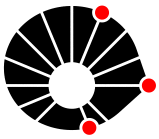
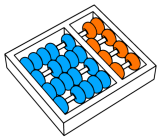




Análise de Modelos Baseados em Redes Neurais Profundas para Lesões de Pele Negra

Luana Felipe de Barros *Sandra Eliza Fontes de Avila*

Relatório Técnico – IC-PFG-21-33
Projeto Final de Graduação
2022 – Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Sumário

Análise de Modelos Baseados em Redes Neurais Profundas para Lesões de Pele Negra

Luana Felipe de Barros* Sandra Eliza Fontes de Avila†

Resumo

O melanoma é o tipo mais grave de câncer de pele devido a sua alta capacidade de provocar metástase. É mais comum em pessoas negras, acometendo frequentemente regiões acrais: palmoplantares, extremidades digitais, mucosas e semi-mucosas. As redes neurais profundas têm mostrado um enorme potencial para melhorar o atendimento clínico e o diagnóstico de câncer de pele. Utilizá-las poderia melhorar o acesso a serviços de dermatologia no Brasil, que são muito desiguais. Atualmente, os estudos realizados utilizam bases de dados predominantemente brancas e não reportam resultados dos diagnósticos em uma diversidade de tons de pele de pacientes. Portanto, neste trabalho apresentamos uma avaliação de modelos supervisionados e autosupervisionados em bases de dados com lesões de pele localizadas em regiões acrais, que são prevalentes em pessoas negras, além de apresentar uma base de dados com lesões de pele em regiões acrais. Ainda, avaliamos bases de dados em relação à Escala Fitzpatrick para verificar o desempenho em pele negra. Atualmente, esses modelos não podem ser utilizados de forma generalista, pois tem bons resultados apenas em lesões em pele branca. A criação de modelos específicos devido a uma negligência na criação de bases de dados diversas é inaceitável. As redes neurais profundas têm grande potencial para melhorar o diagnóstico principalmente para populações com menos acesso a dermatologia, mas a inclusão de lesões de pele com características menos comuns é extremamente necessária para que essas populações tenham acesso aos benefícios de uma tecnologia inclusiva.

1 Introdução

O câncer de pele é o mais comum entre todos os tipos de cânceres. O melanoma é um câncer de pele raro, representando apenas 3% dos casos de câncer de pele no Brasil [30]. No entanto, é o tipo mais agressivo, causando cerca de 60% das mortes pela doença devido à alta probabilidade de provocar metástase. Por isso, o diagnóstico precoce tem sido fundamental para melhorar as taxas de sobrevida de pacientes e garantir um bom prognóstico.

O melanoma é cerca de 20 vezes mais comum em pessoas de pele branca [7]. Acredita-se que, em comparação com pessoas de pele negra, essa diferença se deve à fotoproteção

*Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP.
†Instituto de Computação, Universidade Estadual de Campinas, Campinas, SP.

conferida pela melanina na pele de pigmentação escura [24]. Contudo, a melanina não protege o indivíduo completamente de desenvolver câncer de pele. Inclusive, o melanoma acral (subtipo de melanoma com características próprias que ocorre nas regiões plantares, palmares e subungueais) é o tipo mais raro e mais agressivo de melanoma e costuma se desenvolver com maior intensidade na população negra, cerca de 70% dos casos são entre afrodescendentes [1].

Além disso, o Brasil possui uma incidência de melanoma acral, ou acrolentiginoso, superior à média mundial, muito provavelmente devido à miscigenação da população. O melanoma acral não está relacionado à exposição solar, já que costuma se desenvolver nas palmas das mãos, plantas dos pés e nas unhas [29]. Logo, os pacientes de pele negra enfrentam um prognóstico ruim, com o aumento da morbidade e mortalidade, que geralmente é resultado do diagnóstico tardio nessa população [24]. Além disso, as regiões acrais, especialmente os pés, são muitas vezes negligenciadas por dermatologistas nas avaliações físicas [17].

Esse problema pode estar relacionado ao viés de que a pigmentação escura é completamente protetora do câncer de pele por parte dos médicos e a população. Ainda, o diagnóstico tardio do melanoma em pele negra indica que essa população requer mais atenção da dermatologia. Atualmente, a maioria dos livros didáticos que servem como roteiros para o diagnóstico de doenças de pele não incluem imagens de doenças de pele como aparecem em pessoas negras, ou quando incluem, esse número não passa de 10% [37]. Isso causa severos problemas no diagnóstico de doenças em pele negra, visto que uma mesma lesão pode ter características diferentes dependendo da cor da pele. Um exemplo disso é a localização do melanoma em pele negra ser predominante em regiões acrais, que por não serem expostas ao sol, podem passar despercebidas pelos dermatologistas no momento da avaliação do paciente levando a diagnósticos errôneos. Portanto, é comum que o melanoma seja confundido pelo paciente com micoses, machucados ou outras condições benignas [29].

Isso ocorre porque dermatologistas são treinados a reconhecer os padrões que são utilizados a classificar uma lesão. No entanto, como esse é um processo visual, o viés racial está acoplado neste padrão porque a cor da pele ao redor da lesão impacta também na cor e características da lesão. Logo, se um dermatologista apenas aprende a diagnosticar uma lesão pelas suas características em pele branca, pode não conseguir classificá-la em pele negra já que os padrões podem ser diferentes [37].

Portanto, a falta de diversidade de exemplos de lesões em pele negra na formação médica pode causar consequências graves para a população negra, o que é um grande problema principalmente para o Brasil em que 54% da população é negra [36]. Esse fato pode ser piorado ao considerar a formação de médicos não especialistas em dermatologia: estes são responsáveis por cerca de 60% dos atendimentos de pacientes com afecções de pele, sendo 90% deles não são adequadamente diagnosticados [26]. Ainda, as hipóteses de diagnóstico feitas por médicos generalistas e dermatologistas divergem — esse fato levou a uma mudança no currículo da Faculdade de Ciências Médicas da Universidade Estadual de Campinas (Unicamp), com a inclusão de temas teóricos de dermatoses prevalentes, como tumores benignos, onicopatias, discromias, orientações e tratamentos básicos na dermatologia [9]. Logo, se existe uma dificuldade de diagnóstico em câncer de pele negra

por parte dos dermatologistas, é esperado que este cenário piore por parte de médicos generalistas.

Além disso, a distribuição dos dermatologistas é desigual no Brasil, visto que 63,5% deles se concentram na Região Sudeste, que corresponde a 41,6% da população brasileira [26]. Consequentemente, regiões remotas possuem dificuldade de acesso a consultas especializadas em dermatologia. Todas essas questões contribuem para um diagnóstico tardio.

