Research, Society and Development, v. 11, n. 9, e31411931868, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i9.31868

Efeito da deterioração em imagens por ressonância magnética sobre redes neurais profundas

Effect of decay in magnetic resonance imaging on deep neural networks

Efecto de la decadencia en las imágenes de resonancia magnética en las redes neuronales profundas

Recebido: 17/06/2022 | Revisado: 29/06/2022 | Aceito: 01/07/2022 | Publicado: 10/07/2022

Carlos Leandro Silva dos Prazeres
ORCID: https://orcid.org/0000-0002-3028-728X
Hospital Universitário de Sergipe, Brasil
E-mail: c.leandro@outlook.com
Perseu Lúcio Alexander Helene de Paula
ORCID: https://orcid.org/0000-0002-5927-4546
Hospital Universitário de Sergipe, Brasil
E-mail: plahp@uol.com.br
Mozart Nolasco Monte
ORCID: https://orcid.org/0000-0002-3458-4521
Hospital Universitário de Sergipe, Brasil
E-mail: mozartmonte@hotmail.com
Marcela Costa Alcântara Estácio
ORCID: https://orcid.org/0000-0001-7077-1415
Hospital Universitário de Sergipe, Brasil
E-mail: marcelacostaalcantara@gmail.com
Esdras Adriano Barbosa dos Santos
ORCID: https://orcid.org/0000-0002-3621-1913
Universidade Federal de Sergipe, Brasil
E-mail: esdras.adriano@gmail.com
Laélia Campos
ORCID: https://orcid.org/0000-0001-5701-9851
Universidade Federal de Sergipe, Brasil
E-mail: lpbcampos@gmail.com

Resumo

Nas últimas décadas, tarefas de classificação e segmentação de achados clínicos com uso de redes neurais convolucionais cresceram bastante na esfera do diagnóstico por imagem e, mais precisamente, na modalidade de imagem por ressonância magnética. Porém, pouco se sabe a respeito do comportamento dessas arquiteturas quando confrontadas com fatores que degradam a resolução espacial e a resolução de contraste, uma vez que a maioria dos modelos é treinada com imagens de alta qualidade, o que não é condizente com o cotidiano geral. Por isso, faz-se necessário analisar a performance das redes neurais pré-treinadas, sob condições em que haja deterioração da imagem de entrada. Neste trabalho, foram avaliados os efeitos da degradação de ambas as resoluções, tanto em tarefas de classificação quanto de segmentação de tumores cerebrais, para três arquiteturas: Mobilenet, Vgg16 e SEResNeXt50. Os resultados obtidos demonstraram que as tarefas executadas são muito afetadas pelas distorções na qualidade das imagens, em especial nos casos em que as deteriorações se tornam mais intensas.

Palavras-chave: Imageamento por ressonância Magnética; Tumores cerebrais; Aprendizado profundo.

Abstract

In the last decades, tasks of classification and segmentation of clinical findings using convolutional neural networkshave grown significantly in the sphere of radiology, and more precisely in the modality of magnetic resonance imaging. However, little is known about the behavior of the proposed architectures when faced with factors that degrade the spatial resolution and contrast resolution, since most models are trained with high quality images, which is not consistent with the general daily life. Therefore, it is necessary to analyze the performance of pre-trained neural networks under conditions in which there is deterioration of the input image. In this work, the effects of degradation of the resolutions were evaluated, both in classification and segmentation tasks of brain tumors, for three architectures: Mobilenet, Vgg16 and SEResNeXt50. The results obtained showed that the tasks performed are greatly affected by image quality distortions, especially in cases where the deteriorations become more intense.

Keywords: Magnetic resonance; Brain tumors; Deep learning.

Resumen
En las últimas décadas, las tareas de clasificación y segmentación de hallazgos clínicos mediante redes neuronales convolucionales han crecido significativamente en el ámbito del diagnóstico por imagen y, más precisamente, en la modalidad de resonancia magnética. Sin embargo, poco se sabe sobre el comportamiento de estas arquitecturas ante factores que degradan la resolución espacial y la resolución de contraste, ya que la mayoría de los modelos son entrenados con imágenes de alta calidad, lo que no es acorde con la vida cotidiana general. Por lo tanto, es necesario analizar el desempeño de las redes neuronales pre-entrenadas, bajo condiciones en las que existe deterioro de la imagen de entrada. En este trabajo se evaluaron los efectos de la degradación de ambas resoluciones, tanto en tareas de clasificación como de segmentación de tumores cerebrales, para tres arquitecturas: Mobilenet, Vgg16 y SEResNeXt50. Los resultados obtenidos mostraron que las tareas realizadas se ven muy afectadas por distorsiones en la calidad de las imágenes, especialmente en los casos en que los deterioros se vuelven más intensos.
Palabras clave: Resonancia magnética; Tumores del cerebro; Aprendizaje profundo.

