Análise de domínios no aprendizado por transferência para classificação de nódulos tireoidianos

Karine Mendes Tavares[1], **Alexei Manso Correa Machado**[1]

[1]Instituto de Ciências Exatas e Informática – Pontifícia Universidade Católica de Minas Gerais (PUC Minas)
Belo Horizonte – MG – Brasil

karine.m.tavares@hotmail.com, alexeimcmachado@gmail.com

Abstract. *With the increase number of detections of thyroid nodules, it is important to more assertively identify the type of nodule and, in addition, to avoid the use of invasive techniques, such as Fine Needle Thyroid Aspiration (FNAB) and surgeries. In view of the limitation of publicly available medical images, this article performs a comparative study of the use of pre-trained models ImageNet, a generic database, and RadImageNet, a medical database, in the classification of these nodules classified as TI-RADS 4, with the hypothesis that the results obtained through the weights of the medical base are superior to those of ImageNet. The validation results showed that the RadImageNet weights generate superior and statistically relevant values for the DenseNet-121 architecture, but that the same scenario does not occur for the InceptionResNetV2 architecture.*

Resumo. *Com o aumento da quantidade de detecções de nódulos de tireoide, é importante realizar a identificação mais assertiva do tipo de nódulo e, além disso, evitar a utilização de técnicas invasivas, tais como Punção Aspirativa de Tireoide por Agulha Fina (PAAF) e cirurgias. Tendo em vista a limitação de imagens médicas disponíveis publicamente, este artigo realiza um estudo comparativo da utilização de modelos pré-treinados ImageNet, uma base de dados genérica, e RadImageNet, uma base de dados médica, na classificação desses nódulos classificados como TI-RADS 4, com a hipótese de que os resultados obtidos através dos pesos da base médica são superiores aos ImageNet. Os resultados de validação mostraram que os pesos RadImageNet geram valores superiores e estatisticamente relevantes para a arquitetura DenseNet-121, mas que o mesmo cenário não ocorre para a arquitetura InceptionResNetV2.*

1. Introdução

A tireoide é uma glândula localizada na parte frontal do pescoço e é responsável por dois hormônios, a tiroxina (T4) e a tri-iodotironina (T3), que regulam o metabolismo do corpo humano. Devido a sua importância, o diagnóstico e tratamento de doenças da tireoide se mostra fundamental. Como mostrado em outros estudos, um problema comumente reportado na região da tireoide é a presença de nódulos que causam o câncer de tiroide [Nguyen et al. 2020].

A incidência de nódulos palpáveis de tireoide na população adulta é de cerca de 67% e 10% deles podem ser malignos. Técnicas de imagens modernas, tais como tomografia e ultrassonografia são utilizadas para detectar, diagnosticar e gerenciá-los. As

imagens de ultrassonografia modo B são menos invasivas e apresentam o melhor custo-benefício. Entretanto, para interpretá-las corretamente é necessário um alto nível de expertise, treinamento e, além disso, a leitura das imagens é altamente afetada pelos ruídos, manchas e operador responsável pelo ultrassom. Uma captura imprecisa do nódulo pode resultar desnecessariamente na aspiração com agulha (biópsia) [Pedraza et al. 2015].

A hipo-ecogenicidade, ausência de halo, microcalcificações, solidez, fluxo intranodular e formato dos nódulos de tireoide são características sonográficas identificadas por radiologias como sugestivas para malignidade [Gaitini et al. 2011]. Baseadas nelas, foi desenvolvido o Sistema de Relatórios e Dados de Imagens da Tireoide (TI-RADS) para categorizar os nódulos e estratificar o risco de malignidade. As pontuações TI-RADS 1, 2, 3, 4a, 4b, 4c e 5 se referem respectivamente à "tireóide normal", "benigno", "provavelmente benigno", "um elemento suspeito", "dois elementos suspeitos", "três ou mais elementos suspeitos"e "provavelmente maligno"[Kwak et al. 2011].

O principal desafio dos métodos de diagnóstico por meio de imagens ultrassonográficas é a sensibilidade ao ruído e a baixa acurácia devido à extração de características desnecessárias. Com o desenvolvimento de sistemas de diagnóstico assistido por computador (CAD) e a utilização de aprendizado profundo, a acurácia do diagnóstico do câncer aumentaria consideravelmente [Wang et al. 2019]. Para isso, as Redes Neurais Convolucionais (CNNs) vem se provando eficientes em várias tarefas de aprendizado, incluindo os problemas de classificação de imagens [Zhu et al. 2017], aprendendo automaticamente as características de textura úteis para o problema de detecção/classificação em vez de usar métodos de extração manuais (fixos) [Nguyen et al. 2020].