Por conta destas dificuldades, algumas iniciativas vêm sendo implementadas, como a teledermatologia e o uso de sistemas de informação baseados em inteligência artificial (IA) para auxiliar nos diagnósticos das afecções de pele. Como o diagnóstico é baseado em aspectos visuais, o uso de algoritmos de aprendizado profundo para analisar padrões em imagens dermatoscópicas vêm se mostrando promissores. As redes neurais, por exemplo, tornaram-se rapidamente uma metodologia de escolha para a análise de imagens já que esta reduz a tarefa de engenharia de características, que é muito importante para a análise automática de vários tipos de imagens, incluindo imagens médicas. Embora esta técnica tenha resultados interessantes, com alta acurácia, ainda existem desafios a serem enfrentados, como conjuntos de dados com poucas imagens de lesões de pele e viés racial.

Daneshjou et al. [21] realizaram uma revisão de estudos nos quais bases de dados de imagens na dermatologia estavam sendo utilizadas por algoritmos de IA, com o intuito de avaliar se as bases foram descritas de forma adequada e identificar possíveis fontes de viés. Mostrou-se que, dentre 70 estudos, apenas 20% dos estudos incluíram descrições da etnia ou raça do paciente e apenas 10% incluíram informações sobre o tom de pele dos pacientes, em pelo menos 1 conjunto de dados utilizado. Além disso, dentre os estudos que envolviam o diagnóstico de câncer de pele, apenas 64% atenderam o padrão ouro na anotação dos rótulos das lesões. A pesquisa ainda ressaltou falta de transparência na descrição dos dados, como a falta de informações de diversidade sobre os pacientes, e padronização e confiabilidade dos rótulos das imagens.

A capacidade de generalização de modelos de AI para resolução de problemas reais depende fortemente da diversidade, quantidade e qualidade dos dados nos quais foram treinados. Portanto, é necessário entender o desempenho dos modelos no diagnóstico de lesões voltadas para a população negra. Entretanto, devido a falta de informação relacionada ao tom de pele dos pacientes na maioria das bases de dados disponíveis, optamos por realizar uma avaliação do desempenho de alguns modelos — que têm alcançado bons resultados no diagnóstico de lesões de pele — em regiões acrais, já que o melanoma acral é predominante nessas regiões em pessoas não brancas [24].

O objetivo deste trabalho consiste em avaliar modelos baseados em redes neurais profundas em lesões de pele benignas e malignas localizadas em regiões acrais: palmas das mãos, solas dos pés e unhas. Enfatizamos que, ao início deste trabalho, não encontramos bases de dados com informações sobre a diversidade racial dos pacientes. No entanto, recentemente foi disponibilizada uma base de dados chamada *Diverse Dermatology Images* (DDI) [22] com lesões de pele comprovadas por biópsia com diversas representações de tons de pele. Portanto, incluímos-a na avaliação de forma isolada no final do estudo, sendo considerada necessária devido ao avaliar os desempenhos dos modelos de forma mais direta em relação aos tons de pele dos pacientes.

As principais contribuições deste trabalho são:

- Construção de uma base de dados com imagens clínicas e dermatoscópicas de lesões de pele localizadas em regiões acrais;
- Avaliação de modelos baseados em redes neurais profundas previamente treinados de forma auto-supervisionada e supervisionada para diagnóstico de melanoma e lesões benignas nos seguintes conjuntos:
  - Base de dados proposta (imagens selecionadas de bases de dados existentes considerando regiões acrais)
  - *Diverse Dermatology Images* (DDI) [22]

O restante do texto está organizado da seguinte forma. Na Seção 2, os fundamentos principais do trabalho são abordados brevemente. Na Seção 3, é descrita a metodologia proposta, incluindo as bases de dados (ISIC Archive, Derm7pt, PAD-UFES-20, Atlas Dermatológicos) e o *pipeline* de avaliação. Na Seção 4, são apresentados os resultados para os modelos supervisionados e auto-supervisionados. Na Seção 5, a base de dados *Diverse Dermatology Images* (DDI) é descrita e os resultados. Na Seção 6, são apresentadas as conclusões e direções futuras.

2 Fundamentação Teórica

As redes neurais profundas (*deep neural networks*, DNNs) são o estado da arte para a análise de lesões de pele. Muitos algoritmos médicos se baseiam em padrões visuais para classificação de lesões de pele, como a regra ABCD (Assimetria, Borda, Cor e Diâmetro das lesões) [16] e a regra de 7 pontos [8]. Devido a grande capacidade de extração de padrões de forma automatizada das DNNs, elas estão sendo utilizadas em diversas tarefas de visão computacional. Análises de lesões de pele estão sendo beneficiadas dessa tecnologia em diversas frentes como classificação [10, 19, 33, 34, 40] e segmentação de lesões [38], análise de vieses [12, 13, 15] e expansão de dados [14] através da síntese de imagens [11].

O treinamento de DNNs requer uma grande quantidade de dados para alcançar uma boa extração de padrões. Muitas DNNs utilizam a estratégia de aprendizado supervisionado (*supervised learning*, SL). Nesta abordagem, a rede aprende a extrair padrões com supervisão de dados anotados. Logo, se queremos uma rede capaz de classificar uma lesão como benigna ou melanoma, devemos realizar o processo de treinamento do modelo. Para tal, um conjunto de imagens e suas respectivas classes são passadas como entrada para a DNN. De forma automatizada, padrões serão extraídos a fim de diferenciar as duas classes. Ao final, espera-se que o modelo possa ser utilizado para classificar imagens ainda não vistas no processo de treinamento, ou seja, que tenha adquirido a capacidade de generalização do problema baseado na qualidade e quantidade das amostras.

No entanto, o processo de coleta e anotação de dados é caro. Consequentemente, na área médica temos escassas bases de dados anotadas. Por isso, a técnica de *transfer learning* [33] vem sendo utilizada em *deep learning*. Essa técnica consiste em pré-treinar uma

rede neural de forma supervisionada em um grande conjunto de dados não relacionado à tarefa de interesse, como forma de criar um bom extrator de padrões, e posteriormente, realizar um ajuste fino (*finetuning*) com uma rede de classificação em um menor conjunto de dados de interesse.

Normalmente, a base de dados ImageNet [23] é utilizada para pré-treinar os modelos, já que possui 1.2 milhões de imagens e 1000 categorias diversas. Apesar desta técnica apresentar resultados promissores, ainda existe o risco que as representações aprendidas pela rede não se adaptem adequadamente à tarefa de interesse. Por conta disso, uma nova estratégia chamada aprendizado auto-supervisionado (*self-supervised learning*, *SSL*) [19] tem se mostrado uma alternativa a esse problema.