1. Introdução

Nas últimas décadas, muito tem se falado a respeito de inteligência artificial, e esta vem crescendo e se espalhando, cada vez mais, em diversos setores da sociedade, em especial, na área da saúde (Lobo, 2017). Tarefas de classificação e segmentação de patologias fazendo uso de redes neurais profundas (ou *deep neural networks*) cresceram bastante em diversas aplicações, incluindo na radioterapia e no radiodiagnóstico, mais precisamente, na modalidade de imagem por ressonância magnética (Sahiner et al., 2019).

Em função disso, um grande esforço vem sendo feito para que os achados de imagem possam ser automaticamente classificados e segmentados, gerando assim, uma grande visibilidade para as redes neurais convolucionais (ou *convolutional neural networks*, CNNs). Basicamente, esse tipo de rede é uma classe especializada em lidar com imagens, em razão de ser composta por camadas convolucionais, camadas de *pooling* e funções de ativação, as quais são responsáveis por filtrar, reduzir dimensionalidade e, em geral, ajudar no aprendizado de padrões complexos de dados através da extração de uma série de conjuntos de características (Cui et al., 2020).

Não bastando a variedade de classes de redes neurais, as CNNs apresentam variantes, como é o caso dos *autoencoders* convolucionais, que são constituídos por camadas de codificação e decodificação, responsáveis por codificar a informação de entrada em um espaço menor e, depois, decodificar para sua forma original ou para outras, visando à execução de tarefas voltadas para a reconstrução de imagens (Munir et al., 2019).

Não somente a arquitetura dessas camadas apresentadas é um fator determinante para o sucesso em uma tarefa, como também a informação visual que é entregue para a camada de entrada da rede neural (Thambawita et al., 2021), visto que, na prática clínica de um setor de imagem por ressonância, são frequentemente necessários ajustes em parâmetros de resolução, principalmente, quando a adequação de protocolos deve ser feita (McRobbie et al., 2017), e assim, a qualidade da imagem produzida se torna, então, um fator cada vez mais importante e decisivo.

Por isso, é necessário avaliar como a qualidade da imagem influencia a performance das redes, levando em consideração degradações de resolução, já que, na maioria dos casos, estas são treinadas e validadas com imagens de alta qualidade (Dodge & Karam, 2016). Além disso, este estudo deve ser conduzido de forma quantitativa, posto que os algoritmos de visão computacional operam de forma diferente do sistema visual humano (SVH). Nesse caso, enquanto a análise e o reconhecimento humano é resiliente à baixa qualidade das imagens, saber como os algoritmos são afetados em termos de performance (Koziarski & Cyganek, 2018) se torna relevante.

Neste sentido, o presente trabalho se propõe a avaliar os efeitos da degradação na resolução espacial e na resolução de contraste em tarefas de segmentação e classificação de tumores cerebrais, considerando três arquiteturas do estado da arte: Mobilenet (Howard et al., 2017), Vgg16 (Simonyan & Zisserman, 2014) e SEResNeXt50 (Hu et al., 2018).

