SARAH LUSTOSA HAIEK

USO DE DEEP LEARNING PARA IDENTIFICAÇÃO DE SEGMENTOS ANATÔMICOS EM RADIOGRAFIAS

Monografia apresentada à Universidade Federal de São Paulo para a obtenção do Certificado de Conclusão do Programa de Residência Médica.

Coautores: Julia Tagliaferri e Paletta
Lucas Teixeira Diniz
Vinícius Ribeiro

Orientador: Prof. Dr. Nitamar Abdala
Co-orientador: Dr. Felipe Campos Kitamura

SÃO PAULO
2020

UNIVERSIDADE FEDERAL DE SÃO PAULO

ESCOLA PAULISTA DE MEDICINA

DEPARTAMENTO DE DIAGNÓSTICO POR IMAGEM

Chefe do Departamento:

Prof. Dr. Henrique Carrete Junior

Supervisor da Residência Médica:

Prof. Dr. Rodrigo Regacini

Sumário

Lista de abreviaturas e símbolos

CNN	Convolutional neural network / Rede neural convolucional
DICOM	Digital imaging and communications in medicine / Imagem digital e comunicações em medicina
F1 score	Escore F1
PACS	Picture archiving and communication System / Sistema de comunicação e arquivamento de imagens

1 INTRODUÇÃO

As técnicas de *deep learning* tem sido usadas de forma crescente para a classificação de imagens. Uma dificuldade comum na aplicação de redes neurais a imagens radiológicas é o acesso a quantidades suficientes de dados de imagem. Outro grande desafio é a curadoria das imagens clínicas a serem usadas como dados.

Mesmo que o formato DICOM tenha a capacidade de armazenar imagens e metadados associados, como informações sobre a modalidade de imagem e segmento anatômico analisado, essas informações muitas vezes não estão presentes ou não são confiáveis, apresentando descritores errôneos.

Metadados facilmente identificados por um radiologista a partir de análise superficial da imagem, como a região analisada pela radiografia (*Body Part*), são particularmente inconsistentes.

É essencial para o funcionamento de modelos de *machine learning* que esses descritores estejam corretos (*Body Part, View Position*). O motivo disso é que uma rede neural treinada para analisar radiografia de tórax terá como possíveis saídas categorias como: pneumonia, pneumotórax, derrame pleural, nódulo. Se colocarmos uma radiografia de joelho para ser avaliada por essa rede neural, elá irá predizer alguns desses achados, o que seria um erro grosseiro. Portanto, é fundamental sabermos qual a parte do corpo foi examinada na radiografia, permitindo o direcionamento para o algoritmo correto.

Ferramentas para identificar valores corretos desses metadados facilitariam o uso de grandes conjuntos de imagens.

Neste estudo, testamos a eficácia de diferentes CNNs na identificação de segmentos anatômicos contemplados nas radiografias. Uma classificação automática e precisa permitiria o direcionamento automático de exames para redes neurais específicas, evitando erros grosseiros.

2 MÉTODOS

Este estudo retrospectivo foi aprovado pelo comitê de ética e pesquisa e o consentimento informado foi dispensado.

Imagens de radiografias realizadas entre novembro e dezembro de 2017 foram selecionadas do PACS do Hospital São Paulo, e os dados dos pacientes foram anonimizados.

Para anotação, foram utilizados descritores para os exames radiográficos mais comumente utilizados, usando a ferramenta MD.ai (MD.ai, New York), exemplificada na Figura 1.