Na SSL, também utiliza-se a técnica de *transfer learning* para pré-treinar um modelo que consiga extrair boas representações em um conjunto de dados maior e depois é feito um *finetuning* em um conjunto menor. No entanto, a rede é treinada de forma auto-supervisionada, ou seja, sem a supervisão de dados anotados. Em vez disso, a mesma é treinada para realizar alguma tarefa *pretexto* com alguma anotação sintetizada. Por exemplo, podemos aleatoriamente rotacionar imagens da ImageNet [23] e treinar a rede para predizer os ângulos de rotação dessas imagens [25]. A tarefa pretexto é apenas utilizada com o objetivo de estimular a rede a criar transformações nas imagens e aprender as melhores representações no espaço de características que as descrevem. Desta forma, temos uma rede extratora de características poderosa que pode ser utilizada em alguma outra tarefa de interesse.

Chaves et al. mostraram que o uso de modelos auto-supervisionados no problema de classificação de lesões de pele é benéfico, especialmente em cenários com escassez de dados de treino. Portanto, neste trabalho avaliamos seis modelos treinados pelos autores para classificação de lesões de pele em regiões acrais, nas classes melanoma e benigno. Em todas abordagens, sendo uma supervisionada e cinco auto-supervisionadas, utilizamos a rede ResNet-50 como a rede extratora de características (*backbone*). As abordagens auto-supervisionadas diferenciam-se justamente em técnicas relacionadas à escolha da tarefa pretexto, sendo elas: BYOL (*Bootstrap Your Own Latent*) [27], InfoMin [39], MoCo (*Momentum Contrast*) [28], SimCLR (*Simple Framework for Contrastive Learning of Visual Representations*) [20], e SwAV (*Swapping Assignments Between Views*) [18].

3 Metodologia

3.1 Base de Dados

Para construir a base de dados utilizada neste trabalho, iniciou-se uma busca por bases de dados e atlas dermatológicos na Internet que possuíam anotação sobre a localização da lesão. Analisamos 17 bases de dados, listadas em `https://www.medicalimageanalysis.com/research/skinia`. Posteriormente, filtramos bases nas quais haviam imagens em que a lesão era localizada em regiões acrais: palmas das mãos, solas dos pés e unhas. Com isso, restaram-se três bases muito conhecidas na literatura — *International Skin Imaging Collaboration* (ISIC *Archive*) [6], 7-*Point Checklist Dermatology Dataset*

(Derm7pt) [31], e PAD-UFES-20 [35] —, e três atlas dermatológicos — *Interactive Dermatology Atlas* (dermatlas.net) [2], DermIS [3] e dermnetNZ [4]. Como o nosso objetivo é avaliar um classificador de lesões como benignas ou melanoma, além do critério de localização, procuramos imagens com lesões benignas e excluímos lesões malignas que não fossem melanoma. Detalhes da base construída podem ser vistos na Tabela 1.

Algumas amostras das bases de dados podem ser observadas na Figura 1.

Após a obtenção das imagens, fizemos a organização das imagens separando-as em pastas referentes à base e à classe do diagnóstico: benigno ou melanoma. Além disso, as bases passaram por um processo de limpeza como a padronização de anotação da localização da lesão e o nome da lesão. Ainda, houve a construção de variáveis indicadoras de informações, como o tipo da imagem (clínica ou dermatoscópica) e o rótulo do diagnóstico (melanoma ou benigno).

Criamos um arquivo CSV de metadados contendo informações comuns a todas as bases, já que elas continham informações diferentes disponibilizadas. O arquivo possui informações como nome da imagem, localização da lesão, diagnóstico da lesão, dados sobre o tipo da imagem, a base de dados e rótulo da imagem.

Apresentamos na sequência os passos realizados para obtenção de dados em cada base e atlas.

3.1.1 ISIC Archive

Para coletar de dados nesta base, acessamos o ISIC Archive [6] e aplicamos os seguintes filtros:

1. **Atributos clínicos**: Filtramos as imagens pela localização da lesão, selecionando palmas e solas (773 imagens resultantes);

2. **Atributos de diagnóstico da lesão**: Das 773 imagens resultantes, descartamos as classificadas como carcinoma ou desconhecidas, restando 400 imagens de lesões benignas e melanoma na palma da mão e sola do pé;

3. Como Chaves et al. treinaram modelos utilizando dados do ISIC [6], excluímos as imagens utilizadas em seu conjunto de treinamento, totalizando 149 imagens. Esse passo é necessário para impedir que os resultados sejam erroneamente otimistas devido a uma contaminação de dados de treino e teste.

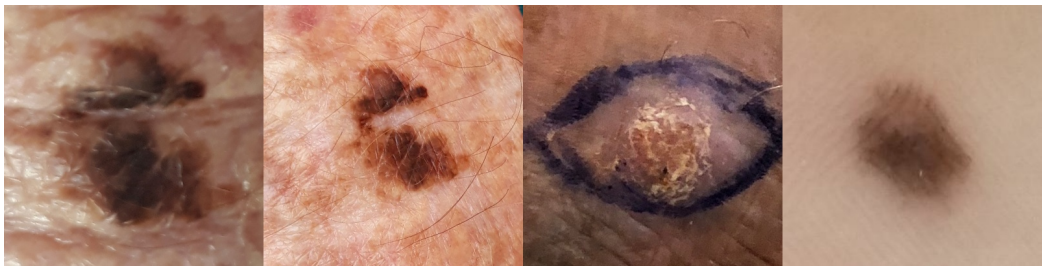
3.1.2 Derm7pt

A base de dados Derm7pt [31] é largamente conhecida na literatura, com 1011 imagens para cada lesão (uma versão clínica e uma dermatoscópica). Além disso, a Derm7pt apresenta metadados interessantes, como presença ou ausência de padrões visuais, localização da lesão, sexo do paciente, nível de dificuldade da lesão, pontuação na regra dos 7 pontos.

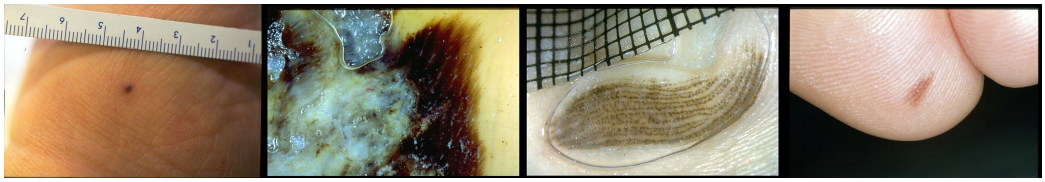
Avaliamos imagens clínicas e dermatoscópicas de forma independente, nomeando o conjunto de dados como *derm7pt-clinic* e *derm7pt-derm*, respectivamente.

