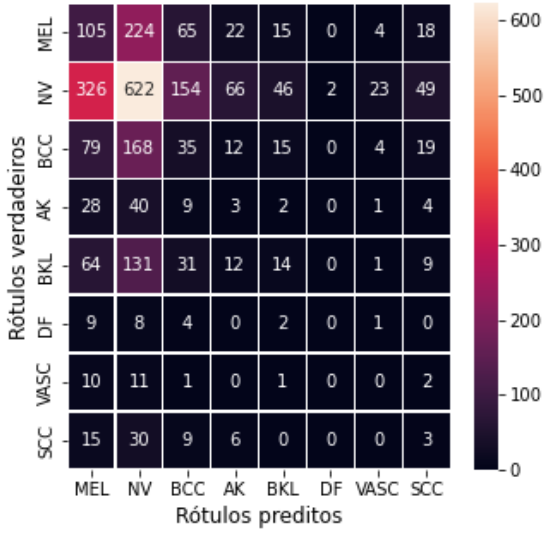

Figura 7. Gráfico da matriz de confusão do modelo SGD com Nesterov para o conjunto de teste.

Métricas do modelo treinado com otimizador Adam

Figura 8. Gráfico da perda do modelo treinado com otimizador Adam - sem cortes - ao longo do treinamento. Dado o valor de perda de 1420,41 durante a época 10 se fez necessário o corte para melhor visualização.

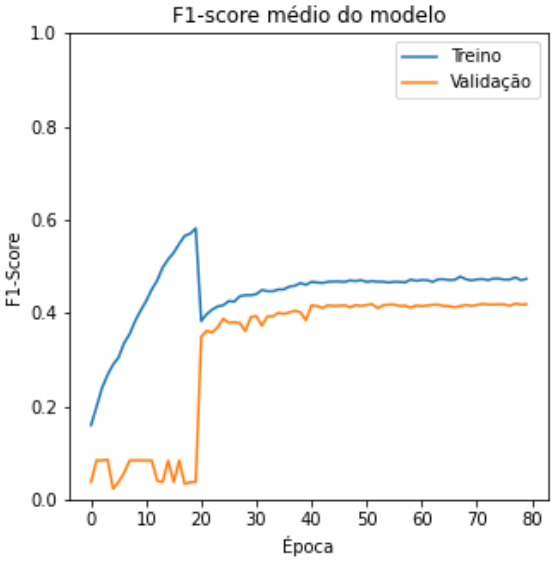

Figura 10. Gráfico do F1-score do modelo Adam (média aritmética entre as classes) de treino e validação pelas épocas - durante o treinamento.

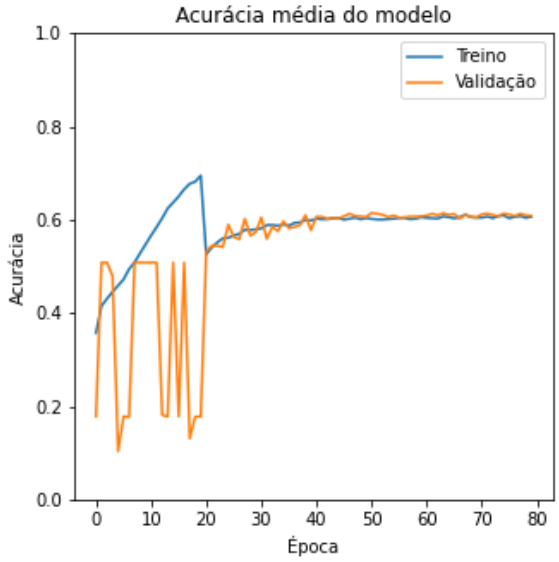

Figura 9. Gráfico da acurácia do modelo treinado com Adam (média aritmética entre as classes) de treino e validação pelas épocas - durante o treinamento.

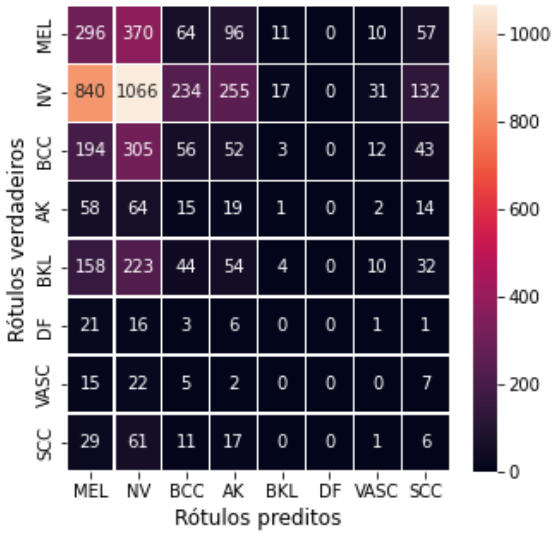

Figura 11. Gráfico da matriz de confusão do modelo Adam para o conjunto de validação.

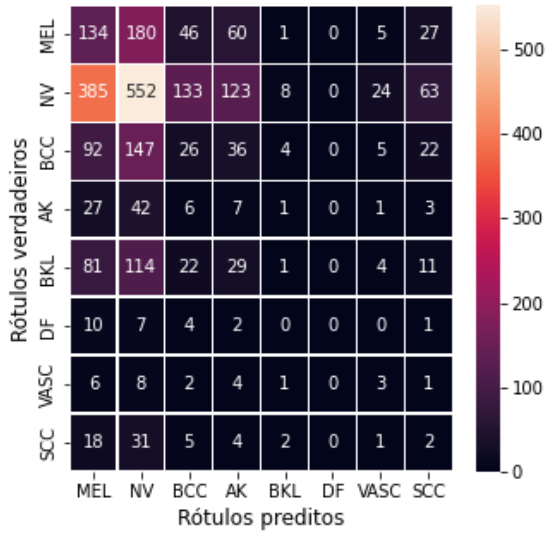

Figura 12. Gráfico da matriz de confusão do modelo Adam para o conjunto de teste.