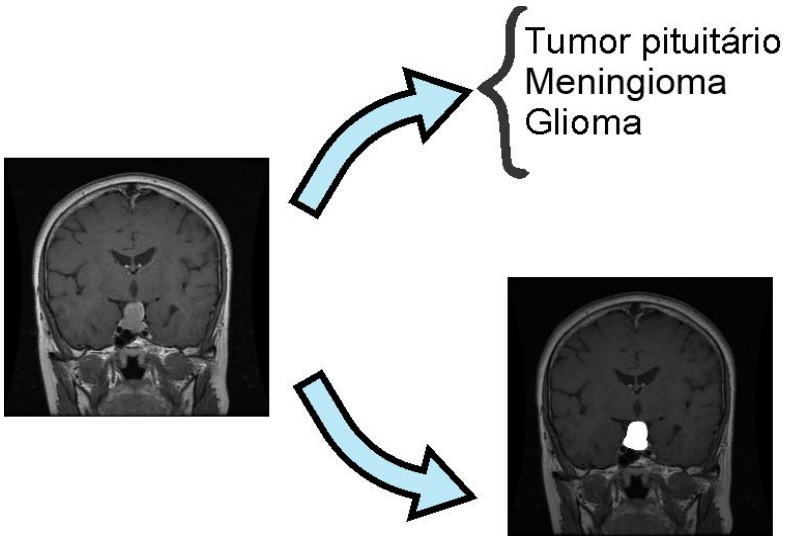
2. Metodologia

Neste estudo, foi utilizado um conjunto de dados (*dataset*) de imagens diagnosticadas com tumores cerebrais, contendo 3064 imagens por ressonância magnética, ponderadas em T1, com administração de contraste (Cheng et al., 2015; 2016). As fatias (*slices*) coletadas contêm 708 meningiomas, 1426 gliomas e 930 tumores pituitários, de 233 pacientes que fizeram os exames no Nanfang Hospital, Guangzhou, China e no General Hospital, Tianjing Medical University, China. Elas possuem matriz de 512x512, tamanho de pixel de 0,49×0,49 mm2, espessura de corte de 6 mm e *slicegap* de 1 mm.

No âmbito do pré-processamento dos dados, tanto na tarefa de classificação quanto de segmentação (Figura 1), a dimensão das imagens teve que ser reduzida para 256x256, devido à capacidade computacional utilizada. Além disso, esses dados também foram normalizados para que os pixels tivessem uma distribuição similar, favorecendo, assim, a redução do tempo de cálculo (Sola & Sevilla, 1997).

Em especial, para a tarefa de segmentação, a expansão de dados (*data augmentation*) – considerando rotação, ampliação, deslocamento, contraste e distorções geométricas – foi utilizada com a finalidade de diversificar as amostras e evitar sobreajuste (*overfitting*), aumentando, assim, o número de amostras de treino em 50% (Bu et al., 2022).

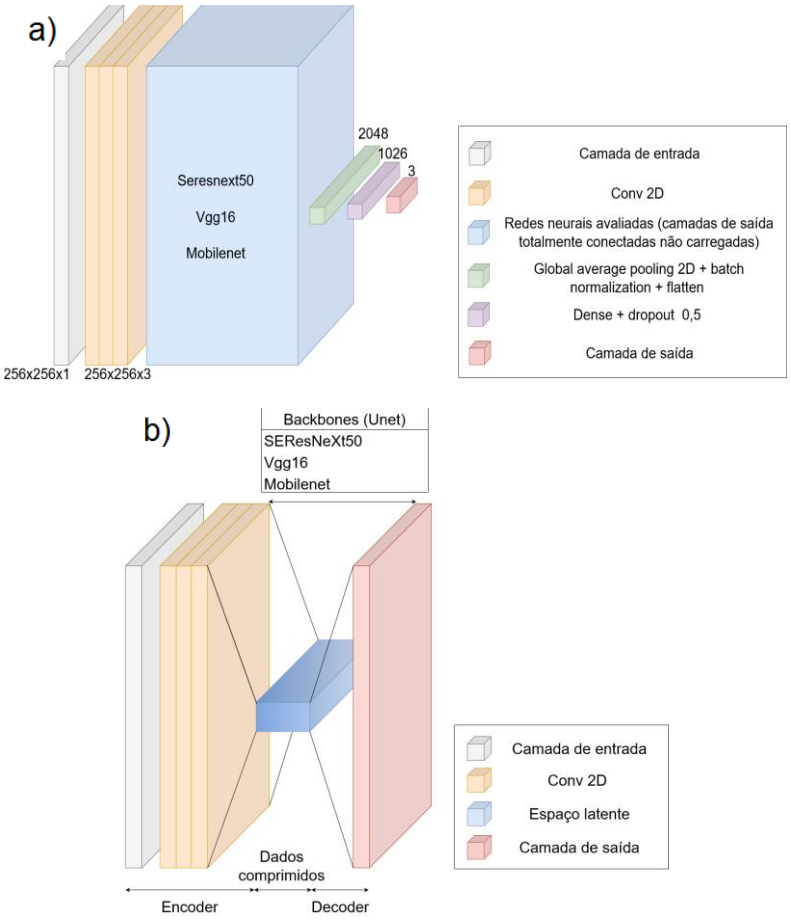
Figura 1 – Tarefas de classificação e segmentação semântica.



Fonte: Autores (2022).

Para executar as tarefas propostas, foram utilizadas três arquiteturas, aplicadas à classificação e à segmentação, como mostra a Figura 2. Além disso, foi utilizado o recurso de aprendizado por transferência, a fim de tornar este estudo exequível e viabilizar o aproveitamento do aprendizado de características de baixo nível em domínios diferentes. Vale a pena ressaltar que essas redes neurais foram desenvolvidas com uso da linguagem Python 3.10 (Van Rossum & Drake, 2009), da biblioteca Tensorflow 2.8.0 (Abadi et al., 2016), com sua API Keras (Chollet et al., 2015), e sendo executadas em Unidades Gráficas de Processamento (GPU) e Unidades de Processamento de Tensores (TPU).