Tabela 1: Descrição e número de lesões benignas e melanoma de cada base de dados utilizadas neste trabalho.

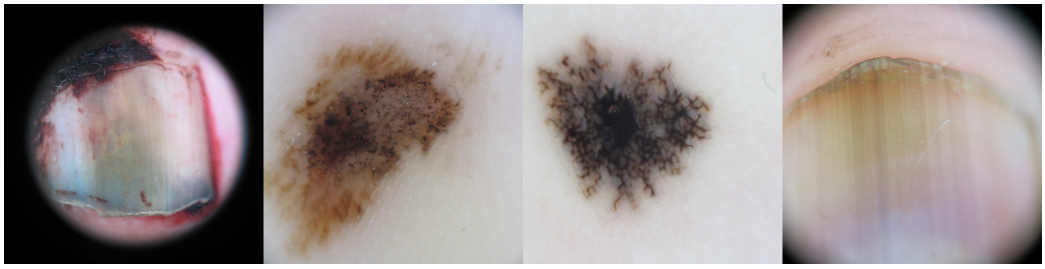
Base de Dados	#Total	#Mel.	Diagnósticos (benignos)	Localização	Tipo
ISIC	149	72	nevus, lesão vascular, lentigo, ceratose seborreica	palma, sola	dermato
derm7pt-clinic	62	3	nevus, lesão vascular	acral	clínica
derm7pt-derm	62	3	nevus, lesão vascular	acral	dermato
PAD-UFES-20	98	2	ceratose actínica, ceratose seborreica	mão, pé	clínica
dermatlas.net	8	1	nevus, lentigo, queratoacantoma	dedo, mão, pé	clínica
DermIS	12	10	linfagioma	dedo, pé	clínica
DermnetNZ	34	34	–	pé, unha	clínica e dermato



a) PAD-UFES-20



b) Derm7pt



c) ISIC Archive

Figura 1: Imagens de Lesões de Pele extraídos das bases PAD-UFES-20 [35], Derm7pt [31] e ISIC Archive [6]. Pode-se observar a presença de artefatos, como réguas, gel, pelos e bordas escuras nas imagens.

Para coletar imagens da base Derm7pt, filtramos imagens pela localização da lesão, que nesta base era equivalente ao atributo “acral”, totalizando 62 imagens. Esse filtro foi suficiente para que obtivéssemos apenas lesões benignas ou melanoma. Na Tabela 2, mostramos a quantidade de lesões por dificuldade do diagnóstico (baixo, médio, ou alto) pela classe da lesão (melanoma ou benigna).

Tabela 2: Número de lesões agrupadas pelo nível de dificuldade para cada classe para a base de dados Derm7pt [31].

	Número de Imagens	
Dificuldade	Melanoma	Benigna
Baixo	1	31
Médio	0	21
Alto	2	7
Total	3	59

3.1.3 PAD-UFES-20

A base de dados PAD-UFES-20 [35] é composta por 2298 imagens clínicas de pacientes coletadas por smartphones. Além disso, traz como parte dos metadados a Escala Fitzpatrick: uma classificação dos fototipos cutâneos de 1 a 6, a partir da capacidade de cada pessoa em se bronzear e a sensibilidade e vermelhidão quando exposta ao sol [5].

O critério de filtragem das imagens foi a princípio o atributo de localização da lesão: ”mão e pé”, restando 142 imagens. Em seguida, excluímos as imagens em que as lesões eram classificadas como carcinoma (malignas), restando 98 imagens. A Tabela 3 mostra o número de imagens correspondente à escala de Fitzpatrick no conjunto de dados filtrado.

3.1.4 Atlas Dermatológicos

A base de dados também foi composta por imagens de atlas dermatológicos obtidas na Internet, como o dermatlas.net [2], DermIS [3] e DermNetNZ [4]. Nestes sites, filtramos imagens pela localização da lesão utilizando os seguintes termos de busca: *hand* (mão), *hands* (mãos), *foot* (pé), *feet* (pés), *acral* (acral), *finger* (dedo), *nail* (unha), *nails* (unhas). Selecionamos manualmente as imagens que atendiam o diagnóstico da lesão ser melanoma ou alguma lesão benigna.

Imagens obtidas por atlas dermatológicos são minoria da construção da nossa base de dados (veja Tabela 1), visto que as imagens não são padronizadas e não tem confiabilidade garantida na anotação das lesões das imagens.

Tabela 3: Quantidade de imagens em relação à escala Fitzpatrick dentre as 98 imagens filtradas da base PAD-UFES-20. É importante ressaltar que não possuímos imagens referentes às escalas 5 e 6.

	Número de Imagens	
Escala	Melanoma	Benigna
1	2	4
2	0	25
3	0	12
4	0	1
5	0	0
6	0	0
Desconhecido	0	54
Total	2	96

3.2 *Pipeline* de Avaliação

O pipeline para avaliação dos modelos de classificação de imagens de lesões de pele está dividido em duas etapas principais: pré-processamento e inferência do modelo. O pipeline é representado na Figura 2.

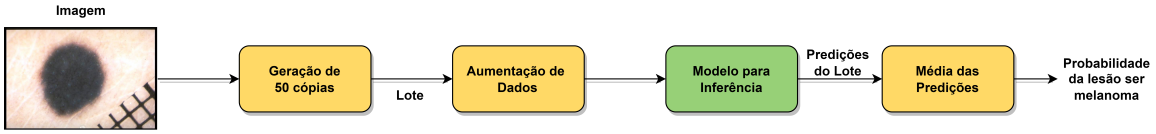


Figura 2: *Pipeline* de avaliação dos seis modelos para as bases de dados citadas anteriormente.

A etapa de **pré-processamento** consiste na aplicação da técnica de aumentação de dados (*data augmentation*) nos dados de teste, que tem se mostrado benéfica para melhorar o desempenho de problemas de classificação [40]. Para isso, avaliamos o conjunto de teste em lotes/*batches*. Logo, para cada imagem do conjunto foi criado 50 cópias, que caracterizada um *batch*, em que transformações variadas foram aplicadas, como: redimensionamento, espelhamento, rotações, mudanças de cores e também uma normalização utilizando a média e desvio padrão da ImageNet, no qual os modelos foram pré-treinados, a fim de manter a coerência com as transformações aplicadas nos dados de treino.

Após esse pré-processamento, temos a etapa de **inferência do modelo**. Nesta etapa, o *batch* foi passado para ser avaliado por um modelo. Obtemos, então, representações/*features* que dependem do modelo utilizado. Essas representações foram então passadas para uma camada *softmax*, em que obtemos um vetor de probabilidades que indica a probabilidade de cada cópia do *batch* ser classificada como melanoma e benigna. Com isso, obtemos o valor da probabilidade da lesão ser melanoma — já que é a classe de interesse/positiva —, e obtemos a média das probabilidades obtidas dentre todas as 50 cópias. O valor obtido

é a confiança na predição do modelo em relação à classe melanoma.

