**Universidade de Brasília**
**Faculdade de Tecnologia**
**Departamento de Engenharia Mecânica**

# Modelo de aprendizagem não supervisionado baseado em saliência visual para segmentação automática da região pulmonar em imagens de raio-X

Pedro Aurélio Coelho de Almeida

DISSERTAÇÃO DE MESTRADO

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS MECATRÔNICOS

Brasília

2023

**Universidade de Brasília**
**Faculdade de Tecnologia**
**Departamento de Engenharia Mecânica**

# Modelo de aprendizagem não supervisionado baseado em saliência visual para segmentação automática da região pulmonar em imagens de raio-X

Pedro Aurélio Coelho de Almeida

Dissertação de Mestrado submetida ao Departamento de Engenharia Mecânica da Universidade de Brasília como parte dos requisitos necessários para a obtenção do grau de Mestre

Orientador: Prof. Dr. Díbio Leandro Borges

Brasília
2023

Coelho de Almeida, Pedro Aurélio.
C769m Modelo de aprendizagem não supervisionado baseado em saliência visual para segmentação automática da região pulmonar em imagens de raio-X / Pedro Aurélio Coelho de Almeida; orientador Díbio Leandro Borges. -- Brasília, 2023.
66 p.

Dissertação de Mestrado (Programa de Pós-Graduação em Sistemas Mecatrônicos) -- Universidade de Brasília, 2023.

1. Segmentação semântica. 2. DeepLabV3. 3. CNN. 4. Convolução. 5. Imagens de raio-X de pulmão. 6. Inteligência Artificial. 7. IA. I. Leandro Borges, Díbio, orient. II. Título

**Universidade de Brasília**
**Faculdade de Tecnologia**
**Departamento de Engenharia Mecânica**

# Modelo de aprendizagem não supervisionado baseado em saliência visual para segmentação automática da região pulmonar em imagens de raio-X

Pedro Aurélio Coelho de Almeida

Dissertação de Mestrado submetida ao Departamento de Engenharia Mecânica da Universidade de Brasília como parte dos requisitos necessários para a obtenção do grau de Mestre

Trabalho aprovado. Brasília, 25 de agosto de 2023:

---

**Prof. Dr. Díbio Leandro Borges, UnB/CIC**
Orientador

---

**Profa. Dra. Suélia Rodrigues Fleury Rosa, UnB/FGA**
Examinador interno

---

**Prof. Dr. Hélio Pedrini, Unicamp/IC**
Examinador externo

Brasília
2023

*Este trabalho é dedicado a todos que me auxiliaram durante a jornada: meus pais que sempre me apoiaram em todas as minhas decisões; minha namorada Natália que é extremamente prática e realista, sempre me apoiando e me lembrando de manter os pés no chão; meu orientador, professor Díbio, cuja paciência, conhecimento e incentivo me guiaram pelos caminhos não lineares da pesquisa; minha psicóloga, Luciana, sem a qual não haveria estabilidade mental para trabalhar.*

# Agradecimentos

Agradecimentos são realizados ao Prof. Díbio Leandro Borges por todo o apoio e dedicação na árdua tarefa de orientar pesquisas, ao PPMEC por todo suporte e celeridade na resolução de demandas institucionais, à UnB por desempenhar um papel de lar para todos os estudantes e funcionários e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) pela bolsa de pesquisa concedida. Há também de se mencionar o Laboratório de Imagens, Sinais e Áudio (LISA) e os departamentos de Ciência da Computação e Engenharia Mecânica pela estrutura acadêmica que é a base para qualquer pesquisa.

# Resumo

Separar automaticamente regiões com propriedades semelhantes em uma imagem, também chamada de segmentação, é uma tarefa desafiadora para sistemas computacionais, mas também capaz de economizar esforço e evitar erros por fadiga de seres humanos. Uma das áreas do conhecimento que pode se beneficiar de métodos de segmentação automática é a de imagens médicas. Nesse aspecto, raios-X torácicos compõem uma aplicação extremamente valiosa de segmentação automática de imagens, devido ao seu baixo custo de implementação e captura de informações ligadas a doenças pulmonares. Métodos computacionais automáticos atuais com grande aplicação na segmentação pulmonar de raios-X necessitam de dados previamente rotulados para 'aprender' a realizar essa tarefa. Uma alternativa a esses métodos é a aprendizagem profunda não supervisionada, que precisa somente da imagem de raio-X. Considerando o aspecto visual de segmentação e o destaque visual da área pulmonar dentro de uma imagem de raio-X, propõe-se combinar aprendizagem profunda não supervisionada com a área de saliência visual, que busca estimar as porções da imagem que mais atraem a atenção visual humana, para segmentar a região pulmonar de raios-X. O método de saliência visual é comparado a outros trabalhos de aprendizagem não supervisionada e também supervisionadas destinados à segmentação de regiões corporais em imagens em escala de cinzas. Os resultados com as métricas Dice, Jaccard, precisão e revocação nas bases de dados JSRT e MC indicam que a melhoria de desempenho do modelo baseado em saliência é estatisticamente significativa quando comparado às técnicas não supervisionadas. Se analisado a partir de abordagens supervisionadas, o método baseado em saliência parece ser adequado como substituto das suas contrapartes, tendo em vista a flexibilidade obtida pela independência de rótulos manualmente definidos. Trabalhos futuros incluem segmentar a área cardíaca e identificar anomalias em imagens de raios-X de forma não supervisionada.

**Palavras-chave**: Segmentação semântica. DeepLabV3. CNN. Convolução. Imagens de raio-X de pulmão. Inteligência Artificial. IA.

# Abstract

Automatically dividing an image into regions of similar properties, named segmentation, is a challenging task for computers, and it can avoid human errors induced by fatigue. One area that may be greatly benefited from automatic segmentation methods is medical imaging analysis. Within it, chest X-rays are amongst the cheapest and most widely available type of medical images. Provided they can be used for diagnosing lung related diseases, they are an excellent target for automatic image segmentation methods. The current state-of-the art image segmentation relies on manual labels defined *a priori* to 'learn' the necessary features for this task. Deep unsupervised learning stands as an interesting alternative to supervised methods, since it only requires the input (e.g. the image X-ray) for training. Due to the visual nature of image segmentation and the standout aspect of the lungs on an X-ray, the combination of unsupervised learning and visual saliency (i.e. the attempt to model human visual attention) is tested for the lung segmentation on X-ray images. The saliency method is compared to state-of-the-art supervised and unsupervised models designed for grayscale medical image segmentation. Results using the Dice, Jaccard, precision and recall scores on JSRT and MC datasets indicate that the saliency method enhanced performance over other unsupervised approaches is statistically significant. When compared to supervised models, the saliency method appears to adequately substitute them given the flexibility achieved by the independence from manual labels. Future work includes segmenting the cardiac area and identifying anomalies on X-ray images in an unsupervised fashion.

**Keywords**: Semantic segmentation. DeepLabV3. CNN. Convolution. Chest X-ray images. Artificial Intelligence. AI.

# Lista de ilustrações

# Lista de tabelas

# Sumário

# 1 Introdução

Este capítulo apresenta o problema de segmentação da região de imagens de raios-X a partir de um modelo profundo não supervisionado. Uma descrição desse problema é feita na Seção 1.1. As perguntas que nortearam o desenvolvimento desse estudo se encontram na Seção 1.2. O objetivo geral do trabalho é exposto na Seção 1.3, enquanto que os objetivos específicos são listados na Seção 1.4. A Seção 1.5 traz as contribuições científicas da proposta. A Seção 1.6 resume brevemente o conteúdo dos demais capítulos.

## 1.1 Caracterização do problema

Raio-X torácico é uma modalidade de imagem médica barata e de amplo acesso, cujo uso pode auxiliar no diagnóstico de pneumonias, tuberculose, câncer pulmonar, dentre outras condições. Por ser um trabalho repetitivo e que envolve diferenças sutis de intensidade entre casos normais e atípicos, abordagens computacionais podem auxiliar os profissionais da saúde, indicando regiões de interesse que poderiam passar despercebidas pelo olhar humano. Uma das tarefas do fluxo geral de diagnóstico assistido por computador (CAD, na sigla em inglês) é a segmentação da região pulmonar, uma etapa fundamental para habilitar a extração de atributos utilizados na análise das imagens (Qin et al., 2018).

Tradicionalmente, diversas áreas de processamento de sinais, como, por exemplo, estimação de saliência visual utilizaram conjuntos de atributos fixos a fim de realizar tarefas como, por exemplo, classificação (Wei et al., 2012). Com o advento de modelos parametrizáveis capazes de ajustar seus parâmetros a partir dos dados, chamados de *neural networks* (em inglês ou redes neuronais em tradução livre), tornou-se possível delegar ao modelo a escolha e criação de atributos a partir dos dados de entrada e, com isso, atingir desempenho superior às abordagens tradicionais (Goodfellow et al., 2016, pp. 1-7). Grandes quantidades de dados permitem que os modelos tenham desempenho comparável à de humanos em várias tarefas. Porém, a construção dessas grandes bases de dados rotuladas é custosa, o que pode implicar escassez ou indisponibilidade (Qin et al., 2018 e Borji et al., 2019).

Como forma de mitigar a dependência de bases rotuladas, linhas de pesquisa baseadas em conceitos de aprendizagem não supervisionada como, por exemplo, o *K-means* (James et al., 2023, pp. 515-519), em inglês, ou $K$-médias, em tradução livre, começaram a ser desenvolvidas. Trabalhos como os de Kim e Ye (2020) e Chen et al. (2021b) exemplificam técnicas de agrupamento, semelhantes ao $K$-médias, utilizadas para imagens coloridas e em escalas de cinza. As vantagens em se combinar redes neuronais e aprendizagem não supervisionada estão na capacidade de aprender atributos complexos dos dados de entrada sem precisar de rótulos manualmente definidos, o que abre um leque para aplicação

em uma quantidade maior de dados, cujos atributos podem mudar dinamicamente. Essa flexibilidade pode estar associada a um desempenho menor do que aquele obtido por modelos supervisionados (Croitoru et al., 2019).

Uma abordagem não supervisionada para segmentação pulmonar em imagens de raios-X, alternativa ao uso de funções de agrupamento como o $K$-médias, pode se basear no conceito de saliência visual. Essa área propõe modelos para estimar a atenção visual humana, ou seja, regiões na imagem que atraem o olhar das pessoas (Borji et al., 2019 e Wolfe e Horowitz, 2017). Uma vez que, para as pessoas em geral, a área pulmonar é visualmente distinta e contrastante do restante da imagem, modelos que buscam aproximar os mecanismos que guiam o olhar humano podem também conseguir realizar segmentação pulmonar de raios-X. Uma das vantagens de modelos baseados em saliência com relação àqueles baseados em agrupamento por cor está na possibilidade de aprender critérios baseados em orientações, texturas e outras características salientes como, por exemplo, aquelas descritas por Wolfe e Horowitz (2017).

## 1.2 Perguntas de pesquisa

As perguntas que nortearam a pesquisa apresentada foram:

- Conceitos de saliência visual podem ser utilizado para treinar um modelo profundo não supervisionado para segmentação pulmonar em imagens médicas?
- O modelo não supervisionado consegue manter seu desempenho quando uma distribuição de dados diferente daquela utilizada para treinamento é apresentada a ele?

## 1.3 Objetivo geral

O objetivo é demonstrar a viabilidade do uso da aprendizagem profunda não supervisionada em conjunto com princípios de saliência visual para aplicações nas áreas de imagens médicas, fornecendo evidências substanciais.

## 1.4 Objetivos específicos

- Adaptar o método de aprendizagem profunda não supervisionada para estimação de saliência visual proposto por Zhang et al. (2018) a um contexto de segmentação pulmonar de imagens médicas;
- Comparar a proposta baseada em saliência com abordagens de aprendizagem profunda não supervisionada baseada em agrupamento por cores ou níveis de intensidade descritas por Kim e Ye (2020) e Chen et al. (2021b);

- Testar a abordagem baseada em saliência em diferentes bases de raios-X torácicos para verificar se o modelo obtém desempenhos semelhantes em um base de dados com distribuição de dados diferente daquela utilizada para treinamento.

## 1.5 Contribuições científicas

Como contribuições para o processo científico, pode-se mencionar a descrição de uma abordagem para segmentação pulmonar em imagens de raios-X que é não supervisionada e se baseia em princípios de atenção visual.

## 1.6 Organização do trabalho

O Capítulo 2 apresenta uma revisão de literatura, o Capítulo 3 mostra os materiais e métodos adotados, o Capítulo 4 apresenta e discute os resultados e o Capítulo 5 conclui o texto e sugere trabalhos futuros.

# 2 Revisão de Literatura

Conceitos importantes relacionados aos temas de segmentação, redes neuronais e processamento de imagens serão abordados nesse capítulo. A Seção 2.1 apresenta a definição de raios-X como ondas eletromagnéticas, sua relação com a análise da região pulmonar e como são armazenadas em computadores. As técnicas de equalização de histograma e correção gamma são definidas na Seção 2.2. O conceito de saliência visual e exemplos de métodos heurísticos tradicionais que estimam a saliência são encontrados na Seção 2.3. Métodos de aprendizagem de máquinas, em especial aqueles relacionados à aprendizagem profunda, são descritos na Seção 2.4. Caso o leitor tenha familiaridade com esses temas, pode pular para o Capítulo 3 que apresenta os materiais e métodos utilizados.

## 2.1 Raio-X

A visão humana é o componente responsável por captar luz do ambiente que circunda uma pessoa e transformá-la em um sinal elétrico que será processado pelo cérebro. Estudos mostraram que a luz pode ser entendida como uma onda composta por partículas sem massa (fótons) que carregam energia. Além disso, as cores que os humanos conseguem ver são determinadas pelas frequências dessa luz, dentro de um espectro chamado de luz visível. Pesquisas com essas ondas mostraram que existem frequências que são invisíveis para o olhar humano, mas podem ser detectadas por elementos sensíveis à determinada frequência. Assim, pode-se subdividir o espectro de luz em sete grandes regiões: frequências maiores que a porção visível (raios gamma, raios-X e ultravioleta), espectro visível e frequências menores que a porção visível (infravermelho, micro-ondas e ondas de rádio), como conceituado por Gonzalez e Woods (2018, pp. 50, 54–56). A Figura 1, adaptada de Villate (2019, p. 262), ilustra as diferentes bandas do espectro de luz.

Os exames de raio-X usam ondas de luz com frequências de ordem de magnitude de $10^{18}$ Hertz. Essas ondas são geradas pela colisão de elétrons em altas velocidades contra núcleos atômicos e, caso passem por um filme sensível a essa radiação, geram imagens que ajudam a tornar visíveis partes do corpo humano, como ossos e pulmões. Para serem analisadas por computadores, essas imagens precisam ser digitalizadas, processo que envolve transformar os valores contínuos de intensidade do raio-X em valores discretos (Gonzalez e Woods, 2018, pp. 24-26). A Figura 2 mostra um exemplo de raio-X torácico após o processo de digitalização.

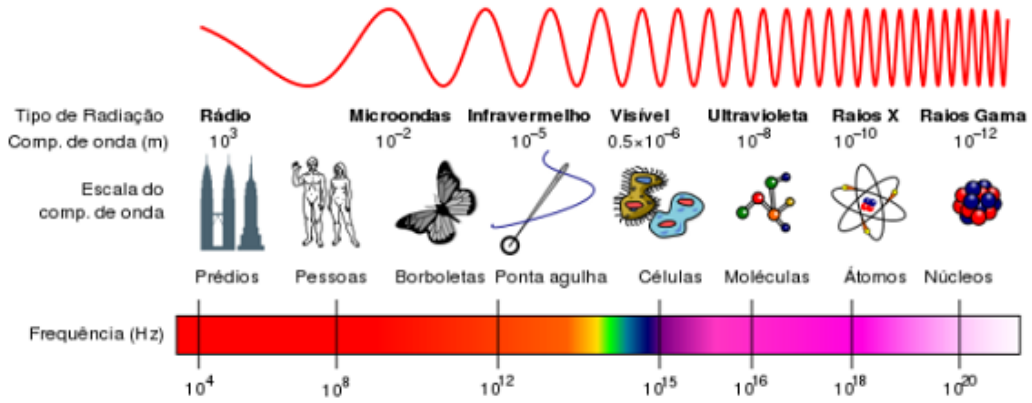



Figura 1 – Representação das diferentes frequências e comprimentos de onda do espectro luminoso, bem como objetos com escala semelhante a cada grupo de ondas (rádio, micro-ondas, infravermelho, visível, ultravioleta, raios-X e raios gamma).

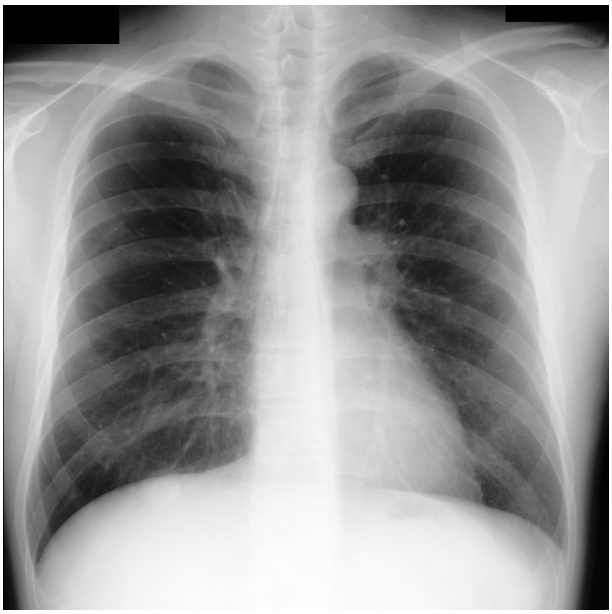


Figura 2 – Raio-X torácico após digitalização extraído da base de dados JSRT (Shiraishi et al., 2000). Caso a imagem seja ampliada, pode-se perceber que ela é composta por elementos discretos (pixels) e, por isso, contém transições não contínuas de valores, diferentemente da versão contínua impressa sobre papel sensível à onda de raio-X.

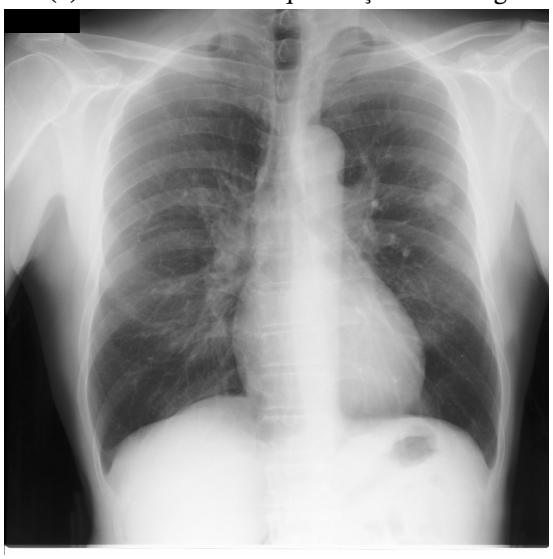
## 2.2 Processamento de imagens em escala de cinza

Observando a Figura 2, pode-se perceber que as cores associadas ao espectro visível de luz mostradas na Figura 1 não estão presentes nas imagens de raio-X. Em vez delas, vê-se uma escala de luminosidade que pode ir do preto (mínima intensidade) ao branco (máxima intensidade) e que apresenta diversos níveis de cinza entre esses extremos. A esse tipo de imagem, dá-se o nome de escala de cinza. Por seu uso em várias áreas do conhecimento, técnicas de processamento de imagens foram criadas para melhorar ou extrair atributos desse tipo de imagem. Dentre essas, pode-se mencionar a equalização de histograma e a correção gamma (Gonzalez e Woods, 2018, pp. 57, 134–140, 125–128).

