[PSI5790-2025 aula 5 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, no Brasil a partir de 50 anos, uma vez a cada 2 anos, para detectar precocemente o câncer de mama. Inteligência artificial poderia auxiliar o médico na tarefa de analisar as mamografias.

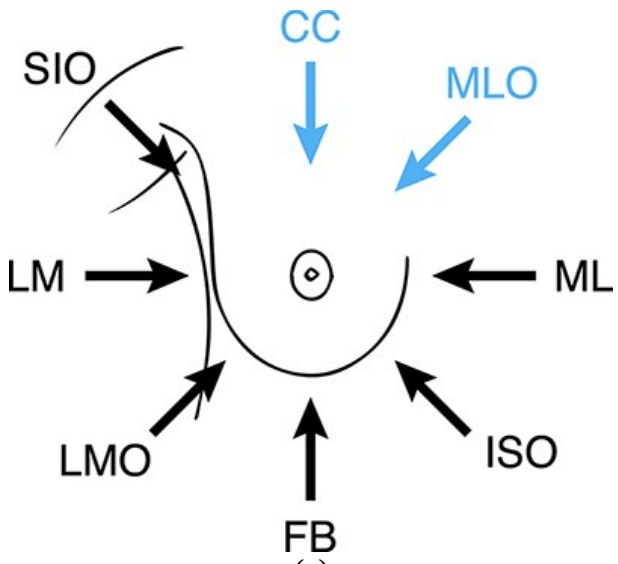
O problema principal que queremos resolver é: "Dada uma mamografia, dizer se a paciente tem câncer ou não". Porém, há problemas secundários:

1. Localizar onde está o câncer dentro da mamografia.
2. Segmentar a lesão.
3. 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.
4. Predizer o risco da paciente (que não tem câncer) vir a ter câncer no futuro.

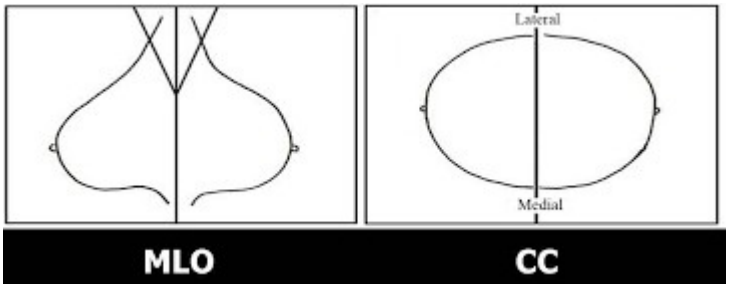
Dizer se tem câncer ou não em mamografia é um problema típico de classificação de imagens.

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.

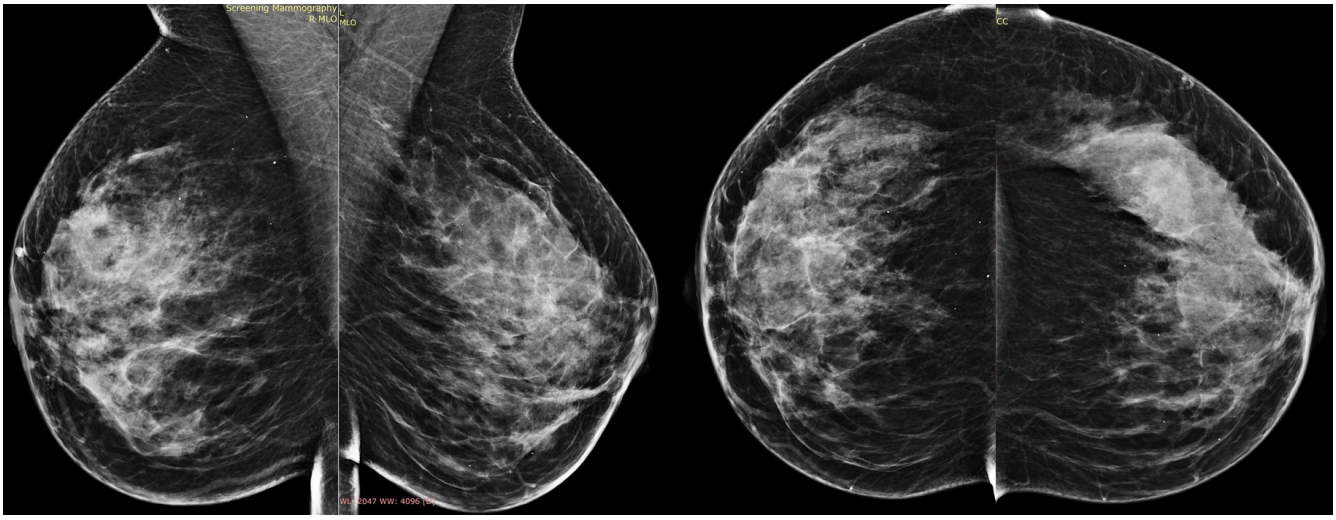
Além disso, costuma-se tirar duas incidências para cada mama: CC (craniocaudal) e MLO (mediolateral-oblíqua) (figura 1).