O tamanho da base de dados ao treinar a rede neural interfere diretamente na qualidade do aprendizado e resultados finais, por isso os modelos pré-treinados são comumente utilizados, onde obtém-se os pesos de uma rede neural treinada em um domínio semelhante e uma base de dados grande. Isso diminui consideravelmente o tempo e esforço necessário para obter-se bons resultados ao treinar a rede para o problema desejado. Neste estudo é analisado a performance na classificação de nódulos TI-RADS 4 ao utilizar-se modelos pré-treinados ImageNet [Deng et al. 2009], uma base de dados composta de milhões de imagens do mundo natural e que é aplicada em diversos estudos, e os modelos pré-treinados na base de dados médica, RadImageNet [Mei et al. 2022], que busca trazer uma maior similaridade entre as imagens finais e as de treinamento, com a hipótese de que seus resultados de classificação são superiores aos ImageNet.

O restante deste artigo está organizado assim: na Seção 2 são citados trabalhos relacionados com o tema deste artigo; a Seção 3 descreve a base de dados, como foi realizado a seleção das imagens, métodos e métricas utilizadas; a Seção 4 apresenta os resultados e discussões e, por fim, a seção 5 trata das considerações finais.

2. Trabalhos Relacionados

Um número considerável de estudos vem adotando o aprendizado profundo para resolver problemas de classificação. Rehman et al. [Rehman et al. 2021] aborda uma técnica automatizada para detecção e segmentação de nódulos de tireoide utilizando imagens de ultrassom. Foi utilizado um modelo de aprendizado profundo com uma rede neural convolucional e um backbone VGG-16 para melhorar a acurácia comparado a um modelo simples. Zhu et al. [Zhu et al. 2017] aborda como as redes neurais vem provando sua

eficiência em várias tarefas de aprendizagem, incluindo problemas de classificação. Contudo, treinar uma rede neural desde o inicio requer um número enorme de imagens e as imagens médicas são geralmente mais difíceis de coletar e mais complicadas de processar devido às suas particularidades. Além disso, os métodos tradicionais de aumento da base de imagens apresentam riscos de eliminar a região primordial da imagem por corte aleatório, tal como o tumor nas imagens de ultrassom. Para resolver esse problema é proposto a construção de uma pequena rede neural, composta somente por camadas convolucionais, para gerar novas imagens baseadas nas originais.

Sob outro enfoque, Nguyen et al. [Nguyen et al. 2020] levanta uma limitação em comum presente em outros estudos: não considerar totalmente os problemas associados com os métodos baseados em aprendizado profundo, tais como desequilíbrio das amostras de imagem de treinamento, a profundidade da rede e a variação do tamanho dos objetos. Para superar essa limitação é proposto a modificação da função de perda de uma CNN convencional e uma combinação de várias redes para melhorar a capacidade de aprendizado. Como cada modelo tem sua própria arquitetura e diferentes métodos de aprendizado das características das imagens de entrada, o uso de vários modelos baseados em redes neurais pode ajudar a extrair informações mais ricas em comparação com o uso de um modelo individual. Gomes Ataide et al. [Gomes Ataide et al. 2020] busca reduzir a subjetividade no atual processo de diagnóstico, utilizando características geométricas e morfológicas (G-M), que representam as características visuais dos nódulos da tireoide, para fornecer aos médicos suporte à decisão. O desempenho das características extraídas foi avaliado usando um classificador random forest (RFC) e os resultados obtidos do RFC foram comparados com outras técnicas de extração e classificação.

Hang [Hang 2021] ilustra um método de ponta a ponta que envolve a combinação das características profundas com as características convencionais para formar um espaço de características híbrido e, para a classificação, compara-se os resultados da rede ResNet18 com a Res-GAN, que supera a primeira. Com o objetivo de aumentar o número de amostras analisadas, Chi et al. [Chi et al. 2017] e Song et al. [Song et al. 2020] utilizam uma base de dados adicional, além da DDTI (*Digital Database Thyroid Image*), apresentando, respectivamente, um sistema completo de classificação de imagens de ultrassom da tireoide baseado em um modelo GoogLeNet ajustado e uma CNN híbrida que realiza a extração de características e recortes, possibilitando uma melhor distinção entre os nódulos benignos e malignos. Buscando classificar outras categorias além da binária, Seixas e Machado [Seixas and Machado 2022] realizam também a classificação entre os 3 grupos TI-RADS 4 e individual dos TI-RADS. São comparados os resultados de CNNs e *Support Vector Machines* (SVMs) e variações das imagens, com resolução 360x360 com o fundo verde e quadrados com dimensões 160, 256 e 272 pixels.

Mei et al. [Mei et al. 2022] demonstra o valor do pré-treinamento com milhões de imagens radiológicas em comparação com imagens fotográficas ImageNet para aplicações médicas ao utilizar aprendizado profundo. Apesar da alta performance apresentada com os modelos pré-treinados com ImageNet, os modelos treinados em bases de dados médicas podem alcançar uma performance melhor. Uma vez que em trabalhos passados [Xie and Richmond 2018, Parakh et al. 2019, Ghesu et al. 2022] foi mostrado uma melhora na performance dos modelos pré-treinados em bases médicas. É apresentada, então, a base de dados RadImageNet e, a partir dela, são gerados novos modelos pré-

treinados exclusivamente a partir de imagens médicas para serem usadas em aplicações médicas.

