UNIVERSIDADE FEDERAL DE SÃO PAULO
INSTITUTO DE CIÊNCIA E TECNOLOGIA
GRADUAÇÃO EM ENGENHARIA BIOMÉDICA

Avaliação do Potencial da ResNet-50 para Distinção de Tumores Neurais em Imagens Patológicas

Aluno: Giovanna Calabrese dos Santos
Coorientadora: Profª. Drª. Anna Luíza Damaceno Araújo
Orientador: Prof. Dr. Matheus Cardoso Moraes

Trabalho apresentado ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

São José dos Campos, São Paulo
2023

FOLHA DE APROVAÇÃO

GIOVANNA CALABRESE DOS SANTOS

Avaliação do Potencial da ResNet-50 para Distinção de Tumores Neurais em Imagens Patológicas

Trabalho apresentado ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

São José dos Campos, 01 de Julho de 2023.

Banca Avaliadora

Prof. Dr. Henrique Mohallem Paiva
(Presidente - Universidade Federal de São Paulo)

Profª. Drª. Regina Célia Coelho
(Membro 1 - Universidade Federal de São Paulo)

Profª. Drª. Thaína Aparecida Azevedo Tosta
(Membro 2 - Universidade Federal de São Paulo)

Trabalho apresentado ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

Calabrese dos Santos, Giovanna

Avaliação do Potencial da ResNet-50 para Distinção de Tumores Neurais em Imagens Patológicas / Giovanna Calabrese dos Santos. – São José dos Campos, São Paulo, 2023.
8 f. : il.
Orientador: Matheus Cardoso Moraes
Coorientadora: Anna Luíza Damaceno Araújo

Monografia (Curso de Graduação em Engenharia Biomédica) – Instituto de Ciência e Tecnologia, Universidade Federal de São Paulo, 2023.

1. Classificação. 2. Tumores Neurais. 3. ResNet-50.
I. Cardoso Moraes, Matheus. II. Damaceno Araújo, Anna Luíza. III. Instituto de Ciência e Tecnologia. IV. Universidade Federal de São Paulo. V. Avaliação do Potencial da ResNet-50 para Distinção de Tumores Neurais em Imagens Patológicas.

Trabalho apresentado ao Instituto de Ciência e Tecnologia da Universidade Federal de São Paulo como requisito parcial para obtenção do título de Graduação em Engenharia Biomédica.

Avaliação do Potencial da ResNet-50 para Distinção de Tumores Neurais em Imagens Patológicas

Giovanna Calabrese dos Santos[1], Anna Luíza Damaceno Araújo[2], Matheus Cardoso Moraes[1]

[1] Instituto de Ciência e Tecnologia - Universidade Federal de São Paulo / Engenharia Biomédica
[2] Faculdade de Medicina - Universidade de São Paulo / Departamento de Cabeça e Pescoço

***Resumo*—Tumores neurais são difíceis de diferenciar apenas com base na celularidade e, geralmente, requerem coloração imuno-histoquímica para auxiliar na identificação da linhagem celular. Os tumores podem ser assintomáticos e benignos, e, ainda assim gerar complicações e afetar a qualidade de vida do paciente, especialmente se localizados em regiões vitais do corpo. É fundamental um diagnóstico preciso e precoce para garantir o melhor tratamento, o que muitas vezes não ocorre devido ao alto custo das metodologias tradicionais de diagnóstico. Este artigo aborda a implementação da Rede Neural Convolucional ResNet-50 para diagnóstico precoce a partir da diferenciação dos três tipos mais comuns de tumores neurais - neurofibroma, perineurioma e schwannoma - pela análise anatomopatológica da lesão. A partir da arquitetura da ResNet-50 foram gerados, treinados e validados dois modelos derivados (A e B) para a classificação. Em ambos, foi utilizado o mesmo banco de dados de 30 pacientes, com o total de 106.782 *patches* divididos entre os conjuntos de treinamento, validação e teste, com estratégias para evitar *overfitting* e *data-leakage*. O Modelo A apresentou uma acurácia de 69% e uma perda de 4,3, enquanto o Modelo B obteve, respectivamente, 66% e 3,7. Ambos apresentaram resultados satisfatórios para a diferenciação de apenas duas das três classes de entrada, alcançando cerca de 90% e 80% como verdadeiro positivo no Modelo A, e 84% e 71% no Modelo B. No geral, os resultados foram prejudicados pelo baixo volume de dados e pequena variabilidade em uma das classes do problema. Como próximos passos, almeja-se a melhoria das métricas a partir do aumento do banco de dados, melhoria na proporção de divisão dos conjuntos, inclusão de outros tipos de tumores neurais para classificação e aplicação e avaliação de outras redes neurais para o mesmo tipo de classificação.**

Abstract*— *Neural tumors are difficult to distinguish based solely on cellularity and often require immunohistochemical staining to assist in identifying the cell lineage. Tumors can be asymptomatic and benign yet still generate complications and affect the patient's quality of life, mainly if located in vital body regions. Accurate and early diagnosis is essential to ensure the best treatment, which often does not occur due to the high cost of traditional diagnostic methodologies. This article addresses the implementation of the Convolutional Neural Network ResNet-50 for early diagnosis based on the differentiation of the three most common types of neural tumors - neurofibroma, perineurium, and schwannoma - through anatomopathological analysis of the lesion. Two derived models (A and B) were generated, trained, and validated for classification using the ResNet-50 architecture. Both used the same database of 30 patients, with 106,782 patches divided among the training, validation, and testing sets, with strategies to avoid overfitting and data leakage. Model A achieved an accuracy of 69% and a loss of 4.3, while Model B achieved 66% and 3.7, respectively. Both showed satisfactory results for differentiating only two of the three input classes, reaching approximately 90% and 80% as true positive in Model A, and 84% and 71% in Model B. Overall, the results were hindered by the low volume of data and small variability in one of the problem classes. In the following steps, is an aim to improve the metrics by increasing the database, improving the proportion of set division, including other types of neural tumors for classification, and applying and evaluating other neural networks for the same type of classification.