(a)



(b)



(c)

Figura 1: Vistas MLO e CC.

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 2): 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.

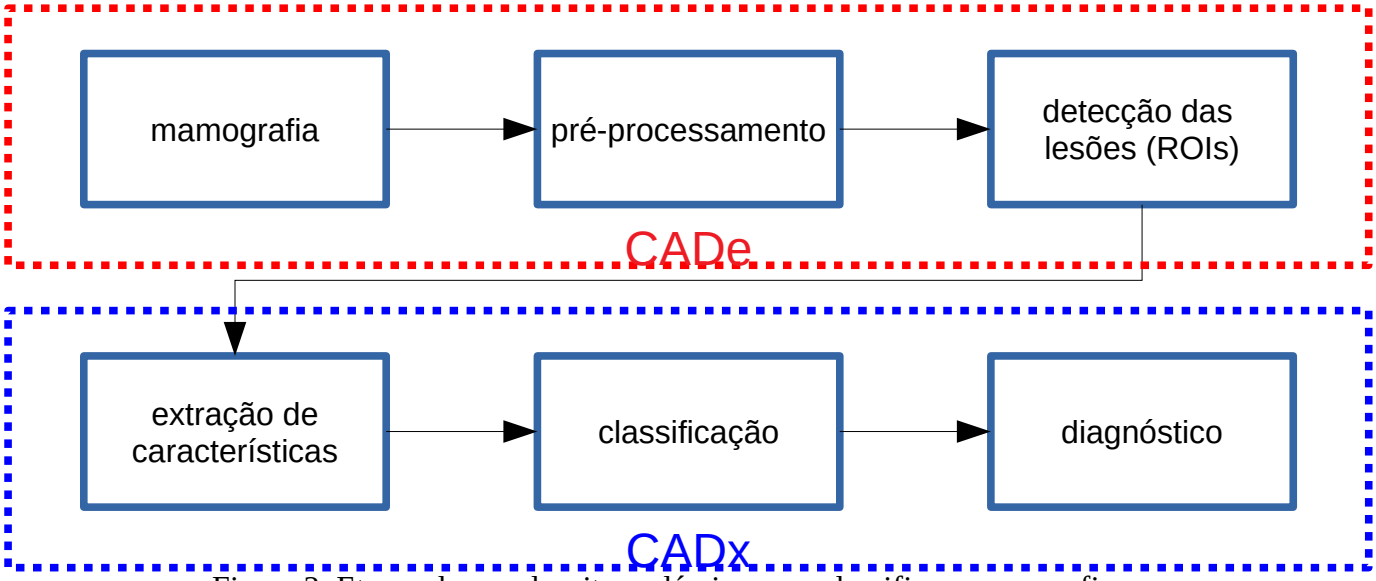
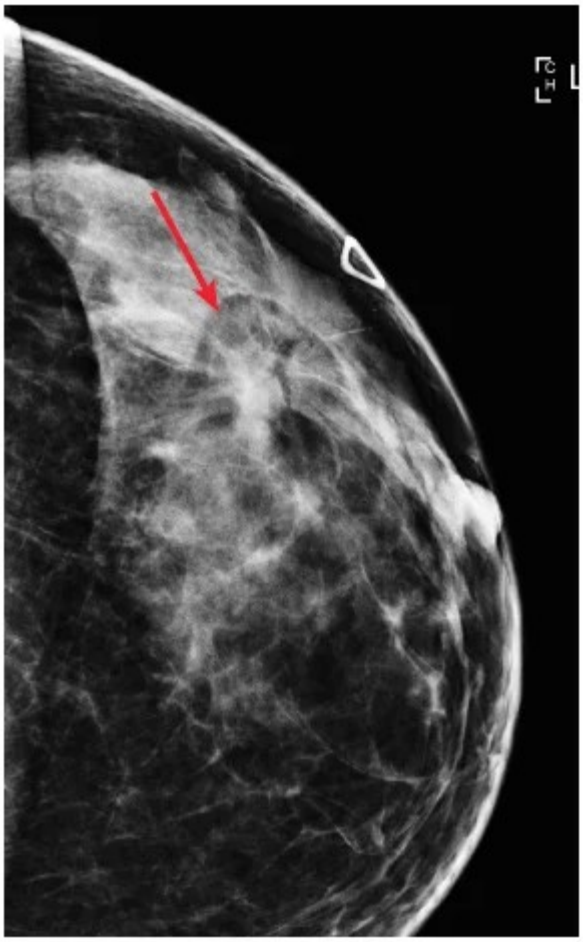


