[PSI5790 aula 7, parte 1, início]

Aplicações de Aprendizado Profundo em Processamento de Imagens

Exemplo de uso de aprendizado profundo em processamento de imagens: análise de mamografia

Vou descrever rapidamente o uso de inteligência artificial na análise de mamografia, para servir como exemplo de como o uso de deep learning consegue resolver algumas tarefas consideradas impossíveis há alguns anos atrás.

1. Análise de mamografia por computador

Mamografia é o raio-x da mama. Recomenda-se que todas as mulheres façam mamografia, a partir de uma certa idade, para detectar precocemente o câncer de mama. Inteligência artificial poderia auxiliar o médico na tarefa de analisar as mamografias.

O problema principal que queremos resolver é: "Dada uma mamografia, dizer se a paciente tem câncer ou não". Porém, há problemas secundários:

a) Localizar onde está o câncer dentro da mamografia.

b) Segmentar o câncer.

c) Ordenar as mamografias pela probabilidade de ter câncer. Isto poderia fazer o radiologista olhar com mais urgência as mamografias com alta probabilidade de ter câncer.

Dizer se tem câncer ou não em mamografia é um problema típico de classificação de imagens. Só que mamografia possui ≈ 4000×3000 = 12 milhões de pixels com 4096 níveis de cinza! A lesão muitas vezes ocupa somente uma pequena área. Note que uma imagem "normal" só possui 256 níveis de cinza (3 bandas se for colorida) e o objeto de interesse ocupa uma grande parte da imagem. Assim, é possível reconhecer o objeto mesmo se redimensionar a imagem para baixa resolução (da ordem de 224×224 pixels). Se reduzir mamografia para a resolução 224×224 pixels, torna-se impossível visualizar muitas lesões.

1. Métodos clássicos não tinham a pretensão de classificar a mamografia inteira em câncer/não-câncer. A ideia era ajudar o médico radiologista, apontando as regiões suspeitas da mamografia. Dividiam o problema em dois sub-problemas (figura 1): CADe (detecção das ROIs - regiões de interesse) e CADx (classificação das ROIs). Os dois problemas eram resolvidos tipicamente utilizando técnicas desenvolvidas manualmente.
2. A rede neural convolucional profunda aprende automaticamente, a partir dos exemplos, quais são as características que mais ajudam a classificar em câncer/não-câncer. Depois, classifica a mamografia inteira em câncer/não-câncer usando essas características.