Figura 2 – Redes estudadas. a) Rede neural convolucional com diferentes arquiteturas pré-treinadas usando ImageNet b) *Autoencoder* convolucional baseado na rede U-net e estruturado com diferentes arquiteturas pré-treinadas usando ImageNet.



Fonte: Autores (2022).

Como dito anteriormente, foram utilizadas 3 arquiteturas para as redes neurais com diferentes tamanhos, como descrito na Tabela 1.

Tabela 1 – Arquiteturas usadas na etapa de validação (10x10 *stratified k-foldcross-validation*) usando *early stop* e TPU.

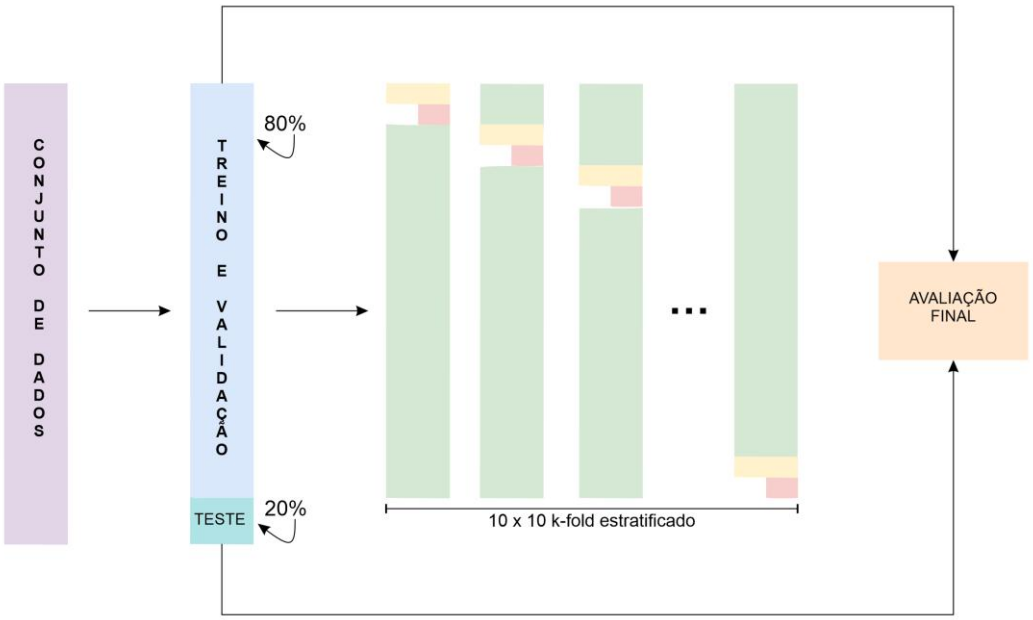
Arquiteturas	**Total de parâmetros**	**Nº médio de épocas por treino**	**Tempo médio de treino (hh:mm:ss)**
SEResNeXt-50 (classificação)	27.692.673	36	00:14:33
SEResNeXt-50 (segmentação)	34.594.183	36	00:23:18
Vgg16 (classificação)	15.246.161	29	00:02:50
Vgg16 (segmentação)	23.752.279	43	00:07:08
Mobilenet (classificação)	4.287.697	29	00:02:34
Mobilenet (segmentação)	8.336.343	44	00:06:17

Fonte: Autores (2022).

Para que o estudo fosse realizado de forma a obter modelos generalizáveis, por conseguinte, menos sobreajustados, o *dataset* foi dividido em conjuntos de treino/validação e teste, com o objetivo de treinar, avaliar e validar o funcionamento do modelo.

Em termos de confrontação entre modelos e validação estatística, foi utilizada a estratégia de 10x10 *stratifiedk-foldcross-validation*, uma adaptação da 10x10 *k-foldcross-validation* em que as partições contêm as mesmas proporções das classes que, nesse caso, são os achados nas imagens por ressonância magnética, como pode ser visto na Figura 3. Vale a pena ressaltar também que as métricas utilizadas nas tarefas de classificação foram: área sob a curva ROC (AUC) e F1, enquanto para as tarefas de segmentação foi usada F1, também conhecida como Dice. AUC foi escolhida porque ilustra o desempenho de um classificador em todos os seus limiares, F1 por ser uma média harmônica entre precisão e revocação, e Dice por ser um índice associado à sobreposição espacial.

Figura 3 – Esquema de divisão do *dataset*.



