



UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA E TECNOLÓGICA

UNIVERSIDADE FEDERAL DO PARANÁ

THIAGO GUILHERME GONÇALVES

RELATÓRIO FINAL

PROGRAMA	<input type="checkbox"/> PIBITI	<input type="checkbox"/> Fundação
DE IC:	MODALIDADE:	Araucária
<input checked="" type="checkbox"/> PIBIC	<input type="checkbox"/> CNPq	<input checked="" type="checkbox"/> Voluntária
<input type="checkbox"/> PIBIC Af	<input type="checkbox"/> UFPR TN	
<input type="checkbox"/> PIBIC EM		

TÍTULO DO PLANO DE TRABALHO

Relatório apresentado à Coordenação de Iniciação Científica e Tecnológica da Universidade Federal do Paraná como requisito parcial da conclusão das atividades de Iniciação Científica ou Iniciação em desenvolvimento tecnológico e Inovação - Edital 01/2023
Orientador(a): Prof.(a) Fabiano Yokaichiya
Título do Projeto: Redes neurais aplicadas a características radiômicas extraídas de imagens médicas de sarcoma de partes moles para previsão de metástase pulmonar

CURITIBA

2023

RESUMO

Nesse estudo, dispomos de uma coletânea de imagens médicas de sarcoma de tecidos moles das extremidades para 51 pacientes com histologia comprovada, extraídas através de diferentes técnicas de ressonância magnética (MRI), como T1-weighted (T1), T2-weighted fat-suppression sequences (T2FS) e Short Tau Inversion Recovery (STIR); essas imagens encontram-se em uma base de dados pública. O Sarcoma de tecido fino, de acordo com a OMS, representa cerca de 0.7% dentre os tumores malignos incidentes em adultos nos Estados Unidos, no geral, a maioria dos novos casos são de tumores intermediários ou avançados, e o caso mais comum de ocorrência é nas extremidades, mesmo assim, existem diversas formas de tratamento para o controle desta patologia, entretanto cerca de 25% dos pacientes diagnosticados com essa doença também desenvolvem metástase distante (o percentual aumenta para 50% quando o tumor está mais avançado), e a região mais comum (80% dos casos) é o pulmão. A taxa de sobrevivência de um paciente com metástase no pulmão é de 9% após 5 anos do prognóstico (alta letalidade), e com isso, nosso estudo busca acelerar o prognóstico de metástase pulmonar acompanhando as imagens de sarcoma das partes moles das extremidades logo na região onde está o tumor, para cada técnica individualmente, e também através da combinação das imagens obtidas por diferentes técnicas, T1, T2FS e STIR, através de técnicas de aprendizagem de máquina supervisionada (redes neurais artificiais), onde já é sabido o histórico clínico de cada paciente.

INTRODUÇÃO

O sarcoma de partes moles constitui um grupo de neoplasias malignas de origem nas células mesenquimais, com mais de 50 subtipos reconhecidos pela Organização Mundial da Saúde (OMS). Nos Estados Unidos, eles representam cerca de 0,7% das novas malignidades em adultos. A maioria dos casos são tumores de grau intermediário ou alto que podem ocorrer em vários locais, sendo as extremidades a localização mais comum. Geralmente, a terapia proporciona um excelente controle local para os sarcoma de partes moles das extremidades (STS), mas cerca de 25% dos pacientes com STS desenvolvem metástase, e este número aumenta para 50% nos casos de tumores de grau mais alto. Os pulmões são o principal local de metástase, afetando aproximadamente 80% dos casos metastáticos nos STS. Os

pacientes diagnosticados com metástase pulmonar geralmente têm um prognóstico ruim, com uma taxa de sobrevivência, em 3 anos, de 46% após a ressecção cirúrgica e 17% sem ela. Terapias sistêmicas aprimoradas são necessárias para os STS das extremidades com risco de metástase pulmonar, possivelmente envolvendo regimes de quimioterapia mais agressivos ou terapias adaptadas à histopatologia do tumor. Dessa forma, a avaliação precoce do risco de desenvolvimento de metástase pulmonar durante o manejo do STS é fundamental, oferecendo tratamentos potencialmente mais direcionados e uma melhor taxa de sobrevivência desses pacientes.

