UFCSPA
Universidade Federal de Ciências da Saúde de Porto Alegre

Trabalho Final - Análise de Metadados DICOM

Vinícius Franceschi

2018

1 Banco de imagens utilizado

O Banco de imagens utilizado foi retirado do Cancer Imaging Archive, sendo denominado Lung Image Database Consortium image collection (LIDC-IDRI). Portanto, consiste em imagens de câncer de pulmão. Neste dataset estão presentes imagens de CT (tomografia computadorizada), DX (radiografia digital) e CR (radiografia computadorizada) categorizadas em cânceres malignos e benignos de 1.010 pacientes, com um total de 254.527 imagens e 124 GB ao todo.

O banco está disponível no seguinte link:

<https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI>

2 Imagens selecionadas

Foram selecionadas 50 imagens dos primeiros 50 pacientes, ou seja, uma imagem para cada paciente. Por coincidência, eu havia utilizado este mesmo dataset para um trabalho anterior dessa disciplina que utilizava uma rede neural convolucional para classificar a presença ou ausência de tumor. No dataset, cada paciente possui uma série de imagens com diferentes cortes, então foi necessário que se escolhesse um corte específico de CT para cada paciente. Após uma pesquisa sobre imagens de diagnóstico de câncer de pulmão, constatei que o corte transversal é bastante usado para o diagnóstico de tumor no pulmão, como mostra a imagem abaixo:

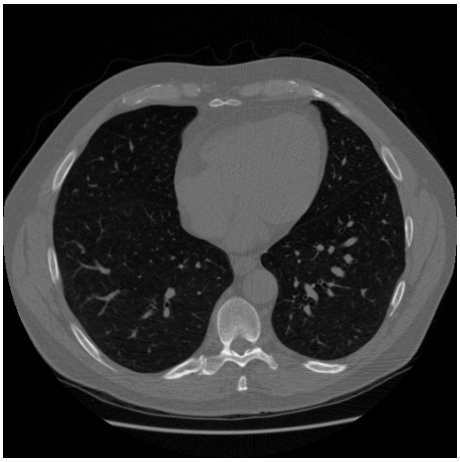


Figura 1 – Exemplo de imagem selecionada

Este corte pode ser encontrado nas imagens com número **000040.dcm** em todos os pacientes. Sabendo disso e de que a pasta na qual este corte estava presente em um subdiretório específico, realizei uma extração computacional utilizando as bibliotecas **os** e **shutil**, para navegar pelos diretórios e copiar o arquivo desejado para a pasta **dataset**. Desta forma, não irei disponibilizar todas as imagens, baixadas em meu computador pois excederia o limite do upload, mas somente a pasta **dataset** com os arquivos já extraídos nessa etapa.

A consideração de somente 50 pacientes para a escolha das 50 imagens foi feita devido à inviabilidade do download de todo o dataset devido ao seu tamanho, uma vez que não é possível baixar apenas as imagens de interesse, mas somente todo o dataset (124 GB).

3 Características selecionadas

A primeira etapa para a extração das características foi a execução e o entendimento do código disponibilizado pela professora (disponível em: <https://www.kaggle.com/gpreda/visualize-ct-dicom-data>), utilizando seu próprio dataset com imagens DICOM e TIFF.

Após, foram selecionadas as imagens que fariam parte do dataset final e executado o código utilizando o arquivo de metadados do LIDC-IDRI. Contudo, o arquivo contém apenas alguns atributos significativos e repetidos com aqueles já codificados no exemplo visto em aula, como **Modality e Manufacturer**, como mostra a tabela abaixo:

	ID	Manufacturer	Modality	path
0	LIDC-IDRI-0001	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
1	LIDC-IDRI-0002	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
2	LIDC-IDRI-0003	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
3	LIDC-IDRI-0004	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
4	LIDC-IDRI-0005	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
5	LIDC-IDRI-0006	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
6	LIDC-IDRI-0007	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
7	LIDC-IDRI-0008	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
8	LIDC-IDRI-0009	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
9	LIDC-IDRI-0010	GE MEDICAL SYSTEMS	CT	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...