### 2.2.1 Equalização de histograma

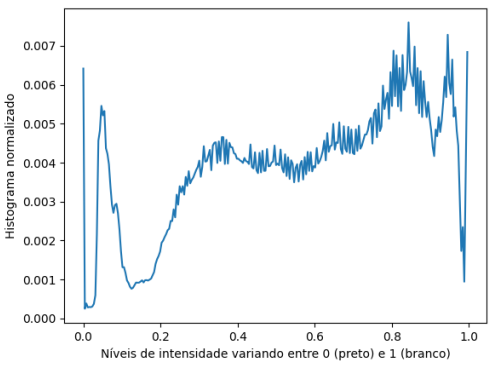
A equalização de histograma é uma técnica utilizada para aproximar o histograma de uma imagem a uma distribuição uniforme. Visualmente, tem efeito de ampliar o contraste de uma imagem, tornando os detalhes mais fáceis de serem visualizados (Gonzalez e Woods, 2018, pp. 134-140). A Figura 3 mostra uma imagem de raio-X antes e após a equalização de histograma, bem como os histogramas em cada fase desse processo. Note o maior contraste da região pulmonar com relação à área corporal adjacente da versão com histograma equalizado (Figura 3 b) com relação a sua versão não equalizada (Figura 3 a). Além disso, o histograma, após aplicação da normalização de histograma (Figura 3 d), é mais próximo de uma distribuição uniforme do que a versão não equalizada (Figura 3 c).

(a) Raio-X antes da equalização de histograma



(b) Raio-X após a equalização de histograma

(c) Histograma de imagem não equalizada




(d) Histograma de imagem equalizada

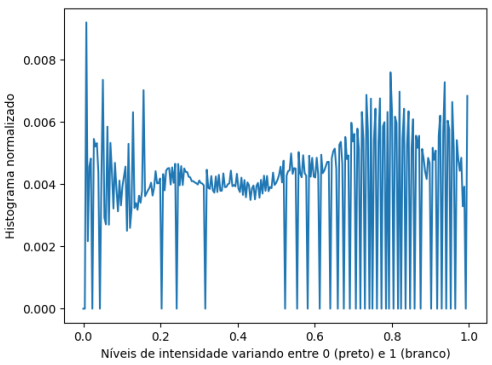



Figura 3 – Raio-X, retirado da base JSRT (Shiraishi et al., 2000), antes (a) e após (b) a aplicação da equalização de histograma, bem como dos respectivos histogramas (c) e (d).

### 2.2.2 Correção gamma

A correção gamma consiste em elevar cada valor de intensidade de uma imagem a um fator $\gamma$. Nesse procedimento, a escala de cinzas varia em um intervalo $[0, 1]$. Logo, utilizar $gamma \in (0,1)$ vai mapear as intensidades para patamares mais claros (mais próximos a 1), enquanto que $\gamma \in (1, \infty)$ leva os níveis de cinza para intensidades mais escuras (mais próximos a 0), como ilustrado nas Figuras 4 e 5 (Gonzalez e Woods, 2018, pp. 125-128). Note que os valores extremos 0 e 1 não são afetados pela correção gamma, o que preserva a amplitude de níveis de intensidade para imagens que apresentam esses extremos.

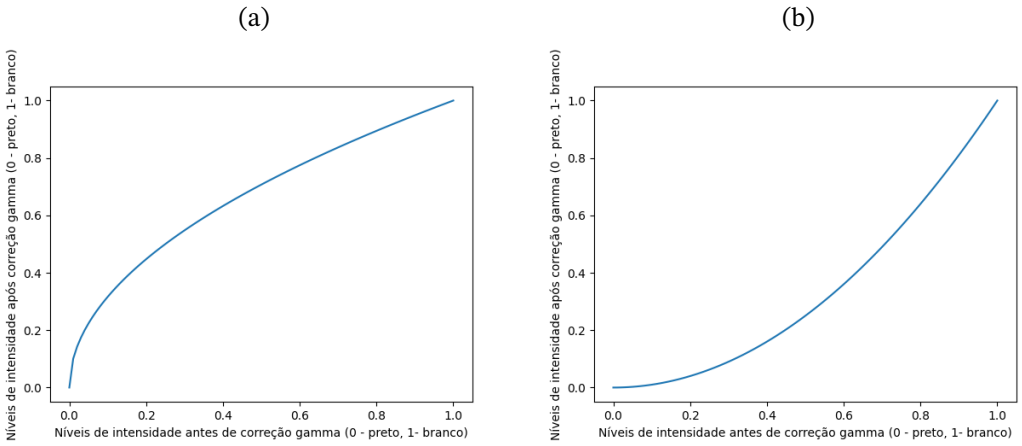



Figura 4 – Correção gamma com diferentes intervalos de $\gamma$: (a) entre $(0, 1)$ ($\gamma = 0.5$), tem-se os seguintes mapeamentos 0->0, 0.25->0.5, 0.5 ->0.71, 0.75->0.87, 1->1; (b) entre $(1, \infty)(\gamma = 0.5)$, tem-se os seguintes mapeamentos 0->0, 0.25->0.06, 0.5 ->0.25, 0.75->0.56, 1->1

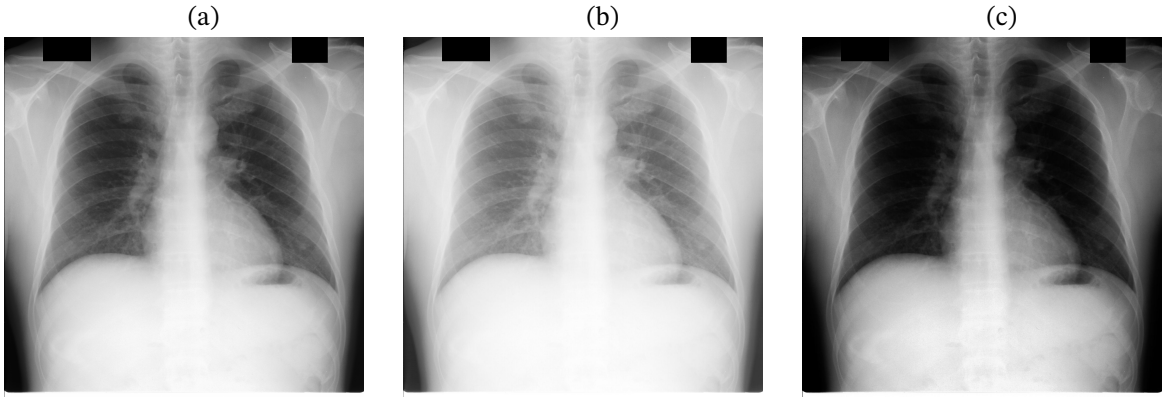



Figura 5 – Aplicações de diferentes valores de $\gamma$ em Raio-X extraído da base JSRT (Shiraishi et al., 2000): (a) imagem sem processamento, (b) processamento com $\gamma = 0.5$, (c) processamento com $\gamma = 2$.

## 2.3 Saliência visual

O campo de visão cotidiano das pessoas pode englobar centenas ou milhares de objetos diferentes, mas a atenção visual é um recurso cerebral limitado. Para interagir com o

ambiente à sua volta em tempo real, o cérebro precisa filtrar rapidamente a informação recebida para direcionar sua atenção a elementos potencialmente importantes. A essas porções se dá o nome de regiões, ou objetos, salientes. Vários fatores parecem ser levados em conta na definição de saliência visual, como contraste de cores e orientações de objetos, formatos, contexto da cena, histórico visual e valor atribuído a alguma característica específica da imagem (Wolfe e Horowitz, 2017). A Figura 6 exemplifica esses fatores.

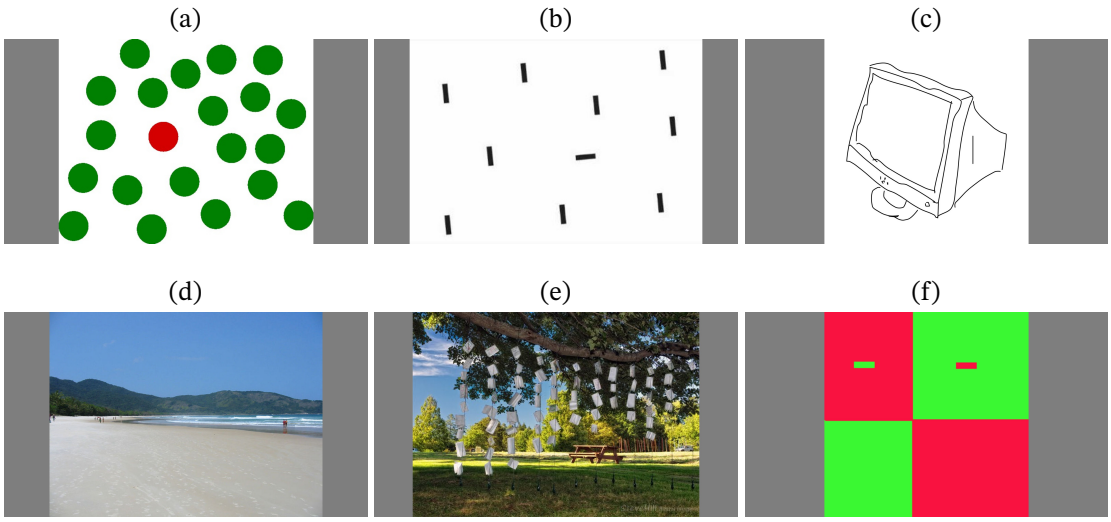



Figura 6 – Ilustrações, retiradas da base de dados CAT2000 (Borji e Itti, 2015), exemplificando os fatores que guiam saliência: a) contraste de cor ressalta o círculo vermelho; b) orientação destaca a barra mais horizontal; c) informações de contorno que definem formas são suficientes para identificar objetos; d) caso uma pessoa fosse procurar por pessoas na imagem, ela olharia para a região da praia e do mar, evitando o céu devido ao contexto; e) se alguém precisou identificar bancos em parque recentemente, o histórico visual facilitará encontrar o objeto em situações semelhantes; f) se retângulos vermelhos dentro de quadrados brancos forem recompensadas, facilita-se a busca desses objetos na imagem.

Considerando que, em imagens de raio-X, a região pulmonar tem destaque visual para um observador padrão, pode-se pensar que modelos de saliência visual consigam aproximar a região pulmonar. Desse modo, não seria necessário utilizar conhecimento médico especializado, já que o objetivo é somente realizar a segmentação automática para auxiliar outros sistemas computacionais e não prover um diagnóstico detalhado. Assim, o sistema de segmentação pode ser totalmente automatizado e disponível em casos totalmente novos.

Devido às características de compressão de informação e extração de objetos, os mecanismos de saliência visual foram modelados por diferentes abordagens. Uma das formas de saída de modelos de saliência é um mapa de valores contínuos indicando a probabilidade de cada pixel pertencer a um objeto saliente. Assim, o objetivo não é modelar os pontos aos quais o olho humano vai prestar atenção, mas sim o(s) objeto(s) correspondente(s) (Borji et al., 2019).

Tradicionalmente, os atributos utilizados para cálculo de saliência eram fixos e

não eram extraídos a partir dos dados. Em regra, essas abordagens consideravam aspectos *bottom-up* (em inglês), englobando diferenças de cores, orientações, tamanhos, dentre outras heurísticas. Entre os diversos trabalhos que utilizam um conjunto de atributos fixos pré determinados, cita-se, como exemplos, a abordagem de Zhu et al. (2014), baseada na hipótese de que objetos salientes raramente encostam nas bordas da imagem, e a de Wei et al. (2012), que adiciona a essa hipótese a característica de que objetos salientes em geral são mais heterogêneos do que os não salientes. A Figura 7 mostra exemplos de mapas de saliência *Robust Background Detection* (Zhu et al., 2014), RBD na sigla em inglês, e *Geodesic Saliency* (Wei et al., 2012), GS na sigla em inglês. Observe que, na Figura 7, ambos os mapas capturam parte da região saliente composta pelas garrafas e taça, porém apresentam ruídos que incluem as pedras ou excluem parte de uma das garrafas.

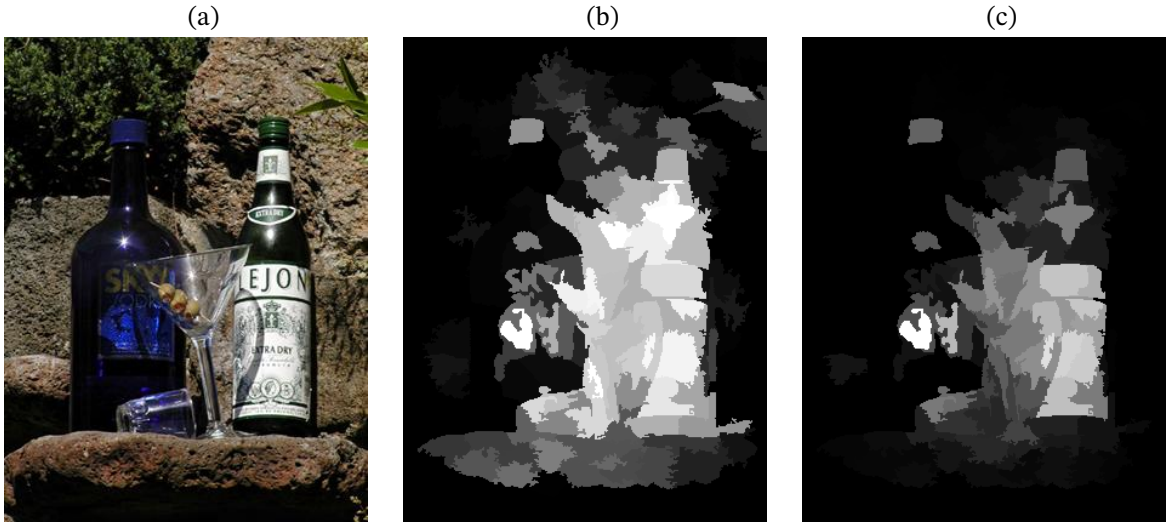



Figura 7 – Imagem de entrada (a), extraída da base MSRA-B Liu et al. (2011), e os respectivos mapas de saliência contínuos produzidos pelo modelos GS (b) e RBD (c).

Os trabalhos de Wei et al. (2012) e Zhu et al. (2014) utilizam o conceito de superpixel, o qual se baseia em agrupar pixels de uma imagem que apresentem alguma propriedade semelhante como, por exemplo, cores e texturas. Devido a essa propriedade, pode-se considerar que cada superpixel está dentro de somente um objeto de uma imagem, geralmente preservando os contornos desses objetos (Veksler et al., 2010). Estudos como o de Felzenszwalb e Huttenlocher (2004) geralmente produzem superpixels com tamanhos variados, enquanto que abordagens como as de Achanta et al. (2012) procuram uniformizar os tamanhos dos superpixels. Esses trabalhos estão ilustrados na Figura 8.

Com a alta eficácia alcançada por técnicas de aprendizagem profunda em problemas de visão computacional, como as redes residuais propostas por He et al. (2015), pesquisadores começaram a aplicar esses métodos para estimação de mapas de saliência. A principal vantagem da aprendizagem profunda é poder automaticamente extrair, a partir dos dados de entrada fornecidos na etapa de treinamento, os atributos necessários para realizar uma

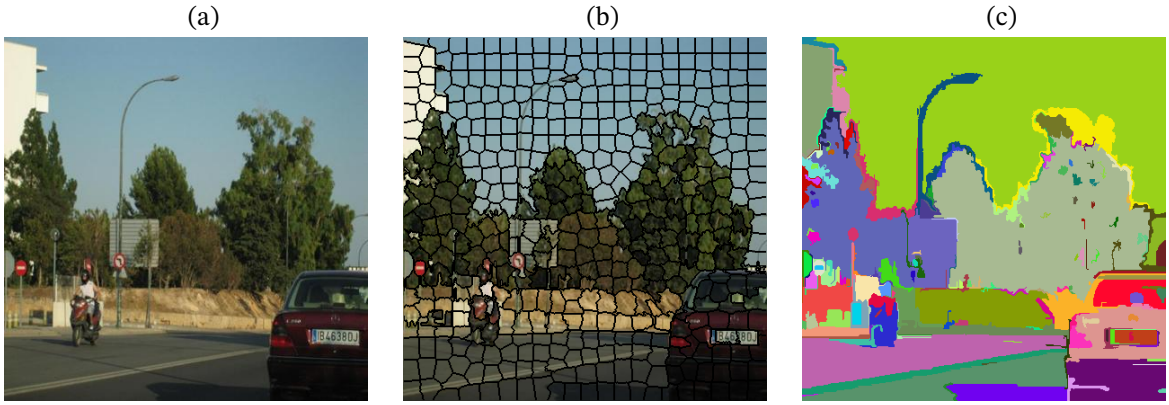



Figura 8 – Imagem de entrada (a), extraída da base criada por Koehler et al. (2014) , e as respectivas segmentações em superpixels produzidas pelos métodos de Achanta et al. (2012) (b) e Felzenszwalb e Huttenlocher (2004) (c).

tarefa, o que permite alcançar um desempenho maior do que os métodos tradicionais que utilizavam atributos fixos e pré-determinados (Borji et al., 2019).

## 2.4 Aprendizagem de máquina

O campo de pesquisa de inteligência artificial (IA) busca, em essência, fazer máquinas imitarem ou simularem o comportamento humano. Para esta finalidade, abordagens utilizando bases de conhecimento foram propostas com o objetivo de programar um sistema computacional com um conjunto de regras formais, de modo que esse conseguisse processar informações com base em regras de inferência lógica. Esses conjuntos de informações formais auxiliaram computadores a resolver problemas que as pessoas consideram difíceis ou mentalmente cansativos como, por exemplo, jogar xadrez em um nível competitivo. Porém, em tarefas cotidianas como reconhecimento de faces ou objetos, mostrou-se um problema muito difícil para técnicas como as bases de conhecimento, uma vez que essas atividades são muitas vezes feitas de forma intuitiva e sem uma definição formal das regras e procedimentos adotados (Goodfellow et al., 2016, pp. 1-2).

A partir da complexidade em definir as regras e relações necessárias para modelar conceitos relacionados ao comportamento humano, surge uma subárea da IA conhecida como aprendizagem de máquina. O foco passa a ser definir atributos relevantes sobre os dados e deixar o sistema computacional estabelecer relações entre as variáveis a partir de dados fornecidos para ajustar o modelo, ou seja, para treiná-lo (Goodfellow et al., 2016, pp. 2-4). Métodos como árvore de decisão, máquinas de vetores suporte (tradução livre para a expressão em inglês *support vector machines*), $K$-médias e regressão logística são exemplos de técnicas de aprendizagem de máquina (James et al., 2023, pp. 335-338, 379–383, 515–519, 133–135).

O modelo de árvore de decisão utiliza uma estrutura de dados de árvore para repre-

sentar os limiares de separação para definir cada classe e, por isso, sua tomada de decisão é fácil de ser interpretada, como ilustrado na Figura 9. As máquinas de vetores suporte utilizam transformações não lineares para linearizar um problema de classificação, processo ilustrado na Figura 10 [1], e, desse modo, traçar hiperplanos de separação dos dados. O algoritmo de $K$-médias atribui aos dados de entrada uma classe aleatória no intervalo $[1, k]$, calcula o centroide de cada classe e reatribui os dados à classe pertencente ao centroide mais próximo, repetindo esse ciclo de cálculo do centroide e reclassificação até a estabilização dos centroides. Já a regressão logística utiliza a função logística $\sigma(x) = \frac{e^x}{1+e^x}$ geralmente substituindo o parâmetro $x$ pelo produto escalar de atributos $a$ e seus respectivos pesos $w$, ou seja, $x = a_1w_1 + a_2w_2 + ... + a_nw_n$ (James et al., 2023, pp. 335-338, 379–383, 515–519, 133–135).