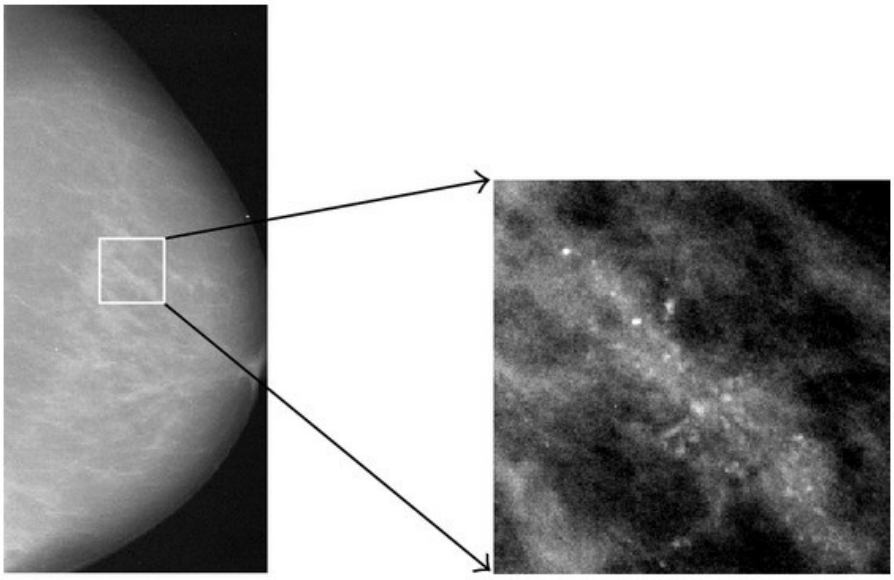
Figura 2: Etapas de um algoritmo clássico para classificar mamografias.

2. Análise de mamografia clássica

Há dois tipos principais de câncer de mama: “massa” e “microcalcificação” (figura 3).



(a) Lesão tipo massa.



(b) Lesão tipo microcalcificação.
Figura 3: Principais tipos de lesões.

Na figura 4, 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 (porcentagem de pacientes sadias que é identificada como sadia) desses CADe era muito baixa, gerando por volta de um falso-positivo por vista. Assim, esses sistemas ajudavam a melhorar a sensibilidade (porcentagem de pacientes doentes que é identificada como doente) 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%).