Os tumores não são homogêneos, mas consistem em múltiplas sub populações clonais de células cancerígenas, formando sistemas complexos e em rápida evolução devido às diversas intervenções terapêuticas. Em cânceres como os STS, essa heterogeneidade se manifesta em diversos níveis, incluindo genes, proteínas, microambientes celulares, tecidos e pontos de referência anatômicos dentro dos tumores. No entanto, estudar a heterogeneidade tumoral por meio de biópsias histopatológicas invasivas é desafiador, visto que as amostras podem variar dependendo da seção do tumor.

Para enfrentar esse desafio, o campo emergente da radiômica se concentra na extração e análise de informações substanciais de imagens médicas por meio de análise avançada de variáveis quantitativas. Deste modo, as imagens diagnósticas de diferentes tipos de ressonância magnética da região em que incide o STS (tridimensionais e avaliada em camadas) podem oferecer informações prognósticas valiosas sobre o risco de desenvolvimento da metástase pulmonar, que é o alvo de estudo neste trabalho, onde faremos uma abordagem utilizando redes neurais artificiais para validar as possibilidades de avaliação desse prognóstico.

MATERIAIS E MÉTODOS

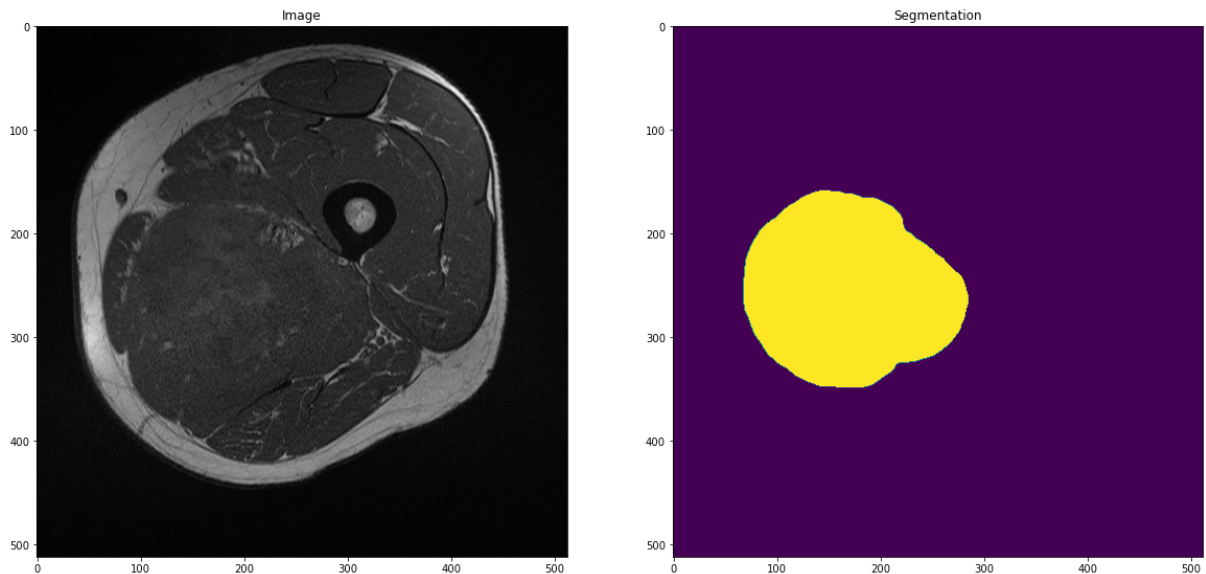
Utilizamos a biblioteca open-source pyradiomics em Python para extrair as variáveis radiômicas das imagens de ressonância magnética obtidas por diferentes técnicas e armazená-las de forma estruturada no formato csv, as imagens estão salvas localmente no formato .nrrd, de modo que para cada imagem, existe seu respectivo rótulo segmentando somente a região do tumor. Nas configurações de extração das