(a)

| Dia | Recomendação do filme | Interesse sobre o tema | Vontade de passear | Ir ao cinema |
|---|---|---|---|---|
| D1 | Pouco interessante | Alto | Baixa | Não |
| D2 | Pouco interessante | Alto | Alta | Sim |
| D3 | Interessante | Baixo | Alta | Não |
| D4 | Muito interessante | Baixo | Baixa | Sim |
| D5 | Interessante | Alto | Alta | Sim |
| D6 | Muito interessante | Alto | Baixa | Sim |

(b)

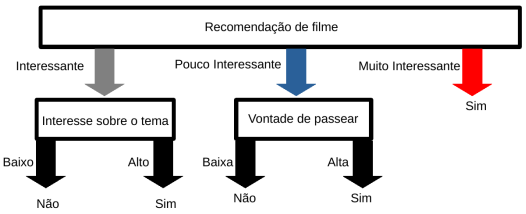



Figura 9 – (a) Tabela de atributos (Recomendação do filme, Interesse sobre o tema e Vontade de passear) utilizados para prever o resultado da saída Ir ao cinema e (b) árvore de decisão resultante, onde os nós folhas representam a saída do modelo. Perceba que as relações entre os atributos são descobertas automaticamente pelo computador. Exemplo baseado em Mitchell (1997, p. 53).

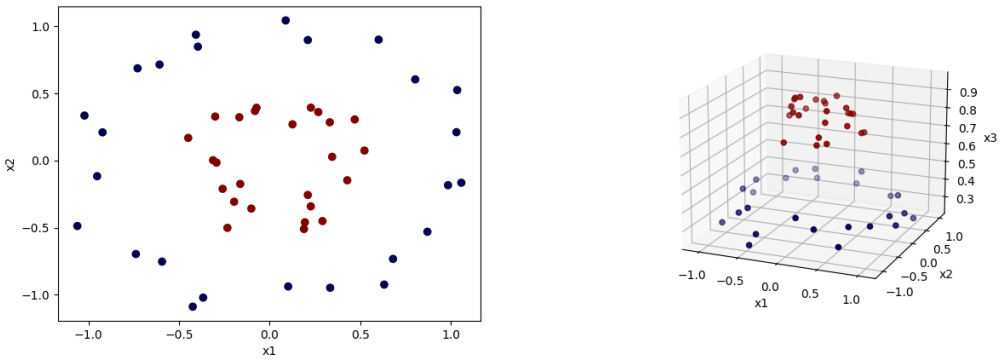



Figura 10 – À esquerda: separação não linear entre duas classes representadas pelas cores azul e vermelho, dados dois atributos x1 e x2. À direita: utilizando uma função não linear $r(x1, x2) = e^{(x1^2+x2^2)}$, pode-se criar uma separação linear entre as classes.

Para amenizar a dependência da escolha de atributos que consigam fornecer uma separação dos dados em diversas classes, surge o campo de aprendizagem profunda. Nele,

[1] Imagens adaptadas de exemplo online (https://jakevdp.github.io/PythonDataScienceHandbook/05.07-support-vector-machines.html, último acesso em 28/06/2023)

os modelos passam a criar representações internas automaticamente a partir dos dados fornecidos. Para isso, modelos conexionistas como os da Figura 11 são construídos, sendo que os nós mais à esquerda da imagem costumam representar conceitos mais simples como texturas ou orientações, enquanto nós mais à direita geralmente retratam conceitos mais abstratos como partes de objetos (Goodfellow et al., 2016, pp. 5-7).

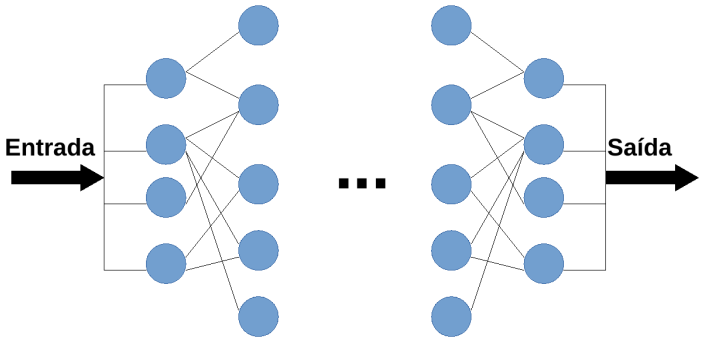



Figura 11 – Conceito geral de aprendizagem profunda: redes conexionistas que transformam os dados de entrada em dados de saída a partir de aplicação de funções e operações matemáticas (convolução, regressão logística, dentre outros), cuja estrutura pode ser vista como um grafo onde os nós são as funções e as arestas realizam a composição de funções.

### 2.4.1 Aprendizagem profunda

O conhecimento obtido a partir de estudos sobre o funcionamento do cérebro por áreas como a neurociência é uma importante fonte de inspiração para a criação de arquiteturas de aprendizagem profunda. Modelos iniciais no campo de aprendizagem profunda chamaram de neurônio a unidade básica de cálculo que associa uma matriz de pesos $W = [w_1, w_2, ..., w_n]$ a dados de entrada $X = [x_1, x_2, ..., x_n]$, cuja saída era calculada como $f(X,W) = WX^T$. Essas funções lineares apresentavam certas limitações como, por exemplo, não conseguirem capturar o comportamento do operador lógico OU Exclusivo (XOR na sigla em inglês). Para simular comportamentos mais complexos, funções não lineares passaram a ser combinadas com a função linear inicialmente proposta. Um exemplo de não linearidade é aplicar a função de regressão logística à $f(X,W)$, obtendo $f'(X,W) = \sigma(WX^T)$ (Goodfellow et al., 2016, pp. 13-14, 165–172).

Além dos avanços teóricos relacionados à aprendizagem profunda, é importante mencionar o desenvolvimento de sistemas computacionais que possibilitaram o treinamento de redes neuronais com cada vez mais parâmetros e com maiores quantidades de dados de entrada. No caso, pode-se mencionar a melhoria nas *Central Processing Units* (CPUs, em inglês ou unidades centrais de processamento em tradução livre), o aumento de capacidade de armazenamento dos computadores e o advento de *Graphical Processing Units* (GPUs em inglês ou unidade de processamento gráfico em tradução livre) com capacidades de processamento paralelo cada vez maiores (Goodfellow et al., 2016, pp. 438-441).

#### 2.4.1.1 Camadas convolucionais e de *pooling*

O aperfeiçoamento tecnológico disseminou o uso de redes capazes de realizar aprendizagem profunda, o que motivou a criação de diversas arquiteturas de camadas (as colunas de nós na Figura 11 que calculam operações não lineares) nessas redes, cada uma criada para um tipo de problema. Para tarefas cujas entradas são imagens, camadas convolucionais e de *pooling* (em inglês) como proposto por Lecun et al. (1998) são amplamente utilizadas (Goodfellow et al., 2016, p. 326).

As operações de convolução nessas camadas utilizam uma matriz de parâmetros $W$, também chamada de filtro, que é compartilhada com todos os pixels da imagem como ilustrado na Figura 12 (a). Esse compartilhamento reduz o número de parâmetros da rede e confere ao sistema invariância translacional, uma vez que o padrão capturado pelo filtro passa por toda a imagem (Lecun et al., 1998). Para conservar o tamanho da matriz após a aplicação da camada convolucional, como feito por Lecun et al. (1998), pode-se utilizar uma matriz expandida com zero nas adjacências externas da matriz original como ilustrado na Figura 12 (b).

A operação de *pooling*, por sua vez, consiste em utilizar uma janela, com ou sem parâmetros treináveis, para realizar operações de agregação como o valor máximo ou a média dos valores da janela como exemplificado na Figura 13. Geralmente é colocada após a saída não linear de camadas convolucionais e pode realizar subamostragem, como feito por Lecun et al. (1998).

Outro conceito importante relacionado a redes neuronais convolucionais (aquelas que utilizam camadas convolucionais) é o de rede totalmente convolucional (em tradução livre ou *Fully Convolutional Network* - FCN - em inglês). Essa consiste em utilizar operações de convolução ou *pooling* na rede toda, sendo capaz de gerar saídas com duas ou mais dimensões que podem ser usadas para tarefas como segmentação de imagens (Long et al., 2015).

#### 2.4.1.2 Redes recorrentes

Redes recorrentes são empregadas para trabalhar com séries temporais e dados que precisem armazenar ou processar sequências de estados. A estrutura geral segue a ideia de compartilhamento de parâmetros presente nas redes convolucionais (Lecun et al., 1998), sendo as variáveis treináveis relacionadas ao estado da rede compartilhadas entre iterações consecutivas da rede. Esse compartilhamento permite à rede se ajustar e processar informações que são apresentadas em pontos diferentes de uma sequência. Note que o cálculo das variáveis de estado e de saída da rede pode ser feito com qualquer tipo de camada parametrizável (Goodfellow et al., 2016, pp. 367-368). A Figura 14 ilustra o esquema geral de uma rede recorrente e exemplifica sua operação durante duas iterações.

(a)

Entrada

| | | | |
|---|---|---|---|
| X1 | X2 | X3 | X4 |
| X5 | X6 | X7 | X8 |
| X9 | X10 | X11 | X12 |
| X13 | X14 | X15 | X16 |

Pesos

| | | |
|---|---|---|
| W1 | W2 | W3 |
| W4 | W5 | W6 |
| W7 | W8 | W9 |

Saída

| | |
|---|---|
| Y1 | Y2 |
| Y3 | Y4 |

(b)

Entrada

| | | | |
|---|---|---|---|
| X1 | X2 | X3 | X4 |
| X5 | X6 | X7 | X8 |
| X9 | X10 | X11 | X12 |
| X13 | X14 | X15 | X16 |

Entrada expandida

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | X1 | X2 | X3 | X4 | 0 |
| 0 | X5 | X6 | X7 | X8 | 0 |
| 0 | X9 | X10 | X11 | X12 | 0 |
| 0 | X13 | X14 | X15 | X16 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 |

Figura 12 – a) Matrizes envolvidas em uma operação de convolução entre uma entrada e os pesos convolucionais, procedimento que pode ser entendido como alinhar a matriz de pesos W de forma que o peso W5 multiplique as entradas X6, X7, X10 e X11 e o resultado das multiplicações de matrizes sejam as respectivas saídas Y1, Y2, Y3 e Y4. b) Para que a matriz de saída tenha o mesmo tamanho da entrada, é necessário adicionar à entrada uma borda artificial que geralmente é preenchida com zeros.

#### 2.4.1.3 Arquitetura codificador-decodificador

A arquitetura codificador-decodificador baseia-se em mapear uma entrada para um espaço de atributos capaz de identificar as principais informações ou características dos dados através de um codificador e reconstruir esse 'código' para um espaço com mesma dimensão da entrada ou até mesmo para reconstruir a entrada (Badrinarayanan et al., 2017 e Ranzato et al., 2007). Para evitar que a estrutura codificador-decodificador aprenda a função de identidade, pode-se definir restrições para o processo de codificação (Goodfellow et al., 2016, p. 499). Uma dessas, ilustrada na Figura 15, é reduzir o tamanho do espaço de atributos gerado pelo codificador e utilizar o decodificador como um interpolador para retornar à dimensão original.

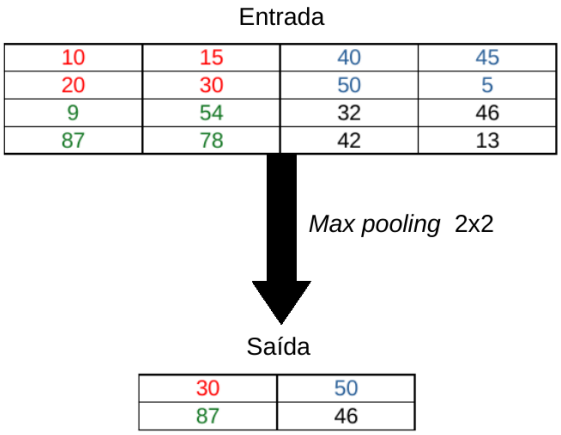



Figura 13 – Na operação de *max pooling* $NxM$, observa-se o maior número dentro de uma janela deslizante (células coloridas da entrada), o qual se torna a saída (células coloridas). Operações como *average pooling* seguem o mesmo princípio, diferenciando-se somente na função de aglutinação (méida no caso de*average pooling*).

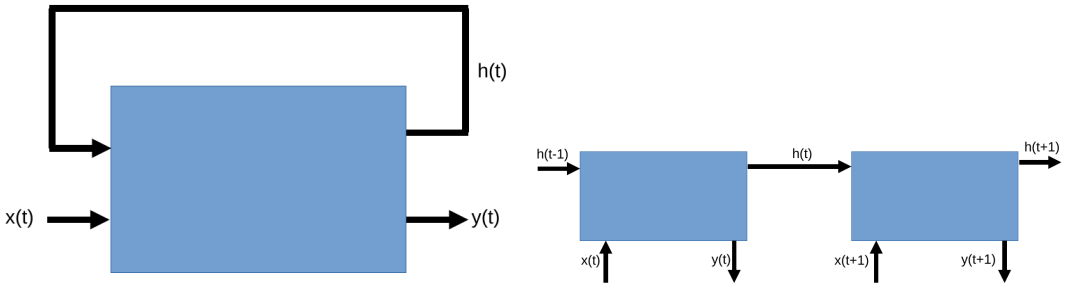



Figura 14 – À esquerda: esquemático geral de uma rede recorrente, na qual uma saída de estado h(t) é reto-alimentada na rede; à direita: transferência da variável de estado h(t) durante duas iterações consecutivas.

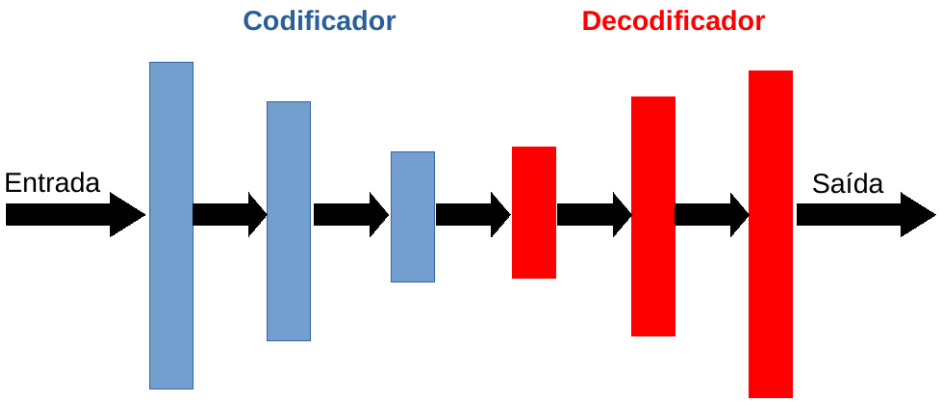



Figura 15 – Estrutura geral da arquitetura codificador-decodificar. Note que a resolução espacial é reduzida no codificador, forçando o modelo a obter uma representação compacta dos dados, expandida no decodificador para retornar ao tamanho original.

#### 2.4.1.4 *Transformer*

Vaswani et al. (2017) definiram uma nova estrutura de codificador-decodificador, denominada *Transformer*, para realizar tradução de textos sem empregar convoluções ou redes recorrentes. Para isso, Vaswani et al. (2017) utilizaram um módulo de atenção no qual três vetores controlam a importância de cada palavra dado uma sequência (contexto) textual: *query* $Q$, *key* $K$ e *value* $V$ (em inglês, ou busca, chave e valor em tradução livre). A saída de cada módulo de atenção pode ser entendida como $Softmax(QK^T)V$. O módulo de atenção do codificador recebe como parâmetros $Q = XW^Q$, $K = XW^K$ e $V = XW^V$, funcionando com auto-atenção e podendo controlar quais palavras da sequência de entrada focar em cada momento. Já o decodificador recebe *value* e *key* do codificador e utiliza como *query* a sequência de palavras já traduzidas. Vaswani et al. (2017) adicionaram à sequência de palavras de entrada e saída (traduções) uma codificação espacial para informar à rede a posição de cada palavra na frase.

#### 2.4.1.5 Arquitetura *DeepLab*

A primeira versão do DeepLab (Chen et al., 2015) introduziu o conceito de convolução *atrous* (do francês *à trous* ou "com buracos", em tradução livre) para substituir as últimas camadas VGG-16 (Simonyan e Zisserman, 2015) e aumentar a resolução espacial nessas camadas, quando comparadas à resolução original da rede VGG-16, para a tarefa de segmentação semântica. A convolução *atrous*, ilustrada na Figura 16, consiste em espaçar cada peso dos filtros convolucionais tradicionais, aumentando assim o campo de visão da rede sem aumentar o número de parâmetros. Esse processo de captar informações espacialmente mais distantes pode substituir o efeito causado pelo uso de *pooling* sem reduzir a resolução da matriz após a convolução. Como essa rede do DeepLab causava uma redução espacial, a imagem era interpolada bilineamente e a técnica de *Conditional Random Fields* (Krahenbuhl e Koltun, 2011), CRFs na sigla em inglês, refinava o resultado de segmentação.

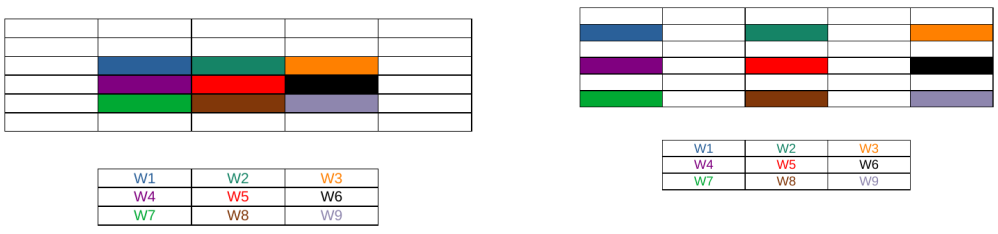



Figura 16 – À esquerda: convolução tradicional, onde a multiplicação da matriz de pesos com a entrada está marcada com cores distintas. À direita: convolução *atrous*, onde há uma dilatação dos pesos e se nota que, com a mesma quantidade de parâmetros, obtém-se um campo de visão maior que a convolução tradicional.

Com o advento de DeepLabV2 (Chen et al., 2017), as redes VGG-16 (Simonyan e

Zisserman, 2015) e ResNet (He et al., 2015) foram testadas como possíveis modelos de base. Chen et al. (2017) também introduziram o conceito de *Atrous Spatial Pyramid Pooling* (ASPP em inglês) que consiste em utilizar convoluções *atrous* com múltiplas dilatações para capturar informações de uma saída convolucional em paralelo. As diferentes informações obtidas por cada convolução *atrous* do ASPP foram unidas em uma saída final e refinadas por CRFs (Krahenbuhl e Koltun, 2011).

Em sua terceira iteração, a versão DeepLabV3 (Chen et al., 2019) empregou a arquitetura ResNet (He et al., 2015) como base, utilizou diferentes dilatações nas convoluções *atrous* que não pertencem ao módulo ASPP e juntou informações globais, obtidas por meio de *global average pooling* (em inglês) da última saída de convolução *atrous*, com os mapas do ASPP. Esses dois fluxos (global e ASPP) foram concatenados e uma convolução 1x1 foi responsável por gerar a segmentação final. Como a interpolação utilizada pelo DeepLabV3 retornou a saída para a resolução original da imagem de entrada, o módulo de CRFs (Krahenbuhl e Koltun, 2011) não precisou ser utilizado.

Para refinar os detalhes de segmentação durante o processo de interpolação de imagens feito pelo modelo DeepLabV3 (Chen et al., 2019), Chen et al. (2018) utilizaram a saída da rede DeepLabV3 como um codificador e adicionaram duas operações de convolução em um módulo de decodificador para realizar a interpolação para o tamanho da imagem de entrada com refinamento de bordas, criando assim a arquitetura DeepLabV3+.