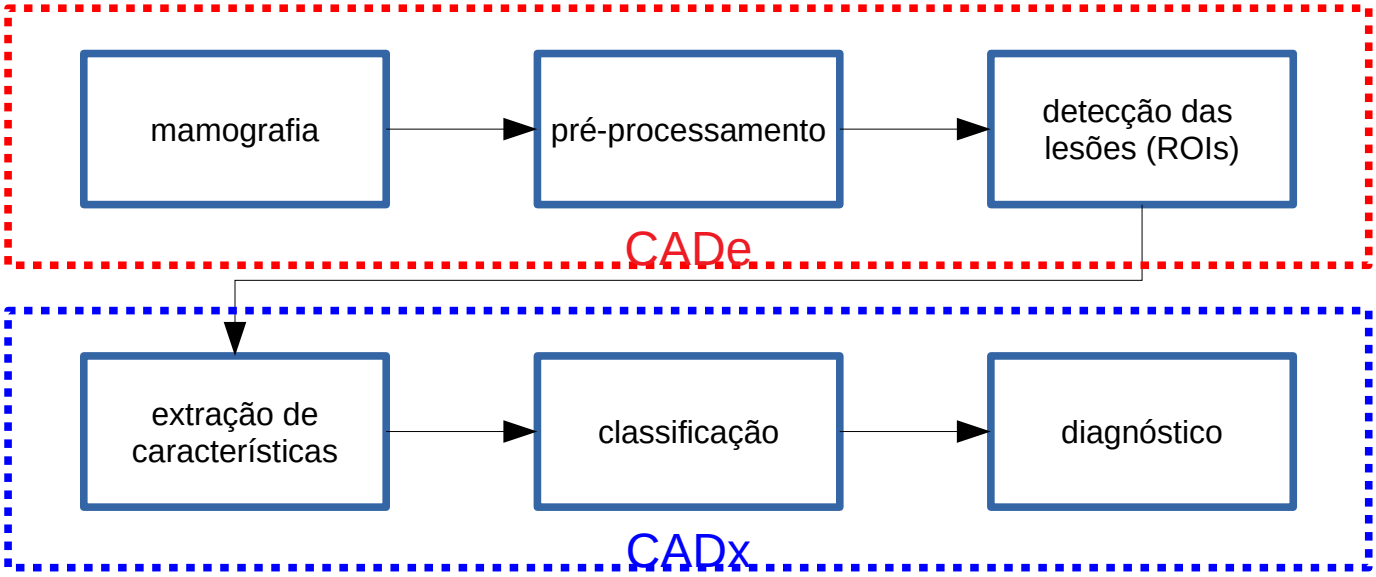

Figura 1: Etapas de um algoritmo clássico para classificar mamografias.

2. Análise de mamografia clássica

Há dois tipos principais de câncer de mama: “massa” e “microcalcificação” (figuras A e B).

a

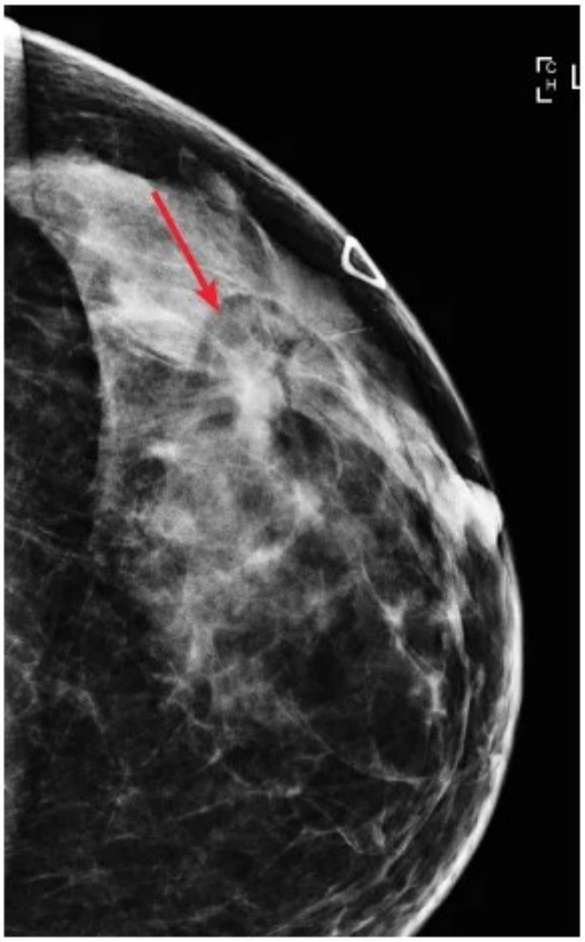

Figura A: Lesão tipo massa.