propriedades radiômicas, extraímos com base nas imagens originais, e também com a aplicação de filtros, como Wavelet e Laplacian of Gaussian (LoG) para $\sigma=3$ e $\sigma=5$. Também selecionamos diferentes classes de variáveis para serem extraídas dos diferentes tipos das imagens originais e com filtros aplicados, para esse estudo foram selecionadas as classes: shape, firstorder, glcm, glrlm, glszm, glldm, ngtdm. Ao final da configuração, logo antes de extrair as variáveis, fizemos a normalização dos tons de cinza das imagens para reduzir as diferenças de sinal de MRI, aplicamos algumas transformações para que a grande maioria dos valores de cinza abaixo da média não fiquem negativos (somente outliers). Estamos utilizando uma imagem (docker) do jupyter notebook contendo a biblioteca tensorflow com gpu (instalada no gerenciador de pacotes do python) para o desenvolvimento do modelo de previsão, utilizando a arquitetura MLP (multilayer perceptron) através do módulo keras, e mais uma série de módulos do scikit-learn para a cálculo de métricas de classificação, otimização dos hiperparâmetros e pesquisa (simplificada) pela arquitetura mais assertiva da rede neural MLP. Também fazemos uso de algumas bibliotecas em Python para visualização e manipulação de dados, tais como pandas e matplotlib.

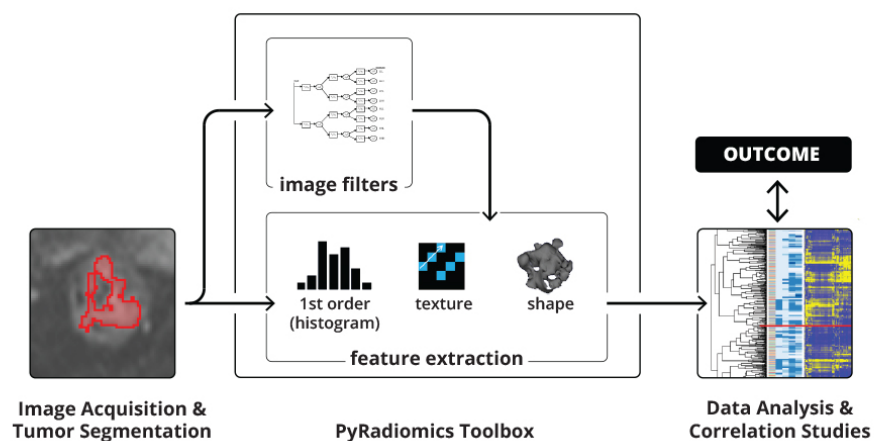
RESULTADOS E DISCUSSÃO

Com base na base de dados pública do The Cancer Imaging Archive (TCIA) [1], obtivemos, para cada um dos 51 pacientes, 2 tipos de MRI. Imagens do tipo T1-weighted (T1) eram comum a todos os pacientes, e dessa forma, também estavam presentes 25 imagens do tipo T2-weighted fat-suppression sequences (T2FS) e 26 imagens do tipo Short Tau Inversion Recovery (STIR), totalizando 51 pares de imagens para cada paciente.

As variáveis radiômicas foram extraídas, conforme as especificações descritas no item “Materiais e Métodos”, a partir dos pares de MRI do STS e segmentação (label) para cada tipo de MRI obtido de cada um dos 51 pacientes disponíveis na base de dados, conforme é possível observar pela imagem abaixo, que representa uma camada de um dos tipos de MRI.



A imagem abaixo, descreve de forma geral as etapas para a aquisição das features radiômicas de saída, que foram obtidas usando a biblioteca do python, pyradiomics, dado a coleção dos pares de MRI e label no formato .nrrd junto a aplicação de filtros de transformações sobre essas imagens, que entramos mais em detalhes na seção anterior.