#### 2.4.1.6 Operador gradiente

O operador de gradiente $\nabla$ é um conceito matemático amplamente utilizado em problemas de aprendizagem profunda. Ele consiste em calcular a derivada, ou taxa de variação instantânea, de uma função $f$ com relação a seus parâmetros $X = \{x_1, x_2, ..., x_n\}$ de maneira que $\nabla f(X) =< \frac{df(X)}{dx_1}, \frac{df(X)}{dx_2}, ..., \frac{df(X)}{dx_n} >$. As leituras de Thomas et al. (2012a, pp. 117-121) e Thomas et al. (2012b, pp. 245-248) são sugeridas, respectivamente, para definições matemáticas formais de derivada e gradiente.

Sua principal utilidade para o problema de aprendizagem profunda é o fato de que ele indica a direção de maior crescimento de uma função, ou seja, qual direção deve ser tomada no espaço definido por $X$ para aumentar o valor de $f(X)$. O negativo de $\nabla f(X)$ informa ao usuário qual é a taxa de menor crescimento (maior decréscimo) de $f(X)$. Caso $f(X)$ represente o erro cometido por uma rede neuronal, o operador gradiente fornece indicativos de como reduzir o erro cometido (Goodfellow et al., 2016, pp. 80-81).

#### 2.4.1.7 Atualização dos parâmetros da rede

Os parâmetros ou pesos das redes de aprendizagem profunda são comumente ajustados por meio do negativo do gradiente seguindo a expressão

$$W_{novo} = W_{atual} - \lambda \nabla E(W) \tag{2.1}$$

, onde a matriz de pesos $W$ é atualizada com base no gradiente de uma função de erro, também chamada de função de custo, com relação aos parâmetros, sendo $\lambda$, também chamada de taxa de aprendizagem, uma constante que controla o passo dado a cada atualização (Goodfellow et al., 2016, pp. 82-84).

A Equação 2.1 pode ser calculada utilizando-se o conjunto completo de dados de entrada, caso em que o algoritmo de otimização dos parâmetros é chamado de *gradient descent*, em inglês, ou por meio de amostras aleatórias desse conjunto, cujo algoritmo é chamado de *stochastic gradient descent* (Goodfellow et al., 2016, pp. 80-84, 149–150) ou SGD na sigla em inglês. Outro conceito importante relacionado aos algoritmos de otimização é o de época, que corresponde ao ajuste de parâmetros obtido após o modelo ser treinado com todos os dados presentes no conjunto de treinamento.

O SGD consiste, como mencionado anteriormente, em calcular a média dos gradientes em uma porção amostrada aleatoriamente do conjunto de dados (também chamada de *minibatch* em inglês). As principais vantagens do SGD são não aumentar o custo computacional de cálculo do gradiente ao aumentar a base de dados, já que o tamanho do *minibatch* é menor que a coleção completa, e poder convergir antes mesmo de passar por todos os dados se a base for grande o suficiente. Outro conceito relacionado a essa otimização é o conceito de *momentum*, o qual consiste em somar ao gradiente atual uma média móvel com decaimento exponencial dos gradiente anteriores, processo que pode acelerar a convergência em situações de gradientes pequenos ou ruidosos (Goodfellow et al., 2016, pp. 290-296).

Outro algoritmo de otimização comumente usado é o Adam, proposto por Kingma e Ba (2017). Esse utiliza momentos de primeira (estimação da média) e segunda ordem (estimação da variância), os quais são exponencialmente ponderados no processo de cálculo da média móvel. Uma característica importante sobre esse algoritmo é que ele ajusta o enviesamento inserido ao considerar os momentos iniciais como 0. Além disso, como argumentado por Kingma e Ba (2017), a atualização dos parâmetros é limitada a uma região próxima dos valores atuais, o que previne mudanças bruscas de parâmetros, processo que pode ocorrer em abordagens como o SGD caso o valor do gradiente seja elevado.

#### 2.4.1.8 Técnicas para treinamento de redes profundas

Treinar redes neuronais profundas é desafiador devido à complexidade provocada pela cascata de operações não lineares, uma vez que elas podem gerar gradientes extremos.

Técnicas como *batch normalization*, em inglês, e conexões residuais são exemplos que buscam auxiliar o treinamento dessas redes.

Visando mitigar o efeito deletério da mudança de distribuição de saída das camadas de redes neuronais, a qual pode levar à saturação de funções como a regressão logística e resultar em uma convergência lenta, Ioffe e Szegedy (2015) criaram uma nova camada chamada de *batch normalization* (BN na sigla em inglês). Ela normaliza as ativações de uma camada da rede de modo que as variáveis tenham média zero e desvio padrão igual a um. Desse modo, as demais camadas da rede podem operar sem saturação e o sistema como um todo ser menos sensível à escolha da taxa de aprendizagem $\lambda$. Os valores de média e desvio padrão podem ser calculados em cada *minibatch* durante a fase de treinamento e podem ser substituídos pela média e desvio padrão de todo o conjunto de dados durante inferência.

A camada de BN pode ser parametrizada com uma função afim que fornece à rede a capacidade de reter a expressividade da entrada não normalizada - Ioffe e Szegedy (2015) mencionam o exemplo de sair da zona linear de certas funções - e tornar o aprendizado da média mais fácil, uma vez que ela se torna dependente apenas dos parâmetros da camada de BN e não das camadas anteriores a ela como explicado por Goodfellow et al. (2016, pp. 313-317).

He et al. (2015) exploram o problema de degradação do aprendizado de redes profundas quando mais camadas são adicionadas. A solução, proposta por eles, baseia-se em utilizar conexões residuais que levam entradas de camadas inciais da rede para aquelas mais profundas. He et al. (2015) argumentam que, caso a função a ser modelada pela rede seja linear, as conexões com função de identidade podem aproximar melhor esse comportamento quando comparadas a composições de funções não lineares. Em casos nos quais a aproximação é não linear, as conexões de identidade podem fornecer um condicionamento à rede, auxiliando a modelar partes aproximadamente lineares. A Figura 17 ilustra a arquitetura de conexões residuais.

Com o conceito de conexões residuais, He et al. (2015) foram capazes de utilizar redes mais profundas, chamadas de ResNets, do que trabalhos anteriores e de atingir o estado-da-arte em tarefas como classificação de imagens.

#### 2.4.1.9 Modelos supervisionados

A aprendizagem supervisionada é uma das formas de treinamento de modelos de aprendizagem de máquinas na qual as respostas desejadas são fornecidas por pessoas. Essas podem ser, por exemplo, classes de objetos em uma imagem (pássaro, avião, cachorro, dentre outras) ou imagens bidimensionais com rótulos identificando regiões de interesse. O objetivo em utilizar informações manualmente produzidas é replicar o comportamento humano em uma determinada tarefa durante a fase de inferência, a qual é feita de forma automática sem a presença dos referidos rótulos (Goodfellow et al., 2016, pp. 102-103, 137–138).

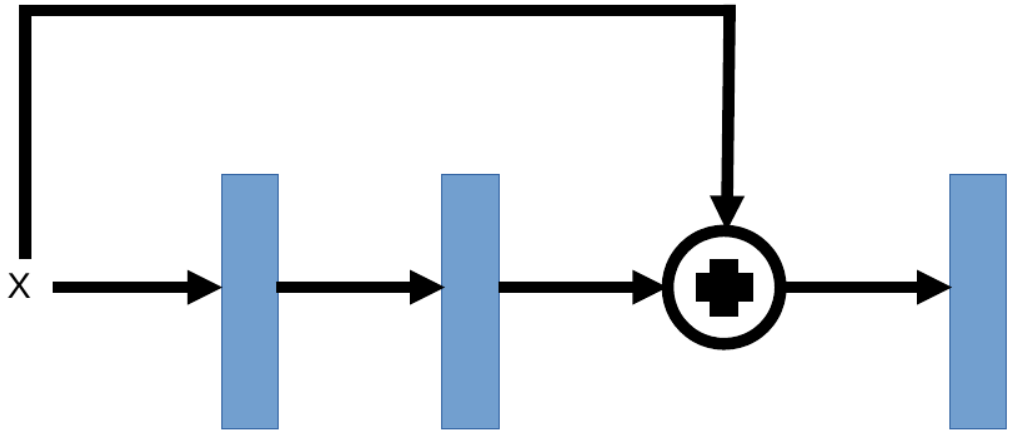



Figura 17 – As conexões residuais adicionam uma entrada anterior ao processamento das camadas de redes neuronais (retângulos) à saída dessas, facilitando estimações lineares em redes profundas.

No campo de detecção de objetos salientes, diversos autores propuseram mecanismos de aprendizagem profunda para estimar a segmentação de objetos salientes a partir de rótulos previamente estabelecidas por um conjunto de pessoas. Como exemplos, Ke e Tsubono (2022) e Luo et al. (2017) exploraram, respectivamente, as hipóteses de que adicionar os contornos dos rótulos de referência à função de custo de reconstrução desses e fundir atributos de camadas profundas, responsáveis por capturar aspectos globais da imagem, com atributos de camadas inciais, responsáveis por capturar informações locais, podem resultar em uma performance comparável ao estado-da-arte. Já Wang et al. (2022) buscaram reduzir a carga de trabalho de rotulagem, criando um modelo profundo capaz de estimar a segmentação de objetos salientes exclusivamente a partir da informação de que a imagem de entrada continha ou não pelo menos um objeto saliente.

A segmentação de imagens médicas por meio de aprendizagem profunda supervisionada é abordada por trabalhos como os de Maity et al. (2022), Oh et al. (2020), Nishio et al. (2021), Singh et al. (2021) e Chen et al. (2021a), os quais exploram arquiteturas codificador-decodificador, frequentemente empregadas nessa área. Utilizando a arquitetura DeepLabV3+ (Chen et al., 2018), Singh et al. (2021) realizaram segmentação pulmonar a partir de imagens de raio-X sem pré-processamento, porém aplicando pós-processamento no resultado de segmentação para reduzir falsos positivos. Com imagens de raios-X pré-processadas com equalização de histograma, Nishio et al. (2021) utilizaram a rede U-Net (Ronneberger et al., 2015) para segmentação pulmonar. Visando diagnóstico automático de COVID-19 a partir de raios-X, Oh et al. (2020) segmentaram a região pulmonar com a rede FC-DenseNet (Jegou et al., 2017) que recebia, como entrada, raios-X pré-processados com equalização de histograma e correção gamma ($\gamma = 0.5$). Maity et al. (2022) processaram imagens de raio-X com as transformações *Top-Bottom Hat* (em inglês) seguida de *Contrast Limited Adaptive*

*Histogram Equalization* (CLAHE na sigla em inglês) e as utilizaram como entrada de rede inspirada na U-Net++ (Zhou et al., 2018) para segmentar a área pulmonar. Chen et al. (2021a) combinaram a arquitetura U-net (Ronneberger et al., 2015) e *Transformer* (Vaswani et al., 2017) para segmentação de órgãos abdominais a partir de tomografia computadorizada e também realizando experimentos em segmentação cardíaca em imagens de ressonância magnética.

Os resultados obtidos por modelos profundos supervisionados os colocam como o estado-da-arte em termos de performance. Esse sucesso, contudo, está atrelado a grandes bases de dados que requerem um grande esforço humano para rotulação, especialmente em tarefas como segmentação de imagens, cujos rótulos costumam ser definidos pixel a pixel. Questiona-se sobre a possibilidade de realizar as mesmas tarefas de forma não supervisionada (expressão usada no sentido de não haver supervisão humana para que o treinamento ocorra). Dessa forma, pode-se treinar modelos com uma quantidade ainda maior de dados ou em situações novas para as quais dados manualmente fornecidos ainda não estão disponíveis.

#### 2.4.1.10 Modelos não supervisionados

Modelos não supervisionados podem ser definidos como aqueles que não possuem rótulos pré-definidos ou cujos rótulos de treinamento foram fornecidos por mecanismos automáticos, como outros modelos computacionais. Tarefas como agrupamento e auto-codificação são comumente relacionadas à aprendizagem não supervisionada (Goodfellow et al., 2016, pp. 102-103, 142–144).

No campo de saliência visual, trabalhos como os de Zhang et al. (2018), Croitoru et al. (2019) e Zhang et al. (2021) combinaram redes neuronais com modelos heurísticos para separar a saliência do ruído de cada rótulo. Zhang et al. (2018) utilizaram a rede DeepLabV2 (Chen et al., 2017) para estimar saliência visual a partir de múltiplos rótulos ruidosos. Zhang et al. (2018) argumentam que, como redes profundas podem ser propensas a aprender ruídos, um módulo de ruído teve que ser criado para guiar o modelo a aprender somente aspectos relacionados à saliência e não ao ruído. Zhang et al. (2021) partiram da hipótese de que redes neuronais primeiro se ajustam à função não ruidosa e depois aprendem ruído para definirem uma maneira alternativa para treinar a rede DeepLabV2 (Chen et al., 2017) utilizando apenas um rotulador ruidoso. Croitoru et al. (2019) empregaram várias redes profundas distintas em um procedimento professor-aluno no qual redes aprendiam a segmentar e ensinavam a próxima geração que rotulava com maior performance que a anterior.

Objetivando realizar segmentação pulmonar em imagens de raio-X com base em saliência visual, o modelo proposto por Zhang et al. (2018) foi escolhido como base de trabalho. Comparado ao descrito por Croitoru et al. (2019), a abordagem de Zhang et al. (2018) utiliza somente uma rede neuronal e não precisa armazenar múltiplas iterações. Quando comparado a Zhang et al. (2021), o emprego de múltiplos rotuladores ruidosos

sugerido por Zhang et al. (2018) aumenta o escopo de heurísticas que aproximam a saliência visual humana, o que pode auxiliar a rede a realizar a tarefa de segmentação. Vale lembrar que, diferente das abordagens tradicionais de saliência visual, não há informações de cores disponíveis em raios-X, o que reforça a hipótese de que o mecanismo de atenção necessário nesse cenário pode envolver múltiplas heurísticas.

Para a área de imagens médicas, o trabalho de Kuang et al. (2020) explorou a consistência temporal de imagens de ressonância magnética da coluna vertebral para treinar uma rede convolucional a partir de vértebras definidas por um algoritmo baseado em regras de decisão. As diversas imagens, que, ao longo do tempo, de forma repetida, foram automaticamente denotadas como vértebras e constituíram os rótulos para treinamento da rede. Kim e Ye (2020) propuseram uma função de custo de agrupamento baseado no funcional Mumford-Shah (Mumford e Shah, 1989) e utilizaram uma U-Net (Ronneberger et al., 2015) para aprendizagem semi-supervisionada a fim de detectar tumores no fígado e no cérebro. Kim e Ye (2020) também se valeram da função de custo proposta para segmentação semântica não supervisionada de imagens naturais. Chen et al. (2021b) focaram em usar uma arquitetura de rede convolucional recorrente (Liang e Hu, 2015) para segmentação de lesões ósseas a partir da função de custo de agrupamento baseada em *Robust Fuzzy C-Means* (Pham, 2001), RFCM na sigla em inglês, e compararam essa função de custo com aquela definida por Kim e Ye (2020).

A vantagem da aprendizagem não supervisionada se encontra na rapidez em que os modelos podem ser treinados, uma vez que não há necessidade de esperar por bases rotuladas. Assim, eles podem se adequar de forma mais dinâmica a diversas situações a custo de um desempenho, em geral, menor do que modelos supervisionados (Croitoru et al., 2019).

# 3 Materiais e Métodos

As bases de dados, modelos utilizados e métricas para avaliação de desempenho serão detalhadas nas seções deste capítulo. As bases de dados de raios-X utilizadas para treinamento e teste são descritas na Seção 3.1. Na Seção 3.2, encontra-se as etapas de pré-processamento da imagem de entrada que foram utilizadas nos experimentos. As porções dos dados definidas para treinamento e para teste são definidas na Seção 3.3. O modelo baseado em saliência visual proposto é definido na Seção 3.4. Na Seção 3.5, um modelo não supervisionado proposto por Chen et al. (2021b), criado para segmentação de imagens médicas, é descrito como abordagem para ser comparada àquela baseada em saliência. Os experimentos que analisaram diferentes hiperparâmetros, bem como aqueles fixos são expostos na Seção 3.6. As métricas de avaliação de desempenho são definidas na Seção 3.7. A linguagem de programação, as bibliotecas de código para operar modelos de aprendizagem profunda e os endereços eletrônicos referentes aos repositórios dos códigos referentes aos modelos utilizados se encontram na Seção 3.8. A Seção 3.9 descreve o *hardware* utilizado para pesquisa.

É importante ressaltar que as comparações descritas neste capítulo, cujos resultados estão no Capítulo 4, não é completa, pois o foco delas é apresentar uma alternativa aos métodos supervisionados sem necessariamente superá-los em desempenho.

## 3.1 Bases de dados

Três conjuntos de imagens de raio-X pulmonar foram utilizados para treinamento e teste do modelo de saliência proposto, bem como dos comparativos que utilizam os níveis de intensidade da imagem:

- Banco de dados de nódulos pulmonares fornecidos pela *Japanese Society of Radiological Technology* (JSRT) e descrito por Shiraishi et al. (2000). Contém duzentas e quarenta e sete radiografias posterior-anterior, das quais cento e cinquenta e quatro possuem um nódulo pulmonar e noventa e três não tem nódulo pulmonar. Os dados foram digitalizados para 12 níveis de escala de cinza com tamanho 2048x2048 pixels e cada pixel corresponde a 0,175mm. As imagens foram empregadas nas fases de treino e teste como detalhado na Seção 3.3. As imagens foram baixadas de um repositório Kaggle [1].

- Os dados advindos da *Segmentation in Chest Radiographs* (SCR) e descritos por Ginneken et al. (2006) apresentam a segmentação esperada dos pulmões relativos às imagens

[1] https://www.kaggle.com/datasets/raddar/nodules-in-chest-xrays-jsrt - Último acesso em 9 de setembro de 2022

presentes na base JSRT. É importante ressaltar que a base SCR foi usada somente na fase de teste.

- A base *Montgomery County chest X-ray* (MC) descrita por Jaeger et al. (2014) consiste em cento e trinta e oito imagens torácicas, sendo oitenta normais e cinquenta e oito com sinais de tuberculose. Dados quantizados em escala de cinza de 12 bits com tamanhos de 4020x4892 ou 4892x4020 pixels. As imagens e suas respectivas segmentações foram utilizadas para testar a consistência do modelo com relação a diferentes bases.

## 3.2 Pré-processamento

A Figura 18 mostra o fluxo geral de processamento de imagens que servirão como entrada para um modelo de aprendizagem profunda não supervisionada para realizar segmentação da região pulmonar. Os processamentos adicionais estão descritos na Seção 3.6.

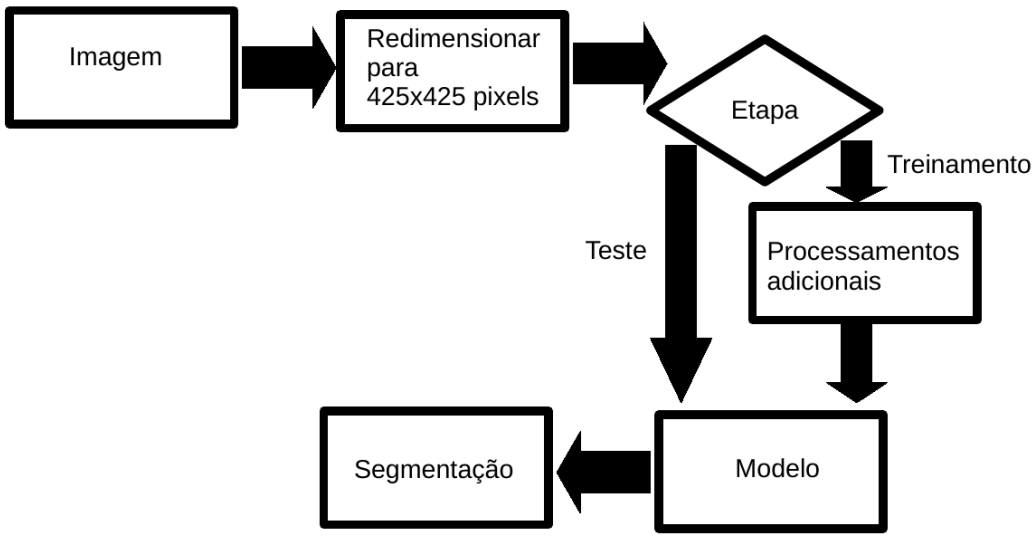



Figura 18 – Fluxo geral de processamento de dados.