***Index Terms*—Classificação, Tumores Neurais, ResNet-50, Biópsia, Diagnóstico.**

I. Introdução

A. Problema

Os tumores neurais classificam-se como neoplasias primárias dos nervos periféricos, ou seja: lesões nos prolongamentos dos neurônios, responsáveis pela transmissão de estímulos nervosos ao restante do corpo. A ocorrência destes é comum na região da cabeça e pescoço (CP), podendo acometer tecidos moles e apresentando grande incidência em tecidos com maiores chances de trauma, como a língua [1][2].

Tumores de origem neural são difíceis de diferenciar apenas com base na celularidade e, geralmente, requerem coloração imuno-histoquímica para auxiliar na identificação da linhagem celular. Isso ocorre porque a morfologia destes pode se sobrepor a de outros tumores, como os fusocelulares de origem endotelial, fibroblástica e miofibroblástica. Clinicamente, a presença desses tumores neurais pode ser silenciosa e, portanto, assintomática. No entanto, indivíduos com a suspeita patológica podem relatar dormência na região afetada, sensibilidade ou dor [1].

Dentro dessa categoria, portanto, há uma grande sobreposição de características e proximidade entre os diagnósticos. Dessa forma, além de tardios, esses diagnósticos podem ser feitos erroneamente, o que contribui para um desfecho mais desfavorável [2].

1) Tipos de Tumores: Os tumores neurais são divididos entre não-encapsulados e encapsulados (nódulos cercados por capsula fibrosa). O neurofibroma e o perineurioma representam diagnósticos diferenciais importantes relativos ao primeiro grupo, pois podem apresentar estroma mucoso ou mixóide; enquanto o neuroma encapsulado em paliçada (NEP) e o schwannoma são os diagnósticos diferenciais mais comuns para a segunda classe e podem apresentar algum grau de paliçada (feixes compactos e paralelos de fascículos de células de Schwann).

Neurofibromas afetam a bainha do nervo periférico, composta por uma mistura de células de Schwann, células perineurais, e fibroblastos intraneuronais. Podem estar associados à síndrome genética neurofibromatose tipo I, na qual se desenvolvem múltiplos neurofibromas [3].

O perineurioma é um tumor de tecidos moles composto por células fusiformes e células *fibroblast-like* arranjadas em padrão estoriforme (células orientadas em diversas direções). O termo "perineurioma" é tradicionalmente usado para tumores nos quais a grande maioria das células apresenta diferenciação perineural [4][5].

Dentre os encapsulados, o NEP é caracterizado como uma sólida proliferação e expansão das células de Schwann e áxons de um nervo periférico cutâneo. É formado por nódulos encapsulados na derme profunda e tecidos subcutâneos, apresentando-se, portanto, na pele ou também em mucosas oral, nasal ou peniana [6].

E, por fim, o schwannoma (ou neurilemoma) trata-se de um tipo de tumor benigno que afeta as células de Schwann. Constitui-se no padrão de paliçada alternada e as células tumorais são fusiformes com extensões citoplasmáticas alongadas, conferindo um aspecto ondulado a espiralado. Apesar de comum na região de CP, esta patologia representa apenas 8% dentre os tipos de tumores do sistema nervoso central (SNC). Além disso, também pode se originar nos nervos fora da medula espinal e gerar complicações relativas à pressão exercida nesta, como fraqueza, perda sensorial ou problemas de controle intestinal ou da bexiga, comumente vistos em casos de lesões como para- ou tetraplegia [7].

A Figura 1 mostra lâminas de biópsias com as características histológicas de cada um dos tumores apresentados. É com base na análise destas que o diagnóstico e distinção da patologia é realizado.

Em resumo, essas neoplasias primárias costumam ser raras e benignas, com crescimento lento (exceto schwannoma) e praticamente assintomáticas. No entanto, é importante destacar que, apesar desses aspectos, tais patologias podem gerar complicações e sintomas que afetam a qualidade de vida do paciente, especialmente se estiverem localizadas em regiões importantes do corpo. Portanto, é fundamental que o diagnóstico seja feito correta e precocemente, a fim de garantir o melhor tratamento.