De forma geral, foram extraídas 1037 características radiômicas de cada tipo de segmentação do STS por MRI, T1-weighted (T1), T2-weighted fat-suppression sequences (T2FS) e Short Tau Inversion Recovery (STIR). As variáveis de modo geral não precisaram de pré-processamento pois estavam em formato numérico, tal como os dados clínicos de cada paciente que indicavam o desenvolvimento de metástase pulmonar (1) ou não (0).

O nosso estudo focou no desenvolvimento, treino e otimização do algoritmo de redes neurais MLP para prevenção de metástase pulmonar, testando a técnica LOOCV (leave one out cross validation) utilizando o modelo que apresentou ao longo de várias buscas, no grid de parâmetros, a melhor arquitetura/conjunto de hiperparâmetros encontrados durante a etapa de otimização. Isso se deu devido à baixa disponibilidade de imagens de STS presentes neste banco de dados, sendo portanto uma abordagem inusitada, visto que, ao mesmo tempo em que há escassez de dados coletados (pois a incidência da doença é baixa) é expresso na literatura de aprendizagem de máquina que redes neurais requerem uma grande quantidade de dados disponíveis para reconhecimento efetivo de padrões, e assim o estudo necessitou de várias adaptações e aperfeiçoamentos para gerar resultados consistentes sem técnicas de reamostragem (que para uma base de 51 amostras não fariam tanta diferença quando trata-se de aplicações de redes neurais MLP).

Os dados radiômicos extraídos foram separados em 5 conjuntos: T1, T2FS, STIR, T1+T2FS e T1+STIR. Os conjuntos T1+T2FS e T1+STIR são simplesmente a união dos dados radiômicos comuns aos mesmos pacientes, mas que apenas foram obtidos via métodos de MRI distintos, e assim, T1+T2FS e T1+STIR contam ambos com 2074 features (já que as features dizem a respeito de determinada forma de extração), e respectivamente, 25 e 26 amostras (da mesma forma que T2FS e STIR respectivamente).

O pós-processamento foi realizado utilizando o método de seleção de variáveis *SelectKBest* com a função de score $f_classif$ (calculando o ANOVA F-value para todas as amostras), reduzindo as features extraídas para somente as 150 mais relevantes para cada um dos 5 conjuntos de amostras, e assim, todos os conjuntos estão nivelados em termos de variáveis.

A etapa de otimização foi feita com uma arquitetura pré estabelecida de MLP, variando somente a quantidade de neurônios (64, 128, 256, 512, 1024) nas camadas ocultas (hidden layers) decaindo pela metade a cada camada (testando ao total 3 camadas de *hidden layers* e uma camada final de apenas 1 neurônio, e portanto a previsão em probabilidade). Tomamos a quantidade fixa de 100 épocas, e fizemos a busca dos hiperparâmetros de *dropout_rate* entre as camadas (0.4, 0.5,