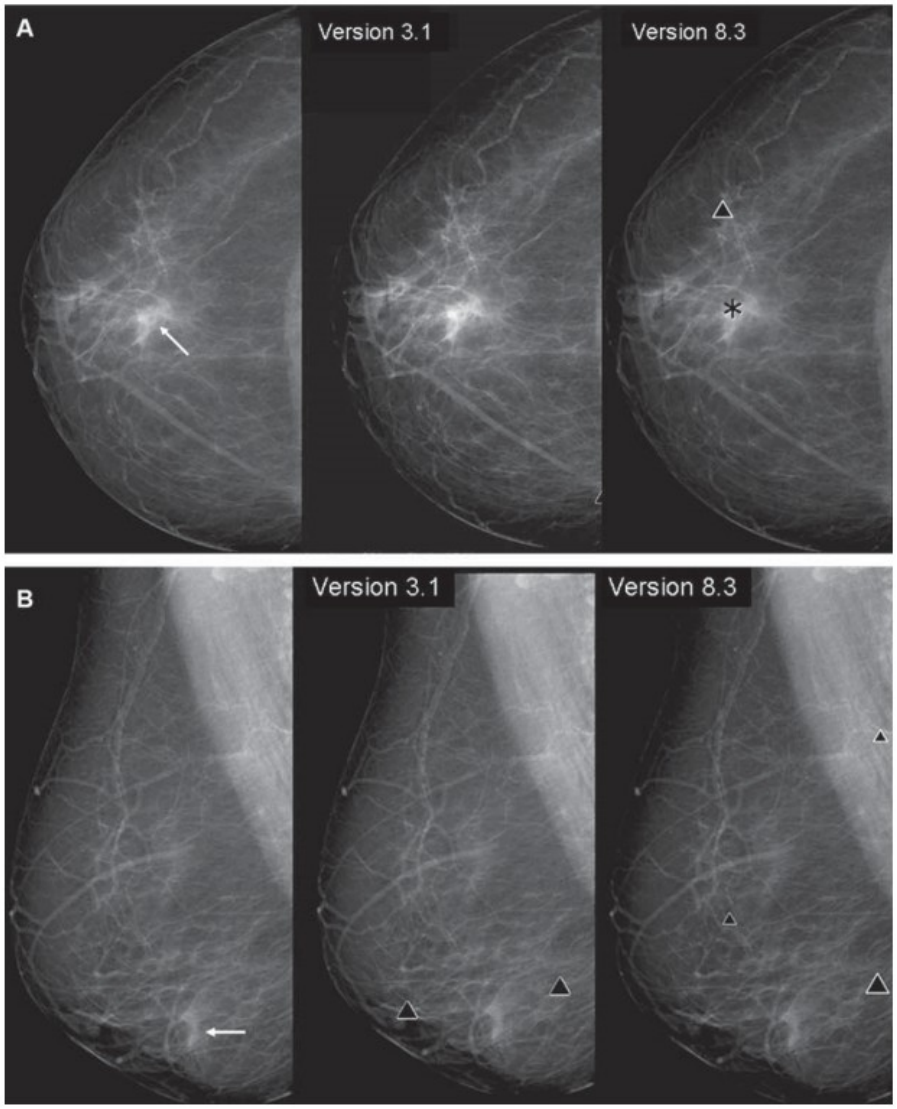


Figura 4: Flecha indica lesão tipo massa verdadeira. CADe indica com * lesão tipo massa e com Δ lesão tipo microcalcificação detectados pelo sistema. Figura retirada de [Kim2010].

Concretizando mais, para detectar microcalcificação usava as técnicas clássicas:

- Contraste local e nível de cinza local.
- Transformada de wavelet para detectar regiões com alta frequência.
- Laplaciano de gaussiana para detectar “blob”.
- Operadores morfológicos.

Para segmentar lesão tipo massa usava as técnicas clássicas:

- Crescimento de região (crescimento de semente).
- Contorno ativo.
- Level sets

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 (isto é, extrair um conjunto de números que permitem classificar ROI em câncer/não-câncer). O trabalho [Elter2009] apresenta algumas características usadas para essa tarefa: forma, densidade, textura, etc.