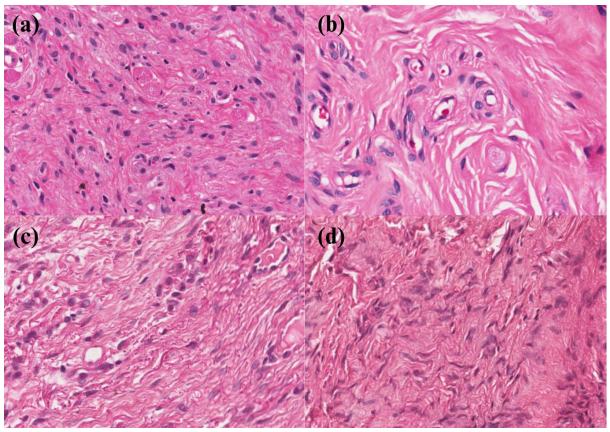

Figura 1: (a) Neurofibroma, (b) perineurioma, (c) NEP e (d) schwannoma. Fonte: Laboratório de Patologia Oral da Faculdade de Odontologia de Piracicaba (FOP-UNICAMP).

B. Solução Atual

Em princípio, o diagnóstico de lesões neurais é conduzido a partir de biópsia da lesão seguida por análise histopatológica através da observação de lâminas coradas em hematoxilina e eosina (H&E) e, eventualmente, por meio de reações de imuno-histoquímica.

Tipicamente, os tumores neurofibroma, schwannoma e NEP são facilmente caracterizados com base na morfologia. No entanto, alguns casos com características histológicas atípicas, como menor grau de estroma paliçado e mixóide, geralmente requerem imuno-histoquímica para caracterizar a diferenciação dos elementos neurais.

De acordo com o estudo de Araújo et al. [8], a alta frequência de uso de um grupo específico de marcadores imuno-histoquímicos está diretamente relacionada com a grande quantidade de lesões neurais diagnosticadas no respectivo serviço, as quais foram as lesões que mais frequentemente exigiram caracterização imuno-histoquímica para ajudar na discriminação de linhagem celular epitelial, neural, melanocítica, de músculo liso, endotelial, e fibroblásticas ou miofibroblásticas.

Adicionalmene, este levantamento demonstrou que, em treze anos, apenas uma pequena porcentagem de casos recebidos no Laboratório de Patologia Oral da FOP-UNICAMP requisitou coloração imuno-histoquímica e que apenas 28% das reações realizadas foram essencialmente necessárias para se chegar ao diagnóstico definitivo. Isso significa que, além

de ser uma etapa demorada e financeiramente custosa, a realização de reações imuno-histoquímicas é indispensável em apenas uma minoria de casos e que, mesmo com esse advento, nos casos em que se faz necessário o estudo imuno-histoquímico, o diagnóstico a nível histológico nem sempre é simples [8].

Em contrapartida, algoritmos de visão computacional se apresentam, então, como a possibilidade de um sistema de suporte ao diagnóstico mais acessível, rápido e menos oneroso.

C. O Estado da Arte em Computer-Aided Diagnosis (CAD)

Com a evolução da tecnologia e aumento de suas aplicações no setor da saúde, está cada vez mais difundida a utilização de inteligência artificial (IA) para diagnóstico e classificação de patologias. Primeiramente com uso de *Machine Learning* (ML) e depois de *Deep Learning* (DL), inovações vêm sendo aplicadas para detecção precoce de câncer por exames de imagem, sendo possível sua identificação tanto no cérebro até na pele e outros órgãos [9].

No caso dos tumores nos nervos periféricos, Mazal et al. [10] fizeram um estudo para diferenciação destas patologias em relação a estrutura da massa tumoral, treinando um modelo de rede neural convolucional, do inglês *Convolutional Neural Network* (CNN), para classificação de massas identificadas em exames de ressonância magnética (RM) entre benignas ou malignas. Das duas redes desenvolvidas na pesquisa, os resultados foram satisfatórios e também semelhantes às interpretações de radiologistas especialistas com relação aos mesmos exames testados nas redes (Tabela I).

Tabela I: Precisões e AUCs de modelos CNN e radiologistas. Fonte: [10].

	Acurácia	**AUC**
CNN 1 (fsT2W)	87%	0,89
Radiologista 1	73%	0,83
Radiologista 2	93%	0,83
CNN 2 (fsT2W + T1W)	93%	0,94
Radiologista 1	71%	0,81
Radiologista 2	71%	0,70

Dessa forma e com base na literatura disponível até o momento de escrita deste trabalho, tem-se que a investigação aqui presente ainda não foi realizada para imagens de lâminas patológicas, já que o estudo referenciado analisou imagens de RM para classificação dos tumores em nervos periféricos.

D. Objetivo

Desenvolver e aplicar modelos de inteligência artificial (IA) baseados em *Representation Learning* (RL), mais especificamente DL, para diagnóstico microscópico de tumores neurais, focando, em princípio, na diferenciação das classes neurofibroma, perineurioma e schwannoma.

II. Materiais e Métodos

A. Amostra

As imagens, advindas da parceria com a FOP-UNICAMP, foram obtidas a partir do escaneamento de lâminas histológicas utilizando o *Aperio Digital Pathology System* (*Leica Biosystems*, Wetzlar, Alemanha), com uma amostragem espacial de 0,47μm por pixel, foco e ampliação automatizadas em 20x.

Ao todo foram incluídos 30 pacientes que haviam sido submetidos a biópsia incisional com anestésico local (lidocaína 2%). Após a biópsia e ressecção dos tumores, os espécimes foram colocados em formol tamponado a 10%. A amostra foi selecionada, retrospectivamente, pelo levantamento de lesões compatíveis com o diagnóstico histológico dentro das categorias neurofibroma (n_0 = 9), perineurioma (n_1 = 7) e schwannoma (n_2 = 14).