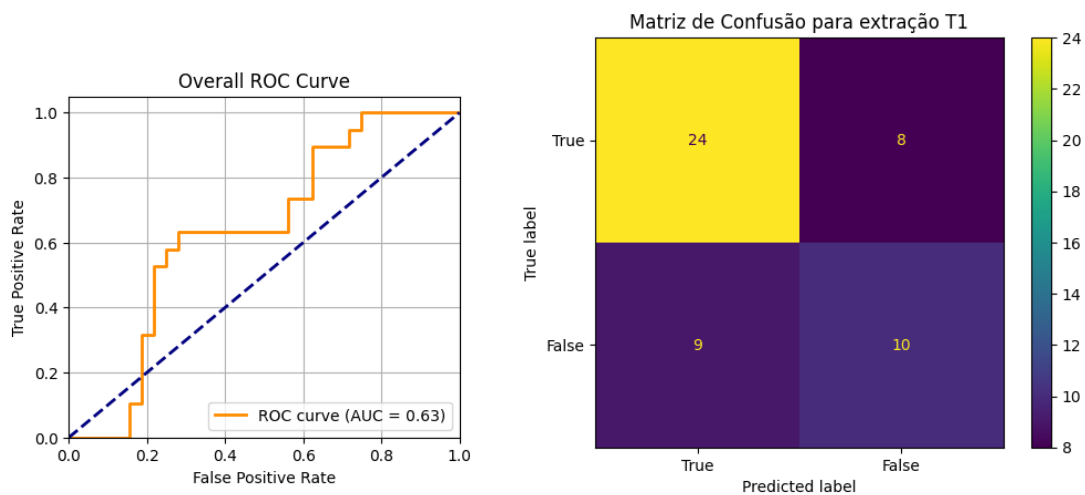
0.6), *learning rate* (0.001, 0.005, 0.01, 0.05, 0.1) e *batch_size* (1, 2, 4, 8, 16). O modelo foi compilado, para todos os testes, sob uma ótica vantajosa e estabelecida na literatura para classificação binária, tomando *Adam* como função de otimização e *binary crossentropy* como função perda, utilizando entre as camadas a função de ativação *RELU* (para trazer não linearidade ao aprendizado, e portanto, maior potencial de aprendizado), exceto pela última camada, que contém somente 1 neurônio e função de ativação *sigmoid*. Foi utilizada a função *RandomizedSearchCV* com 100 iterações, 3 folds para validação cruzada, e uma função de scoring personalizada com pesos para os valores de diferentes métricas de classificação conforme a seguinte equação: $0.6 * accuracy + 0.15 * precision + 0.25 * recall$. Assim, priorizando principalmente a acurácia, mas também precisão e recall (métrica sensível para prevenção de metástase, visto que ela incorpora numericamente a ideia de priorizar ocorrências de falsos positivos, sendo assim mais “conservadora” para um prognóstico tão delicado).

A otimização foi realizada somente para o conjunto T1-weighted (T1), pois o custo computacional desta etapa é alto, e o conjunto dispõe do maior número de amostras. Os valores para os *hiperparâmetros* e número de *hidden layers* que atingiram o maior critério de pontuação, conforme a equação personalizada de *score* descrita no parágrafo acima, foram consistentes para 5 tentativas de busca aleatória, e assim, utilizados para instanciação dos modelos em cada *fold* do treinamento via LOOCV para cada um dos 5 conjuntos. Durante cada *fold*, o modelo é instanciado (ou seja, reinicializado), treinado conforme a divisão realizada pelo LOOCV, e prevê a probabilidade de desenvolvimento de metástase, o qual usamos um threshold de 0.5 para classificar a previsão, os valores target são guardados, a cada fold, junto com o valor previsto convertido em binário, para que possamos calcular as métricas de classificação do problema.

As métricas calculadas para cada um dos conjuntos foram as seguintes:

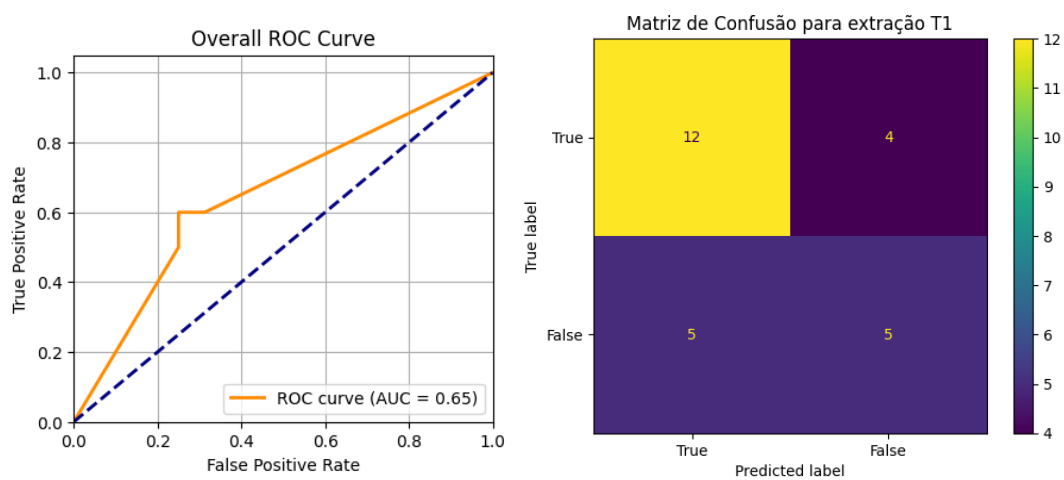
1. T1-weighted (T1)