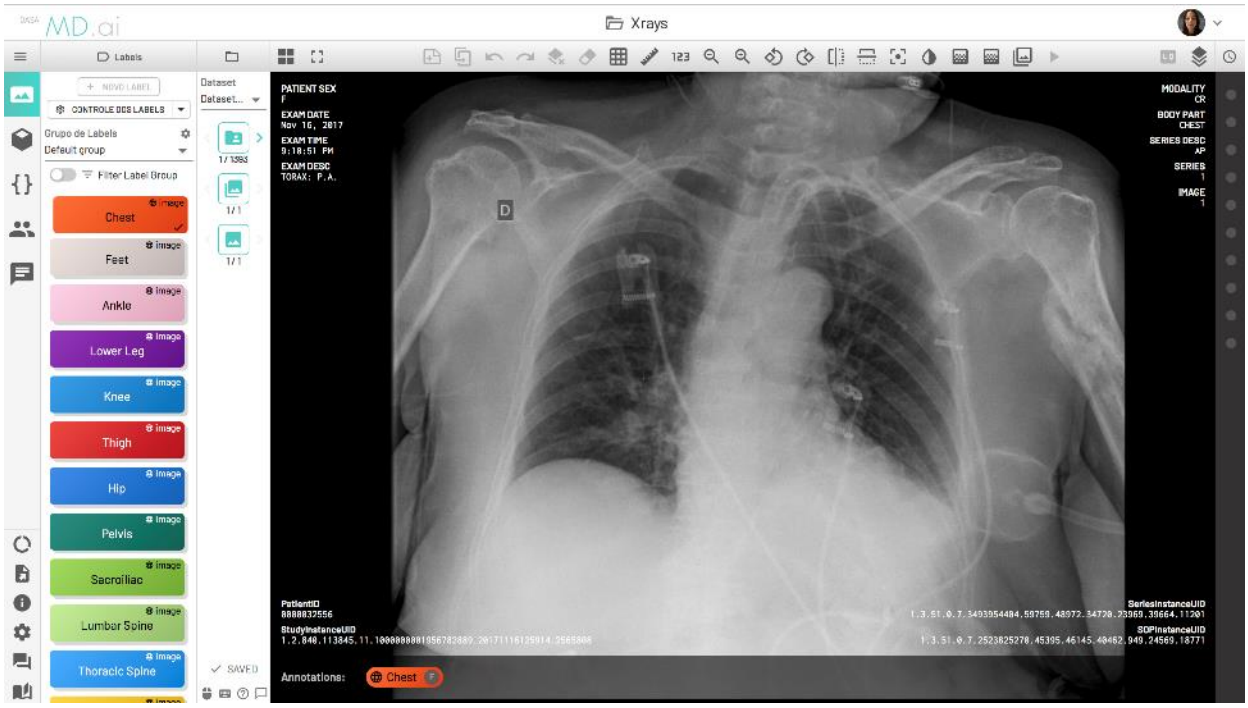

Figura 1: interface da ferramenta de anotação MD.ai.

Os descritores (classes) escolhidos estão listados na Tabela 1. Exames radiográficos menos comumente realizados e exames não abrangidos por este estudo, como densitometria óssea e exames contrastados, foram agrupados com o descritor "*Others*" (outros).

Descritores				
Clavicles	*Wrist*	*Skull*	*Pelvis*	*Ankle*
Shoulder	*Hand*	*Cervical Spine*	*Hip*	*Feet*
Arm	*Finger*	*Thoracic Spine*	*Thigh*	*Chest*
Elbow	Cavum	*Lumbar Spine*	*Knee*	*Abdomen*
Forearm	*Sinus*	*Sacroiliac*	*Lower leg*	*Others*

Tabela 1: descritores utilizados.

Algumas imagens podem abranger mais de uma anatomia, sendo utilizado mais de um descritor, conforme a Figura 2, onde foram utilizados os descritores "*Hand*" (mão) e "*Wrist*" (punho).