A literatura científica não dispõe de uma orientação específica referente ao número amostral para a etapa de treinamento de modelos de IA, mas sugere que as amostras sejam heterogêneas quanto à origem, para que haja maior variabilidade de padrões que possibilitem a amplificação das amostras *input*, e grandes o suficiente, para a correta amostragem entre os conjuntos de treinamento, validação e teste, com propósito de evitar *data-leakage*, que pode causar superestimação dos resultados de validação externa do modelo [11].

B. Processamento das Imagens e Treinamento dos Modelos de IA

1) Anotação das Imagens: Em aprendizado supervisionado, a primeira etapa do pré-processamento das imagens é caracterizada pela anotação manual das imagens das lâminas, assim denominada pois depende de um operador/patologista experiente capaz de discriminar visualmente áreas de interesse (parênquima tumoral) das áreas de estroma (fibroblastos, células endoteliais, vasos sanguíneos, músculos, gordura e glóbulos vermelhos que são comuns a todas as lesões estudadas) e que não colaboram para o diagnóstico. Nessa etapa é possível evitar anotações de regiões de dobra de tecido e regiões brancas (ausência de tecido histológico) como uma etapa de remoção manual de ruído. As regiões de interesse foram anotadas por dois patologistas (D.G.R. e

S.S.S.N.) da FOP-UNICAMP e uma patologista (A.L.D.A.) da Faculdade de Medicina da Universidade de São Paulo (FMUSP). Os patologistas conduziram as anotações com auxílio do software de visualização *Aperio Image Scope* (*Leica Biosystems*, Wetzlar, Alemanha) e mesa digitalizadora *Huion Inspiroy H1060P Graphics Drawing Tablet*, utilizando o mesmo *workstation* no mesmo ambiente e sob as mesmas condições de iluminação.

2) Segmentação e Fragmentação - Criação de Patches: A partir da anotação das lâminas digitalizadas foi possível segmentar as regiões de interesse e, posteriormente, fragmentar essas imagens em *patches* de tamanhos padronizados de acordo com a entrada dos *kernels* da CNN utilizada (299 x 299 pixels). Esta etapa é necessária para mitigar o custo computacional de processamento pela CNN. Uma Unidade de Processamento Gráfico (do inglês, *Graphics Processing Unit* - GPU) dedicada foi utilizada para o treinamento da CNN em um *cluster*.

Cada *patch* foi salvo de acordo com a classificação inicial da imagem patológica anotada (neurofibroma, perineurioma ou schwannoma), com diferenciação e separação dos pacientes por pastas no diretório de trabalho para a criação de dois *datasets* distintos. Tais etapas estão ilustradas na Figura 2.

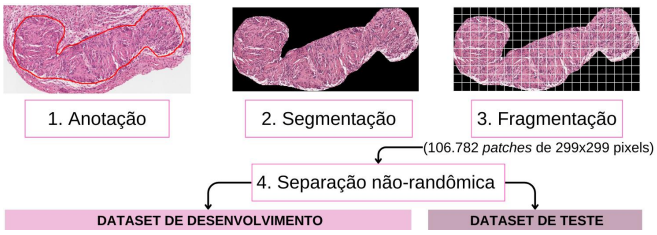

Figura 2: Fluxo de processamento das imagens histopatológicas anotadas.

3) Arquitetura da CNN: Uma Rede Neural Residual (ResNet) é uma Rede Neural Artificial (do inglês, *Artificial Neural Network* - ANN) que empilha blocos residuais um sobre o outro para formar uma rede. A ResNet tem muitas variantes que funcionam com o mesmo conceito, mas têm diferentes números de camadas. Em específico, "ResNet-50" é o título usado para denotar a variante que funciona com 50 camadas de rede neural. Este foi o modelo de CNN escolhido pois demonstrou um melhor desempenho reportado na literatura para classificação a partir da identificação de padrões [12].

Os parâmetros no início do processamento dos dados foram definidos empiricamente como: *learning rate* igual a 10^{-5}; uso do otimizador Adam; a medida de acerto do modelo no grupo validação utilizou a regra do *categorical crossentropy*, retornando a acurácia de acertos a cada época.

4) Treinamento, validação e teste: Para a definição dos grupos treinamento, validação e teste, e seguindo a ideia de amostra heterogênea dentro dos grupos para aumentar a variabilidade, tentou-se seguir a proporção de 8:1:1, respectivamente, e foram criados dois modelos de classificação com a rede ResNet-50:

- Modelo A: mistura de pacientes devido seleção randômica 8:1 para os grupos treinamento e validação;
- Modelo B: sem mistura de pacientes entre grupos.

Na construção dos modelos, a divisão dos *patches* foi realizada de acordo com a Tabela II para o Modelo A e de acordo com a Tabela III para o Modelo B.

Tabela II: Separação de *patches* por classe para Modelo A.

Classes	Treinamento-Validação	Teste
0 - Neurofibroma	21.925	2.347
1 - Perineurioma	21.227	1.694
2 - Schwannoma	53.756	5.833

Tabela III: Separação de *patches* por classe para Modelo B.