Acurácia: 0.67	Precisão: 0.56	Recall: 0.53	F1-score: 0.54
----------------	----------------	--------------	----------------



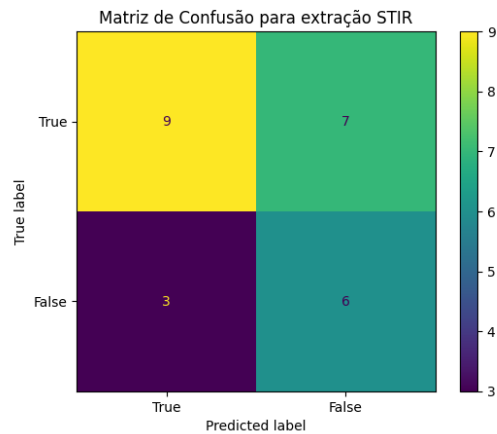
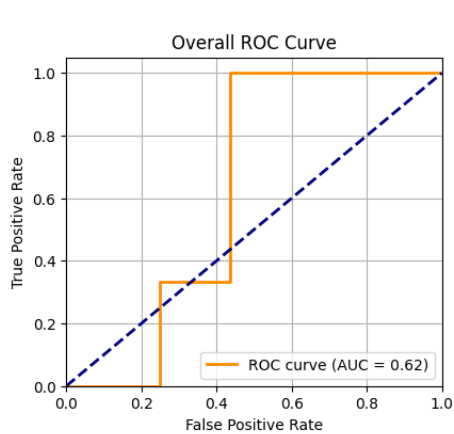
2. T2-weighted fat-suppression sequences (T2FS)

Acurácia: 0.65 Precisão: 0.56 Recall: 0.50 F1-score: 0.53



3. Short Tau Inversion Recovery (STIR)

Acurácia: 0.60 Precisão: 0.46 Recall: 0.67 F1-score: 0.55



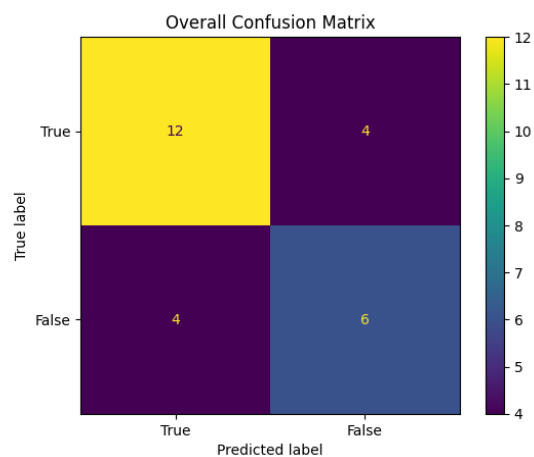
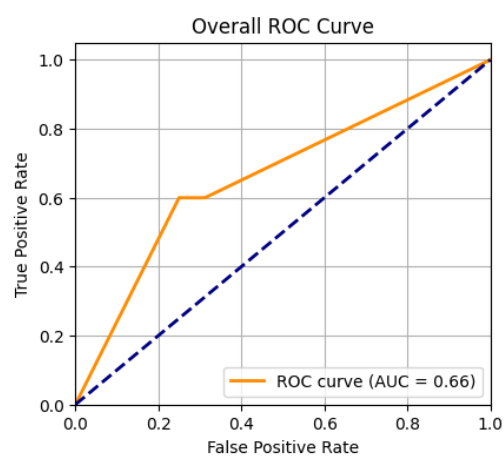
4. T1-weighted (T1) & T2-weighted fat-suppression sequences (T2FS)