Figura 2 – Metadados como ID, Manufacturer e Modality

Mais abaixo no código, na seção “More about DICOM data”, percebi que na verdade os arquivos DICOM possuíam cerca de 90 metadados internos, de modo que eu poderia avaliá-los e escolher aqueles que fossem mais relevantes para a caracterização da imagem.

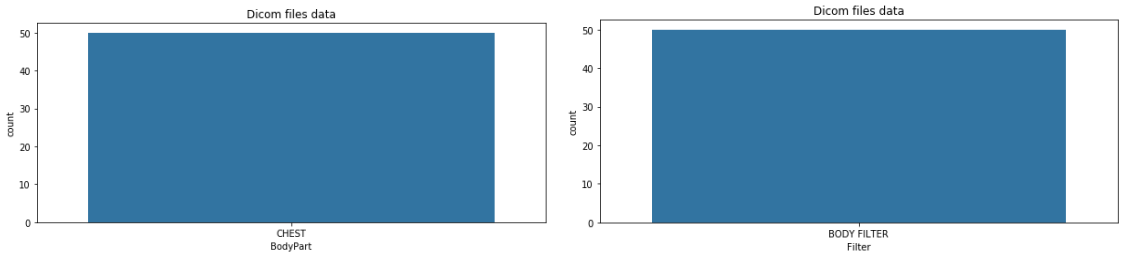
A partir dessa lista de características, resolvi selecionar **Body Part Examined, Exposure Time, Filter Type, Convolution Kernel e Slice Location**, pois imagino que sejam importantes para o conhecimento total da técnica utilizada na CT, para o conhecimento das condições do paciente e para a comparação entre diferentes técnicas radiológicas em busca de melhores resultados.

Abaixo, a tabela mostrando os novos metadados extraídos:

	BodyPart	ConvKernel	Exposure	Filter	ID	Modality	SliceLocation	path
0	CHEST	STANDARD	570	BODY FILTER	LIDC-IDRI-0001	CT	-180.000000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
1	CHEST	STANDARD	478	BODY FILTER	LIDC-IDRI-0002	CT	-192.000000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
2	CHEST	STANDARD	1160	BODY FILTER	LIDC-IDRI-0003	CT	-196.500000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
3	CHEST	STANDARD	690	BODY FILTER	LIDC-IDRI-0004	CT	-266.250000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
4	CHEST	STANDARD	570	BODY FILTER	LIDC-IDRI-0005	CT	-180.044998	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
5	CHEST	STANDARD	1160	BODY FILTER	LIDC-IDRI-0006	CT	-171.500000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
6	CHEST	LUNG	912	BODY FILTER	LIDC-IDRI-0007	CT	-279.500000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
7	CHEST	STANDARD	912	BODY FILTER	LIDC-IDRI-0008	CT	-185.720001	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
8	CHEST	STANDARD	872	BODY FILTER	LIDC-IDRI-0009	CT	-171.250000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...
9	CHEST	STANDARD	500	BODY FILTER	LIDC-IDRI-0010	CT	-276.250000	C:\Users\vinif\Desktop\LIDC-IDRI\dataset\LIDC-...

Figura 3 – Novos metadados extraídos dos arquivos DICOM

O conjunto de dados foi gerado pelo mesmo fornecedor e os dados foram examinados para identificação basicamente de câncer de pulmão, então não foram verificadas diferenças em características como parte do corpo examinada, tipo de filtro utilizado na imagem e kernel de convolução utilizado (somente em um paciente foi usado o filtro LUNG, enquanto em 49 foi usado o STANDARD). Essas Essas informações são mostradas nos gráficos abaixo, nos quais o eixo x representa a característica e o y a quantidade de pacientes nos quais suas imagens indicavam tal característica presente:



(a) Parte do corpo examinada: peito/tórax

(b) Tipo de filtro: BodyFilter

Figura 4 – Características repetidas para todos os 50 pacientes

Quando são analisadas outras características, por exemplo, tempo de exposição e localização do “corte”, verifica-se uma grande diferença entre os pacientes. Na figura abaixo, podemos ver que houveram 13 tempos de exposição diferentes para os 50 pacientes, sendo 570 segundos (cerca de 10 minutos) o tempo mais recorrente (n=12). Já 14 pacientes tiveram seu CT concluído em menos de 570 segundos e 24 pacientes em mais de 570 segundos. O tempo máximo de exposição foi realizado em 11 pacientes, alcançando quase

20 minutos de exposição. O maior tempo de exposição possivelmente esteja relacionado à identificação de tumor e a tentativa de confirmação por meio da realização de novas fatias (imagens).