Os trabalhos encontrados propõem diversas novas técnicas de segmentação, pré-processamento e classificação, porém grande parte dos trabalhos se limitam somente a classificação binária dos nódulos. Os tópicos a seguir realizam um estudo comparativo da classificação de nódulos categoria 4, que são casos mais críticos e que geram mais confusão, e exploram a utilização de modelos pré-treinados em uma base médica para obter melhores resultados.

3. Materiais e Métodos

3.1. Descrição da Base de dados

A base de dados DDTI (*Digital Database Thyroid Image*) utilizada neste estudo foi disponibilizada publicamente pela Universidade Nacional da Colômbia [Pedraza et al. 2015], contendo imagens de ultrassom de quase 300 pacientes válidos (vários pacientes possuem mais de um ultrassom) e totalizando 480 imagens. Estas foram classificadas em pontuações TI-RADS 2, 3, 4a, 4b, 4c e 5, sendo 61 como não cancerígenas (TI-RADS 2 e 3) e 288 como cancerígenas (TI-RADS 4a, 4b, 4c e 5). O restante das imagens não foi classificada e não foi fornecido nenhuma imagem da pontuação TI-RADS 1. Entre as amostras TI-RADS 4, 96 imagens são classificadas como TI-RADS 4a, 79 TI-RADS 4b e 68 TI-RADS 4c. A relação do número de imagens por categoria pode ser vista na tabela 1.

Tabela 1. Número de imagens por classe

TI-RADS	Nº Imagens
2	42
3	19
4a	96
4b	79
4c	68
5	45

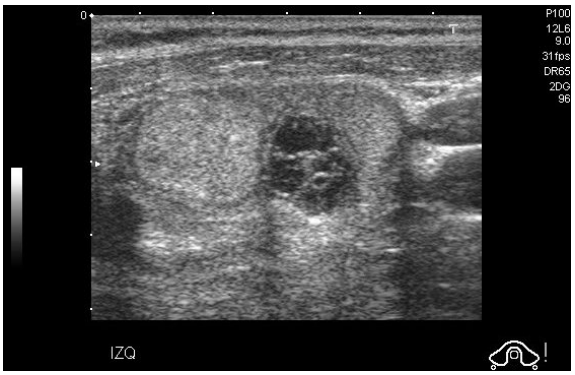
As imagens possuem a resolução de 560x360, com 8-bit de 3 canais e, para cada uma, é fornecido um arquivo XML com anotações completas das imagens, descrição diagnóstica de lesões suspeitas da tireoide, utilizando o sistema de classificação de nódulos tireoidianos TI-RADS, realizada por pelo menos dois radiologistas especialistas, número do caso analisado, idade e sexo dos pacientes, e informações de composição, ecogenicidade, margens e calcificação do nódulo.

3.2. Seleção das Imagens

Para realizar os experimentos foram selecionadas somente as imagens de categoria 4, separadas em TI-RADS 4a, 4b e 4c. O objetivo dessa escolha foi verificar a classificação quando se tem uma base mais homogênea e em nódulos que possuem mais características em comum e geram maior dúvida durante o diagnóstico. Posteriormente foi realizada a validação cruzada estratificada de 10 dobras, em que o processo de validação cruzada é repetido dez vezes e logo após é calculada a média e desvio padrão dos resultados

Figura 1. Exemplo das imagens de ultrassom presentes na base DDTI

(a) TI-RADS 2



(b) TI-RADS 4a

obtidos. Há pacientes que possuem mais de um ultrassom, por isso a separação da base foi realizada de forma que as imagens de um mesmo paciente estejam sempre no mesmo conjunto.

3.3. Segmentação, Aumento de Dados e Pré-processamento

Os métodos utilizados foram baseados nos experimentos de Seixas e Machado [Seixas and Machado 2022], onde através do XML fornecido na base de dados, as imagens foram cortadas de acordo com a sua região de interesse, adicionadas em um fundo verde e salvas na resolução de 360x360. É possível ver o resultado final na Figura 2.