As imagens de entrada foram convertidas de escala de cinzas para RGB, reduzidas para 425x425, utilizando-se a interpolação bi-cúbica, e o intervalo [0, 255] de imagens RGB foi normalizado para [0, 1].

As máscaras utilizadas como referência para avaliação do modelo foram também redimensionadas para 425x425, porém a interpolação do vizinho mais próximo foi utilizada para esse procedimento.

Após a etapa inicial, um pré-processamento adicional de normalização de histograma e correção gamma ($\gamma = 0{,}5$), descrito em Oh et al. (2020), foi testado paralelamente a usar as imagens sem essa etapa adicional. Verificou-se desse modo a possibilidade de melhorias no desempenho do modelo com os realces promovidos. Além dessas duas operações, remapearam-se pixels com intensidade 0 (preto) para 1 (branco), a fim de evitar que fitas ou

demais objetos escuros presentes na digitalização pudessem ser confundidos com a região pulmonar. A Figura 19 ilustra raios-X provenientes das bases JSRT (Shiraishi et al., 2000) e MC (Jaeger et al., 2014) antes e após o pré-processamento.

(a) Entrada sem pré-processamento

(b) Entrada pré-processada

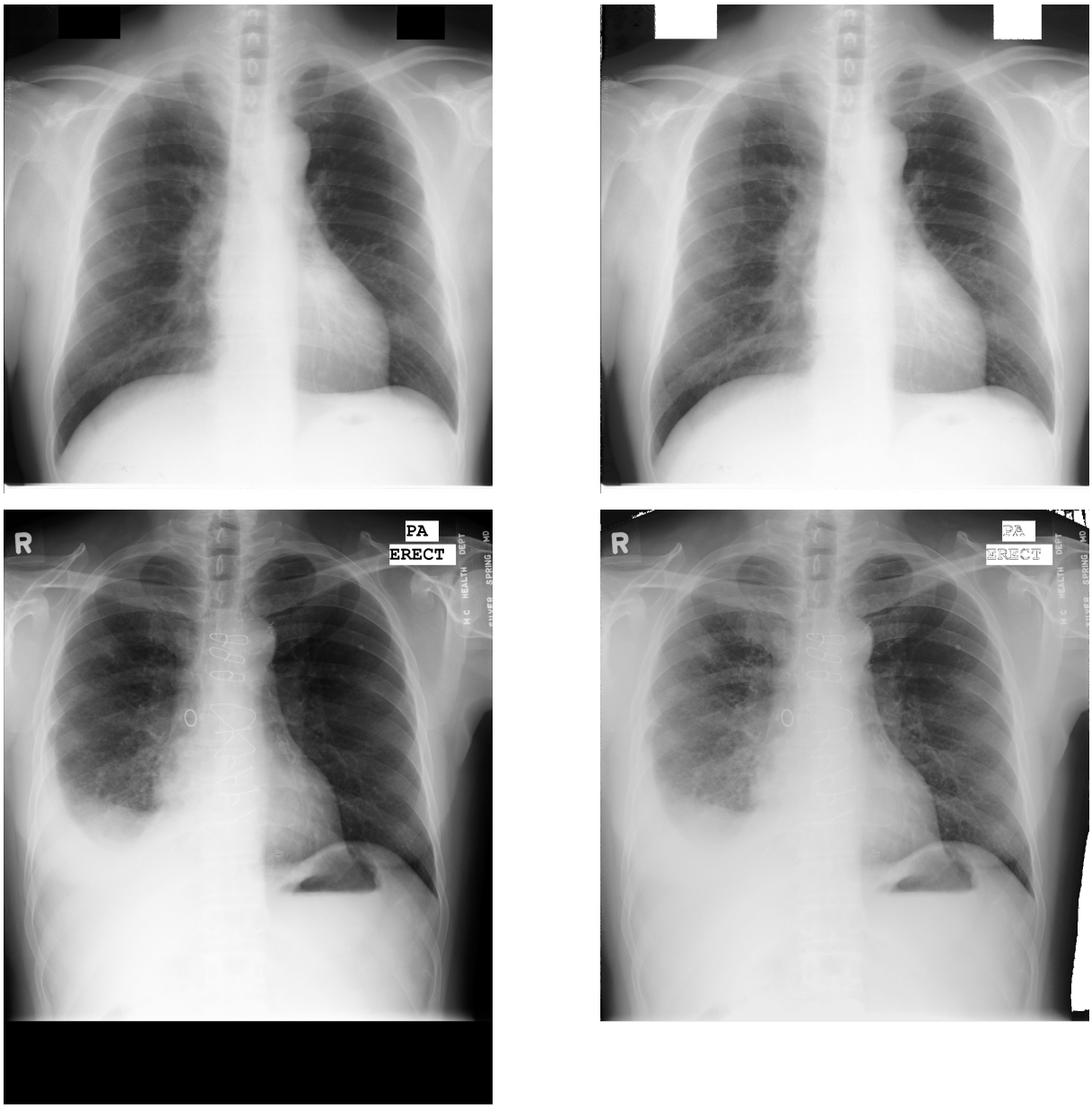



Figura 19 – Efeitos do pré-processamento utilizado sobre as imagens de entrada das bases JSRT (primeira linha) e MC (segunda linha).

## 3.3 Divisão entre treinamento e teste

Como o foco é segmentação pulmonar de raios-X, imagens saudáveis e não saudáveis de ambas as bases JSRT (Shiraishi et al., 2000) e MC (Jaeger et al., 2014) foram usadas sem distinção entre si. Com relação à JSRT (Shiraishi et al., 2000), 201 imagens foram aleatoriamente escolhidas para treinamento e as 46 restantes para teste. Quanto à base MC

(Jaeger et al., 2014), 42 imagens foram aleatoriamente escolhidas para teste. Uma vez que o objetivo é ter um parâmetro de comparação de desempenho com conjuntos de dados distintos, não se necessitou utilizar o conjunto MC (Jaeger et al., 2014) por completo.

Nenhuma porção de dados foi definida como validação para que a maior quantidade possível deles estivesse disponível para treinamento. Esse passo é importante uma vez que muitas aplicações médicas possuem pequena base de dados, dificultando assim a divisão de dados em treinamento e validação.

Técnicas de *data augmentation* (em inglês ou aumento de dados em tradução livre) não foram utilizadas, uma vez que os pesos pré-estabelecidos para o modelo de saliência se mostraram capazes de promover um treinamento adequado com as imagens de raio-X sem sobre ajuste aparente.

## 3.4 Modelo baseado em saliência

A arquitetura de rede neuronal profunda baseada em saliência visual é inspirada no trabalho de Zhang et al. (2018). Esse consiste em aprender padrões de saliência visual implícitos em modelos computacionais heurísticos utilizando a rede neuronal DeepLabV2 (Chen et al., 2017) criada para segmentação semântica. Para segmentação de imagens médicas, optou-se por utilizar a arquitetura DeepLabV3 (Chen et al., 2019), já que ela não necessita de refinamento com CRFs (Krahenbuhl e Koltun, 2011) e está implementada na ferramenta computacional Pytorch. A Figura 20 ilustra o procedimento proposto.

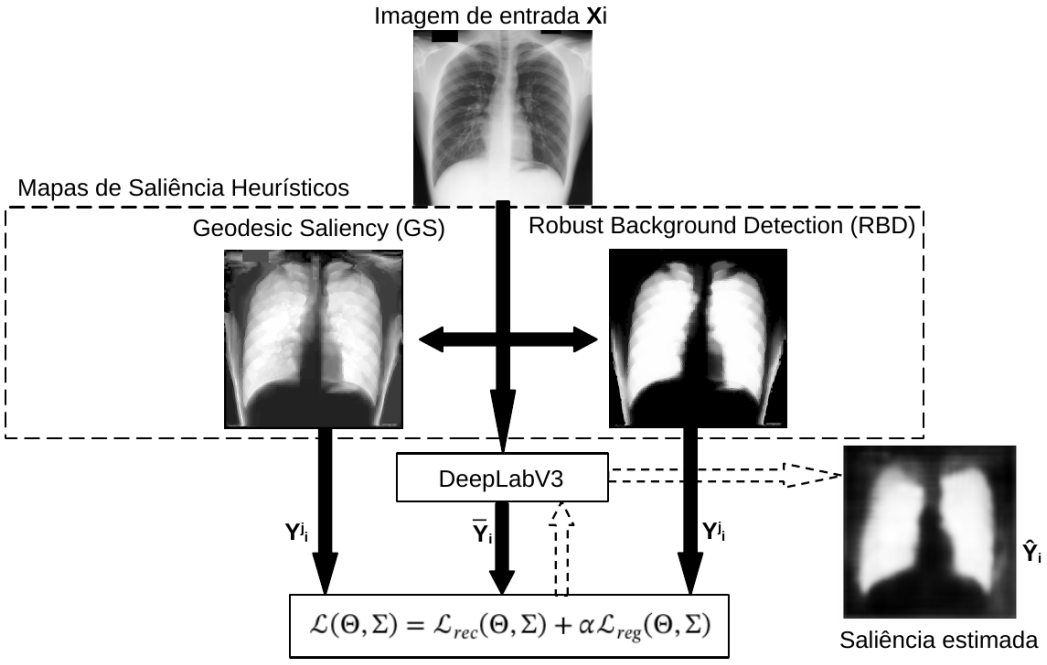



Figura 20 – Esquemático geral para a rede neuronal DeepLabV3 aprender não supervisionadamente a segmentar a região pulmonar a partir de mapas de saliência heurísticos que contêm ruído.

### 3.4.1 Rótulos ruidosos

Dois modelos heurísticos de saliência foram escolhidos para fornecer rótulos "ruidosos"que guiaram a rede neuronal: *Robust Background Detection* (Zhu et al., 2014), RBD na sigla em inglês, e *Geodesic Saliency* (Wei et al., 2012), GS na sigla em inglês. A escolha foi motivada pelas heurísticas que pressupõem que os objetos salientes podem estar fora do centro da imagem. Como os rótulos ruidosos gerados por esses métodos são baseados em superpixel, as imagens de raio-X utilizadas na fase de treino são usadas em sua resolução original (2048x2048) para cálculo desses rótulos. Isso reduz a chance de juntar regiões diferentes adjacentes ao particionar a imagem em super pixeis de tamanho constante. A Figura 21 exibe os mapas ruidosos de saliência obtidos pelos modelos GS e RBD para combinações de entrada com e sem pré-processamento.

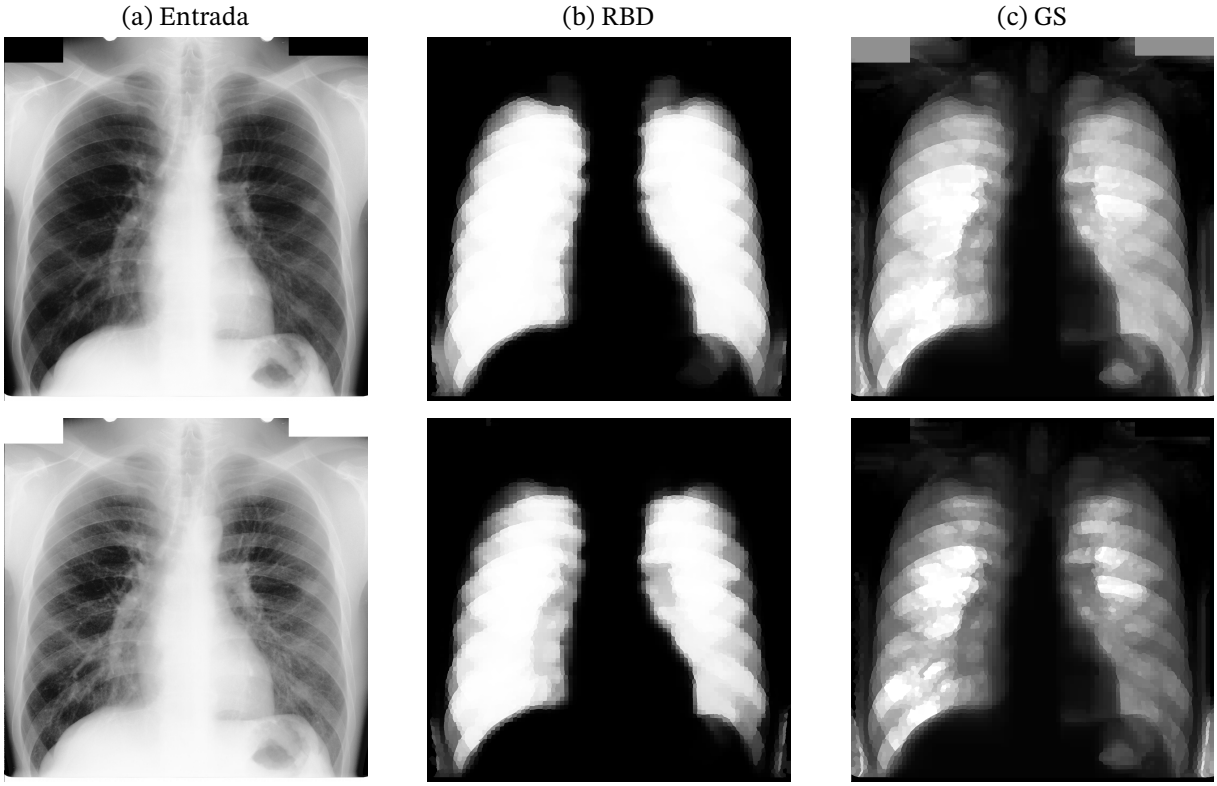



Figura 21 – Exemplos de rótulos ruidosos sem pós-processamento. Colunas (a): Imagens de entrada, sendo a primeira Raio-X sem pré-processamento e a segunda linha a versão após pré-processamento; (b) mapas de saliência obtidos pelo modelo RBD (Zhu et al., 2014); (c) mapas de saliência obtidos pelo modelo GS (Wei et al., 2012) para cada uma das respectivas imagens de entrada.

Assim como os dados de entrada, um teste de pós-processamento dos dados ruidosos foi aplicado usando somente correção gamma ($\gamma = 0{,}5$) para verificar se o aumento não linear do contrate gera benefícios de performance durante o treino. Um teste sem pós-processamento foi utilizado para comparação. A Figura 22 ilustra a operação de pós-processamento de rótulos ruidosos a partir de imagens sem e com pré-processamento.

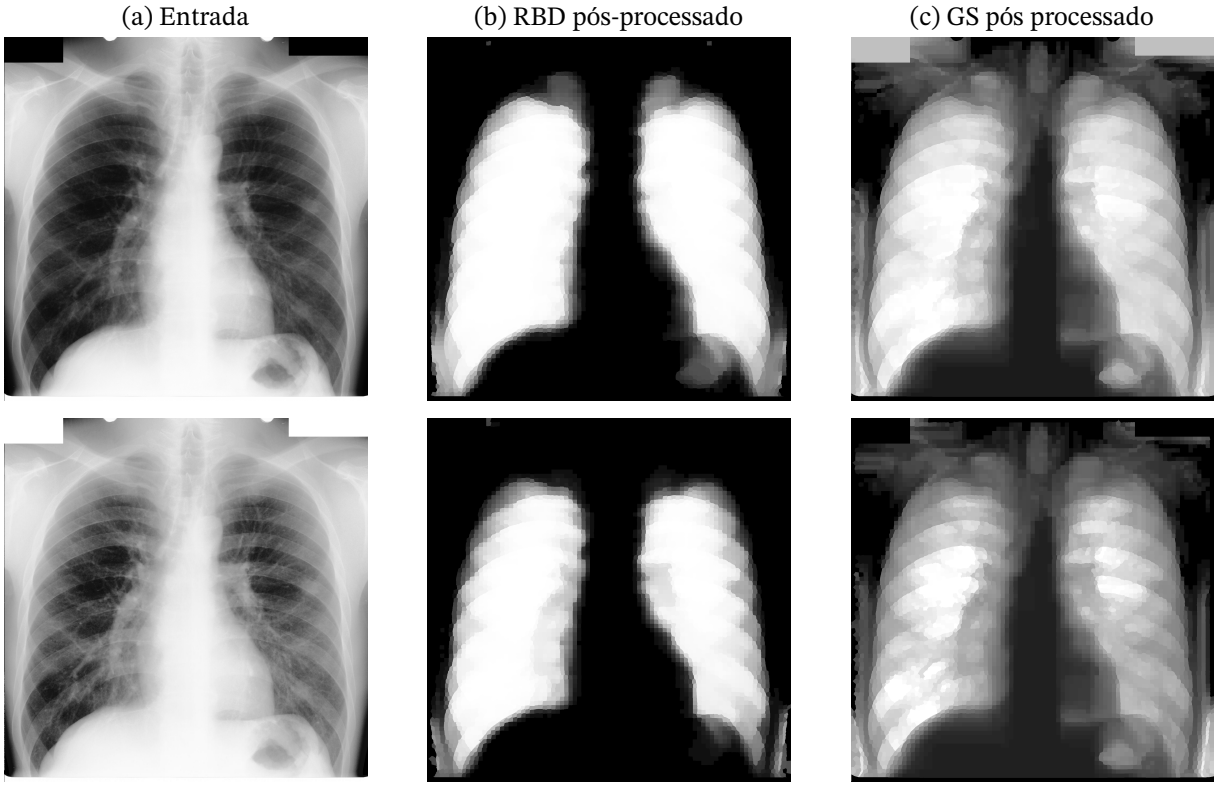



Figura 22 – Exemplos de rótulos ruidosos com pós-processamento. Colunas (a): Imagens de entrada, sendo a primeira Raio-X sem pré-processamento e a segunda linha a versão após pré-processamento; (b) mapas de saliência obtidos pelo modelo RBD (Zhu et al., 2014); (c) mapas de saliência obtidos pelo modelo GS (Wei et al., 2012) para cada uma das respectivas imagens de entrada.

Comparando as Figuras 21 e 22, percebe-se que a operação de pós-processamento realçou a região pulmonar de forma mais acentuada do que o restante do corpo (especialmente para o mapa fornecido pelo RBD). Além disso, o pré-processamento da imagem de entrada (equalização de histograma seguida por correção gamma) parece ter causado uma pequena supressão da caixa torácica nos mapas ruidosos do modelo GS e é possível observar que esses resultados contém a área pulmonar desejada acrescida de ruído que pode incluir regiões não pulmonares ou excluir detalhes da região pulmonar.

### 3.4.2 Função de custo

Dado um conjunto de $N$ imagens de entrada $X = \{x_1, x_2, ..., x_N\}$ e seus respectivos $M$ rótulos ruidosos $Y = \{y_1^1, y_1^2, ...my_N^M\}$, uma rede neuronal definida por um conjunto de parâmetros $\Theta$ pode aprender a função $f$, não ruidosa, comum a todos $M$ rótulos dada por $\bar{Y} = f(X, \Theta)$.