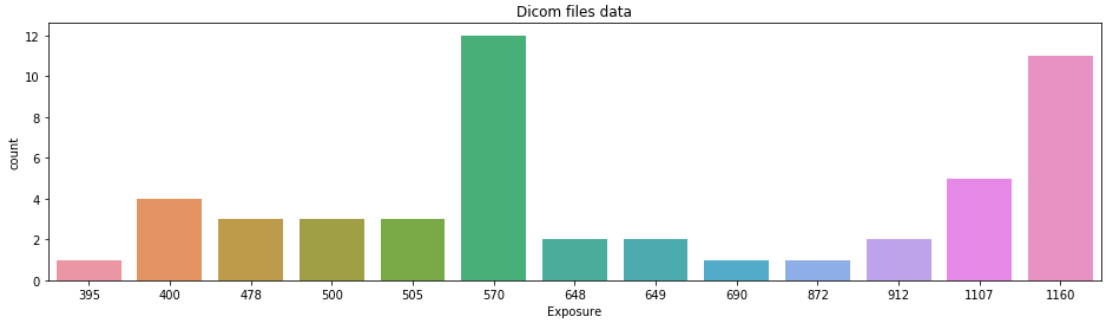


Figura 5 – Tempo de exposição à radiação (em segundos)

Quanto à localização do corte (fatia), definida como a posição relativa do plano da imagem (expressa em mm), observa-se que mesmo que se tenha extraído todas as imagens 40 de CT de cada paciente, a posição espacial do corte realizado foi diferente em quase todos os pacientes (variando de -315 a -15.25), o que acaba sendo normal por se tratarem de pessoas diferentes e que não conseguem ficar exatamente na mesma posição dentro do aparelho.

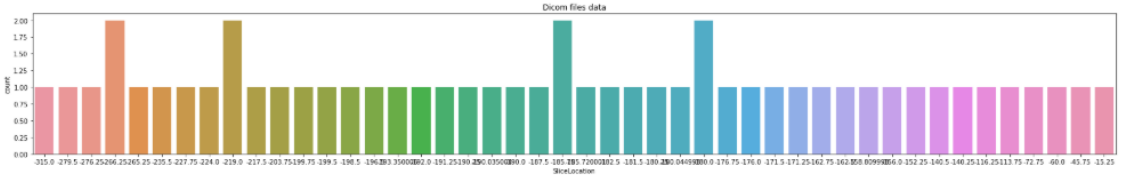


Figura 6 – Localização do corte

Analisando os plots das imagens e os gráficos gerados, podemos ver que este código foi bastante esclarecedor e permitiu que se conhecessem várias características que podem ser utilizadas no auxílio ao diagnóstico pelos radiologistas e físicos médicos.

Contudo, como são dados que relatam mais em que condições as informações foram extraídas do que propriamente os achados clínicos, imagino que uma abordagem computacional completa que auxiliasse no diagnóstico desse tipo de doença deveria agregar uma rede neural artificial para classificação dessas imagens considerando tanto os achados clínicos presentes nas imagens quanto estes metadados que reúnem informações valiosas sobre a coleta e as condições do paciente.

Mas, como o objetivo do trabalho era analisar os metadados DICOM, acredito que o aprendizado foi importante, uma vez que percebi a importância de considerar outras características que não somente a ausência e presença de tumor (identificada basicamente através da imagem), uma vez que na maioria dos casos, principalmente os mais complexos, os metadados podem ser decisivos no diagnóstico e prover um corpo de conhecimento importante para a melhora das práticas radiológicas e a identificação de padrões presentes nos pacientes envolvidos.