Este processo de avaliação foi feito para cada base de dados, para cada um dos 6 modelos: BYOL, InfoMin, MoCo, SimCLR, SwAV e Supervised. Após a obtenção da probabilidade da lesão ser melanoma para todas as imagens de uma base, foram calculadas as métricas: acurácia balanceada, precisão (*precision*), revocação (*recall*) e F1-Score (média harmônica entre *precision* e *recall*). Para calcular a acurácia balanceada, foi considerado um limiar de 0,5. Isso indica que se uma predição tiver probabilidade igual ou maior que 0,5, consideramos a predição como melanoma.

4 Resultados

Nesta seção, apresentamos os resultados das métricas obtidas no pipeline de avaliação. É importante ressaltar que as respectivas bases de dados dermatlas.net, DermIS e DermnetNZ possuem uma quantidade escassa de imagens. Portanto, não foram consideradas na avaliação devido à dificuldade para aplicar as métricas.

As métricas foram obtidas ao se avaliar cada base de dados em cada um dos seis modelos. Apresentamos os resultados das métricas na Tabela 4. Utilizamos a acurácia balanceada devido ao problema de classes desbalanceadas na maior parte dos conjuntos de teste, sendo essa métrica a média da acurácia individual de cada classe. Além disso, dentre todas as classificações de classe positiva/melanoma que o modelo fez, *precision* mede quantas estão corretas. A métrica *recall*, por sua vez, mede, dentre todas as situações de classe positivo/melanoma como valor esperado, quantas estão corretas; positivas/melanoma existentes. Por fim, a métrica F1-Score é a média harmônica de *precision* e *recall*.

Para a base de dados **ISIC**, temos um resultado muito coerente entre acurácia balanceada e F1-Score, em torno de 86% para ambas as métricas. É importante que, ao avaliar um modelo de classificação, a distribuição dos dados de teste seja idealmente a mesma dos dados de treino. Como os modelos foram treinados utilizando a técnica de ajuste fino (*fine-tuning*) na base ISIC 2019, a distribuição dos dados da base de dados ISIC Archive é a mais parecida com as de treino, em relação as outras bases, mesmo que considerando conjuntos disjuntos de treino, validação e teste. Além disso, é interessante observar que na base ISIC 2019 em que os modelos foram treinados, todos os resultados foram superiores a 90% [19]. Isso indica que, mesmo com uma base externa de distribuição mais semelhante a de treino, o resultado com lesões em regiões acrais é muito inferior a outras regiões.

Para as bases **derm7pt-clinic** e **derm7pt-derm**, temos os resultados das métricas considerando as mesmas lesões diferenciando o tipo da imagem: clínica e dermatoscópica. Ao analisar os resultados de F1-Score para as imagens clínicas, vemos que os modelos SwAV, BYOL e Supervisionado erraram todas as classificações da classe melanoma, com *precision* e *recall* iguais a zero. Esse cenário é o mais crítico no contexto de diagnóstico médico, pois acarreta em lesões graves e malignas que não seriam investigadas. Podemos observar que, em geral, as imagens dermatoscópicas tiveram um desempenho melhor comparado às clínicas. Isso se relaciona ao fato dos modelos terem sido treinados com imagens dermatoscópicas da base ISIC 2019. Portanto, é possível notar que o dispositivo de coleta da imagem (dermatoscópio, câmera de celular), pode atenuar ou suavizar

Tabela 4: Métricas obtidas no pipeline de avaliação.

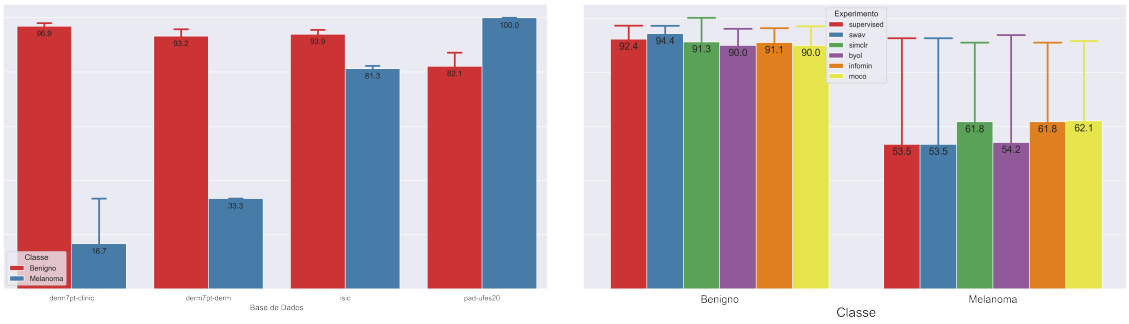
Base de Dados (#Mel./#Benigna)	Modelo	Acurácia Balanceada	*Precision*	*Recall*	F1-Score
ISIC (72/77)	SwAV	88.3	95.1	80.6	87.2
	MoCo	87.1	90.8	81.9	86.1
	SimCLR	88.3	95.1	80.6	87.2
	BYOL	88.4	92.3	83.3	87.6
	InfoMin	86.4	90.6	80.6	85.3
	Supervised	87.0	92.1	80.6	85.9
	Média	87.6	92.7	81.3	86.6
derm7pt-derm (3/59)	SwAV	62.4	16.7	33.3	22.2
	MoCo	63.3	20.0	33.3	25.0
	SimCLR	65.8	50.0	33.3	40.0
	BYOL	61.6	14.3	33.3	20.0
	InfoMin	63.3	20.0	33.3	25.0
	Supervised	63.3	20.0	33.3	25.0
	Média	63.3	23.5	33.3	26.2
derm7pt-clinic (3/59)	SwAV	49.2	0.0	0.0	0.0
	MoCo	65.0	33.3	33.3	33.3
	SimCLR	64.1	25.0	33.3	28.6
	BYOL	48.3	0.0	0.0	0.0
	InfoMin	65.0	33.3	33.3	33.3
	Supervised	49.2	0.0	0.0	0.0
	Média	56.8	15.3	16.6	15.9
PAD-UFES-20 (2/96)	SwAV	95.8	20.0	100.0	33.3
	MoCo	89.1	8.7	100.0	16.0
	SimCLR	88.0	8.0	100.0	14.8
	BYOL	90.1	9.5	100.0	17.4
	InfoMin	91.1	10.5	100.0	19.0
	Supervised	92.2	11.8	100.0	21.1
	Média	91.1	11.4	100.0	20.3

os detalhes das lesões, impactando na qualidade da imagem e consequentemente a quantidade de informação agregada, influenciando assim a distribuição dos dados capturados. Logo, como os modelos foram treinados com imagens dermatoscópicas, imagens de teste que foram capturadas por um dermatoscópio têm distribuição mais próxima às imagens de treino pela natureza do processo de captura. Em suma, as métricas para essa base mostraram F1-Score e acurácia balanceada muito baixos, sendo um resultado insatisfatório,

especialmente em imagens clínicas.