Como argumentado por Zhang et al. (2018), modelos profundos têm tendência a aprender ruído nos dados. Assim, é interessante modelar os mapas de saliência $Y$ como a soma do mapa 'real' não ruidoso $\bar{Y}$ acrescido de ruído $N$. Assim, cada mapa heurístico

$j$ correspondente à imagem $i$ é definido como $y_i^j = \bar{y}_i + n_i^j$ e cada saída da rede neuronal pode ser escrita como $\hat{y} = \bar{y}_i + n_i^j$. Por simplicidade, assume-se que o ruído $n_i^j$ é amostrado de um modelo de ruído definido como uma distribuição Normal independente do rótulo $j$ com média $\mu = 0$ e desvios padrões representados por $\Sigma = \sigma_{i,mn}$ para cada pixel $(m, n)$ da imagem $i$ com dimensões $HxW$ (altura x largura). Logo, a função de custo pode ser definida como

$$\mathcal{L}(\Theta, \Sigma) = \mathcal{L}_{rec}(\Theta, \Sigma) + \alpha \mathcal{L}_{reg}(\Theta, \Sigma) \tag{3.1}$$

O termo $\mathcal{L}_{rec}(\Theta, \Sigma)$ representa a reconstrução dos padrões de saliência intrínsecos aos rótulos ruidosos. É definido como a entropia cruzada entre a saída da rede neuronal e cada mapa ruidoso, ou seja,

$$\mathcal{L}_{rec} = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{m} \sum_{n} \frac{-[y_{i,mn}^j log(\hat{y}_{i,mn}^j) + (1 - y_{i,mn}^j) log(1 - \hat{y}_{i,mn}^j)]}{NMHW} \tag{3.2}$$

A porção $\mathcal{L}_{reg}(\Theta, \Sigma)$ atua como um termo de regularização, restringindo a adequação do modelo aos ruídos inerentes aos $M$ rotuladores. $\mathcal{L}_{reg}(\Theta, \Sigma)$ é definido como a divergência de Kullback-Leibler (KL) entre a distribuição de ruído assumida *a priori* $\sigma_{i,mn}$ e a distribuição *a posteriori* $\hat{\sigma}_{i,mn}$ de ruído, sendo $\hat{\sigma}_{i,mn}$ calculado como o desvio padrão entre os erros $n_i^j = y_i^j - \hat{y}_i$ dos $M$ rotuladores. Assim,

$$\mathcal{L}_{reg} = \sum_{i=1}^{N} log(\frac{\hat{\sigma}_{i,mn}}{\sigma_{i,mn}}) + \frac{\sigma_{i,mn}^2}{2\hat{\sigma}_{i,mn}^2} \tag{3.3}$$

Seguindo os procedimentos descritos em Zhang et al. (2018), o desvio padrão *a priori* é definido como 0. Diferente de Zhang et al. (2018), não há atualização do modelo de ruído, pois não há dados de validação. $\sigma_{i,mn}$ e $\hat{\sigma}_{i,mn}$ são grampeados para o menor valor representável em $(0, 1]$ ao calcular a Equação 3.3 para evitar divisões por 0.

## 3.5 Modelo baseado em níveis de intensidade

A fim de realizar uma comparação de desempenho, escolheu-se o modelo descrito em Chen et al. (2021b) por ser recente, baseado em redes neuronais profundas com aprendizagem não supervisionada e desenhado para um contexto de segmentação de imagens médicas. Chen et al. (2021b) se inspiram em métodos de agrupamento para realizar segmentação de imagens dividindo-as em lesão, osso e plano de fundo.

A rede convolucional recorrente *Recurrent Convolutional Neural Network* (RCNN na sigla em inglês) de Liang e Hu (2015) foi utilizada como modelo para o treinamento das funções baseadas em intensidade testadas por Chen et al. (2021b). Segundo Liang e Hu

(2015), as convoluções recorrentes da RCNN têm como inspiração as conexões recorrentes entre neurônios e, mesmo em entradas estáticas como imagens, permitem que as primeiras camadas adquiram informações contextuais que geralmente estão presente somente em camadas profundas das CNNs. Quando as camadas propostas por Liang e Hu (2015) são mostradas a cada iteração, percebe-se que as convoluções tradicionais são modificadas para também receberem como entradas os resultados da iteração anterior.

Diferente de Liang e Hu (2015), o modelo de Chen et al. (2021b) utilizou cinco camadas convolucionais recorrentes com três iterações cada seguidas de três camadas convolucionais tradicionais e uma camada Softmax para classificação. A *Local Response Normalization* (Krizhevsky et al., 2012), ou LRN na sigla em inglês, usada por Liang e Hu (2015) foi substituída por uma BN no modelo de Chen et al. (2021b). Em ambos os trabalhos (Chen et al., 2021b e Liang e Hu, 2015), a função de ativação utilizada foi a *Rectified Linear Unit* (ReLU na sigla em inglês).

### 3.5.1 Funções de custo

A função de custo proposta por Chen et al. (2021b) tem como alicerce a proximidade de uma classe da rede neuronal com um dos centroides combinada com um termo que suaviza variações brusca de ativação entre pixeis vizinhos. Assim, tem-se

$$\mathcal{L} = \frac{1}{NHW}\left(\sum_{i=1}^{N}\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{k=1}^{C}\bar{y}_{imnk}^{q}|x_{imn}-\nu_{ik}|^2 + \alpha\sum_{i=1}^{N}\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{k=1}^{C}\bar{y}_{imnk}^{q}\sum_{l\in N_{mn}}\sum_{m\in M_k}\bar{y}_{iml}^{q}\right) \quad (3.4)$$

onde $C$ é o número total de classes do modelo profundo, $q \in \mathcal{R}$ é uma constante que determina o grau de pertinência difusa, $N_{mn}$ é a vizinhança 3x3 do pixel corrente $(m,n)$, $M_k$ é o conjunto das classes que exclui a classe atual $k$ e $\nu_{ik}$ pode ser entendido como o centro da classe $k$ para imagem $i$ e é definido como

$$\nu_{ik} = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\bar{y}_{imnk}^{q}x_{imn}}{\sum_{m=1}^{H}\sum_{n=1}^{W}\bar{y}_{imnk}^{q}} \quad (3.5)$$

Chen et al. (2021b) compararam o trabalho deles com outro modelo não supervisionado, também baseado em funções de agrupamento (Kim e Ye, 2020), e, por isso, a função de custo utilizada por Kim e Ye (2020), definida a seguir, é similar à Equação 3.4:

$$\mathcal{L} = \frac{1}{NHW}\left(\sum_{i=1}^{N}\sum_{m=1}^{H}\sum_{n=1}^{W}\sum_{k=1}^{C}\bar{y}_{imnk}|x_{imn}-m_{ik}|^2 + \alpha\left(\left|\frac{d\bar{y}_{imnk}}{dm}\right| + \left|\frac{d\bar{y}_{imnk}}{dn}\right|\right)\right) \quad (3.6)$$

$$m_{ik} = \frac{\sum_{m=1}^{H}\sum_{n=1}^{W}\bar{y}_{imnk}x_{imn}}{\sum_{m=1}^{H}\sum_{n=1}^{W}\bar{y}_{imnk}} \quad (3.7)$$

Nota-se que, na Equação 3.6, a continuidade de ativação em uma vizinhança é feita espacialmente nas direções vertical e horizontal. Além disso, o conceito de lógica difusa não é aplicado.

## 3.6 Experimentos realizados

Os modelos descritos nas Seções 3.4 e 3.5 precisam de hiperparâmetros relacionados à etapa de treinamento, os quais são descritos na Subseção 3.6.1. A aplicação de etapas adicionais de pré-processamento de imagens de entrada (Seção 3.2), pós-processamento de rótulos ruidosos (Seção 3.4) e o número de classes da saída dos modelos baseados em níveis de saliência (Seção 3.5) são hiperparâmetros variáveis que configuraram 12 experimentos distintos, descritos na Subseção 3.6.2.

### 3.6.1 Hiperparâmetros fixos

Para o modelo baseado em saliência visual, a rede DeepLabV3 (Chen et al., 2019) foi treinada utilizando o SGD com uma taxa de aprendizado inicial $\lambda$ de $10^{-3}$ e *momentum* igual a 0.9 durante 20 épocas. A cada época, a taxa de aprendizado foi reduzida seguindo a equação $\lambda = (1 - \frac{epoca}{20})^{10}$. O parâmetro de regularização $\alpha$ da Equação 3.1 foi definido como $10^{-7}$.

Com relação aos modelos baseados em níveis de intensidade, funções de Chen et al. (2021b) e Kim e Ye (2020), o treinamento utilizou Adam (Kingma e Ba, 2017) com taxa de aprendizado $\lambda = 5 * 10^{-4}$ durante 100 épocas. Cabe ressaltar que, diferente do modelo baseado em saliência visual, $\lambda$ permaneceu constante durante todo o treinamento. O fator de regularização $\alpha$ descrito nas Equações 3.4 e 3.6 foi definido como 0.0016 e $10^{-3}$ respectivamente. O valor de $q$ nas Equações 3.4 e 3.5 foi configurado como dois.

Para todos os modelos, um tamanho de *minibatch* de cinco foi utilizado e quatro *minibatches* foram acumulados para cada atualização de parâmetros, gerando um tamanho de *minibatch* efetivo de 20. O último *minibatch* das 201 imagens da base JSRT (Shiraishi et al., 2000), utilizadas para treinamento, foi descartado por ter somente uma imagem, não alcançando com isso o tamanho mínimo de cinco imagens. A média em cada *minibatch* foi calculada para as Equações 3.2, 3.4 e 3.6, enquanto a soma dos valores do *minibatch* foi aplicada para a Equação 3.3.

### 3.6.2 Hiperparâmetros variáveis

Seguindo a lógica de configuração experimental de Nishio et al. (2021), todos os modelos foram testados para os hiperparâmetros: (1) pré-processar a imagem de entrada (PPE) com equalização de histograma seguida por correção gamma ($\gamma = 0.5$); (2) pós-

processar os rótulos de ruidosos (PRR) gerados por GS (Wei et al., 2012) e RBD (Zhu et al., 2014) utilizando correção gamma ($\gamma = 0.5$), aplicável somente ao modelo baseado em saliência; (3) utilizar duas ou três classes $C$ para os modelos fundamentados em níveis de intensidade - RFCM (Chen et al., 2021b) e MS (Kim e Ye, 2020). Assim, doze experimentos foram planejados:

- $S_1$: usar PPE e PRR com modelo de saliência;
- $S_2$: usar PPE e não usar PRR com modelo de saliência;
- $S_3$: não usar PPE nem PRR com modelo de saliência;
- $S_4$: não usar PPE e usar PRR com modelo de saliência;
- $RFCM_1$: não usar PPE e configurar $C = 2$ para o modelo RFCM baseado em níveis de intensidade;
- $RFCM_2$: usar PPE e configurar $C = 2$ para o modelo RFCM baseado em níveis de intensidade;
- $RFCM_3$: não usar PPE e configurar $C = 3$ para o modelo RFCM baseado em níveis de intensidade;
- $RFCM_4$: usar PPE e configurar $C = 3$ para o modelo RFCM baseado em níveis de intensidade;
- $MS_1$: não usar PPE e configurar $C = 2$ para o modelo MS baseado em níveis de intensidade;
- $MS_2$: usar PPE e configurar $C = 2$ para o modelo MS baseado em níveis de intensidade;
- $MS_3$: não usar PPE e configurar $C = 3$ para o modelo MS baseado em níveis de intensidade;
- $MS_4$: usar PPE e configurar $C = 3$ para o modelo MS baseado em níveis de intensidade.

Os testes envolvendo clareamento de imagens (hiperparâmetros PPE e PRR) foram feitos para avaliar se essas operações podem auxiliar os modelos profundos a encontrarem fronteiras de decisão capazes de segmentar as imagens de raio-X com desempenhos semelhantes àqueles obtidos por modelos supervisionados. A configuração $C = 2$ visa analisar se os modelos baseados em intensidade (MS e RFCM) conseguem segmentar a imagem de raio-X em duas grandes regiões: a área pulmonar e o restante do corpo. Já o uso de $C = 3$ explora a possibilidade desses modelos encontrarem uma separação do raio-X em três regiões: pulmões (intensidade mais próxima ao cinza), corpo (intensidade mais próxima ao branco) e fundo da imagem (intensidade próxima ao preto).

Como as combinações desses hiperparâmetros resultaram em um total de doze testes, a otimização bayesiana empregada por Nishio et al. (2021) não foi utilizada, uma vez que a configuração ótima pode ser facilmente detectada de forma manual.

## 3.7 Métricas de avaliação de desempenho

As métricas de precisão, revocação, índices Jaccard e de similaridade Dice, bem como testes estatísticos para avaliar a significância dos resultados, foram utilizadas para quantificar o desempenho dos modelos utilizados. A escolha dessas se deu por causa do uso em trabalhos relacionados à segmentação de imagens médicas como Oh et al. (2020), Larrazabal et al. (2020), Maity et al. (2022) e Chen et al. (2021b).

### 3.7.1 Binarização

O desempenho dos modelos é avaliado a partir de mapas binários. Para o modelo baseado em saliência, definiu-se como 1 (área pulmonar) os pixels cujo valor era superior a 0.5 e como 0 (restante da imagem) os demais, como ilustrado na Figura 23. Com relação aos modelos baseados nas funções MS (Kim e Ye, 2020) e RFCM (Chen et al., 2021b), a classe com maior valor em cada pixel era considerada como 1 e as demais como 0 e, em caso de empate, a primeira classe era escolhida. Esse processo de binarização está ilustrado na Figura 24. Como os modelos baseados em níveis de intensidade MS e RFCM resultam em $C$ mapas binários, aquele com o maior índice Dice durante o teste foi considerado como resultado final do respectivo modelo.

(a) Saída contínua

(b) Saída binária

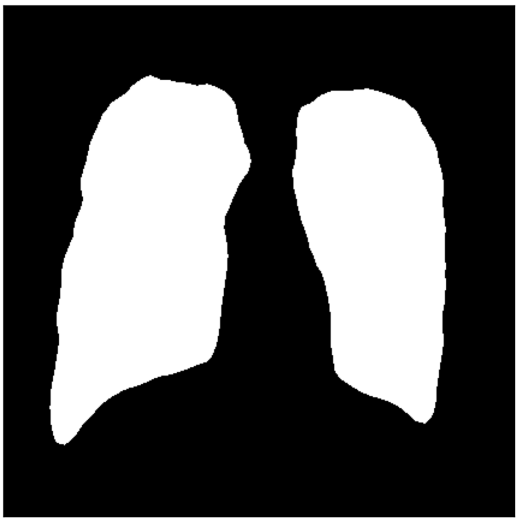


Figura 23 – Para o modelo de saliência, a binarização mapeou pixels com valor maior que 0.5 para 1 (branco) e os demais para 0 (preto).

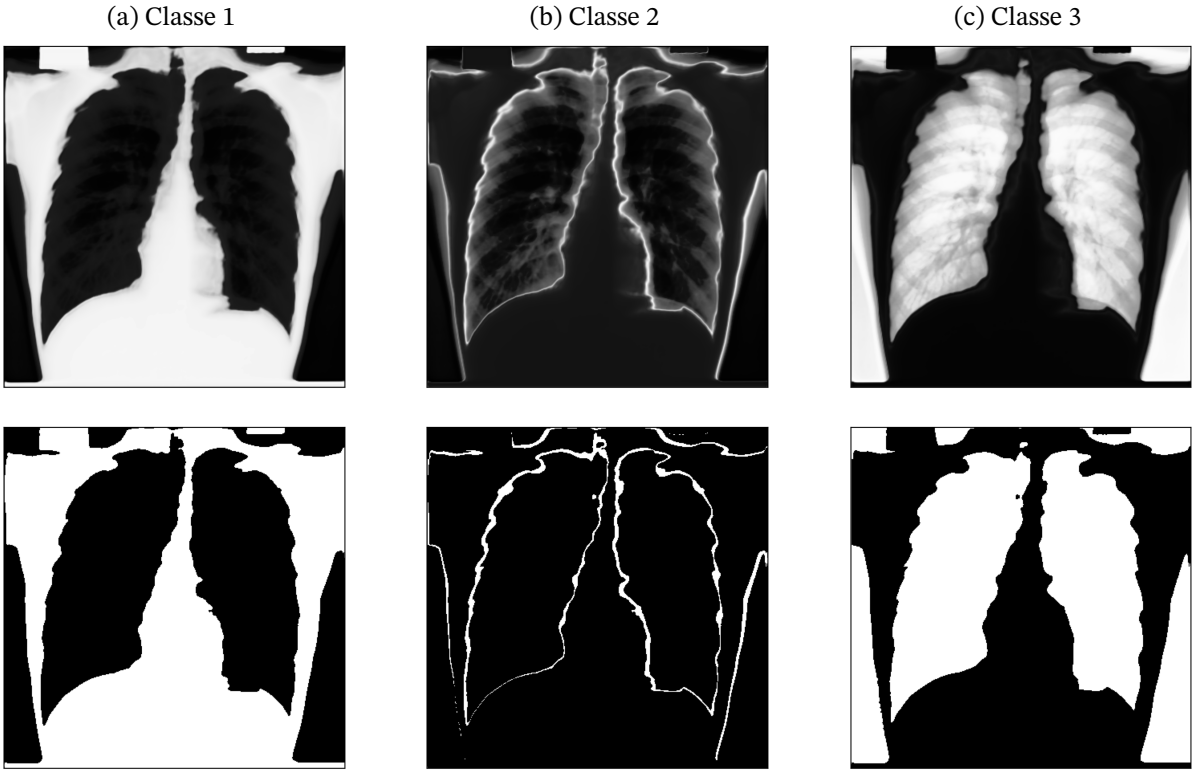



Figura 24 – Para os métodos baseados em níveis de intensidade, o pixel cuja classe tem o maior valor dentre as $C$ saídas (3 no exemplo), recebe o valor 1 (branco), enquanto que o mesmo pixel nas demais recebe 0 (preto). Em casos de empate, o pixel da classe 1 é atribuído a 1 e os demais 0. Para o exemplo, a classe 3 mais se aproxima da segmentação pulmonar, sendo escolhida como resultado final.

Para as definições matemáticas a seguir, consideraram-se mapas binários $P$ (valores calculados pelo modelo), $G$ (valores de referência esperados) e se assumiu que a expressão $|A|$ conta o número de elementos diferentes de zero em $A$.

### 3.7.2 Precisão

Quantifica o percentual de elementos em $P$ corretamente rotulados de acordo com a referência $G$. Assim,

$$Precisão = \frac{|G \cap P|}{|P|} \tag{3.8}$$

### 3.7.3 Revocação

Indica o percentual de rótulos de $G$ identificados por $P$. Logo,

$$Revocação = \frac{|G \cap P|}{|G|} \tag{3.9}$$

### 3.7.4 Índice Jaccard

Também chamado de interseção sobre união (IoU, na sigla em inglês), mede o percentual de interseção entre $P$ e $G$. Portanto,

$$Jaccard = \frac{|G \cap P|}{|G \cup P|} \tag{3.10}$$

### 3.7.5 Índice de similaridade Dice

O índice de similaridade de Dice, também chamado de índice de Dice, é outra métrica que avalia a interseção entre $P$ e $G$, podendo ser calculada a partir da relação precisão-revocação (medida F) ou do índice Jaccard. É escrita na forma a seguir:

$$Dice = 2\frac{|G \cap P|}{|G| + |P|} = 2\frac{PrecisãoRevocação}{Precisão + Revocação} = 2\frac{Jaccard}{Jaccard + 1} \tag{3.11}$$

### 3.7.6 Testes estatísticos

Testes T pareados foram utilizados para avaliar o nível de significância estatística entre os melhores resultados das funções de custo baseadas em saliência, RFCM e MS. O limiar de significância adotado foi 0.05.

Por simplicidade, o índice de Dice foi adotado como medida para classificar o desempenho dos modelos baseados em saliência e intensidade e para calcular os testes estatísticos

## 3.8 Ferramentas de programação

Os modelos foram programados na linguagem de programação Python, versão 3.7.13, usando as bibliotecas de computação tensorial com suporte a processamento paralelo em placas de vídeos Pytorch e Torchvision, versões 1.12.1+cu113 e 0.13.1+cu113 respectivamente. Alguns testes foram feitos nas respectivas versões de Python, Pytorch e Torchvision: 3.10.6, 1.13.1+cu117 e 0.14.1+cu117.