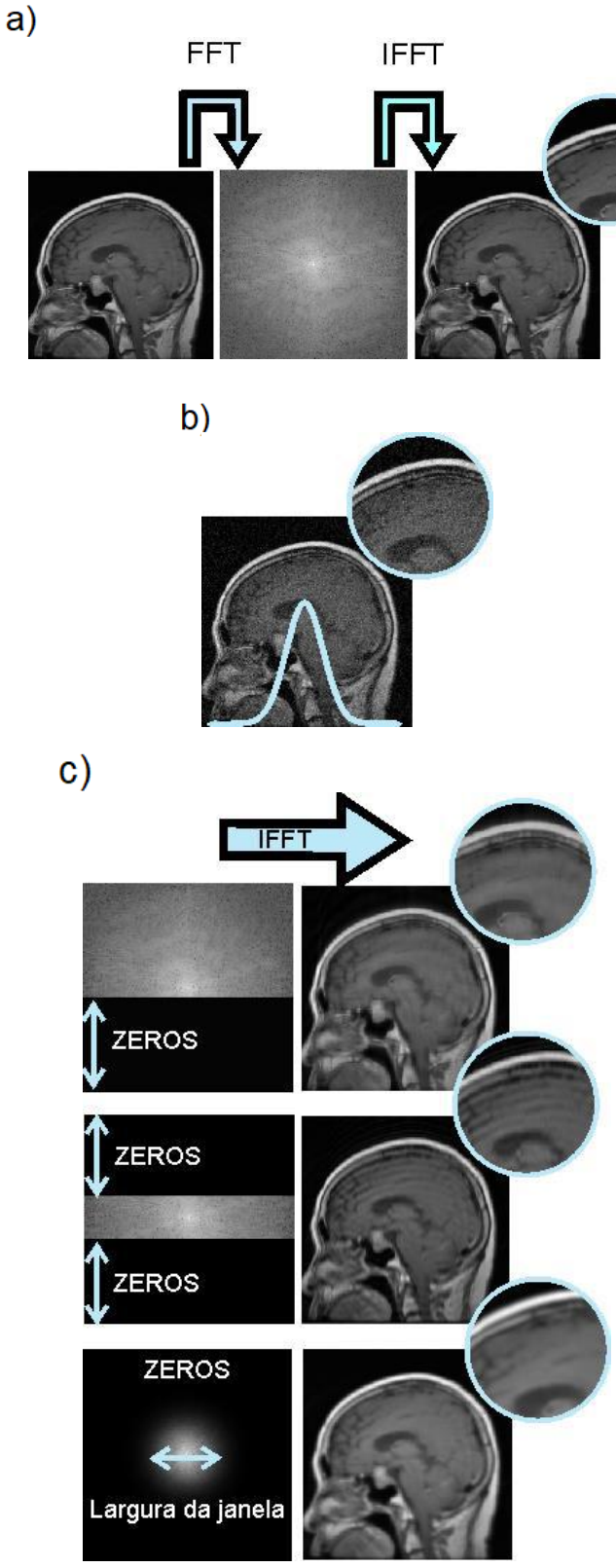
Fonte: Autores (2022).

A comprovação estatística após a rotina de treino e validação consistiu em coletar as médias de cada *cross-validation* para cada uma das dez execuções e, posteriormente, executar a análise de variância entre as médias das métricas coletadas – AUC, F1 e Dice. Dentre os métodos estatísticos para esta análise, foi escolhido o teste de Kruskal-Wallis (Conover, 1999), pois, de forma geral, as amostras tinham variâncias heterogêneas, comprovadas pelo teste de Levene (1961), e distribuições não normais, atestadas pelo método de Shapiro-Wilk (Shapiro & Wilk, 1965).

A posteriori, um novo treino foi executado com os dados de treino e validação utilizados na etapa anterior que, por meio disso, possibilitou aferir o desempenho dos modelos para diferentes graus de deterioração da resolução espacial e de contraste, tanto no domínio espacial quanto no da frequência.

Nesse sentido, foram considerados quatro tipos de distorções aplicados ao grupo de teste: Fourier parcial com zeros substituindo as informações na porção inferior do espaço k, fourier parcial com zeros substituindo as informações na porção inferior e superior do espaço k na mesma proporção, filtro passa-baixa no espaço k, e injeção de ruído gaussiano na imagem final, como é demonstrado na Figura 4 a seguir.

Figura 4 – Distorções aplicadas às imagens do conjunto de teste. a) Transição entre domínios b) Injeção de ruído de distribuição gaussiana c) Transformações no espaço de frequências.



Fonte: Autores (2022).