Para caracterizar ROI tipo microcalcificação:

- Morfologia de calcificações individuais (área, perímetro, circularidade, excentricidade, etc).
- Média e desvio padrão de níveis de cinza na região.
- Distância entre partículas.
- Textura do tecido de fundo.

Para caracterizar ROI de massa:

- Forma (câncer costuma ser espiculado). Circularidade, retangularidade, excentricidade. Momento central. Descritor de Fourier do contorno.
- Análise do gradiente radial.
- Textura.

Depois, os métodos clássicos de aprendizagem (que já estudamos) eram usados para classificá-las: vizinho mais próximo, árvore de decisão, boosting, Bayes, rede neural, SVM, 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, foram mencionadas 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, vamos supor que a chance de uma paciente ter uma certa doença seja 1%. Se um certo algoritmo responder sempre que quem fez o exame não tem doença, 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 5):

a) Verdadeiro positivo (TP): Uma pessoa doente é corretamente classificada como doente.

b) Falso negativo (FN): Uma pessoa doente é classificada erroneamente como sadia.

c) Verdadeiro negativo (TN): Uma pessoa sadia é corretamente 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 doença) 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

O algoritmo que diz que ninguém tem câncer teria acuracidade balanceada de 50%, equivalente a “chute”. Acuracidade balanceada é melhor que acuracidade, mas não resolve completamente o problema de medir o desempenho de um teste.

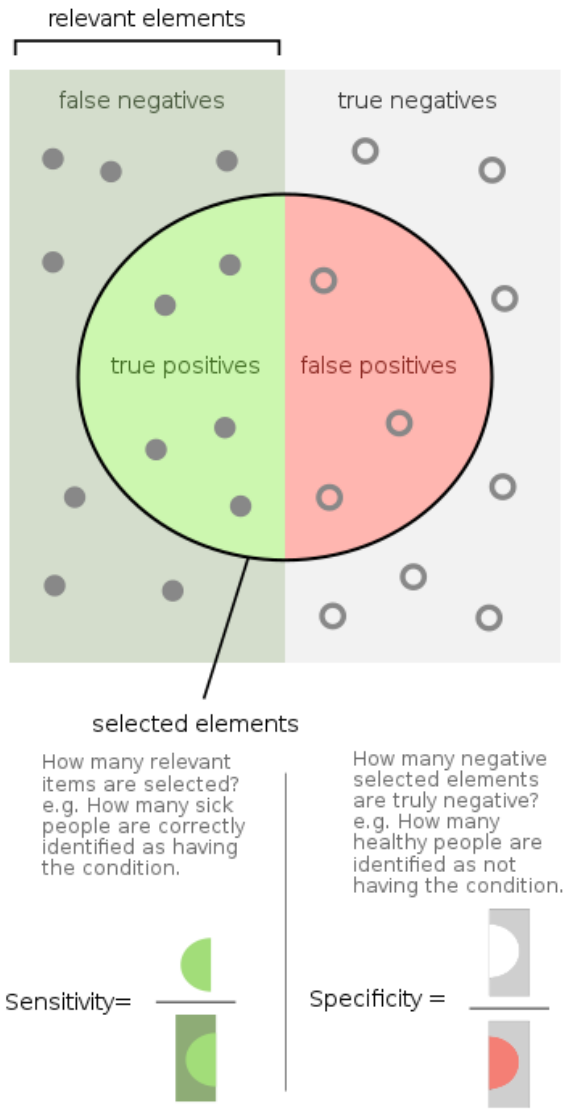


Figura 5: (extraído de Wikipedia)

Na maior parte deste curso, vamos usar simplesmente acuracidade ou taxa de erro (1-acuracidade) como medidas de desempenho, pois o problema estará mais ou menos balanceado (quantidade de exemplos positivos e negativos são semelhantes). Porém, nos exames médicos, não se pode usar acuracidade como métrica, pois normalmente o número de pacientes com doença é muito menor que o número de pacientes sem doença.