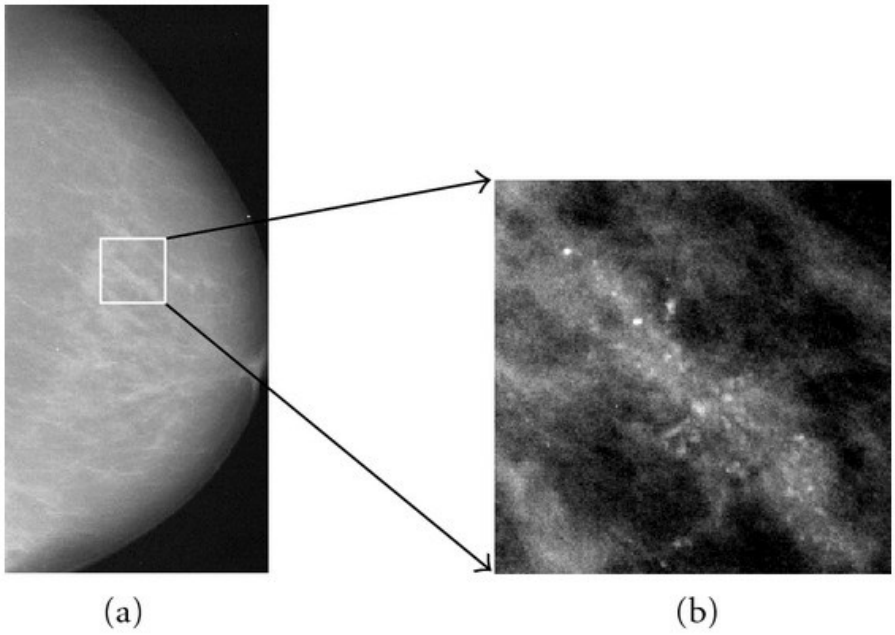

Figura B: Lesão tipo microcalcificação.

Na figura 2, as flechas indicam verdadeiras lesões tipo “massa”. O programa antigo “ImageChecker” marcou com “*” as possíveis lesões “massa” e com “Δ” as possíveis lesões “microcalcificação”. A versão 3 do programa errou todas as marcações, enquanto que a versão 8 acertou uma única lesão “massa” mas gerou vários falsos-positivos [Kim2010, Dromain2013]. Evidentemente, um programa desses não ajuda em nada os radiologistas.

A especificidade desses CADe era muito baixa, gerando por volta de um falso-positivo por vista. Assim, esses sistemas ajudavam a melhorar a sensibilidade do radiologista pois fazia olhar para as regiões suspeitas (de 4% a 15% [Domain2013]) mas pioravam a especificidade pois fazia detectar câncer onde não tinha (de 5% a 35%).