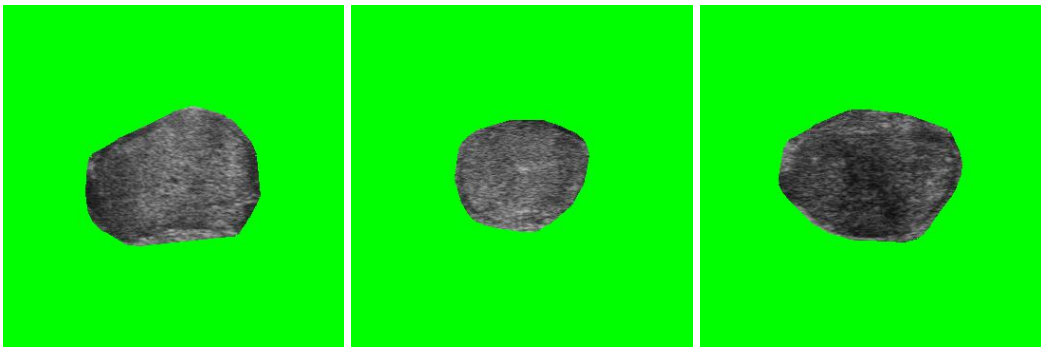
As imagens obtidas após esse processo ainda não são suficientes para treinar as redes, por isso foram aplicadas técnicas de aumento de dados. Todas as imagens foram rotacionadas em 45°, gerando 7 novas imagens para cada exemplar, e logo depois foi realizado o espelhamento de cada uma, dobrando o tamanho da base após a rotação.

Para o treinamento das redes, os pixels de cada imagem foram escalados para valores entre 0 e 1, facilitando o processamento do modelo. E, para obter o melhor tempo de execução, foi utilizado o serviço de nuvem do Google, Google Colaboratory, para treinar a rede na GPU de back-end do Google Compute Engine em Python 3.

3.4. Modelos de treinamento e Arquitetura

Para criação de um modelo de Inteligência Artificial eficiente é necessária uma base de dados grande, possibilitando o aprendizado de todas as características importan-

Figura 2. Exemplo das imagens TI-RADS 4 segmentadas com fundo verde e resolução de 360x360



(a) TI-RADS 4a (b) TI-RADS 4b (c) TI-RADS 4c

tes. Em casos de tamanho de amostra limitado, o aprendizado por transferência [Pan and Yang 2010] é uma abordagem de aprendizado profundo comumente usada, na qual um modelo projetado para um problema pode ser reutilizado para iniciar uma tarefa diferente, mas relacionada.

O aprendizado por transferência com modelos pré-treinados ImageNet é amplamente explorado em aplicações IA de imagens médicas. Contudo, buscando se aproximar mais do domínio médico, os modelos pré-treinados com a rede RadImageNet também são utilizados e tiveram seus resultados comparados. Os pesos RadImageNet foram disponibilizados para as arquiteturas ResNet50, DenseNet-121, InceptionResNetV2 e InceptionV3, por serem amplamente adotadas e usadas em aplicações médicas. A arquitetura DenseNet-121 e InceptionResNetV2 foram escolhidas tendo em vista, respectivamente, os resultados apresentados por Seixas e Machado [Seixas and Machado 2022], na classificação dos nódulos TI-RADS 4, e Mei et al. [Mei et al. 2022], como o melhor modelo de acurácia top-1 das arquiteturas apresentadas.

3.4.1. RadImageNet

O RadImageNet, proposto por Mei et al. [Mei et al. 2022], é um banco de dados de acesso livre composto por 5 milhões de imagens rotuladas de acordo com as tags DICOM (Comunicação de Imagens Digitais em Medicina) e mais de 1 milhão de imagens e diagnósticos de tomografia computadorizada, ressonância magnética e ultrassons de patologia musculoesquelética, neurológica, oncológica, gastrointestinal, endócrina e pulmonar de 500.000 pacientes. O banco contém imagens médicas de 3 modalidades, 11 anatomias e 165 rótulos patológicos. Todas as imagens foram rotuladas por radiologistas americanos clinicamente praticantes do Hospital de Nova York.

A base foi projetada para melhorar o desempenho do aprendizado por transferência em aplicações medicas e, para isso, também é fornecido acesso aos pesos de quatro redes neurais treinadas a partir do zero somente com as imagens médicas.

3.4.2. DenseNet-121

A DenseNet, proposta por Huang et al. [Huang et al. 2018], é uma arquitetura que conecta todas as camadas diretamente, garantindo o fluxo máximo de informações entre elas. Para preservar a natureza *feed-forward* (transmissão de informações somente adiante), cada camada obtém entradas adicionais de todas as camadas anteriores e passa seus próprios mapas de recursos para todas as camadas subsequentes. Esses recursos nunca são combinados por meio de soma antes de serem passados para uma camada; em vez disso, eles são concatenados, aumentando a variação na entrada de camadas subsequentes e melhorando a eficiência. Essa estrutura pode ser observada na figura 3.