Essas transformações foram interessantes, já que puderam simular, mesmo que de forma elementar, deteriorações que

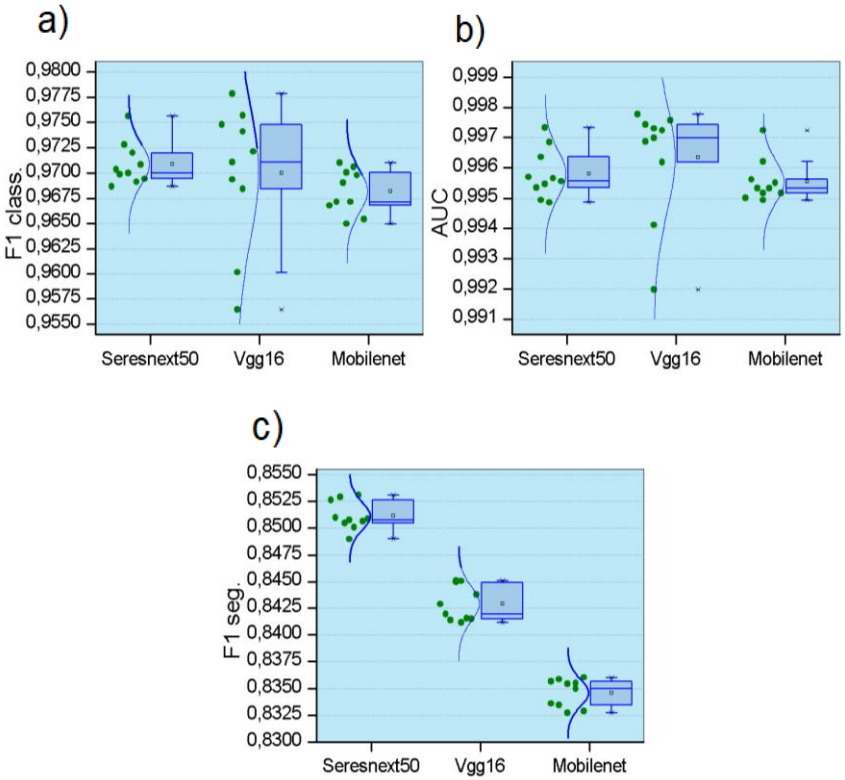
podem acontecer com as imagens por ressonância magnética, principalmente quando se quer acelerar um exame ou quando o protocolo não se encontra otimizado (Stadler et al., 2007).

3. Resultados e Discussão

Antes de apresentar a avaliação de desempenho sob diferentes graus de distorções, foi realizada a validação estatística dos modelos propostos sem distorção, por meio da comparação dos métodos de classificação e segmentação. Em relação a isso, estão sendo evidenciados que os modelos de classificação não apresentam desempenhos significativamente diferentes em termos de AUC e F1, para uma significância de 0,05 ($p>0,05$). Já no caso da tarefa de segmentação semântica, diferenças significativas foram atestadas, considerando a métrica Dice e nível de significância estatística de 0,05 ($p<0,05$).

A Figura 5 mostra a variação das médias para cada tarefa e para cada métrica considerada nas avaliações.