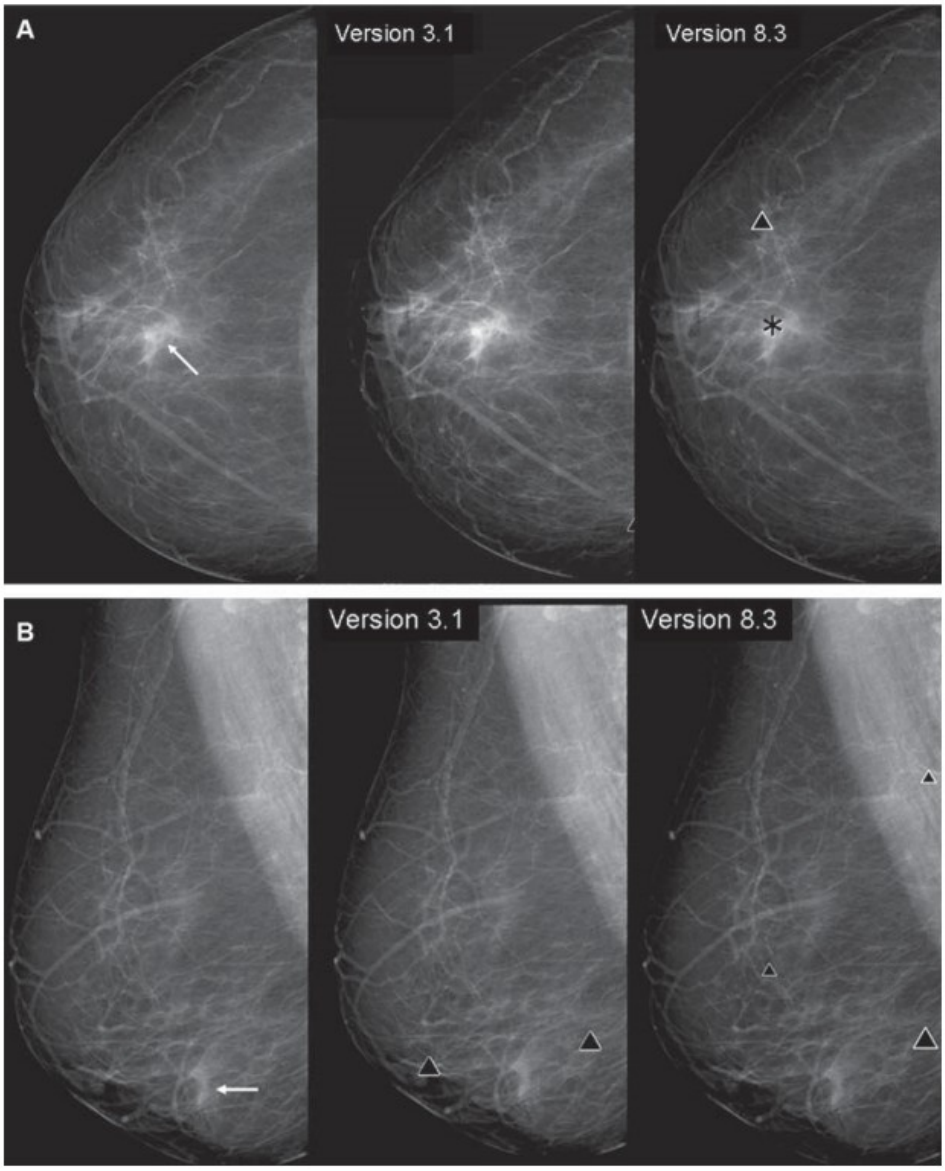

Figura 2: Flecha indica lesão tipo massa verdadeira. CADe indica com * lesão tipo massa e com Δ lesão tipo microcalcificação detectados pelo sistema. Figura retirada de [Kim2010].

Uma vez que as ROIs são localizadas, CADx pode classificá-las em benigno ou maligno. Para classificar ROIs pelo método clássico, é preciso extrair os atributos. O trabalho [Elter2009] apresenta algumas características usadas para essa tarefa: forma, densidade, textura, etc.

Depois, os métodos clássicos de aprendizagem (que já estudamos) eram usados para classificá-las: vizinho mais próximo, árvore de decisão, boosting, Bayes, rede neural, etc.

O trabalho [Ayer2010] mostra tabelas com AUCs de vários CADx em classificar ROIs, cujo resumo está na tabela 1.

Tabela 1: O menor e o maior AUC reportado por [Ayer2010] em CADx para classificar ROIs.

Modalidade	Menor AUC	Maior AUC
Mamografia	0,83	0,965

3. Métrica de desempenho de classificador binário

3.1. Taxa de erro, especificidade, sensibilidade e AUC

Acima, vimos algumas novas medidas de desempenho do algoritmo de aprendizado como sensibilidade, especificidade e AUC. Em exames médicos, não se pode usar a taxa de erro como medida de desempenho. Para entender o por quê, considere que a chance de uma mulher ter câncer de mama é 1%. Se um algoritmo responder que nenhuma mulher tem câncer, a sua taxa de acerto será de 99% e será considerado um ótimo algoritmo (mas completamente inútil).

Um exame médico pode apresentar 4 resultados (figura Y):

a) Verdadeiro positivo (TP): Uma pessoa doente é classificada como doentes.
b) Falso negativo (FN): Uma pessoa doente é classificada erroneamente como sadia.
c) Verdadeiro negativo (TN): Uma pessoa sadia é classificada corretamente como sadia.
d) Falso positivo (FP): Uma pessoa sadia é classificada erroneamente como doente.

A partir das taxas TP, FN, TN e FP, são calculadas sensibilidade e especificidade. Sensibilidade de um exame é a porcentagem de pacientes com doença que são detectadas corretamente pelo exame como tendo doença (TP/(TP+FN)). Especificidade de um exame é a porcentagem de pacientes sem doença que são classificadas corretamente como não tendo doença (TN/(TN+FP)). A taxa de acerto ou acuracidade é a quantidade de elementos classificados corretamente dividido pelo total número de elementos (TP+TN)/(TP+TN+FN+FP).

O exemplo acima (exame que diz que ninguém tem câncer) teria sensibilidade zero, apesar de ter especificidade 100%.

Já vimos que esse algoritmo teria acuracidade 99% levando as pessoas pensarem que o algoritmo é ótimo. Uma forma de evitar este erro é usar acuracidade balanceada:

[https://www.statology.org/balanced-accuracy/]

Acuracidade balanceada é definida como

(sensibilidade+especificidade)/2

E o algoritmo que diz que ninguém tem câncer teria acuracidade balanceada de 50%, equivalente a “chute”.