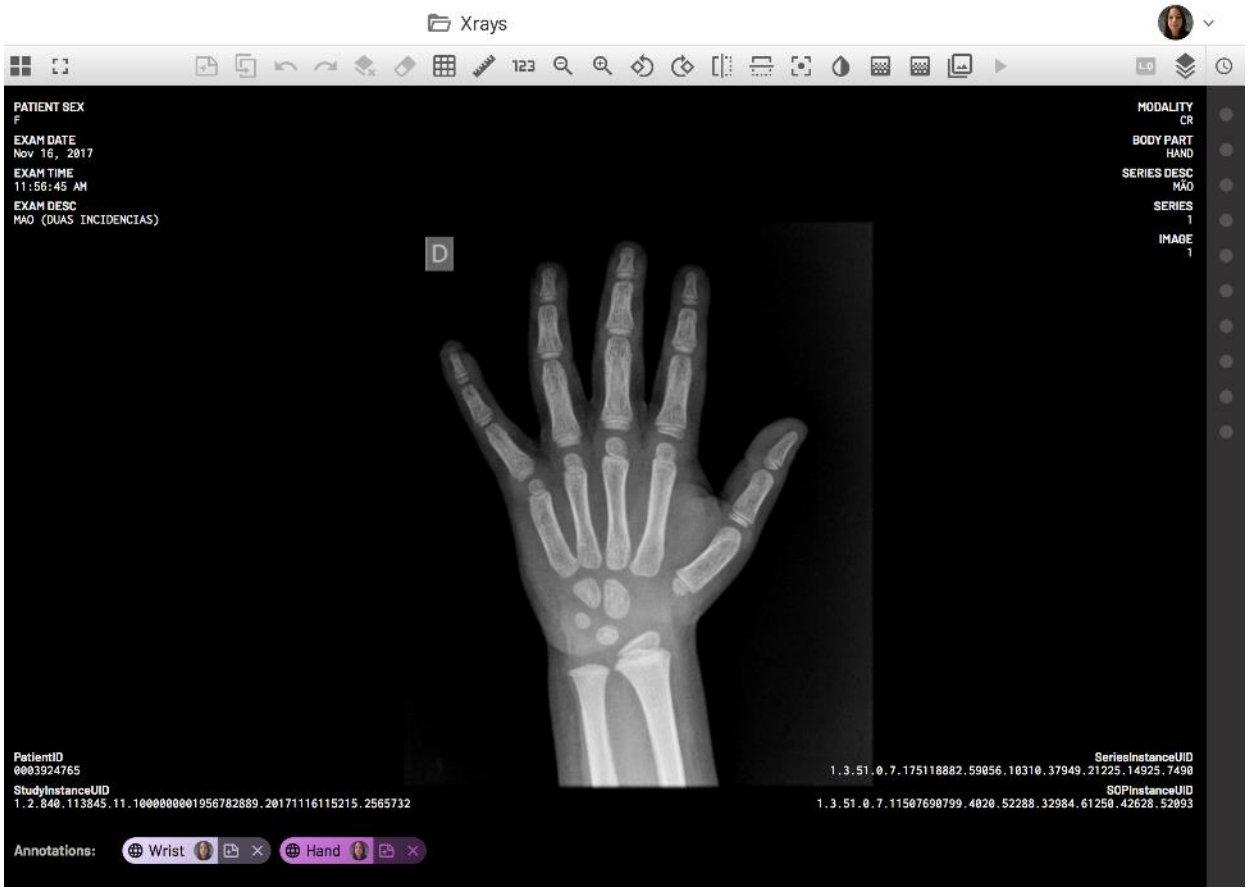

Figura 2: exemplo de imagem contendo mais de uma anatomia.

Optou-se ainda por utilizar os descritores diferenciando apenas os segmentos anatômicos, sem diferenciação entre as incidências radiográficas, conforme exemplo da Figura 3, onde imagens de diferentes incidências foram classificadas igualmente como "*Knee*" (joelho).

Figura 3: exemplo de exame com imagens classificadas com mesmo descritor.

As imagens foram anotadas por quatro médicos com pelo menos 1 ano de experiência em Radiologia.

2.1 *Dataset*

O conjunto de dados utilizado possuía 2.443 imagens de radiografia de 25 classes distintas. Cada imagem poderia conter mais de uma anatomia, configurando um problema *multilabel*.

A Figura 4 demonstra a distribuição de classes no *dataset* completo. Nota-se que houve um grande desbalanceamento entre as classes, sendo algumas extremamente minoritárias. "*Sacroiliac*" (sacroilíaca) possuía apenas duas imagens e "Cavum" apenas uma imagem.