Ao analisar os resultados para a base **PAD-UFES-20**, é possível notar que a acurácia balanceada foi muito boa, acima de 90%. No entanto, analisando o F1-Score, percebemos que está muito baixo porque a *precision* está baixa e o *recall* alto. Ou seja, em geral, os modelos classificaram muitas imagens como melanoma, quando eram benignas. Além disso, o *recall* em 100% induz um aumento na acurácia da classe positiva. Mas, é importante ressaltar que esta só possui 2 amostras, como visto na descrição das bases de dado na Tabela 1.

Nas Figuras 3a e 3a, apresentamos a acurácia balanceada agrupada por classe (benigno ou melanoma) para cada base de dados e modelos. Vemos que, em geral, os modelos obtiveram maiores taxas de acertos da classe benigna que melanoma. Esse resultado já era esperado, visto que imagens de melanoma em regiões acrais são mais raras e possuem padrões visuais diferentes dos vistos nas imagens de treinamento.



a) Acurácia por classe em cada base de dados

b) Acurácia por classe em cada modelo

Figura 3: Acurácia por classe separadas por base de dados e modelo, respectivamente. As barras de erro indicam o desvio padrão.

Para entender melhor o comportamento dos modelos, obtemos também métricas agrupadas por características dos metadados das bases de dados em relação aos modelos ou bases de dados.

4.1 Nível de Dificuldade de Diagnóstico de uma Lesão

A base de dados Derm7pt possui um metadado referente ao nível de dificuldade do diagnóstico de uma lesão definido por médicos, tais como baixo, médio e alto. A Tabela 5 apresenta a média e desvio padrão das métricas acurácia balanceada, *precision*, *recall* e F1-Score agrupadas pelo nível da dificuldade, mostradas separadamente para cada modelo. Ao dividir a base pelo nível de dificuldade, não possuíamos amostras de melanoma de dificuldade média, como mostrado na Tabela 2, impedindo o cálculo de *precision*, *recall* e F1-Score. Para essas lesões, a acurácia mostrada corresponde a acurácia da classe benigna.

Ao analisar a Tabela 2, vemos que a maior parte das imagens consideradas fáceis de diagnosticar pelos médicos eram benignas e as imagens difíceis eram, em sua maioria, lesões de melanoma. Como mostrado na Tabela 5, os modelos resultaram uma média de 86.7%

Tabela 5: Média e desvio padrão das médias referentes ao nível de dificuldade de diagnóstico da lesão calculadas nos modelos.

Dificuldade	Acurácia Balanceada	*Precision*	*Recall*	F1-Score
Baixo	86.7 ± 13.7	70.8 ± 27.7	75.0 ± 27.3	67.8 ± 18.1
Médio	47.0 ± 1.2	0 ± 0	0 ± 0	0 ± 0
Alto	41.7 ± 2.9	0 ± 0	0 ± 0	0 ± 0

acertos em lesões com dificuldade baixa de diagnóstico, 47% acertos em lesões benignas com dificuldade média. Para as imagens com dificuldade alta, os modelos tiveram uma taxa de acerto de 41.7%, considerada insuficiente por ser abaixo de 50%. Esses resultados indicam que os modelos reproduziram o comportamento dos médicos em lidar com a complexidade do problema.

4.2 Escala Fitzpatrick

A base de dados PAD-UFES-20 possui um metadado referente à escala Fitzpatrick, que é uma classificação dos fotótipos de pele, que depende do bronzeamento de pele de uma pessoa. A escala varia de 1 (peles mais claras) a 6 (peles mais escuras). Agrupamos as imagens de acordo com essa escala e calculamos a acurácia por classe mostradas na Tabela 6. Como nem todos os grupos de fotótipos possuíam lesões de melanoma, não foi possível o cálculo das outras métricas.

Para analisar os resultados, é importante notar que, segundo a Tabela 3, apenas o fototipo 1 da escala Fitzpatrick possui imagens de melanoma. Esse desbalanceamento é prejudicial para reportar os resultados, pois a distribuição das amostras é muito diferente das de treinamento e pode gerar interpretações errôneas ao não se analisar a quantidade de imagens testadas. Por exemplo, o fototipo 1 e 4 apresentaram taxa de 100% de acerto nas imagens benignas pois havia uma quantidade escassa de imagens neste grupo.

Portanto, ao analisar a Tabela 6 vemos que não podemos inferir os resultados em relação à classe melanoma, visto que não existe um número suficiente de amostras. Em geral, podemos constatar que para lesões benignas, os modelos tiveram um bom desempenho — em torno de 80%, considerando principalmente grupos mais confiáveis para análise devido a maior quantidade de amostras, como os fotótipos 2 e 3.

Infelizmente, a maior parte dos dados desta base não tem marcação de Fitzpatrick, o que impossibilita uma análise mais assertiva. Por conseguinte, não podemos analisar grupos que seriam o foco da análise do trabalho, com tons de pele mais escuros com fotótipos 5 e 6 da escala.

Tabela 6: Acurácia por classe referentes a cada fototipo de pele presente na base PAD-UFES-20, calculadas pelos modelos.

Escala Fitzpatrick	Benigno	Melanoma
1	66.7 ± 25.8	100.0 ± 0.0
2	77.3 ± 10.9	–
3	88.9 ± 4.3	–
4	100.0 ± 0.0	–

5 Análise Direta: *Diverse Dermatology Images*

A base de dados *Diverse Dermatology Images* (DDI) [22] foi criada intencionalmente para suprir os desafios atuais dos conjuntos de dados disponíveis publicamente, como a escassez de lesões de pele em tons de pele diversificados e comprovadas por biópsia [21]. Portanto, nesta base foi realiza uma curadoria especializada das lesões dos pacientes. Ainda, o tom de pele dos pacientes foi definido pessoalmente por dermatologistas de acordo com a escala Fitzpatrick.