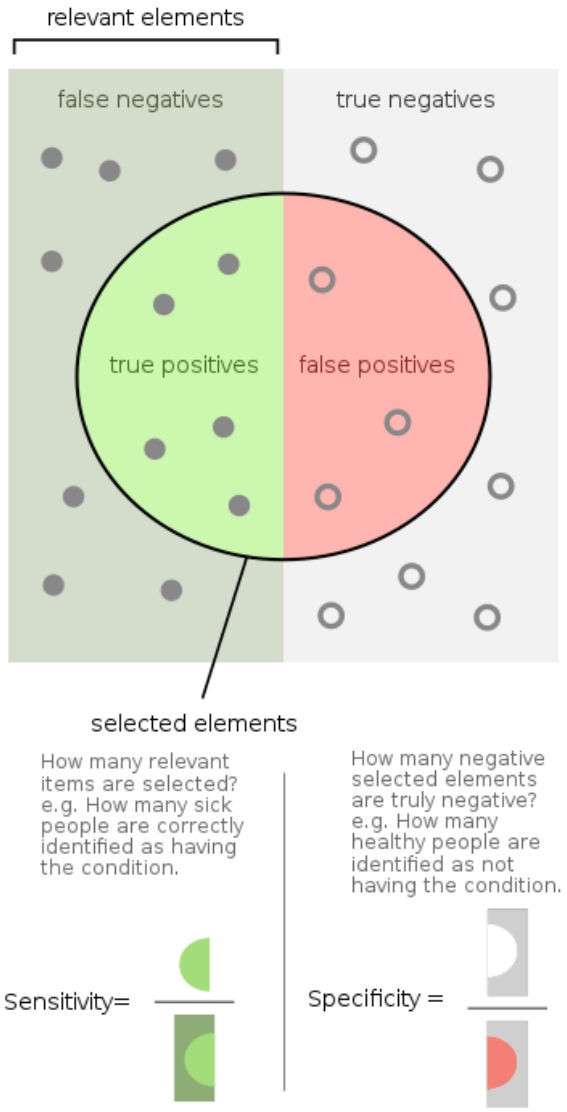

Figura Y: (extraído de Wikipedia)

A taxa de erro é um menos taxa de acerto. Na maior parte deste curso, vamos usar simplesmente a taxa de erro como medida de erro. Porém, no caso de câncer de mama (e outros exames médicos), não se pode usar taxa de acerto ou taxa de erro como medida de erro, pois normalmente o número de pacientes com doença é muito menor que o número de pacientes sem doença. Assim, costuma-se usar sensibilidade/especificidade ou acuracidade balanceada.

O problema é que o computador não costuma forncer uma resposta binária câncer/não-câncer, mas dá uma "nota" entre 0 e 1, uma espécie de "probabilidade" de ter câncer. É necessário limiarizar essa "probabilidade" para se obter a resposta booleana. Assim, sensibilidade/especificidade depende do limiar escolhido. Há um "trade-off" entre as duas, de forma que se mudar o limiar para aumentar a sensibilidade, vai diminuir especificidade (e vice-versa). Consequentemente, não é possível comparar se um método é melhor ou pior que outro baseado em sensibilidade/especificidade/acuracidade.

A métrica de desempenho que não depende da escolha do limiar é AUC (Area Under Curve). AUC mede a área sob a curva ROC (Receiver Operating Characteristic). ROC é a curva de sensibilidade em função de 1-especificidade, obtida variando limiar (figura X). Para cada limiar entre 0 a 1, temos uma sensibilidade e uma especificidade. Traçando a curva com todos os limiares possíveis, temos a curva ROC. Medindo a área embaixo da curva, temos AUC. Um algoritmo com AUC=1 nunca erra. Um algoritmo com AUC=0.5 equivale a um "chute cego". A figura X abaixo exemplifica uma curva ROC.