As implementações dos modelos baseados em saliência e em níveis de intensidade podem ser encontradas em repositórios do github [2] [3].

## 3.9 Configurações computacionais

Os modelos foram treinados utilizando o hardware disponível na versão Pro+ do Google Colab. Alguns testes foram feitos em máquina local com 16 GB de memória RAM, CPU IntelCore i5-10400F e placa de vídeo NVIDIA GeForce RTX 3050.

[2] https://github.com/PedroAcA/Deep-Unsupervised-Saliency-Detection - último acesso em 30/06/2023

[3] https://github.com/PedroAcA/clustering-losses-for-lung-segmentation - último acesso em 30/06/2023

# 4 Resultados e Discussão

Resultados quantitativos dos modelos profundos não supervisionados para as bases JSRT (Shiraishi et al., 2000) e MC (Jaeger et al., 2014) se encontram na Seção 4.1. Resultados qualitativos dos mesmos para esses dados estão na Seção 4.2. Uma comparação entre o modelo de saliência proposto e o estado-da-arte em segmentação pulmonar de imagens médicas é feita na Seção 4.3. Na Seção 4.4 são apresentadas as limitações do modelo baseado em saliência. O artigo originário desta pesquisa é referenciado na Seção 4.5.

## 4.1 Resultados quantitativos

Resultados quantitativos e análise estatística entre os modelos profundos não supervisionados descritos no Capítulo 3 para a base JSRT (Shiraishi et al., 2000) estão na Subseção 4.1.1. A Subseção 4.1.2 apresenta os resultados quantitativos na base MC (Jaeger et al., 2014) para os melhores modelos definidos a partir dos resultados presentes na Subseção 4.1.1.

### 4.1.1 Base JSRT

Os resultados presentes nas Tabelas 1 a 4 apresentam os resultados dos modelos guiados pelas três funções de custo avaliadas, baseadas em saliência a partir do trabalho de Zhang et al. (2018) e em intensidade como em Chen et al. (2021b) e Kim e Ye (2020), na porção de teste da base JSRT (Shiraishi et al., 2000).

O modelo de saliência descrito na Tabela 1 apresentou resultados melhores de acordo com as métricas Dice, Jaccard e revocação quando os rótulos ruidosos foram pós-processados com correção gamma ($\gamma = 0.5$), de acordo com o esperado, visto que esse procedimento aumenta o contraste entre a área pulmonar e o restante do corpo. Diferente do previsto, aplicar pré-processamento na entrada fez um efeito deletério no desempenho. É possível que o nível de contraste original das imagens de entrada já crie uma fronteira não linear que separe os rótulos e que pode ser borrada ou encurtada com as operações de pré-processamento. Porém, estudos mais detalhados são necessários para comprovar essa hipótese.

Tabela 1 – Resultados experimentais na base JSRT para o modelo de saliência. 'N' significa que nenhum pré ou pós-processamento foi aplicado e 'S' simboliza pré ou pós-processamento aplicado. Os melhores resultados estão destacados em negrito, conforme a respectiva métrica, e o melhor modelo é o que apresenta o maior índice Dice.

| Experimento | Entrada pré-processada | Mapas ruidosos pós processados | Dice | Jaccard | Precisão | Revocação |
|---|---|---|---|---|---|---|
| $S_1$ | S | S | 0.86 | 0.76 | 0.92 | 0.82 |
| $S_2$ | S | N | 0.76 | 0.62 | **0.97** | 0.63 |
| $S_3$ | N | N | 0.80 | 0.68 | 0.96 | 0.70 |
| $S_4$ | **N** | **S** | **0.87** | **0.78** | 0.85 | **0.91** |

É interessante notar que, para os resultados utilizando as métricas RFCM e MS mostrados, respectivamente, nas Tabelas 2 e 3, o uso de pré-processamento na imagem de entrada e o número de classes de saída dos modelos não resultaram em diferenças maiores que 0.05 no valor da métrica Dice, com exceção do teste apresentado na terceira linha da Tabela 3 (entrada pré-processada e 3 classes de saída). Esse comportamento pode indicar que os modelos baseados em níveis de intensidade não conseguiram aprender a função que separa os pulmões do restante da imagem.

Tabela 2 – Resultados experimentais na base JSRT para o modelo RFCM. 'N' significa que nenhum pré-processamento foi aplicado e 'S' simboliza pré-processamento aplicado. Os melhores resultados estão destacados em negrito conforme a respectiva métrica e o melhor modelo é o que apresenta o maior índice Dice.

| Experimento | Entrada pré-processada | Classes | Dice | Jaccard | Precisão | Revocação |
|---|---|---|---|---|---|---|
| $RFCM_1$ | N | 2 | 0.65 | 0.49 | 0.50 | **0.93** |
| $RFCM_2$ | S | 2 | 0.66 | 0.50 | 0.53 | 0.91 |
| $RFCM_3$ | N | 3 | 0.67 | 0.51 | **0.56** | 0.86 |
| $RFCM_4$ | **S** | **3** | **0.68** | **0.52** | **0.56** | 0.88 |

Tabela 3 – Resultados experimentais na base JSRT para o modelo MS. 'N' significa que nenhum pré-processamento foi aplicado e 'S' simboliza pré-processamento aplicado. Os melhores resultados estão destacados em negrito, conforme a respectiva métrica e o melhor modelo é o que apresenta o maior índice Dice.

| Experimento | Entrada pré-processada | Classes | Dice | Jaccard | Precisão | Revocação |
|---|---|---|---|---|---|---|
| $MS_1$ | N | 2 | 0.67 | 0.51 | 0.54 | 0.90 |
| $MS_2$ | S | 2 | 0.63 | 0.47 | 0.48 | **0.96** |
| $MS_3$ | N | 3 | 0.52 | 0.36 | 0.56 | 0.49 |
| $MS_4$ | **S** | **3** | **0.68** | **0.52** | **0.57** | 0.85 |

A Tabela 4 apresenta os p-valores dos melhores resultados encontrados nas Tabelas 1 a 3 de acordo com o índice Dice. Percebe-se que o índice Dice do modelo de saliência é maior que os relativos aos modelos de nível de intensidade de forma estatisticamente significativa (p-valor $< 0.05$).

Tabela 4 – Dice, desvio padrão do Dice e p-valores para for os melhores resultados descritos nas Tabelas 1 a 3.

| Modelo | Dice | Desvio padrão do Dice | p-valor comparado com os resultados de saliência |
|---|---|---|---|
| Saliência | 0.87 | 0.04 | Não aplicável |
| RFCM | 0.68 | 0.09 | $1.55(10^{-22})$ |
| MS | 0.68 | 0.09 | $1.55(10^{-22})$ |

### 4.1.2 Base MC

A Tabela 5 exibe os resultados dos melhores modelos destacados nas Tabelas 1 a 3 quando rotulam a base de dados MC (Jaeger et al., 2014). Nota-se que os resultados são

muito similares com relação aos obtidos na base JSRT, o que indica consistência dos modelos com relação a diferentes conjuntos de dados.

Tabela 5 – Dice, Jaccard, precisão e revocação para os modelos de saliência, RFCM e MS na base MC.

| Modelo | Dice | Jaccard | Precisão | Revocação |
|---|---|---|---|---|
| Saliência | **0.88** | **0.79** | **0.86** | **0.91** |
| RFCM | 0.64 | 0.47 | 0.50 | **0.91** |
| MS | 0.64 | 0.48 | 0.51 | 0.89 |

## 4.2 Resultados qualitativos

Os resultados presentes nas Tabelas 1 a 3, referentes à base JSRT (Shiraishi et al., 2000), são ilustrados pela Figura 25 e os da Tabela 5, referentes à base MC (Jaeger et al., 2014), estão na Figura 26. Visualmente, pode-se perceber que ambos os modelos, baseados em saliência e em níveis de intensidade, conseguiram extrair quase toda região equivalente à área pulmonar, o que corresponde a uma alta revocação. Porém, o modelo baseado em saliência consegue ser mais preciso que aqueles baseados em níveis de intensidade ao capturar menos ruído presente no plano de fundo da imagem.

Os modelos baseados em níveis de intensidade mostrados nas Figuras 25 e 26 podem ser utilizados como estágio de pré-processamento para ferramentas computacionais de segmentação pulmonar, uma vez que extraem quase toda a área pulmonar combinada com ruído de plano de fundo. Outra aplicação para esses modelos são como ferramentas rápidas e automáticas de delineamento da área pulmonar que podem ser utilizadas por profissionais da área de saúde, uma vez que esses conseguem remover o ruído de plano de fundo eficientemente.

### 4.2.1 Casos extremos da classificação

As Figuras 27 e 28 exibem as três melhores e piores classificações, utilizando a métrica Dice, para, respectivamente, as bases JSRT (Shiraishi et al., 2000) e MC (Jaeger et al., 2014) considerando o modelo baseado em saliência visual. A partir delas, percebe-se que o modelo tem sua eficácia reduzida quando o contraste entre a região pulmonar e a porção corporal que a circunda é reduzida. Nesse processo, o modelo não identifica pixels que pertencem à área pulmonar. Outra fonte de erro parece estar relacionada ao limiar de binarização adotado (0.5), uma vez que alguns erros foram cometidos quando regiões não pulmonares acima do limiar utilizado, mas com intensidade (nível de ativação) menor que a área dos pulmões, foram erroneamente demarcadas como região pulmonar após a aplicação do limiar.

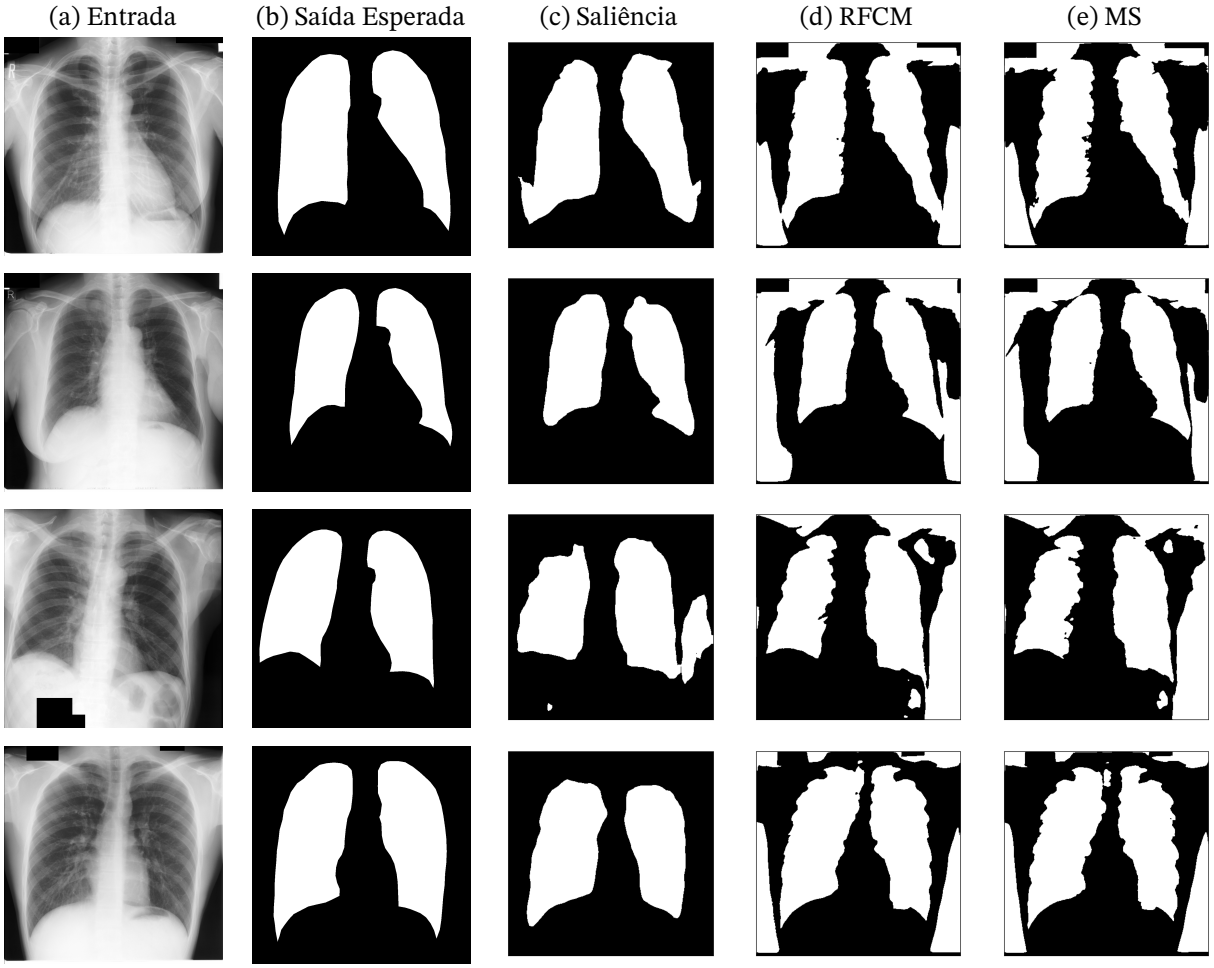



Figura 25 – Resultados qualitativos na base JSRT: (a) Imagem de entrada, (b) Saída esperada, (c) Resposta dada pelo modelo baseado em saliência visual, (d) Resposta dada pelo modelo de nível de intensidade que se baseia na função de custo *Robust Fuzzy C-Means* (RFCM), (e) Resposta dada pelo modelo de nível de intensidade que se baseia na função de custo Mumford-Shah (MS).

## 4.3 Sobre modelos não supervisionados e supervisionados

Os desempenhos obtidos pelos modelos não supervisionados testados no contexto de segmentação semântica mostrados nas Tabelas 1 a 5 e na Figura 25 apresentam evidências que o aprendizado não supervisionado pode ser efetivo para essa tarefa. Dentro dos modelos comparados, aquele baseado em saliência visual obteve os melhores resultados considerando o índice Jaccard como a métrica de comparação.

Pode-se argumentar que a melhoria observada no modelo de saliência é devido ao pré-ajuste de pesos da rede DeepLabV3 (Chen et al., 2019), realizado na base ImageNet (Deng et al., 2009). Porém, é importante notar que houve uma transferência de aprendizado não supervisionada para um contexto totalmente distinto da configuração inicial: substitui-se segmentação de imagens coloridas em ambientes naturais por segmentação de estruturas

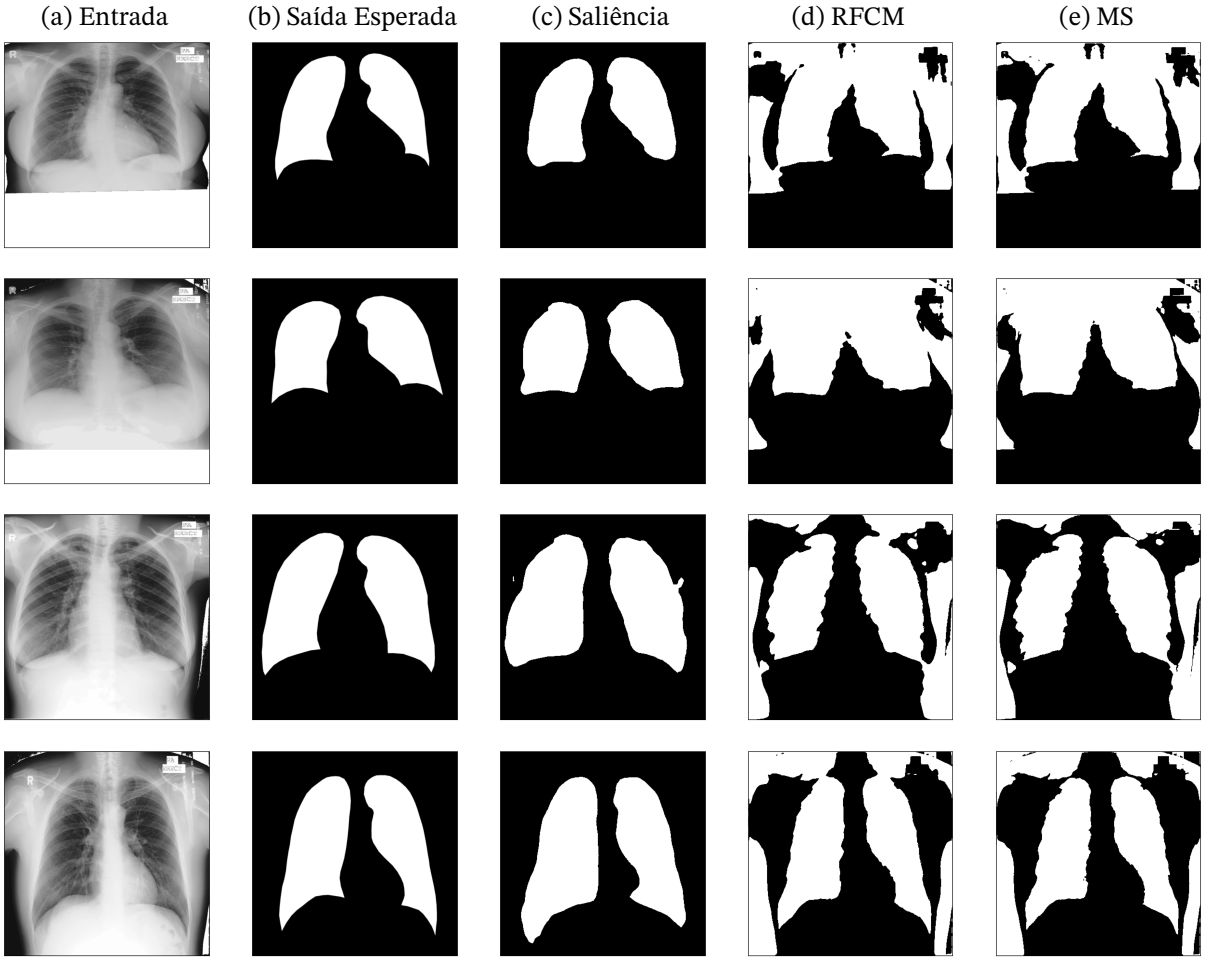



Figura 26 – Resultados qualitativos na base MC: (a) Imagem de entrada, (b) Saída esperada, (c) Resposta dada pelo modelo baseado em saliência visual, (d) Resposta dada pelo modelo de nível de intensidade que se baseia na função de custo *Robust Fuzzy C-Means* (RFCM), (e) Resposta dada pelo modelo de nível de intensidade que se baseia na função de custo Mumford-Shah (MS).

anatômicas em imagens em escala de cinza. Outro aspecto favorável à hipótese de que a inicialização dos parâmetros não é indispensável para a tarefa está no trabalho de Zhang et al. (2021), que obtiveram desempenho semelhante ao modelo pré-treinado na ImageNet (Deng et al., 2009) com base na mesma rede treinada 'do zero' para a tarefa de detecção de objeto saliente.

A Tabela 6 apresenta resultados quantitativos na base JSRT (Shiraishi et al., 2000) para o modelo de saliência proposto e o estado-da-arte. Quando comparados a trabalhos relacionados que empregam aprendizagem supervisionada (Oh et al., 2020, Singh et al., 2021, Maity et al., 2022 e Nishio et al., 2021), o melhor desempenho obtido pela abordagem baseada em saliência teve uma diferença menor que 0.13, utilizando o índice Dice com relação a essas abordagens. A proximidade de desempenho fornece indícios que métodos não supervisionados podem substituir as versões supervisionadas. A vantagem dessa troca está, como argumentado por Croitoru et al. (2019), na capacidade de se adaptarem rapidamente

a mudanças de contexto ou situações nunca antes vistas por não necessitarem de rótulos manualmente fornecidos.