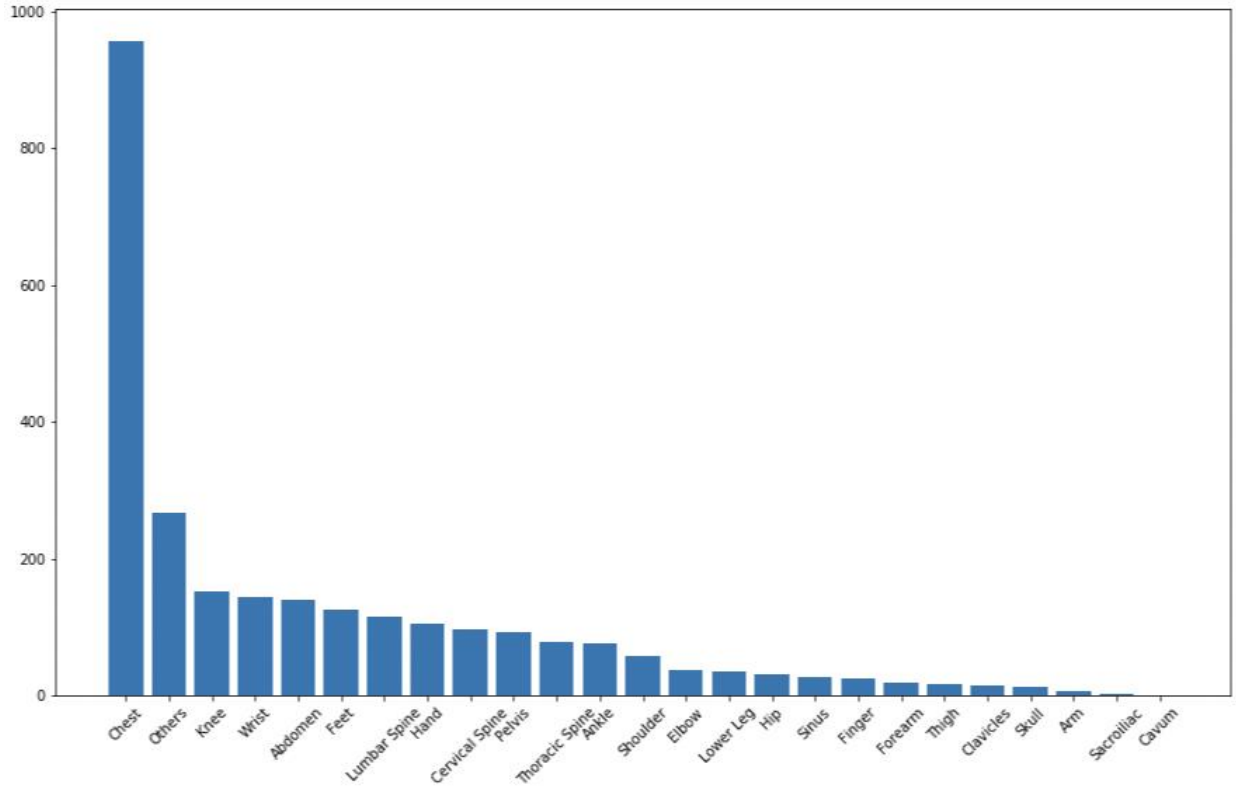

Figura 4: Distribuição de imagens por classe.

O conjunto de dados foi dividido entre treino (70%), validação (15%) e teste (15%). Aplicamos aumentação nas classes minoritárias *off-line* de forma que cada classe foi aumentada em até 10 vezes, de acordo com a proporção em relação à classe majoritária. A aumentação *off-line* foi realizada após a separação dos conjuntos a fim de evitar contaminação entre os subconjuntos.

As aumentações *off-line* aplicadas no conjunto de treino e de validação foram RandomHorizontalFlip e RandomVerticalFlip com probabilidade de 0,5, e RandomRotation de 0 a 360 graus.

Também realizamos aumentações de dados *on-line* no conjunto de treino. As aumentações *on-line* aplicadas foram RandomHorizontalFlip e RandomVerticalFlip com probabilidade de 0,5, e RandomRotation de 0 a 180 graus.

Não aplicamos aumentações de dados no conjunto de teste.

2.2 Modelos

Todos os modelos foram treinados por 100 épocas, com 20 épocas de paciência para *early-stopping*. Aplicamos *Cyclic Learning Rate* como *scheduler*, com a *learning rate* variando entre $1e^{-5}$ e $1e^{-4}$. Utilizamos otimizador *Adam* com fator de regularização de 0.01. A função de perda utilizada foi *Binary Cross Entropy* com balanceamento entre as classes. Avaliamos os modelos utilizando a função de *Hamming*, que fornece o percentual de *labels* mal classificados, o *Exact Match Score*, que fornece o percentual de amostras perfeitamente classificadas, a sensibilidade e a especificidade. A seleção do modelo foi feita de acordo com a melhor sensibilidade.

Testamos cinco modelos: AlexNet, EfficientNet, DenseNet, Inception V4, e ResNet101. Os dois primeiros não apresentaram resultados satisfatórios e foram descartados. Os três últimos apresentaram resultados consistentes e adequados, e foram escolhidos para dar continuidade.