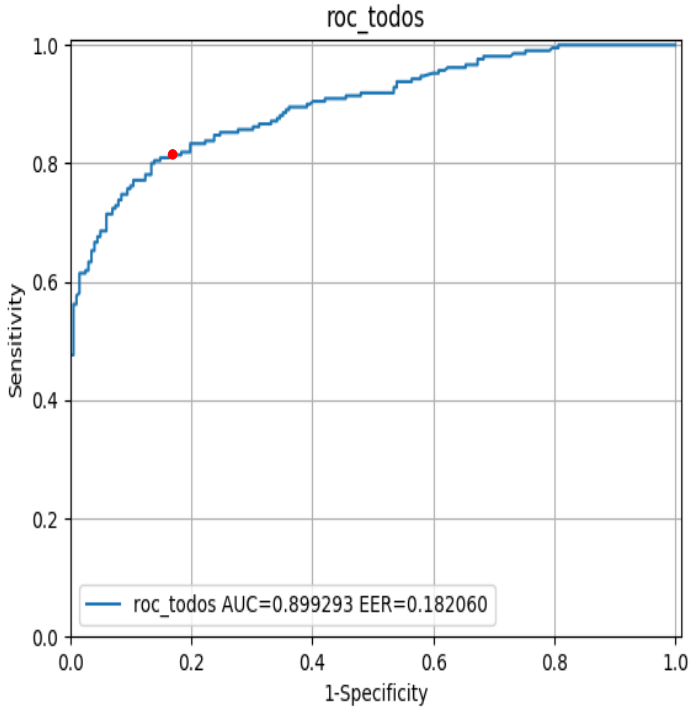

Figura X: Exemplo de curva ROC. O ponto vermelho indica EER (equal error rate).

Há um ponto especial, denominado de EER (equal error rate), onde a acuracidade, sensibilidade e especificidade se tornam iguais (o ponto vermelho na figura X). Este ponto é a intersecção entre a curva ROC e o diagonal principal (y=1-x). Neste ponto especial, é possível calcular acuracidade, sensibilidade e especificidade sem escolher o limiar.

3.2. Padrão ouro

"Ground truth" ou "padrão ouro" é a classificação verdadeira da mamografia. Para saber quanto um sistema de IA ou um radiologista acertou/errou, é necessário conhecer a classificação verdadeira. Normalmente, é muito difícil obter a classificação verdadeira de um exame médico. Não se pode comparar a resposta do sistema de IA com as respostas dos médicos, pois médicos também erram na classificação. Pode-se afirmar que uma mamografia com certeza tem câncer se a paciente foi submetida à biópsia e a análise de tecido indicou câncer. Por outro lado, uma mamografia com certeza não tem câncer se a paciente foi submetida à biópsia e a análise mostrou que a lesão não é cancerígena. Outra possibilidade de descartar câncer é se a paciente fez uma outra mamografia depois de 1 ou 2 anos e não desenvolveu câncer nesse tempo.

3.3. Comparação de IA com ser humano

[A preencher]

4. Análise de mamografia usando aprendizado profundo

Usando aprendizado profundo, o próprio algoritmo de aprendizado extrai as características mais importantes da imagem. Alguns artigos recentes informam sistemas IA com desempenho até superior a especialistas humanos.

[Escreva sobre usar duas vistas. Tomossíntese.]

A grande questão da inteligência artificial atual é como diminuir o número de exemplos de treinamento necessário. Um ser humano consegue aprender a partir de poucos exemplos, enquanto que o computador necessita de milhares ou milhões de exemplos.