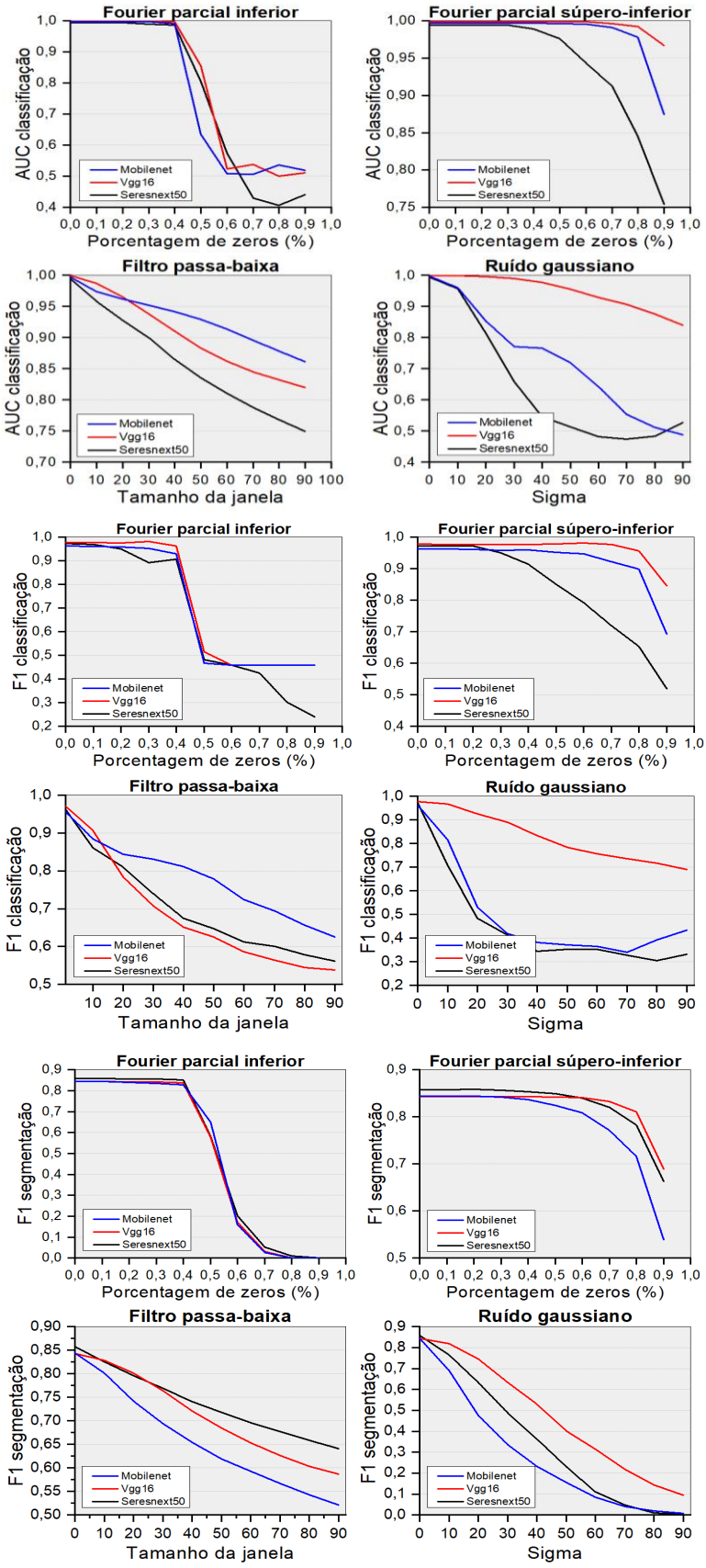
Figura 5 – Variação das métricas para as tarefas propostas. a) AUC aplicada à classificação b) F1 aplicada à classificação c) F1 aplicada à segmentação.



Fonte: Autores (2022).

Os resultados referentes à resiliência de cada rede neural na presença de deteriorações relativas à resolução espacial e à resolução de contraste são apresentados na Figura 6. O intervalo do nível das degradações foi implementado para que análises mais detalhadas pudessem ser feitas para cada caso isoladamente.

Figura 6 – Métricas de avaliação de desempenho sob diferentes graus de distorções.



Fonte: Autores (2022).

De acordo com a Tabela 1, é possível verificar que as redes neurais trabalhadas possuem profundidades diferentes e, consequentemente, tempos de treino diferentes. Correlacionando essa informação aos resultados mostrados na Figura 5, é possível notar, estatisticamente, que nem sempre uma rede com mais parâmetros de treinamento necessariamente vai fornecer um melhor resultado em termos de tarefas relacionadas à classificação de imagens por ressonância magnética. Contudo, em tarefas envolvendo segmentação semântica de achados, a rede com mais parâmetros – SEResNeXt50 – demonstrou um melhor resultado no contexto proposto.

Já analisando a Figura 6, observa-se que as redes estudadas são sensíveis a imagens de baixa qualidade e, consequentemente, obtiveram respostas muito diferentes, a depender do grau de degradação utilizado no conjunto de imagens de teste. Essa redução de performance pode ser explicada pela remoção de textura inerente aos processos executados nas imagens e, consequentemente, pode fazer com que a rede tenha dificuldade em identificar características de baixo nível, podendo propagar respostas distorcidas, conforme o nível da camada aumente (Dodge & Karam, 2016).

Além disso, levando em consideração as tarefas de classificação, percebe-se uma resiliência inicial significativa e uma posterior queda acentuada quando as linhas horizontais de dados são retiradas das imagens, além de uma queda gradual quando o filtro passa-baixo e o acréscimo de ruído gaussiano são gradualmente intensificados. É importante ressaltar que, com exceção do filtro passa-baixo, a arquitetura Vgg16 obteve o melhor grau de resistência a fatores que prejudicam a resolução da imagem, considerando as métricas propostas para avaliar a performance das redes nesse tipo de tarefa.

Considerando a tarefa de segmentação do achado, verifica-se que a Vgg16 também obteve maior resiliência a fatores degradantes, de forma parecida com a tarefa de classificação.