Classes	Treinamento	Validação	Teste
0 - Neurofibroma	19.416	2.779	2.077
1 - Perineurioma	21.227	688	1.006
2 - Schwannoma	47.659	6.049	5.881

Dado o tamanho da base de dados do estudo, não foi possível seguir a proporção a risca, a fim de evitar a inclusão de um mesmo paciente em mais de um grupo de dados, ou seja, evitar *data-leakage*. No entanto, a divisão foi próxima à proporção escolhida, mantendo o conjunto de treinamento como o maior.

C. Avaliação

Para avaliação do teste, foram extraídas as seguintes métricas: matriz de confusão, perda (*loss*), acurácia, precisão, verdadeiro positivo (VP), falso positivo (FP), verdadeiro negativo (VN), falso negativo (FP), sensibilidade (*recall*), especificidade, *F1-score*, curva ROC e área abaixo da curva (AUC), coeficiente de confiança Kappa (κ), predição negativa (VPN) e as taxas de omissão negativa (ON) e omissão positiva (OP) [13]. Um resumo das equações das taxas está pontuado na Tabela IV.

Tabela IV: Métricas extraídas dos Modelos.

Métrica	**Fórmula (%)**
Acurácia	$\frac{Acertos}{Total} \times 100$
Perda	$-\sum Real \times \log(Predict)$
Precisão	$\frac{Desvio\ Padrao}{Media} \times 100$
VP	$\frac{Predict\ Positivo}{Total\ Positivos} \times 100$
FP	$\frac{Predict\ Positivo}{Total\ Negativos} \times 100$
VN	$\frac{Predict\ Negativo}{Total\ Negativos} \times 100$
FN	$\frac{Predict\ Negativo}{Total\ Positivos} \times 100$
Sensibilidade	$\frac{\%VP}{\%VP + \%FN}$
Especificidade	$\frac{\%FP}{\%FP + \%FP}$
F1-score	$\frac{2 \times \%Acuracia \times \%Sensibilidade}{\%Acuracia + \%Sensibilidade}$
VPN	$\frac{\%VN}{\%VN + \%FN}$
ON	$\frac{\%FN}{\%VN + \%FN}$
OP	$\frac{\%FP}{\%VP + \%FN}$

A *pipeline* do algoritmo desenvolvido na linguagem *Python* 3.8, na IDLE *Spyder*, seguiu as principais etapas (1, 2 e 3) elucidadas na Figura 3.

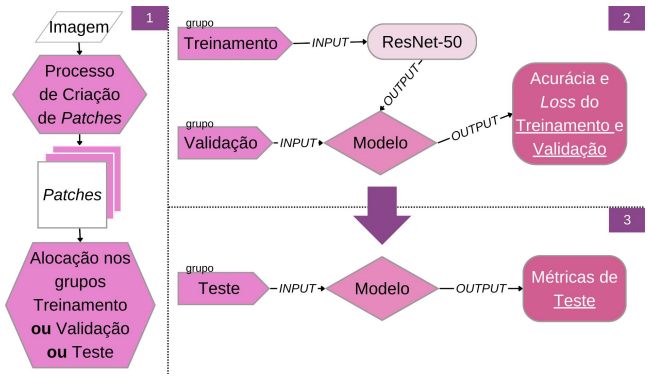

Figura 3: Fluxograma geral da modelagem.

III. Resultados

A. Modelo A

Após o treinamento do Modelo A com o grupo de imagens treinamento, decorreu-se a validação do sistema classificador com imagens distintas, pertencentes ao grupo validação. Nessa etapa, foram geradas as curvas de acurácia ("*accuracy*") e de perda ("*loss*") para o treinamento e a validação, conforme mostram os gráficos da Figura 4, respectivamente.

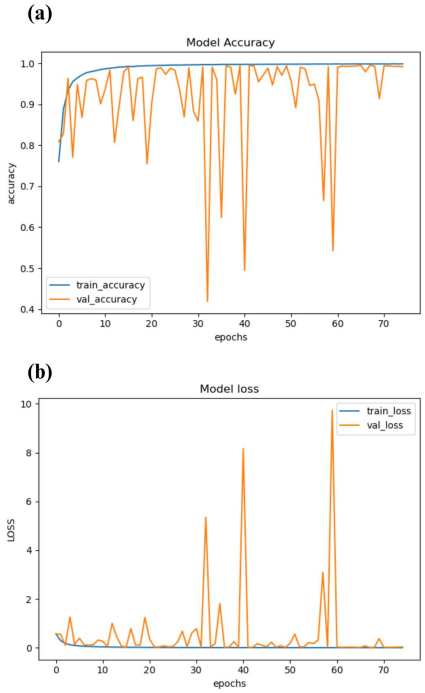

Figura 4: Modelo A: (a) Evolução da acurácia pelas épocas de treinamento e (b) evolução da perda pelas épocas de treinamento.

No processamento seguinte, as imagens do grupo de teste foram finalmente utilizadas para medir a eficácia do sistema de classificação com dados inéditos, ou seja, que não fizeram parte da etapa de treinamento ou de validação, evitando o *overfitting* do modelo aos dados disponíveis. O modelo teve uma acurácia de acertos igual a, aproximadamente, 69% e *loss* de 4,3; já a acurácia balanceada do modelo foi de aproximadamente 58%. As taxas de VP, FP, VN, FN e o índice Kappa (κ) estão dispostos na matriz de confusão normalizada da Figura 5 e as demais métricas retiradas do modelo estão pontuadas na Tabela V.