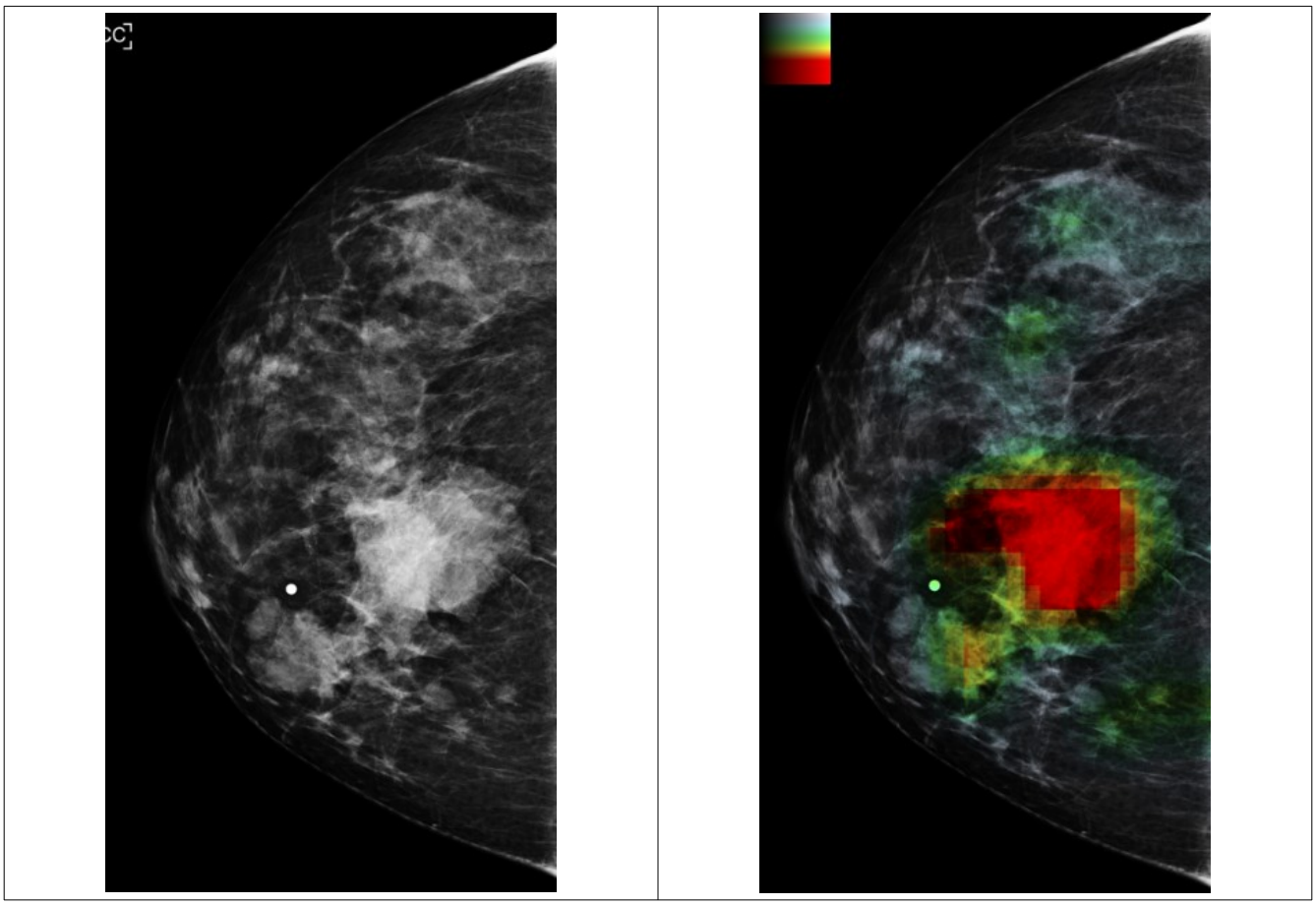

Figura 3: “Heatmap” gerado pelo programa [Wu2020] indicando regiões com maior possibilidade de lesão.

[Kooi2017] T. Kooi, G. Litjens, B. Van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, et al., "Large scale deep learning for computer aided detection of mammographic lesions", Med. Image Anal., vol. 35, pp. 303-312, Jan. 2017.

[Rodriguez2019] A. Rodriguez-Ruiz, K. Lång, A. Gubern-Merida, M. Broeders, G. Gennaro, P. Clauser, et al., "Stand-alone artificial intelligence for breast cancer detection in mammography: Comparison with 101 radiologists", J. Nat. Cancer Inst., vol. 111, no. 9, pp. 916-922, 2019.

[Schaffter2020] T. Schaffter, D. S. Buist, C. I. Lee, Y. Nikulin, D. Ribli, Y. Guan, et al., "Evaluation of combined artificial intelligence and radiologist assessment to interpret screening mammograms", JAMA Netw. Open, vol. 3, no. 3, Mar. 2020.

[McKinney2020] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, et al., "International evaluation of an AI system for breast cancer screening", Nature, vol. 577, no. 7788, pp. 89-94, Jan. 2020.

[Wu2019] N. Wu, J. Phang, J. Park, Y. Shen, Z. Huang, M. Zorin, et al., "Deep neural networks improve radiologists' performance in breast cancer screening", IEEE Trans. Med. Imag., vol. 39, no. 4, pp. 1184-1194, Apr. 2019.

[PSI5790 aula 7, parte 1, fim]