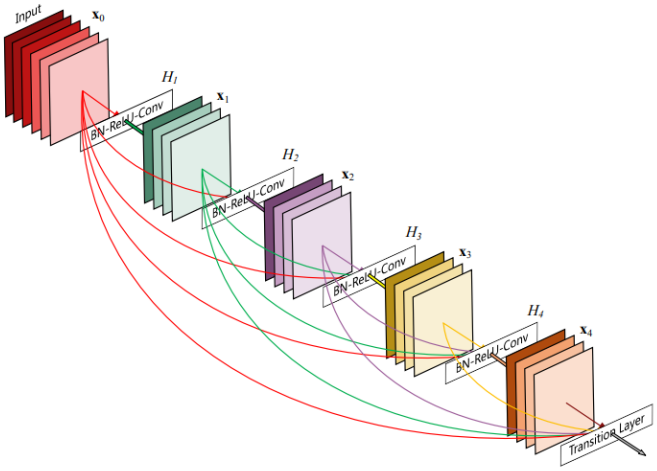


Figura 3. DenseNet de 5 camadas com uma taxa de crescimento de 4. Cada camada leva todos os mapas de recursos anteriores como entrada [Huang et al. 2018].

Dentre as suas variantes, a DenseNet121 possui 121 camadas convolucionais e possui 7381 conexões na CNN, ao invés de somente 121, como ocorre em arquiteturas tradicionais, possibilitando um aprendizado ainda mais profundo.

3.4.3. InceptionResNetV2

A InceptionResNetV2, proposta por Szegedy et al. [Szegedy et al. 2017], possui 164 camadas convolucionais e é uma combinação da estrutura Inception e conexões residuais da ResNet, onde ao invés de simplesmente enviar os dados de entrada adiante em cada camada, é fornecido outro caminho para os dados alcançarem as últimas partes da rede neural, conectando a saída de uma camada convolucional à entrada de outra camada futura através de uma simples soma. Isso soluciona problemas de degradação apresentado em outras estruturas e também reduz o tempo de treinamento. É possível observar seu esquema geral e mais detalhes do seu primeiro nódulo na figura 4.

3.5. Métricas de Avaliação da Qualidade do Modelo

As métricas de Acurácia, Precisão, *Recall* e *F1-score* foram aplicadas para avaliar a qualidade dos modelos. A acurácia indica uma performance geral do modelo, onde dentre todas as classificações, quantas foram classificadas corretamente (predições corretas / todas

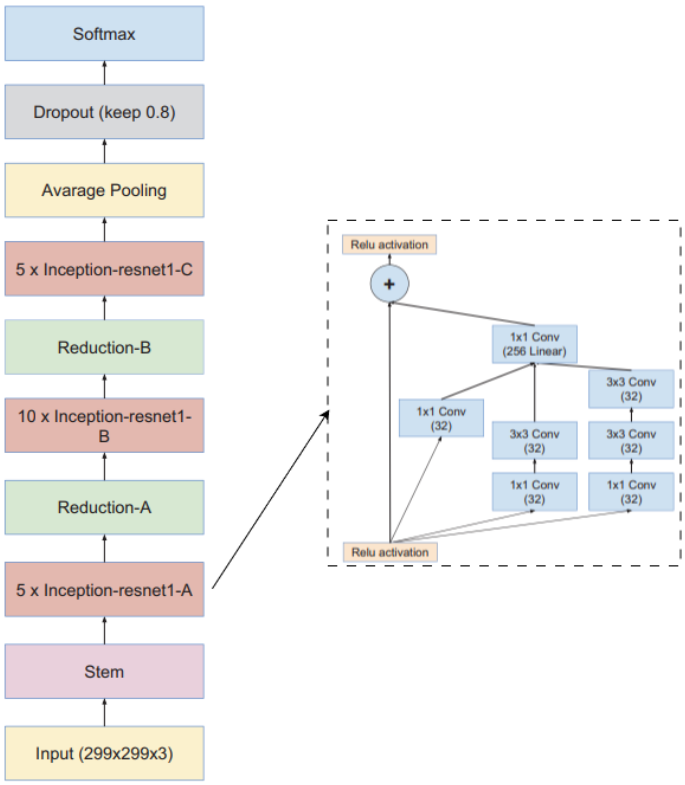


Figura 4. Esquema geral da arquitetura InceptionResNetV2 e do primeiro módulo. Adaptado de [Szegedy et al. 2017].

as predições). A Precisão é a taxa de instâncias corretamente classificadas como pertencentes a uma classe e o total de exemplos classificados como a determinada classe (vp/vp + fp). O *Recall* refere-se a capacidade do método de detectar com sucesso instâncias pertencentes a uma classe (vp/ vp + fn). Em ambas, é aplicada a fórmula para cada classe e retornado a média. Já o *F1-score* é uma média harmônica entre a Precisão e *Recall* (2*precisão**recall*/preciso+*recall*). Onde, vp é verdadeiro positivo, vn é verdadeiro negativo, fp é falso positivo e fn é falso negativo. Todas as médias calculadas são macro, ou seja, não há peso para as classes.