Também foi gerada a curva ROC do procedimento de teste, assim como a determinação da AUC para cada classe, conforme ilustrado na Figura 6.

Ainda, outra métrica extraída do modelo e que expõe seu desempenho associado à sensibilidade da classificação é o *F1-score*. No caso do Modelo A, a média entre as classes gerou um *F1-score* geral de, aproximadamente, 46%.

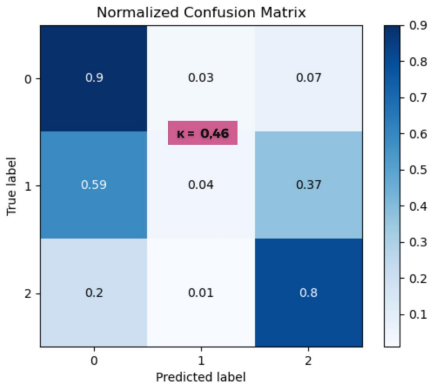

Figura 5: Matriz de confusão normalizada do Modelo A (κ = 0,46), em que a classe 0 é respectiva ao neurofibroma, classe 1 ao perineurioma e a classe 2 ao schwannoma.

Tabela V: Outras métricas de teste do Modelo A por classe.

Métrica	**0**	**1**	**2**
Precisão (%)	43	4,9	71
Sensibilidade (%)	90	4,3	80
Especificidade (%)	63	83	54
F1-score (%)	58	4,6	75
Predição negativa (%)	95	81	65
Omissão negativa (%)	4,8	19	35
Omissão positiva (%)	57	95	28

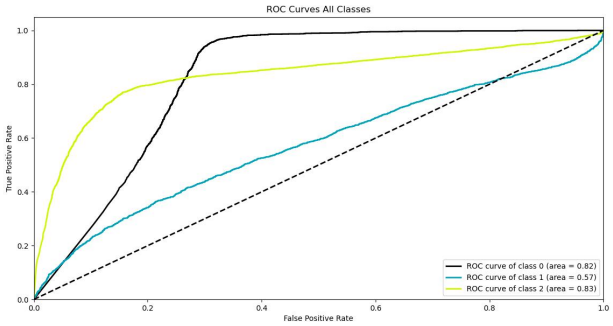

Figura 6: Curva ROC e AUC do teste no Modelo A, em que a classe 0 é respectiva ao neurofibroma, classe 1 ao perineurioma e a classe 2 ao schwannoma.

B. Modelo B

Para o Modelo B seguiram-se os mesmos passos. A acurácia de treino e de validação, além da perda em ambos, tiveram suas curvas geradas conforme as épocas e se encontram ilustradas na Figura 7, respectivamente.

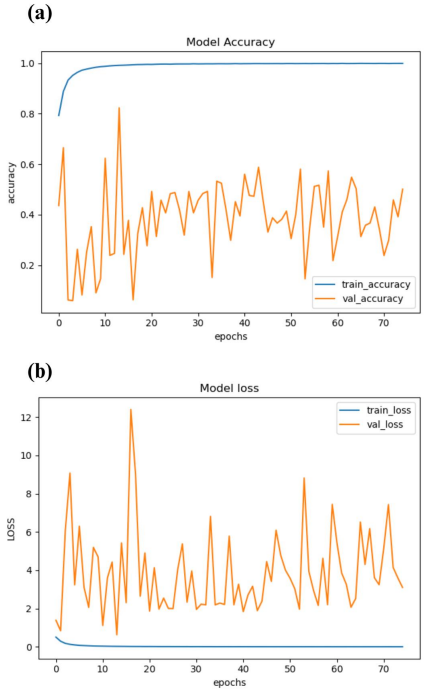

Figura 7: Modelo B: (a) Evolução da acurácia pelas épocas de treinamento e (b) evolução da perda pelas épocas de treinamento.

A acurácia final do modelo na etapa de teste foi de, aproximadamente, 66%, com *loss* igual a 3,7. Em relação a acurácia balanceada devido aos diferentes tamanhos das amostras em cada classe, obteve-se, aproximadamente, 52%. As taxas de VP, FP, VN, FN e o κ resumem-se na matriz de confusão normalizada (Figura 8) e as outras métricas computadas da etapa de teste estão dispostas na Tabela VI.

Além disso, as curvas ROC, além da AUC, para cada classe identificada, estão dispostas na Figura 9.

Quanto ao *F1-score* geral do Modelo B, a média resultante foi de 42%.

IV. Discussão e Conclusão

O sistema de classificação composto pelo Modelo A obteve métricas relativamente boas, com 69% de acurácia no teste (balanceada sendo 58%) e 46% de *F1-score*. Dada a matriz de confusão normalizada apresentada anteriormente (Figura 5), pode-se observar que a taxa de VP foi alta para duas classes: 0 - neurofibroma (90%) e 2 - schwannoma

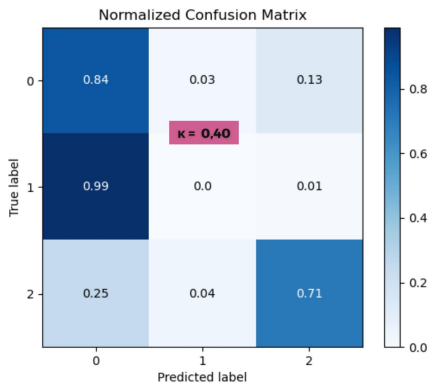

Figura 8: Matriz de confusão normalizada do Modelo B (κ = 0,40), em que a classe 0 é respectiva ao neurofibroma, classe 1 ao perineurioma e a classe 2 ao schwannoma.