Com os três melhores modelos definidos, testamos a realização de *ensemble* com os resultados. As funções escolhidas para *ensemble* foram regressão logística e classificador XGBoost, por serem modelos bem conhecidos na literatura.

3 RESULTADOS

As performances gerais (acurácia e F1 score “macro”) de cada modelo estão listadas na Tabela 2.

Modelo	Acurácia	F1
ResNet101	0,9910	0,9452
DenseNet	0,9914	0,9480
Inception v4	0,9905	0,9433
Regressão Logística	**0,9917**	**0,9491**
Classificador XGBoost	0,9911	0,9454

Tabela 2: Performance dos modelos principais e ensembles.

A seguir, uma análise por classe foi necessária, principalmente para entender o comportamento do modelo nas classes minoritárias. Para essa análise, incluímos apenas o melhor modelo único (DenseNet) e os dois *ensembles* de forma a facilitar a visualização.

As Figuras 5, 6 e 7 apresentam as métricas de acurácia e F1 score “macro” por classe para a DenseNet, o *ensemble* com regressão logística e o *ensemble* usando classificador XGBoost, respectivamente. Além disso, o eixo secundário inclui o número de amostras positivas para cada classe dentro do conjunto de teste. Nota-se que não há exemplos positivos para as classes “*Arm*” (braço) e “Cavum”. Nota-se também que a classe “*Chest*” (tórax) é extremamente presente no conjunto de teste por ser uma classe dominante no conjunto total.

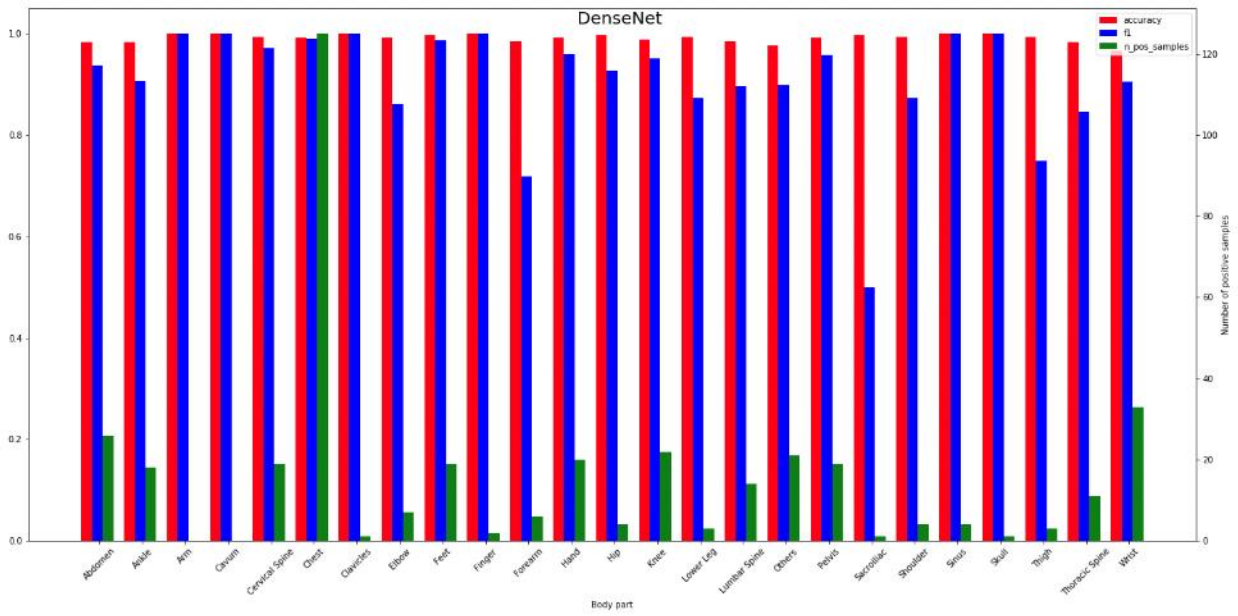

Figura 5: Métricas por classe para a DenseNet.