Na verdade, nenhuma das métricas que vimos (sensibilidade, especificidade, acuracidade e acuracidade balanceada) é adequada. O computador não costuma forncer uma resposta binária câncer/não-câncer, mas dá uma "nota" entre 0 e 1, uma espécie de "probabilidade" da paciente ter câncer. É necessário limiarizar essa "probabilidade" para se obter a resposta booleana.

Assim, todas as métricas que vimos (sensibilidade, especificidade, acuracidade e acuracidade balanceada) dependem do limiar escolhido. Consequentemente, não é possível comparar se um método é melhor ou pior que outro baseado em sensibilidade, especificidade, acuracidade ou acuracidade balanceada, pois precisaria especificar qual foi o limiar escolhido.

A métrica de desempenho que não depende da escolha do limiar é ROC-AUC (Area Under ROC 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 o limiar (figura 6). 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 que nunca erra possui AUC=1 terá curva ROC em forma de Γ. Um algoritmo que equivale a "chute cego" terá AUC=0.5 e a sua curva ROC será $y=x$. A figura 6 exemplifica uma curva ROC.

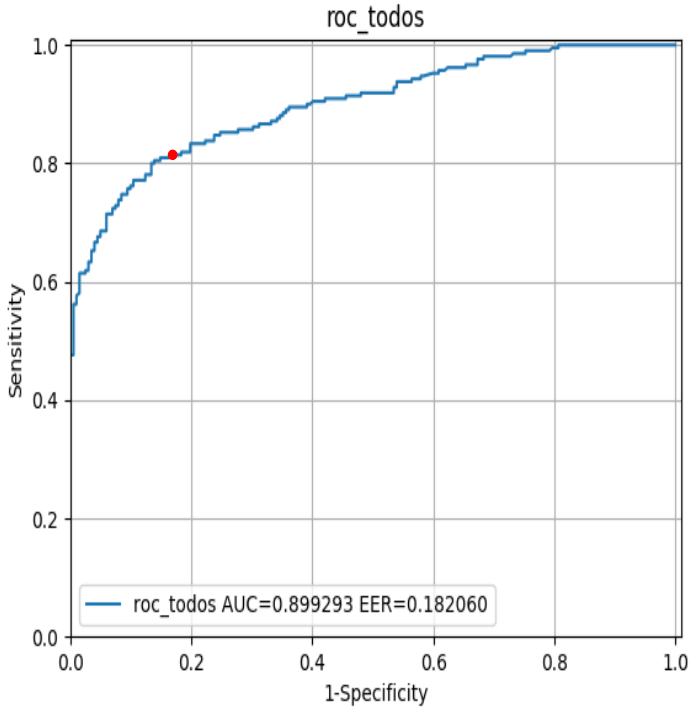


Figura 6: Exemplo de curva ROC. O ponto vermelho indica EER (equal error rate).

Há um ponto especial, denominado de EER (equal error rate), onde a acuracidade, sensibilidade e especificidade se tornam iguais (o ponto vermelho na figura 6). 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 2 anos e não desenvolveu câncer nesse tempo.

3.3. Outras métricas

[A preencher]

[Lição de casa aula 5, #1 (vale 5,0)] ~/haepi/deep/algpi/densa/noisynote/noisynote2.py
Ao classificar os atributos de 5 notas, obteve como resultado:

qp = [[0.32073036] [0.27087152] [0.37556642] [0.2301418] [0.40878805]]

quando a classificação correta seria:

qy = [0.0 0.0 0.0 0.0 1.0]

Calcule taxa de erro, especificidade, sensibilidade para limiar = 0.5 e limiar = 0.4 e limiar=0,3.
Trace a curva ROC, calcule AUC e calcule EER.

4. Análise de mamografia usando aprendizado profundo

Usando aprendizado profundo, o próprio algoritmo de aprendizado extrai os atributos 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.

Outro problema atual da inteligência artificial é fazer IA explicar a decisão tomada. Fazer que IA explique por que classificou uma certa mamografia como cancerígena ou não-cancerígena. Apontar a região com câncer (figura 7) ajudaria o ser humano compreender por que IA classificou uma certa mamografia como cancerígena.