Outrossim, nestas duas tarefas, a Mobilenet e a SEResNeXt tiveram um comportamento deveras notável quando se analisam imagens nas quais as altas frequências são majoritariamente atenuadas.

4. Considerações Finais

Foi observado, a partir dos resultados obtidos neste estudo, que as redes neurais testadas são suscetíveis a degradações na resolução. Também foi demonstrada uma boa resiliência, por parte das arquiteturas estudadas, quando técnicas elementares e artificiais de reconstrução são feitas, além de sensibilidade considerável se forem levados em consideração os efeitos de ruído e suavização. Isso é bastante relevante, pois não é apenas direcionado a uma arquitetura específica ou a uma dada tarefa, mas indistinto para as condições examinadas. Salienta-se também que o modelo Vgg16 apresentou uma robustez superior comparado aos outros na maioria das tarefas.

Por fim, é importante considerar em trabalhos futuros a necessidade por redes que não possuam grande profundidade e sejam menos sensíveis a fatores de degradação da resolução espacial e de contraste, uniformização de *datasets* ou *augmentation*, levando em consideração as deteriorações comuns ao contexto clínico e também comparações entre a resiliência do sistema visual humano e de redes neurais profundas sob mesmas condições de deterioração da qualidade da imagem.

Agradecimentos

Os autores agradecem ao Programa de Residência Profissional em Física Médica subsidiado pelo MEC e à Comissão de Residência Profissional e Multiprofissional em Saúde do Hospital Universitário da Universidade Federal de Sergipe, administrado pela EBSERH.

Referências

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. ArXiv Preprint ArXiv:1603.04467.

Bu, R., Xiang, W., & Cao, S. (2022). COVID-19 Interpretable Diagnosis Algorithm Based on a Small Number of Chest X-Ray Samples. Journal of Shanghai Jiaotong University (Science), 27(1), 81-89.

Cheng, J., Huang, W., Cao, S., Yang, R., Yang, W., Yun, Z., & Feng, Q. (2015). Enhanced performance of brain tumor classification via tumor region augmentation and partition. PloS One, 10(10), e0140381.

Cheng, J., Yang, W., Huang, M., Huang, W., Jiang, J., Zhou, Y., & Chen, W. (2016). Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. PloS One, 11(6), e0157112.

Chollet, F. et al. Keras. (2015). https://keras.io/.

Conover, W. J. (1999). Practical nonparametric statistics. John Wiley & Sons.

Cui, S., Tseng, H. H., Pakela, J., Ten Haken, R. K., & El Naqa, I. (2020). Introduction to machine and deep learning for medical physicists. Medical Physics, 47(5), e127-e147.

Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. In 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX) (pp. 1-6). IEEE.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. ArXiv Preprint ArXiv:1704.04861.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

Koziarski, M., & Cyganek, B. (2018). Impact of low resolution on image recognition with deep neural networks: An experimental study. International Journal of Applied Mathematics and Computer Science, 28(4).

Levene, H. (1961). Robust tests for equality of variances. Contributions to probability and statistics. Essays in Honor of Harold Hotelling, 279-292.

Lobo, L. C. (2017). Inteligência artificial e medicina. Revista Brasileira de Educação Médica, 41, 185-193.

McRobbie, D. W., Moore, E. A., Graves, M. J., & Prince, M. R. (2017). MRI from Picture to Proton. Cambridge University Press.

Munir, K., Elahi, H., Ayub, A., Frezza, F., & Rizzi, A. (2019). Cancer diagnosis using deep learning: a bibliographic review. Cancers, 11(9), 1235.

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.

Sahiner, B., Pezeshk, A., Hadjiiski, L. M., Wang, X., Drukker, K., Cha, K. H., Summers, R. M., & Giger, M. L. (2019). Deep learning in medical imaging and radiation therapy. Medical Physics, 46(1), e1–e36. https://doi.org/10.1002/mp.13264

Shaphiro, S., & Wilk, M. B. J. B. (1965). An analysis of variance test for normality. Biometrika, 52(3), 591-611.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. ArXiv Preprint ArXiv:1409.1556.

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on Nuclear Science, 44(3), 1464-1468.

Stadler, A., Schima, W., Ba-Ssalamah, A., Kettenbach, J., & Eisenhuber, E. (2007). Artifacts in body MR imaging: their appearance and how to eliminate them. European Radiology, 17(5), 1242-1255.

Thambawita, V., Strümke, I., Hicks, S. A., Halvorsen, P., Parasa, S., & Riegler, M. A. (2021). Impact of Image Resolution on Deep Learning Performance in Endoscopy Image Classification: An Experimental Study Using a Large Dataset of Endoscopic Images. Diagnostics, 11(12), 2183.