A base apresentava originalmente um total de 656 imagens clínicas, sendo 208 imagens na escala Fitzpatrick 1–2 (159 benignas e 49 malignas), 241 imagens na escala 3-4 (167 benignas e 74 malignas), e 207 imagens na escala 5–6 (159 benignas e 48 malignas). No entanto, para a avaliar esta base com os modelos pré-treinados utilizando o *pipeline* da Figura 2, foi feita uma filtragem das lesões, excluindo lesões malignas diferentes de melanoma. A quantidade de lesões por classe da lesão após a filtragem é mostrada na Tabela 7. É importante ressaltar que após da exclusão de outras lesões malignas, a quantidade de melanomas por fotótipo diminuiu consideravelmente, resultando em uma base muito desbalanceada. Apesar da base de dados ser mais diversa, ainda mostra uma insuficiência de imagens de melanoma representativa para todos os tons de pele.

Tabela 7: Quantidade de imagens em relação à escala Fitzpatrick dentre as 506 imagens resultantes após a exclusão de lesões malignas diferentes de melanoma, na base *Diverse Dermatology Images*.

	Número de Imagens	
Escala	Melanoma	Benigna
1-2	7	159
3-4	7	167
5-6	7	159

Replicamos as etapas de pré-processamento e inferência dos modelos, utilizadas para as demais bases, como mencionado na Seção 3.2. As métricas obtidas no processo de avaliação são apresentadas na Tabela 8.

Tabela 8: Métricas obtidas no pipeline de avaliação para a base de dados *Diverse Dermatology Images* agrupadas pela escala Fitzpatrick.

Escala Fitzpatrick	Modelo	Acurácia Balanceada	*Precision*	*Recall*	F1-Score
1-2	SwAV	63.0	33.3	28.6	30.8
	MoCo	54.9	12.5	14.3	13.3
	SimCLR	55.3	14.3	14.3	14.3
	BYOL	60.8	15.4	28.6	20.0
	InfoMin	55.3	14.3	14.3	14.3
	Supervised	55.3	14.3	14.3	14.3
	Média	57.4	17.3	19.1	17.8
3-4	SwAV	48.3	0.0	0.0	0.0
	MoCo	44.6	0.0	0.0	0.0
	SimCLR	46.4	0.0	0.0	0.0
	BYOL	44.9	0.0	0.0	0.0
	InfoMin	45.8	0.0	0.0	0.0
	Supervised	47.6	0.0	0.0	0.0
	Média	46.2	0.0	0.0	0.0
5-6	SwAV	47.7	3.2	14.3	5.3
	MoCo	66.6	9.5	57.1	16.3
	SimCLR	53.3	5.4	28.6	9.1
	BYOL	56.1	7.1	28.6	11.4
	InfoMin	62.0	9.1	42.9	15.0
	Supervised	41.8	0.0	0.0	0.0
	Média	54.6	5.7	28.6	9.51

Os resultados mostram que tanto os modelos auto-supervisionados, quanto o supervisionado tiveram um melhor desempenho para imagens da escala Fitzpatrick 1-2, que correspondem a peles brancas, em torno de 57.4%. Para fotótipos 3-4, a acurácia balanceada tem valor médio de 46.2%. Já para fotótipos 5-6, representando peles negras, a acurácia é em média 54.6%. Em suma, para tons de pele mais escuros, vemos que há uma tendência de decaimento na taxa de acerto dos modelos.

Ao analisar os resultados focados nas lesões de melanoma, temos F1-Score em torno de 17.8% para fototipos 1-2. Para fototipos 3-4, F1-Score é igual a zero para todos os modelos. Por fim, temos F1-Score em torno de 9.51% para fototipos 5-6. Esses resultados mostram que os modelos tiveram um desempenho insuficiente para todas as escalas de Fitzpatrick, ao classificar melanoma. No entanto, o resultado foi mais grave para lesões de pele não brancas.

Em geral, ao comparar o desempenho dos modelos auto-supervisionados ao supervi-

sionado vemos que os resultados são parecidos, exceto no fototipo 5-6 no qual o modelo supervisionado foi o único a errar todas as classificações de lesões de melanoma. Isso mostra que os modelos auto-supervisionados tenderam a ser mais robustos na inferência de lesões fora da distribuição original.

6 Conclusão

A avaliação de modelos auto-supervisionados e supervisionados em lesões de pele em regiões acrais retrata uma grande deficiência em robustez e viés nos modelos de *deep learning* para imagens fora do contexto de distribuição, especialmente em peles negras. Em geral, modelos auto-supervisionados se mostraram mais promissores na generalização para classificação de testes fora da distribuição de treinamento, muito provavelmente pela sua maior capacidade de aprender boas representações de forma auto-supervisionada por intermédio de uma tarefa pretexto. Uma possível melhoria seria utilizar técnicas para melhorar a classificação de lesões de pele fora da distribuição original — *out-of-distribution* — nas quais os modelos foram treinados [32].

Os resultados para diagnóstico de melanoma em regiões acrais são insuficientes, podendo causar sérios problemas sociais se usados de forma clínica. O cálculo das métricas e análise dos resultados também foram prejudicados pela quantidade insuficiente de amostras. O poder de generalização de modelos de IA depende fortemente da distribuição dos dados de treinamento. Logo, para que os modelos de IA sejam robustos em relação a diferentes padrões visuais de lesões, faz-se necessário o treinamento com bases de dados representativas do cenário clínico real, com pacientes diversos em características de lesões que englobam também o tom de pele. É preciso uma preocupação e urgência na criação de bases de dados que garantem a transparência dos dados em relação à fonte, processo de coleta e rotulagem das lesões, confiabilidade das descrições dos dados e diversidade étnica e racial dos pacientes, a fim de garantir alta confiança nos diagnósticos feitos pelos modelos.

A situação atual das bases de dados são preocupantes pois impactam no desempenho dos modelos e podem reforçar ainda mais viés no diagnóstico do câncer de pele em pessoas negras. Atualmente, esses modelos não podem ser utilizados de forma generalista, pois tem bons resultados apenas em lesões em pele branca. A criação de modelos específicos devido a uma negligência na criação de bases de dados diversas é inaceitável. As redes neurais profundas têm grande potencial para melhorar o diagnóstico principalmente para populações com menos acesso a dermatologia, mas a inclusão de lesões de pele com características menos comuns é extremamente necessária para que essas populações tenham acesso aos benefícios de uma tecnologia inclusiva. Para que os modelos sejam utilizados de forma prática, os resultados também devem ser promissores para lesões menos conhecidas.

Referências

[1] Pessoas de pele negra têm a forma mais grave do câncer de pele com maior intensidade. URL: https://www.sbcm.org.br/v2/index.php/not%C3%ADcias/1133-sp-1345803603. Accessed: Sep 2021.

[2] Interactive dermatology atlas. URL: https://www.dermatlas.net/,. Accessed: Sep 2021.