Tabela VI: Outras métricas de teste do Modelo B por classe.

Métrica	**0**	**1**	**2**
Precisão (%)	39	0,15	76
Sensibilidade (%)	84	0,30	71
Especificidade (%)	61	75	57
F1-score (%)	54	0,20	73
Predição negativa (%)	93	86	51
Omissão negativa (%)	7	14	49
Omissão positiva (%)	61	100	24

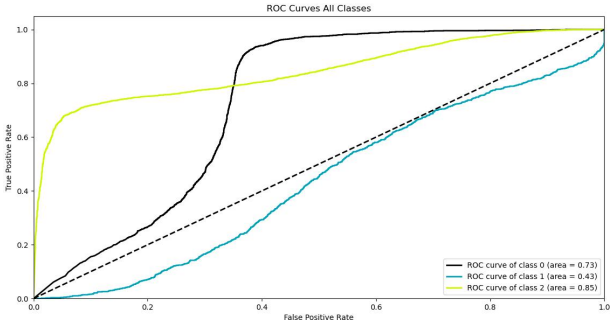

Figura 9: Curva ROC e AUC do teste no Modelo B, em que a classe 0 é respectiva ao neurofibroma, classe 1 ao perineurioma e a classe 2 ao schwannoma.

(80%), as quais possuíam os maiores conjuntos de dados do estudo, ou seja, com maior quantidade de pacientes e, portanto, maior variabilidade de características no tipo de lesão a ser aprendida. Em relação a classe 1 - perineurioma, devido ao banco de dados atualmente limitado, a taxa VP foi baixa (4%), e o classificador distribuiu os diagnósticos desta classe como se fossem das demais, o que, em contrapartida, acabou prejudicando as métricas de avaliação destas. Além disso, neste modelo decorreu-se a mistura de pacientes entre os grupos de treinamento e validação em todas as classes, fazendo com que a avaliação dele fosse superestimada, ou seja, melhor do que o esperado considerando a pequena quantidade de imagens utilizada em todo o projeto, já que o sistema foi treinado com amostras de um paciente e depois foi validado com amostras distintas, mas ainda de um mesmo paciente.

Em complemento, o Modelo B seguiu com o isolamento de cada paciente a um determinado grupo, sem ocorrência de *data-leakage*, o que resultou em métricas não-superestimadas e, portanto, relativamente mais baixas. No geral, a acurácia total final do modelo na etapa de teste foi de 66% (balanceada sendo 52%). Neste classificador, apurou-se a partir da matriz de confusão normalizada (Figura 8) que as classes 0 e 2 continuam com as taxas VP boas, com 84% e 71%, respectivamente, e a classe 1 continua com a métrica baixíssima (aproximadamente 0%), o que novamente afeta as taxas das demais classes. Analogamente ao caso do Modelo A, a principal causa disto ainda é a menor quantidade de imagens da classe.

No geral, apesar das curvas de treinamento e de validação do Modelo A (Figura 4), pode-se afirmar que houve um leve *overfitting* do modelo aos dados de treino devido à baixa acurácia geral obtida na etapa de teste. No entanto, no Modelo B, enxerga-se um *overfitting* mais acenturado já a partir dessas curvas (Figura 7); ou seja, houve o ajuste excessivo do Modelo B aos dados de treino já que as curvas das etapas de treino e validação em cada caso se distanciaram conforme o passar das épocas.

Ainda, o bom desempenho das versões para as classes 0 e 2 também é nítido na observação das curvas ROC (Figura 6 e Figura 9), trazendo AUC igual a, respectivamente: 0,82 e 0,83 (Modelo A), e 0,73 e 0,85 (Modelo B), dado que o valor ideal e máximo possível da AUC é 1,00. Em análise ao coeficiente de confiança Kappa de Cohen (κ), os modelos A e B resultaram em, respectivamente: 0,46 e 0,40, sendo a escala da métrica de -1 a 1. De acordo com a interpretação comum da medida [14], obteve-se uma força de concordância moderada para o Modelo A e uma regular ou suave para o Modelo B; a diferença observada se dá, novamente, devido a superestimação do primeiro.

Enfim, pode-se concluir que a distinção entre as patologias foi relativamente satisfatória nos dois modelos

ao menos para duas das três classes de entrada, justamente levando em consideração a limitação dada pela quantidade de imagens e pacientes disponíveis nesta primeira análise, além da complexidade da CNN escolhida, a ResNet-50, que performa melhor com uma quantidade maior de dados. Todavia, dentre os dois modelos testados, considerando as métricas de avaliação e o nível de *overfitting*, é plausível considerar o Modelo A superior ao Modelo B, sendo favorecido pelo *data-leakage*.

Futuramente, espera-se que o banco de dados a ser trabalhado disponha de maior variabilidade, ou seja, mais casos e mais pacientes, a fim de aumentar as métricas já obtidas e não ter perdas devido ao desbalanceamento entre classes. Além disso, há o interesse em testar outras redes de classificadores para avaliar a melhor solução ao problema de diagnóstico relatado, tanto em termos de tempo de processamento quanto para a obtenção de taxas mais altas na discriminação das patologias, contando também com a adição do NEP como mais um desafio ao sistema classificador a ser desenvolvido.