4. Resultados e Discussão

A validação cruzada estratificada de 10 dobras foi realizada para os pesos ImageNet e RadImageNet, com as arquiteturas DenseNet-121 e InceptionResNetV2, gastando, para cada uma destas, em média 60 horas e 80 horas de execução, respectivamente. Durante o treinamento de cada uma das 10 dobras, são geradas as métricas de Acurácia, Precisão, *Recall* e *F1-score* e o melhor resultado é salvo sempre que há melhora na métrica de *F1-score*. Todos os modelos foram treinados por 50 épocas, com todas as camadas da rede descongeladas, utilizando o otimizador Adam com a taxa de aprendizado de 0,001, para que os pesos dos modelos pré-treinados não sofram uma distorção acelerada e rápida, e função de perda *categorical cross-entropy*. Uma camada de *dropout* de 0,5 e uma camada de saída ativada pela função *softmax* foram introduzidas após a última camada dos modelos pré-treinados. Foi definida uma *seed* em comum em todos os treinamentos, para que os resultados possam ser reprodutíveis.

Os resultados de ambas arquiteturas podem ser observadas nas tabelas 2 e 3. Nelas, é possível ver que a arquitetura DenseNet-121 teve mais sucesso no aprendizado das características dos nódulos e apresenta uma consistência semelhante para ambos os pesos, com diferenças no desvio padrão de 1% na maioria das métricas.

Tabela 2. Média e desvio padrão dos resultados da validação cruzada estratificada de 10 dobras, à esquerda, utilizando pesos ImageNet, e à direita, pesos RadImageNet, ambos com imagens categoria 4 e rede DenseNet-121.

	A.	**Pre.**	**Rec.**	**F1**	**A.**	**Pre.**	**Rec.**	**F1**
Média	56%	56%	56%	54%	59%	60%	59%	57%
DP	5%	6%	6%	6%	7%	7%	7%	8%

Tabela 3. Média e desvio padrão dos resultados da validação cruzada estratificada de 10 dobras, à esquerda, utilizando pesos ImageNet, e à direita, pesos RadImageNet, ambos com imagens categoria 4 e rede InceptionResNetV2.

	A.	**Pre.**	**Rec.**	**F1**	**A.**	**Pre.**	**Rec.**	**F1**
Média	55%	55%	54%	54%	57%	57%	56%	54%
DP	8%	8%	8%	8%	5%	5%	5%	6%

Como em ambas foi variado somente os pesos iniciais e mantido a mesma *seed* e base de dados, é possível realizar a análise de cada métrica através do teste t pareado com nível significância de 0,05, segundo a equação:

$$\mathrm{t} = \frac{\bar{x}_d - \mu_0}{S_d/\sqrt{n}}$$

Considerando:

- $\bar{x}_d$: a média das diferenças de cada par;
- μ_0: diferença esperada entre os dois grupos (zero);
- S_d: desvio padrão das diferenças;
- n: número de pares;

As hipóteses tomadas para cada métrica são:

- Hipótese nula (H_0) = Utilização dos pesos RadImageNet não impacta nos resultados, ou seja, as médias são iguais.
- Hipótese alternativa (H_1) = Utilização dos pesos RadImageNet melhora os resultados, ou seja, média RadImageNet maior que a média ImageNet.

Os valores obtidos para cada métrica se encontram na tabela 4. O valor de p apresentado se refere as probabilidades do valor t encontrado possibilitar a rejeição ou não da hipótese nula. É interessante observar que os resultados de cada arquitetura não foram iguais. Para a DenseNet-121, nas métricas de acurácia, *recall* e *f1-score*, como os valores de p são menores que 0,05, a hipótese nula é rejeitada, ou seja, a diferença das médias amostrais é grande o suficiente para ser estatisticamente significante. Já na métrica de precisão, o valor-p de 0,057 é maior que o nível de significância, o que impossibilita a rejeição da hipótese nula, mesmo tendo um valor bem próximo de 0,05. Para a InceptionResNetV2, todas as métricas apresentam um valor-p maior que 0,05, ou seja, a média