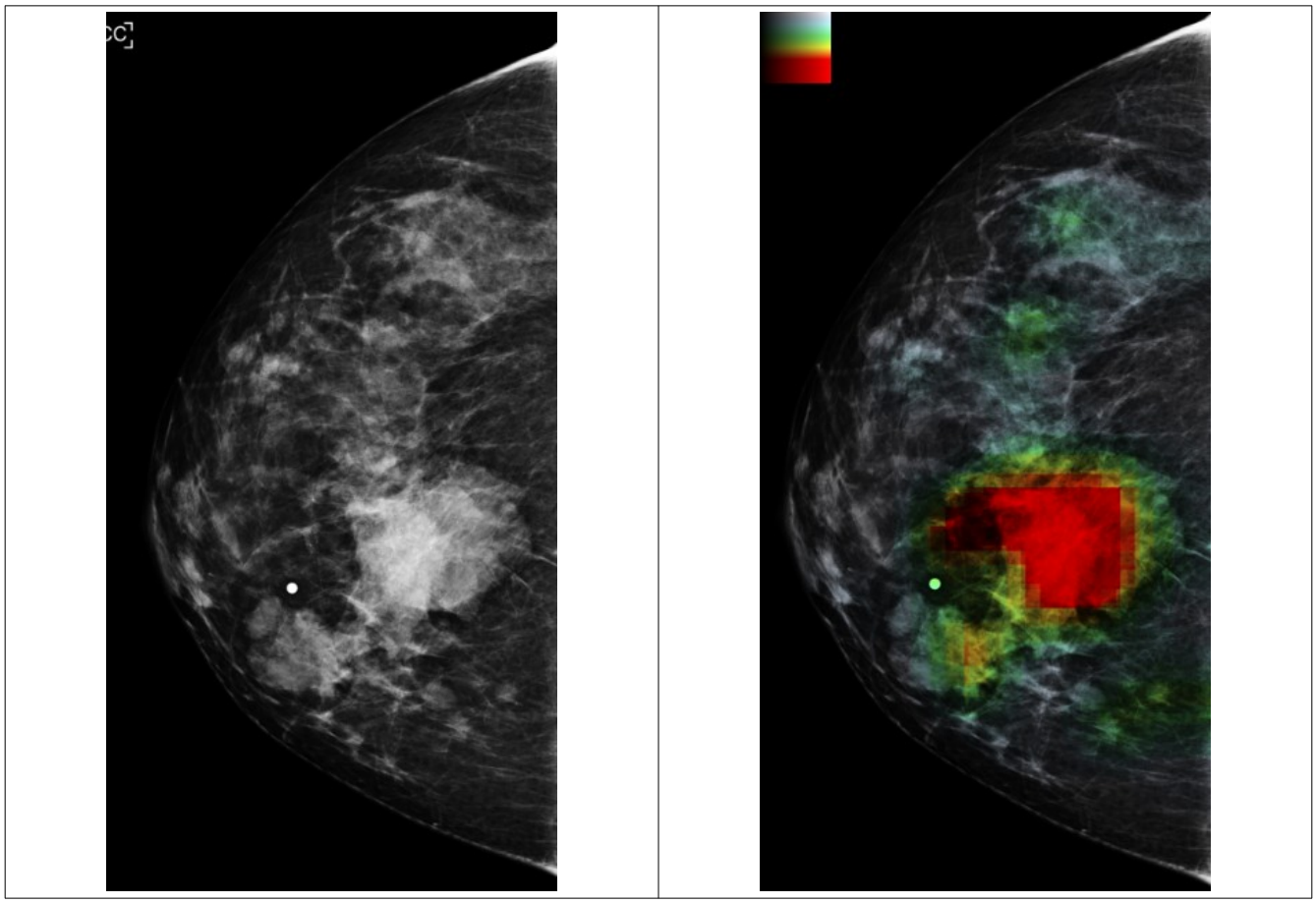


Figura 7: “Heatmap” gerado pelo programa [Wu2020] indicando regiões com maior possibilidade de lesão.

[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.

[PSI5790-2025 aula 5 parte 1. Fim]