[3] Dermis.net: Dermatology information service available on the internet. URL: https://www.dermis.net/dermisroot/pt/home/index.htm,. Accessed: Sep 2021.

[4] Interactive dermatology atlas. URL: https://www.dermatlas.net/,. Accessed: Sep 2021.

[5] Classificação dos fototipos de pele. https://www.sbd.org.br/cuidados/classificacao-dos-fototipos-de-pele/.

[6] Isic archive. URL: https://www.isic-archive.com/. Accessed: Sep 2021.

[7] American Cancer Society. Key statistics for melanoma skin cancer. https://www.cancer.org/cancer/melanoma-skin-cancer/about/key-statistics.html, 2022.

[8] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino. Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis. *Archives of Dermatology*, 134(12):1563–1570, 1998.

[9] C. A. Bernardes. Diagnóstico e condutas dermatológicas em uma unidade básica de saúde. *Revista Brasileira de Educação Médica [online]*, pages 88–94, 2015.

[10] A. Bissoto, F. Perez, V. Ribeiro, M. Fornaciali, S. Avila, and E. Valle. Deep-learning ensembles for skin-lesion segmentation, analysis, classification: RECOD titans at ISIC challenge 2018. *CoRR*, abs/1808.08480, 2018. URL http://arxiv.org/abs/1808.08480.

[11] A. Bissoto, F. Perez, E. Valle, and S. Avila. Skin lesion synthesis with generative adversarial networks. *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 294—302, 2018.

[12] A. Bissoto, M. Fornaciali, E. Valle, and S. Avila. (De)Constructing bias on skin lesion datasets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[13] A. Bissoto, E. Valle, and S. Avila. Debiasing skin lesion datasets and models? not so fast. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 740–741, 2020.

[14] A. Bissoto, E. Valle, and S. Avila. Gan-based data augmentation and anonymization for skin-lesion analysis: A critical review. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1847–1856, 2021.

[15] A. Bissoto, C. Barata, E. Valle, and S. Avila. Artifact-based domain generalization of skin lesion models. In *European Conference on Computer Vision Workshops*, 2022.

[16] A. Bono, S. Tomatis, C. Bartoli, G. Tragni, G. Radaelli, A. Maurichi, and R. Marchesini. The abcd system of melanoma detection. *Cancer*, 85(1):72–77, 1999.

[17] Y. A. Caetano, A. M. Q. Ribeiro, B. R. da Silva Albernaz, I. de Paula Eleutério, and L. F. F. Fróes. Melanoma acral-estudo clínico e epidemiológico. *Surgical & Cosmetic Dermatology*, 12(2):130–134, 2020.

[18] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR*, 2020. URL https://arxiv.org/abs/2006.09882.

[19] L. Chaves, A. Bissoto, E. Valle, and S. Avila. An evaluation of self-supervised pre-training for skin-lesion analysis. In *European Conference on Computer Vision Workshops*, 2022.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. URL https://arxiv.org/abs/2002.05709.

[21] R. Daneshjou, M. P. Smith, M. D. Sun, V. Rotemberg, and J. Zou. Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, 157(11):1362–1369, 11 2021.

[22] R. Daneshjou, K. Vodrahalli, R. A. Novoa, M. Jenkins, W. Liang, V. Rotemberg, J. Ko, S. M. Swetter, E. E. Bailey, O. Gevaert, P. Mukherjee, M. Phung, K. Yekrang, B. Fong, R. Sahasrabudhe, J. A. C. Allerup, U. Okata-Karigane, J. Zou, and A. S. Chiou. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science Advances*, 8(32):eabq6147, 2022.

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

[24] Dermatology Learning Network. Skin cancer in african-americans. https://www.hmpgloballearningnetwork.com/site/thederm/article/2547, 2004.

[25] S. Gidaris, P. Singh, and N. Komodakis. Unsupervised representation learning by predicting image rotations. *ArXiv*, abs/1803.07728, 2018.

[26] T. M. Gomes, A. T. M. S. d. Moura, and A. C. d. Aguiar. Dermatologia na atenção primária: um desafio para a formação e prática médica. *Revista Brasileira de Educação Médica [online]*, pages 125–128, 2012.

[27] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020.

[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9726–9735, 2020.

[29] Instituto Melanoma Brasil. Tipos de melanoma. `https://www.melanomabrasil.org/new-tiposmelanomas`, 2022.

[30] Instituto Nacional de Câncer. Câncer de pele melanoma. `https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/pele-melanoma`, 2022.

[31] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2019.

[32] X. Li, Y. Lu, C. Desrosiers, and X. Liu. Out-of-distribution detection for skin lesion images with deep isolation forest, 2020. URL `https://arxiv.org/abs/2003.09365`.

[33] A. Menegola, M. Fornaciali, R. Pires, F. V. Bittencourt, S. Avila, and E. Valle. Knowledge transfer for melanoma screening with deep learning. In *International Symposium on Biomedical Imaging*, pages 297–300, 2017.

[34] A. Menegola, J. Tavares, M. Fornaciali, L. T. Li, S. E. F. de Avila, and E. Valle. RECOD titans at ISIC challenge 2017. *CoRR*, abs/1703.04819, 2017. URL `http://arxiv.org/abs/1703.04819`.

[35] A. Pacheco, G. Lima, A. Salomão, B. Krohling, I. Biral, G. Angelo, F. Jr, J. Esgario, A. Simora, P. Castro, F. Rodrigues, P. Frasson, R. Krohling, H. Knidel, M. Santos, R. do Espírito Santo, T. Macedo, T. Canuto, and L. Barros. Pad-ufes-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, 08 2020.

[36] E. Prudente. Dados do ibge mostram que 54% da população brasileira é negra. `https://jornal.usp.br/radio-usp/dados-do-ibge-mostram-que-54-da-populacao-brasileira-e-negra/`, 2020.

[37] R. C. Rabin. Dermatology has a problem with skin color. `https://www-nytimes-com.cdn.ampproject.org/c/s/www.nytimes.com/2020/08/30/health/skin-diseases-black-hispanic.amp.html`, 2020.

[38] V. Ribeiro, S. Avila, and E. Valle. Less is more: Sample selection and label conditioning improve skin lesion segmentation. *CoRR*, abs/2004.13856, 2020. URL `https://arxiv.org/abs/2004.13856`.

[39] Y. Tian, C. Sun, B. Poole, D. Krishnan, C. Schmid, and P. Isola. What makes for good views for contrastive learning? In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6827–6839. Curran Associates, Inc., 2020.

[40] E. Valle, M. Fornaciali, A. Menegola, J. Tavares, F. V. Bittencourt, L. T. Li, and S. Avila. Data, depth, and design: Learning reliable models for skin lesion analysis. *Neurocomputing*, 383:303–313, 2020.