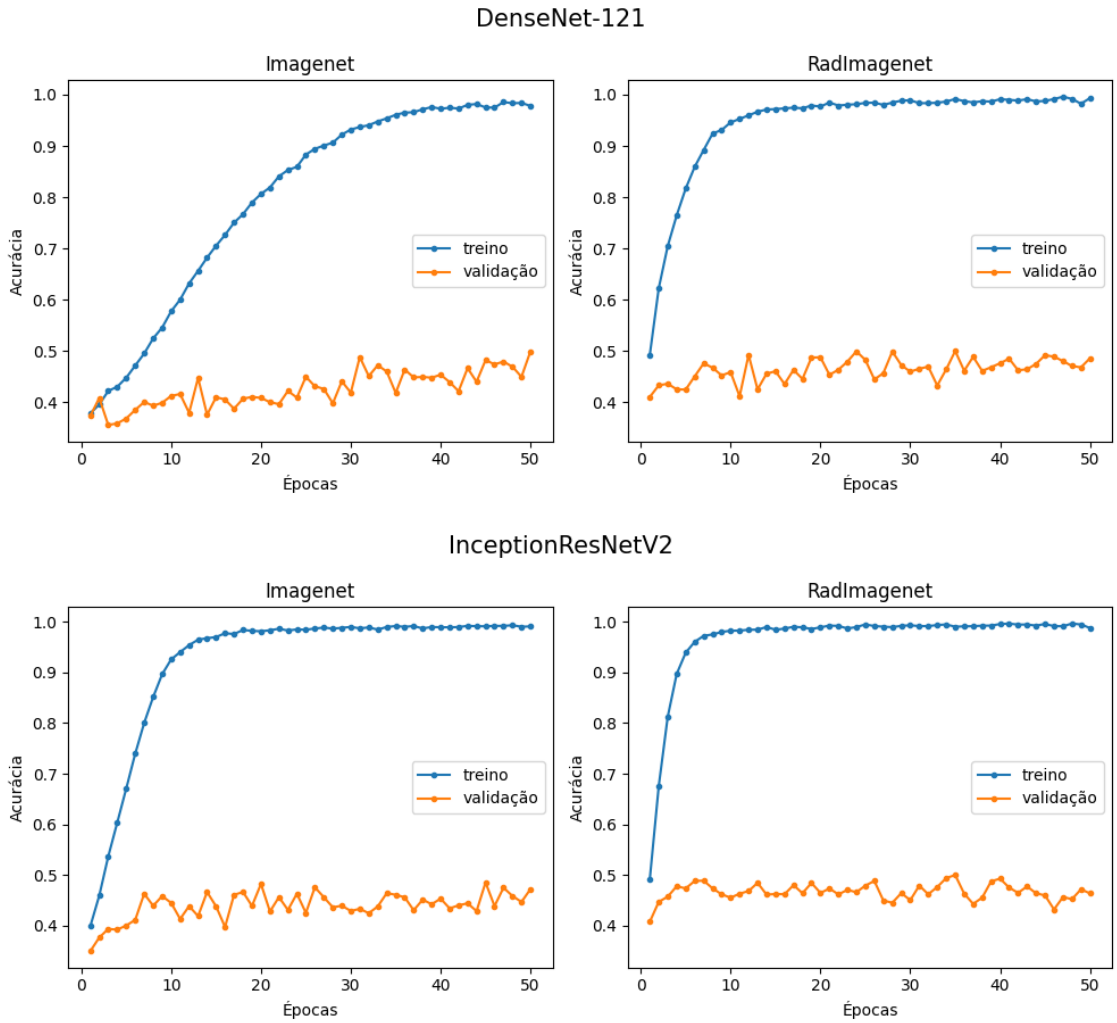
amostral RadImageNet não possui uma diferença grande o suficiente para ser considerada estatisticamente significativa e não é possível rejeitar a hipótese nula. Tem-se, portanto, que iniciar a rede com os pesos RadImageNet apresentou melhoras na classificação de três das quatro métricas utilizadas com a arquitetura DenseNet-121, enquanto para a InceptionResNetV2 não há melhoras estatisticamente significativas.

Tabela 4. Valores de t e p para cada métrica, à esquerda da rede DenseNet-121 e à direita da InceptionResNetV2

	A.	Pre.	Rec.	F1	A.	Pre.	Rec.	F1
t	2,15	1,75	2,26	1,95	0,89	0,88	0,99	0,34
p	0,03	0,057	0,025	0,041	0,2	0,2	0,17	0,37

Para realizar o teste t pareado, os dados analisados são referentes aos melhores modelos de validação salvos em cada dobra. Já nos gráficos da figura 5 são tomadas as médias de acurácia de todas as dobras, por época.

Figura 5. Curvas de treino e validação da métrica de acurácia.



Como os picos de cada dobra são variados, as curvas parecem apresentar um comportamento semelhante, porém há alguns pontos que podem ser levados em consideração. No primeiro gráfico da arquitetura DenseNet-121, é possível notar que tanto a curva de treino quanto a de validação ainda não estabilizaram, por isso seria interessante rodar mais épocas para verificar se a rede ainda pode melhorar ou não, o que não foi possível à limitações de tempo e poder computacional. Enquanto isso, nos demais gráficos, o aprendizado é mais sutil e começa-se a apresentar *overfitting*, principalmente para a arquitetura InceptionResNetV2. O quão rápido as redes RadImageNet aprendem também é outro fator importante que é possível observar, a curva de aprendizado de treino e validação, em ambas arquiteturas, iniciam com valores maiores e convergem mais rapidamente que os ImageNet, o que pode ajudar a diminuir o tempo necessário de treinamento das redes.

Os resultados desse estudo mostram que utilizar modelos pré-treinados em bases médicas apresenta melhoras para a arquitetura DenseNet-121 e que as curvas de aprendizado podem convergir mais rapidamente. Tudo isso focando na classificação dos nódulos TI-RADS 4, que possuem mais características em comum e geram mais confusões entre os médicos.