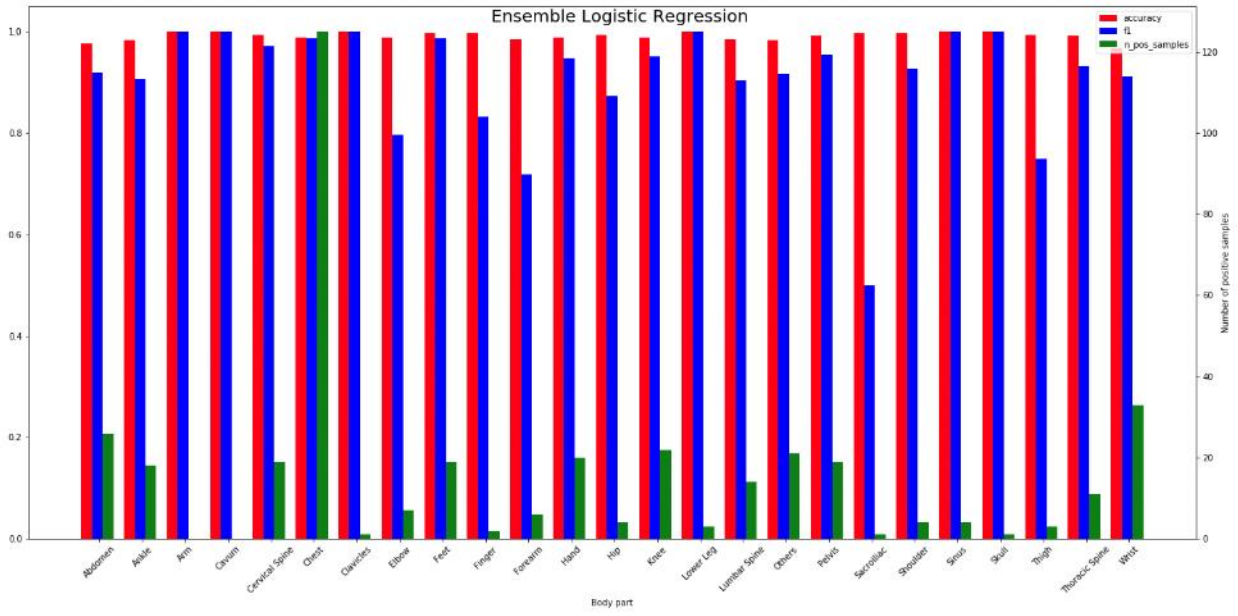

Figura 6: Métricas por classe para o ensemble com Regressão Logística.

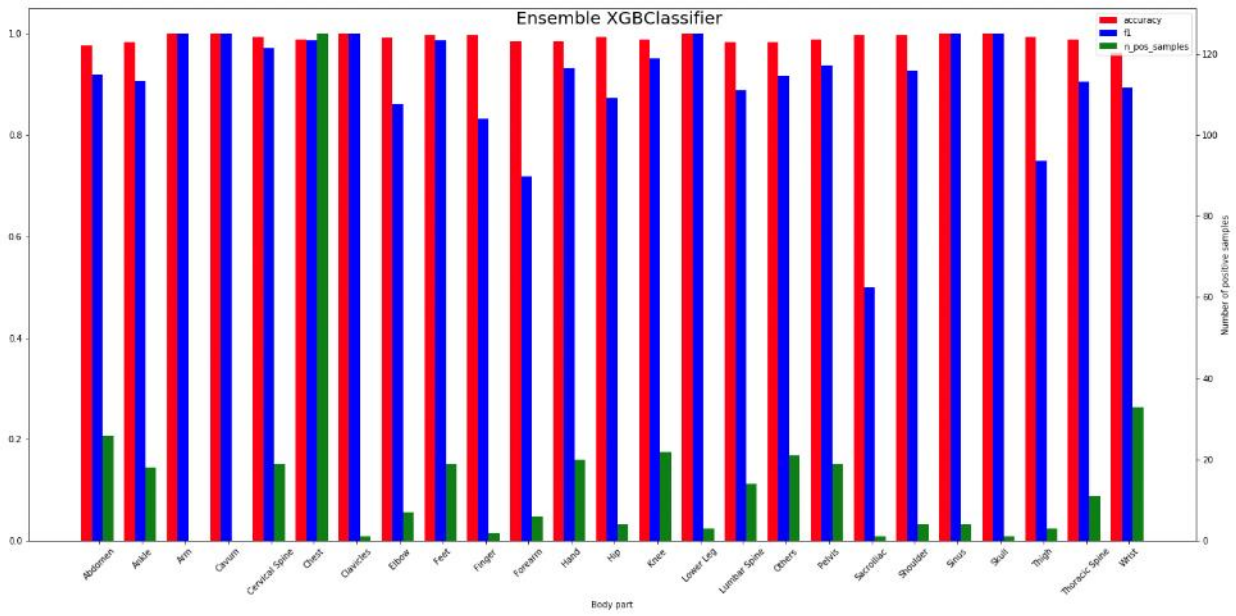

Figura 7: Métricas por classe para o ensemble com classificador XGBoost.