Acurácia: 0.69

Precisão: 0.60

Recall: 0.60

F1-score: 0.60



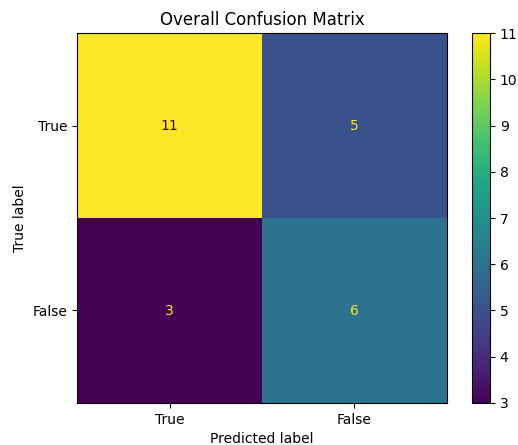
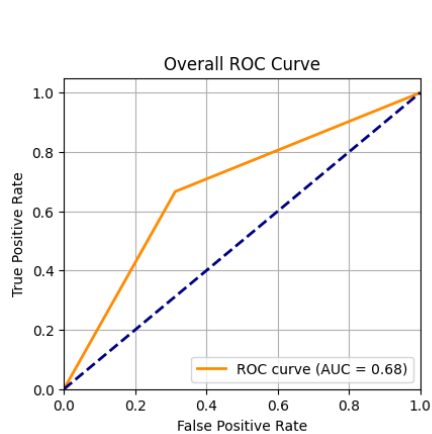
5. T1-weighted (T1) & Short Tau Inversion Recovery (STIR)

Acurácia: 0.68

Precisão: 0.55

Recall: 0.67

F1-score: 0.60



CONCLUSÕES/CONSIDERAÇÕES FINAIS

A união dos valores calculados de características radiômicas obtidas através de diferentes métodos de MRI sobre o sarcoma de partes moles das extremidades (STS) de diferentes pacientes, em todas as tentativas deste estudo, corroboraram para que obtivéssemos, para a mesma rede neural, de arquitetura e hiperparâmetros fixos, instanciadas múltiplas vezes pelo método LOOCV, maiores valores para as métricas de acurácia e recall (métricas de interesse para prognóstico de metástase pulmonar) quando comparado ao resultado obtido por um único método de MRI. Portanto, podemos compreender que, visto a escassa quantidade de imagens disponíveis, quando extraídas as características radiômicas por um certo método de MRI sobre a região em que incide o STS, para um dado paciente, essas características podem ser complementares as características radiômicas obtidas através de outro método de MRI, dentro dos métodos T1-weighted (T1), T2-weighted fat-suppression sequences (T2FS) e Short Tau Inversion Recovery (STIR), no quesito de contribuir com maior assertividade no prognóstico de metástase nas extremidades (pulmão).

REFERÊNCIAS

[1] <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=21266533> A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities (Soft-tissue-Sarcoma)

Vallières M 2013 FDG-PET/MR imaging for prediction of lung metastases in soft-tissue sarcomas of the extremities by texture analysis and wavelet image fusion Master Thesis McGill University, Montreal

Vallières, M. et al 2015 Phys. Med. Biol. 60 5471

HARRISON, M. Machine Learning – Guia de Referência Rápida: Trabalhando com dados estruturados em Python. Novatec Editora, 2019. ISBN 9788575228180.

HARRIS, C. R. et al. Array programming with NumPy. Nature, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020

Thibault G, Fertil B, Navarro C, Pereira S, Cau P, Levy N, Sequeira J and Mari J-L
2009 Texture indexes and gray level size zone matrix: application to cell nuclei
classification Pattern Recognition and Information Processing (PRIP) (Minsk,
Belarus) pp 140–5

Kumar V et al 2012 Radiomics: the process and the challenges Magn. Reson.
Imaging 30 1234–48

Lambin P et al 2012 Radiomics: extracting more information from medical images
using advanced feature analysis Eur. J. Cancer 48 441–6