5. Considerações Finais

O objetivo deste trabalho foi realizar a classificação de nódulos tireoidianos TI-RADS 4 através de aprendizado por transferência ImageNet e RadImageNet e avaliar se há melhoras nos resultados quando se utiliza pesos obtidos de uma base de dados médica, com imagens que se aproximam mais das de treinamento. Para as métricas de acurácia, *recall* e *f1-score*, a arquitetura DenseNet-121 mostrou que a utilização dos pesos RadImageNet apresenta melhoras estatisticamente significantes. Enquanto, com a arquitetura InceptionResNetV2, apesar desta também apresentar melhoras em algumas dobras com pesos da base médica, estes não são considerados estatisticamente significantes. Dessa forma, é possível concluir que nem sempre utilizar pesos obtidos de uma base médica pode resultar em melhores resultados, porém pode trazer benefícios, tais como mais rapidez no treinamento.

Como limitações do trabalho poderíamos citar o uso de apenas uma base de dados de imagens de ultrassom, tendo em vista a dificuldade em conseguir esse tipo de imagens de forma pública, e o longo tempo de treinamento das redes, uma vez que em alguns casos seria interessante treinar por mais épocas. Quanto a trabalhos futuros, é possível realizar testes nas arquiteturas restantes (ResNet50 e InceptionV3), aumentar o número de épocas e testar outras técnicas de pré-processamento e aumento de dados como a apresentada por Chi et al. [Chi et al. 2017].

Referências

Chi, J., Walia, E., Babyn, P., Wang, J., Groot, G., and Eramian, M. (2017). Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of digital imaging*, 30(4):477–486.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Gaitini, D., Evans, R. M., and Ivanac, G. (2011). Chapter 16: thyroid ultrasound. *EFSUMB Course Book*.

Ghesu, F. C., Georgescu, B., Mansoor, A., Yoo, Y., Neumann, D., Patel, P., Vishwanath, R., Balter, J. M., Cao, Y., Grbic, S., et al. (2022). Self-supervised learning from 100 million medical images. *arXiv preprint arXiv:2201.01283*.

Gomes Ataide, E. J., Ponugoti, N., Illanes, A., Schenke, S., Kreissl, M., and Friebe, M. (2020). Thyroid nodule classification for physician decision support using machine learning-evaluated geometric and morphological features. *Sensors*, 20(21):6110.

Hang, Y. (2021). Thyroid nodule classification in ultrasound images by fusion of conventional features and res-gan deep features. *Journal of Healthcare Engineering*, 2021.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. (2018). Densely connected convolutional networks.

Kwak, J. Y., Han, K. H., Yoon, J. H., Moon, H. J., Son, E. J., Park, S. H., Jung, H. K., Choi, J. S., Kim, B. M., and Kim, E.-K. (2011). Thyroid imaging reporting and data system for us features of nodules: a step in establishing better stratification of cancer risk. *Radiology*, 260(3):892–899.

Mei, X., Liu, Z., Robson, P. M., Marinelli, B., Huang, M., Doshi, A., Jacobi, A., Cao, C., Link, K. E., Yang, T., et al. (2022). Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315.

Nguyen, D. T., Kang, J. K., Pham, T. D., Batchuluun, G., and Park, K. R. (2020). Ultrasound image-based diagnosis of malignant thyroid nodule using artificial intelligence. *Sensors*, 20(7):1822.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Parakh, A., Lee, H., Lee, J. H., Eisner, B. H., Sahani, D. V., and Do, S. (2019). Urinary stone detection on ct images using deep convolutional neural networks: evaluation of model performance and generalization. *Radiology: Artificial Intelligence*, 1(4):e180066.

Pedraza, L., Vargas, C., Narváez, F., Durán, O., Muñoz, E., and Romero, E. (2015). An open access thyroid ultrasound image database. In *10th International symposium on medical information processing and analysis*, volume 9287, pages 188–193. SPIE.

Rehman, H. A. U., Lin, C.-Y., and Su, S.-F. (2021). Deep learning based fast screening approach on ultrasound images for thyroid nodules diagnosis. *Diagnostics*, 11(12):2209.

Seixas, I. M. and Machado, A. M. C. (2022). *Uso de redes neurais convolucionais na classificação de nódulos tireoidianos através de ultrassonografia.*

Song, R., Zhang, L., Zhu, C., Liu, J., Yang, J., and Zhang, T. (2020). Thyroid nodule ultrasound image classification through hybrid feature cropping network. *IEEE Access*, 8:64064–64074.

Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. (2017). Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Wang, B., Liu, M., Zhu, M., et al. (2019). Artificial intelligence in ultrasound imaging: Current research and applications.

Xie, Y. and Richmond, D. (2018). Pre-training on grayscale imagenet improves medical image classification. In *Proceedings of the European conference on computer vision (ECCV) workshops*.

Zhu, Y., Fu, Z., and Fei, J. (2017). An image augmentation method using convolutional network for thyroid nodule classification by transfer learning. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, pages 1819–1823. IEEE.