As Figuras 8 e 9 comparam as acurácias e F1 scores por classe, respectivamente, do principal modelo único (DenseNet) e dos dois *ensembles*. Além disso, o eixo secundário inclui o número de amostras positivas por classe dentro do conjunto de teste.

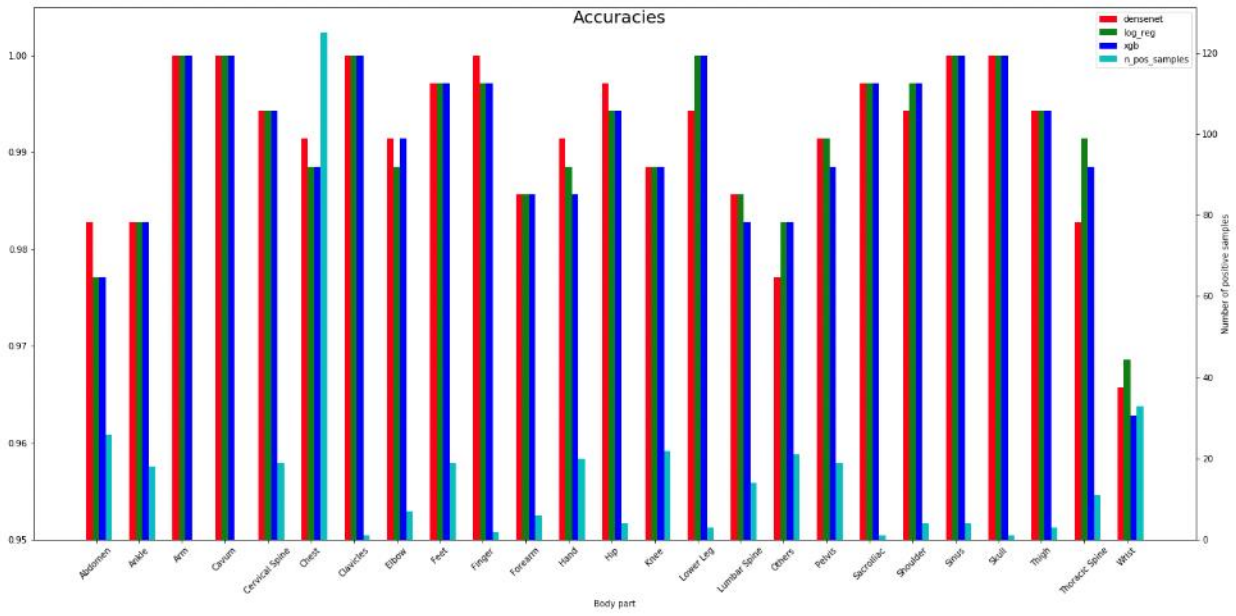

Figura 8: Comparação das acurácias do principal modelo único e dos dois ensembles.