AGRADECIMENTOS

Primeiramente, gostaria de prestar minha gratidão a Deus: sem Ele e sem a força dEle, guiando-me e me mostrando do que sou capaz, eu nada seria e nem estaria presente neste mundo.

Em segundo, gostaria de agradecer a minha família, que, desde minha infância, mostrou-me o meu caminho já me chamando de engenheira quando eu nem sabia o que isso significava. Em especial, deixo aqui marcado minha eterna gratidão aos meus pais, Ivanildo e Liliana; eles sempre estiveram ali por mim e me proporcionaram todas as condições e oportunidades que eu sei que, com minha idade, eles não tiveram. Vocês são os meus heróis e meus amores para todo sempre. Ainda, à minha irmã Bruna: obrigada por sempre me acompanhar e incentivar, para além do aspecto acadêmico. Mais uma vez, essa conquista é por e para vocês.

Não poderia deixar de agradecer também aos meus amigos, aqueles que estiveram do meu lado passando as mesmas dificuldades ou dúvidas e sempre conseguiram ser lar para mim. Para além da escola, da UNIFESP ou da Universidade de Coimbra, os amigos que conquistei pelo caminho são meu grande porto seguro e sempre serei grata por ter vocês em minha vida.

Agradeço também aos professores que passaram por mim e puderam moldar o que sou hoje. Gratidão especial ao meu orientador, Prof. Dr. Matheus Cardoso Moraes, por toda paciência e crença em meu potencial, e a minha coorientadora, Prof.ª Dr.ª Anna Luíza Damaceno Araújo, por toda disponibilidade e proatividade mesmo em momentos difíceis. Vocês todos foram peças indispensáveis em minha trajetória como pesquisadora. Ainda, estendo o agradecimento aos pesquisadores e profissionais da FOP pela contribuição clínica, especialmente a Daniela Giraldo-Roldán e ao Sebastião Silvério de Souza Neto.

Por fim, agradeço a UNIFESP, ao ensino público de qualidade e a todos os funcionários públicos que realmente se dedicam a educação e formação de jovens pelo país.

REFERÊNCIAS

1 NASCIMENTO, A. G. d. A.; ARAÚJO, B. L. d.; SOBRAL, A. P. V. Tumores neurais na cavidade oral: estudo imuno-histoquímico. *Rev. cir. traumatol. buco-maxilo-fac*, p. 7–12, 2019.

2 MARTORELLI, S. B. d. F. et al. Neurofibroma isolado da cavidade oral: relato de caso. *Rev. cir. traumatol. buco-maxilo-fac.*, v. 10, n. 2, p. 43–48, 2010. ISSN 1808-5210.

3 MULLINS, B. T.; HACKMAN, T. Malignant peripheral nerve sheath tumors of the head and neck: a case series and literature review. *Case Rep. Otolaryngol.*, Hindawi Limited, v. 2014, p. 368920, dez. 2014.

4 VARGAS, T. J. d. S. et al. Perineurioma esclerosante: relato de caso e revisão da literatura. *An. Bras. Dermatol.*, FapUNIFESP (SciELO), v. 84, n. 6, p. 643–649, dez. 2009.

5 GERMAIN, M.; SLACK, R. S. Dining in with BCL-2: new guests at the autophagy table. *Clin. Sci. (Lond.)*, Portland Press Ltd., v. 118, n. 3, p. 173–181, out. 2009.

6 CID-O: Classificação Internacional de Doença para Oncologia. [S.l.]: Edusp, 1996.

7 BATISTA, K. T. et al. Treatment strategy for benign nerve tumors. *Rev. Bras. Cir. Plást.*, GN1 Genesis Network, v. 35, n. 1, p. 72–77, 2020.

8 ARAÚJO, A. L. D. et al. The role of immunohistochemistry for primary oral diagnosis in a brazilian oral pathology service. *Appl. Immunohistochem. Mol. Morphol.*, Ovid Technologies (Wolters Kluwer Health), v. 29, n. 10, p. 781–790, 2021.

9 JENA, B. et al. Brain tumor characterization using radiogenomics in artificial intelligence framework. *Cancers (Basel)*, MDPI AG, v. 14, n. 16, p. 4052, ago. 2022.

10 MAZAL, A. et al. Convolutional neural networks accurately predict benign versus malignant status among peripheral nerve sheath tumors. *J. Neuroimaging Psychiatry Neurol.*, United Scientific Group, v. 06, n. 01, 2021.

11 COLLINS, G. S. et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Ann. Intern. Med.*, American College of Physicians, v. 162, n. 1, p. 55–63, jan. 2015.

12 HE, K. et al. Deep residual learning for image recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [S.l.]: IEEE, 2016.

13 HARRISON, M. *Machine Learning–Guia de referência rápida: trabalhando com dados estruturados em Python*. [S.l.]: Novatec Editora, 2019.

14 JÚNIOR, N. O.; LEÃO, M. G. de S.; OLIVEIRA, N. H. C. de. Diagnóstico das lesões do joelho: comparação entre o exame físico e a ressonância magnética com os achados da artroscopia. *Revista Brasileira de Ortopedia*, v. 50, n. 6, p. 712–719, 2015. ISSN 0102-3616. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0102361615000533>.