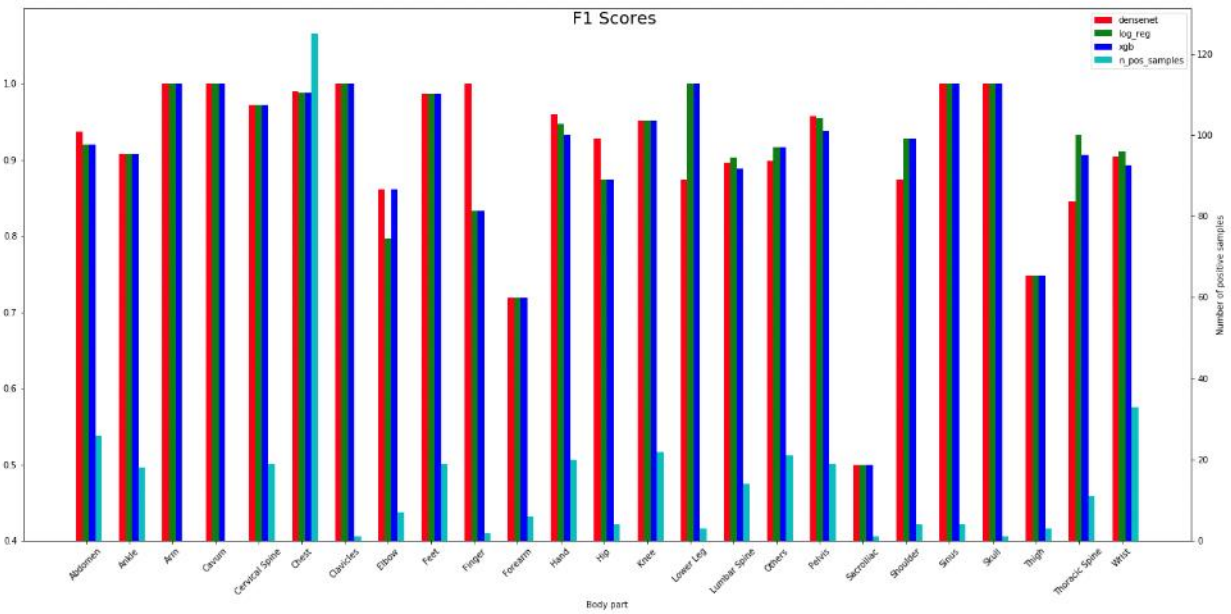

Figura 9: Comparação dos F1 scores do principal modelo único e dos dois ensembles.

4 DISCUSSÃO

Como podemos notar, o *ensemble* utilizando regressão logística apresenta um resultado agregado marginalmente superior à DenseNet, quando não consideramos o resultado por classe. Na análise por classe, as diferenças entre as abordagens não são claras, especialmente para classes minoritárias.

Quando observamos classes majoritárias ("*Chest*" e "*Abdomen*"), a DenseNet é superior ou equivalente aos demais. Por outro lado, quando analisamos classes minoritárias, o resultado é menos evidente. Em classes como "*Finger*" e "*Hip*", a DenseNet possui um resultado muito superior aos demais. Em outras como "*Lower Leg*" e "*Shoulder*", os *ensembles* levam vantagem em relação ao modelo único.

Os resultados apresentados por todos os modelos selecionados, assim como pelos *ensembles*, são muito promissores, possuindo uma performance razoável até mesmo em classes minoritárias.

A grande vantagem de utilizar a DenseNet em relação aos *ensembles* é a simplicidade do modelo. Para utilizarmos os *ensembles*, dependemos de três modelos de *deep learning* mais um modelo de regressão logística treinados previamente, tornando o *pipeline* de inferência mais complexo. Por outro lado, *ensembles* são amplamente utilizados na literatura como forma de aumentar a capacidade de aprendizado e reaproveitando o conhecimento adquirido por diferentes arquiteturas.

Apesar dos resultados atuais serem promissores, resultados mais conclusivos podem ser obtidos coletando mais amostras de classes minoritárias, especialmente aquelas ausentes no conjunto de teste, de forma a validar a capacidade de generalização dos modelos treinados.

Há ainda espaço para explorarmos técnicas mais complexas de aumentação de dados, explorar mais a fundo o ponto de corte dos modelos para cada classe (nesse estudo, utilizamos 0,5 como ponto de corte). Porém nenhum desses caminhos substitui a coleta de novas amostras de classes minoritárias.

Como perspectiva, a classificação da incidência radiográfica é o próximo passo.

5 CONCLUSÃO

O uso de *deep learning* para identificação de metadados referentes aos segmentos anatômicos analisados em radiografias apresentou resultados muito promissores. A metodologia alcançou uma performance razoável até mesmo em classes minoritárias. Resultados mais conclusivos poderão ser obtidos coletando mais amostras destas classes, de forma a validar a capacidade de generalização dos modelos treinados.

6 REFERÊNCIAS

1. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, Kadoury S, Tang A. Deep Learning: A Primer for Radiologists. RadioGraphics 2017; 37:2113-2131.

2. Lakhani P, Sundaram B. Deep learning at chest radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. Radiology 2017; 284:574-582.

3. Rajkomar A, Lingam S, Taylor AG, Blum M, Mongan J. High-Throughput Classification of Radiographs Using Deep Convolutional Neural Networks. J Digit Imaging 2017; 30:95-101.