Tabela 6 – Dice e Jaccard do modelo não supervisionado baseado em saliência visual proposto e abordagens supervisionadas do estado-da-arte para segmentação da região pulmonar em raios-X da base JSRT.

| Modelo | Dice | Jaccard |
|---|---|---|
| Saliência | 0.87 | 0.78 |
| Oh et al. (2020) | Não aplicável | 0.955 |
| Singh et al. (2021) | 0.969 | 0.985 |
| Nishio et al. (2021) | 0.976 | 0.954 |
| Maity et al. (2022) | 0.983 | 0.968 |

## 4.4 Limitações do modelo baseado em saliência visual

A arquitetura de rede neuronal DeepLabV3 (Chen et al., 2019) utilizada para treinamento utilizando estimativas de saliência visual tem uma dependência de pré inicialização de parâmetros para que consiga ser adaptada para contextos com poucos dados e rótulos ruidosos. Essa dependência muito provavelmente advém da quantidade de parâmetros dessa rede. Apesar de já argumentado que trabalhos como os de Zhang et al. (2021) podem fornecer uma inicialização não supervisionada à rede DeepLabV3 (Chen et al., 2019), o uso de redes com menos parâmetros treináveis, como, por exemplo, a rede Unet (Ronneberger et al., 2015) pode aliviar o requisito de grandes bases de dados disponíveis para inicialização do modelo.

## 4.5 Artigo publicado

As ideias expostas neste texto culminaram na publicação do artigo *A deep unsupervised saliency model for lung segmentation in chest X-ray images* (de Almeida e Borges, 2023).

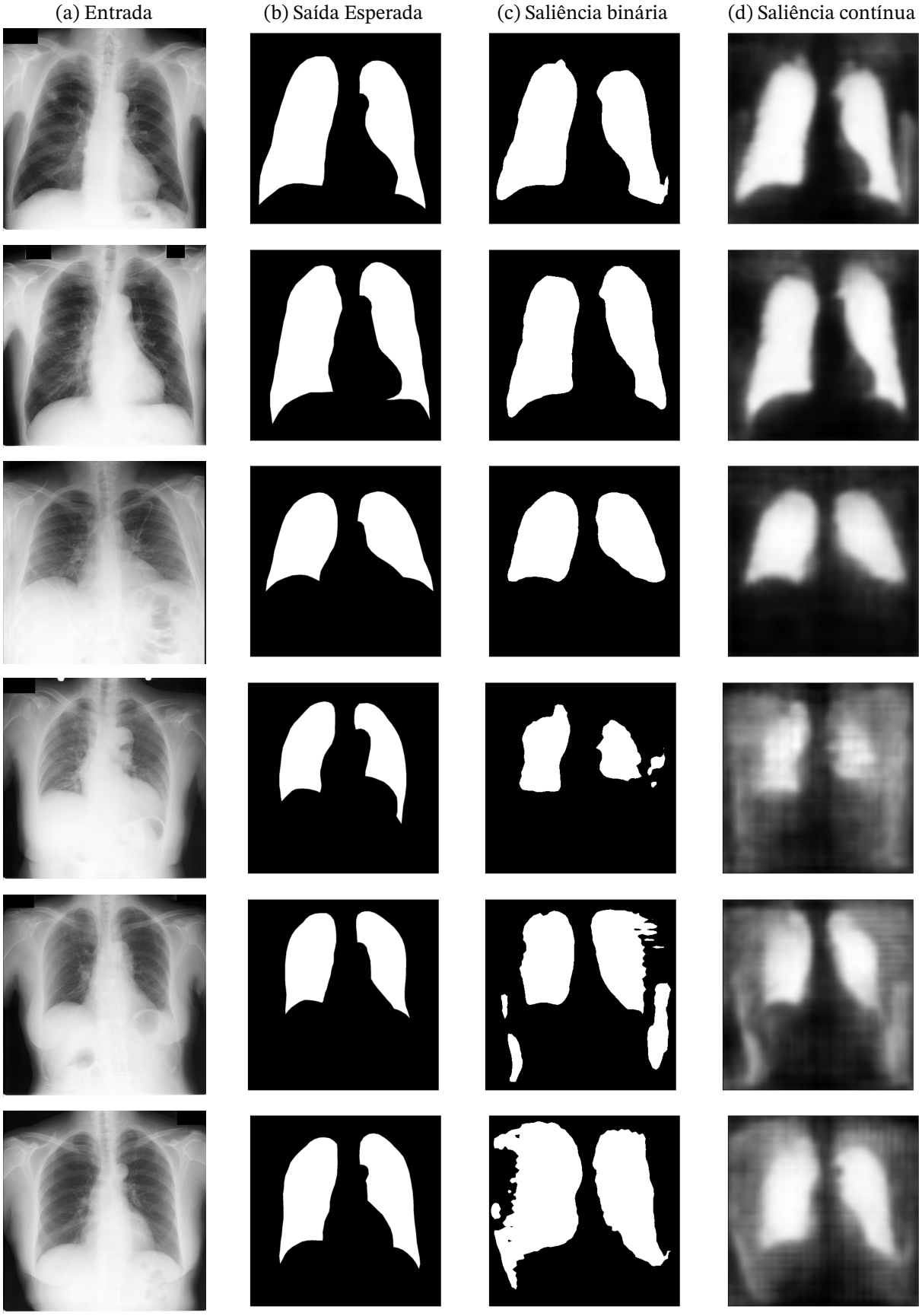



Figura 27 – Resultados qualitativos para os três maiores e três menores valores de desempenho do modelo de saliência, utilizando o índice Dice, na base JSRT: (a) Imagem de entrada, (b) Saída esperada, (c) Resposta binária dada pelo modelo baseado em saliência visual, (d) Resposta contínua dada pelo modelo baseado em saliência visual.

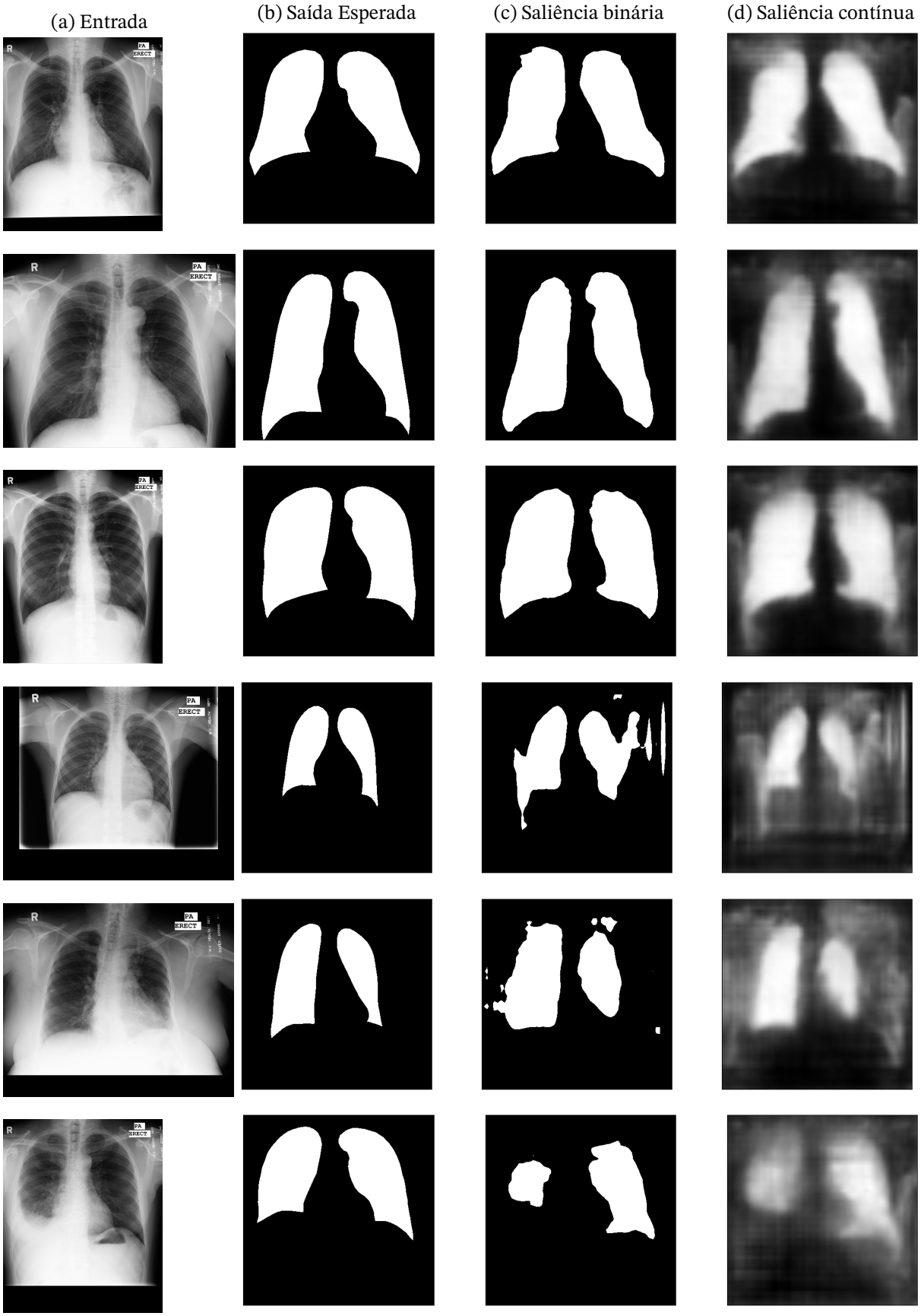



Figura 28 – Resultados qualitativos para os três maiores e três menores valores de desempenho do modelo de saliência, utilizando o índice Dice, na base MC: (a) Imagem de entrada, (b) Saída esperada, (c) Resposta binária dada pelo modelo baseado em saliência visual, (d) Resposta contínua dada pelo modelo baseado em saliência visual.

# 5 Conclusões

Devido ao seu baixo custo e ampla utilização, raios-X torácicos são exames de imagem médica de grande valor para a sociedade. Suas aplicações incluem detecção de câncer pulmonar, pneumonia, tuberculose, dentre outras condições que afetam os pulmões. A análise manual de diversos exames gera fadiga e pode levar a erros. Por isso, ferramentas computacionais que auxiliam os profissionais de saúde com o diagnóstico são alternativas interessantes. O advento de técnicas de aprendizagem profunda motivou a adoção desses modelos para as áreas médicas, aproveitando-se de grandes bases de dados rotulados para alcançar desempenho maiores que os apresentados por serem humanos em algumas tarefas da área médica (Mahomed et al., 2020).

Visando mitigar a necessidade de rotulação manual dos dados de treinamento e explorando princípios gerais que guiam a atenção humana, modificou-se o modelo não supervisionado de saliência visual proposto por Zhang et al. (2018) para a tarefa de segmentação pulmonar. Esse foi treinado em uma porção de testes da base JSRT (Shiraishi et al., 2000) e testado em porções de teste das bases JSRT (Shiraishi et al., 2000) e MC (Jaeger et al., 2014). Uma vez que a base de dados médica utilizada para treinamento é menor do que a base de saliência visual utilizada por Zhang et al. (2018), a atualização do modelo de ruído proposta por Zhang et al. (2018) não foi empregada para permitir maior quantidade de dados para treinamento.

O modelo proposto foi comparado com duas abordagens não supervisionadas avaliadas em contextos de segmentação de imagens médicas em escala de cinzas. Uma delas, descrita por Chen et al. (2021b), baseia-se no princípio de *Robust Fuzzy C-Means* (RCFM na sigla em inglês) de Pham (2001), enquanto que a outra, proposta por Kim e Ye (2020), faz uso do funcional Mumford-Shah (Mumford e Shah, 1989). Esses trabalhos de aprendizagem profunda não supervisionada (Chen et al., 2021b e Kim e Ye, 2020) podem ser considerados como métodos de agrupamento de regiões baseados em níveis de intensidade.

Utilizando métricas como índices Dice e Jaccard, precisão e revocação, combinados com testes T pareados, os resultados sugerem que o modelo baseado em saliência teve um melhor desempenho que os modelos baseados em níveis de intensidade (Chen et al., 2021b e Kim e Ye, 2020). Além disso, os resultados obtidos na base MC (Jaeger et al., 2014), um conjunto de dados diferente daquele utilizado para ajustar os parâmetros, mantêm a performance próxima daquela obtida na base JSRT (Shiraishi et al., 2000).

Comparando o modelo não supervisionado baseado em saliência com abordagens supervisionadas de segmentação de raios-X (Oh et al., 2020, Singh et al., 2021, Maity et al., 2022 e Nishio et al., 2021), nota-se que a proposta baseada em saliência pode servir como

substituta das contrapartes supervisionadas, uma vez que ela tem desempenho próximo ao supervisionado e adiciona o benefício de ser sempre disponível para novo treinamento em situações nas quais rótulos humanos não estão disponíveis.

Trabalhos futuros incluem realizar a segmentação do coração de forma não supervisionada, bem como testar essa classe de aprendizagem para identificar anomalias (doenças ou condições desviantes do esperado). Outra linha de pesquisa tem como foco não precisar de rótulos externos para identificar regiões salientes em uma imagem.

# Referências

ACHANTA, R.; SHAJI, A.; SMITH, K.; LUCCHI, A.; FUA, P.; SÜSSTRUNK, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 34, n. 11, p. 2274–2282, nov. 2012. Citado nas pp. 23, 24.

BADRINARAYANAN, V.; KENDALL, A.; CIPOLLA, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 39, n. 12, p. 2481–2495, 2017. DOI: 10.1109/TPAMI.2016.2644615. Citado na p. 28.

BORJI, A.; CHENG, M.-M.; HOU, Q.; JIANG, H.; LI, J. Salient object detection: A survey. **Computational Visual Media**, v. 5, n. 2, p. 117–150, jun. 2019. DOI: 10.1007/s41095-019-0149-9. Citado nas pp. 15, 16, 22, 24.

BORJI, A.; ITTI, L. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. **CVPR 2015 workshop on "Future of Datasets"**, 2015. arXiv preprint arXiv:1505.03581. Citado na p. 22.

CHEN, J.; LU, Y.; YU, Q.; LUO, X.; ADELI, E.; WANG, Y.; LU, L.; YUILLE, A. L.; ZHOU, Y. **TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation**. 2021a. arXiv: 2102.04306 [cs.CV]. Citado nas pp. 34, 35.

CHEN, J.; LI, Y.; LUNA, L. P.; CHUNG, H. W.; ROWE, S. P.; DU, Y.; SOLNES, L. B.; FREY, E. C. Learning fuzzy clustering for SPECT/CT segmentation via convolutional neural networks. **Medical Physics**, v. 48, n. 7, p. 3860–3877, 2021b. DOI: 10.1002/mp.14903. Citado nas pp. 15, 16, 36, 37, 43–47, 50, 58.

CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 40, n. 4, p. 834–848, 2017. Citado nas pp. 30, 31, 35, 40.

CHEN, L.-C.; PAPANDREOU, G.; KOKKINOS, I.; MURPHY, K.; YUILLE, A. L. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs. In: ICLR (Poster). 2015. Citado na p. 30.

CHEN, L.-C.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. Rethinking atrous convolution for semantic image segmentation. arXiv 2017. **arXiv preprint arXiv:1706.05587**, v. 2, 2019. Citado nas pp. 31, 40, 45, 53, 55.

CHEN, L.-C.; ZHU, Y.; PAPANDREOU, G.; SCHROFF, F.; ADAM, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: FERRARI, V.; HEBERT, M.; SMINCHISESCU, C.; WEISS, Y. (Ed.). **Computer Vision – ECCV 2018**. Cham: Springer International Publishing, 2018. P. 833–851. ISBN 978-3-030-01234-2. Citado nas pp. 31, 34.

CROITORU, I.; BOGOLIN, S. V.; LEORDEANU, M. Unsupervised Learning of Foreground Object Segmentation. **International Journal of Computer Vision**, Springer US, v. 127, n. 9, p. 1279–1302, 2019. DOI: 10.1007/s11263-019-01183-3. Citado nas pp. 16, 35, 36, 54.

DE ALMEIDA, P. A. C.; BORGES, D. L. A deep unsupervised saliency model for lung segmentation in chest X-ray images. **Biomedical Signal Processing and Control**, v. 86, p. 105334, 2023. ISSN 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2023.105334. Disponível em: <https://www.sciencedirect.com/science/article/pii/S174680942300767X>. Citado na p. 55.

DENG, J.; DONG, W.; SOCHER, R.; LI, L.-J.; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: IEEE. 2009 IEEE conference on computer vision and pattern recognition. 2009. P. 248–255. Citado nas pp. 53, 54.

FELZENSZWALB, P. F.; HUTTENLOCHER, D. P. Efficient Graph-Based Image Segmentation. **International Journal of Computer Vision**, v. 59, n. 2, p. 167–181, set. 2004. ISSN 1573-1405. DOI: 10.1023/B:VISI.0000022288.19776.77. Disponível em: <https://doi.org/10.1023/B:VISI.0000022288.19776.77>. Citado nas pp. 23, 24.

GINNEKEN, B. van; STEGMANN, M.; LOOG, M. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. **Medical Image Analysis**, v. 10, n. 1, p. 19–40, 2006. Citado na p. 37.

GONZALEZ, R. C.; WOODS, R. E. Digital Image Processing, Global Edition. In: fourth. New York: Pearson, 2018. Citado nas pp. 18–21.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. http://www.deeplearningbook.org. Citado nas pp. 15, 24, 26–28, 31–33, 35.

HE, K.; ZHANG, X.; REN, S.; SUN, J. **Deep Residual Learning for Image Recognition**. 2015. arXiv: 1512.03385 [cs.CV]. Citado nas pp. 23, 31, 33.

IOFFE, S.; SZEGEDY, C. **Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift**. 2015. arXiv: 1502.03167 [cs.LG]. Citado na p. 33.

JAEGER, S.; CANDEMIR, S.; ANTANI, S.; WÁNG, Y.-X. J.; LU, P.-X.; THOMA, G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. eng. **Quantitative imaging in medicine and surgery**, China, v. 4, n. 6, p. 475–477, dez. 2014. DOI: 10.3978/j.issn.2223-4292.2014.11.20. Citado nas pp. 38–40, 50–52, 58.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An Introduction to Statistical Learning with Applications in R**. 2. ed.: Springer, jun. 2023. Citado nas pp. 15, 24, 25.

JEGOU, S.; DROZDZAL, M.; VAZQUEZ, D.; ROMERO, A.; BENGIO, Y. The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Los Alamitos, CA, USA: IEEE Computer Society, jul. 2017. P. 1175–1183. DOI: 10.1109/CVPRW.2017.156. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPRW.2017.156>. Citado na p. 34.

KE, Y. Y.; TSUBONO, T. Recursive Contour-Saliency Blending Network for Accurate Salient Object Detection. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). 2022. P. 1360–1370. DOI: 10.1109/WACV51458.2022.00143. Citado na p. 34.

KIM, B.; YE, J. C. Mumford-shah loss functional for image segmentation with deep learning. **IEEE Transactions on Image Processing**, IEEE, v. 29, p. 1856–1866, 2020. DOI: 10.1109/TIP.2019.2941265. Citado nas pp. 15, 16, 36, 44–47, 50, 58.

KINGMA, D. P.; BA, J. **Adam: A Method for Stochastic Optimization**. 2017. arXiv: 1412.6980 [cs.LG]. Citado nas pp. 32, 45.

KOEHLER, K.; GUO, F.; ZHANG, S.; ECKSTEIN, M. P. What do saliency models predict? **Journal of Vision**, v. 14, p. 14–14, 3 mar. 2014. ISSN 1534-7362. DOI: 10.1167/14.3.14. Disponível em: <http://jov.arvojournals.org/Article.aspx?doi=10.1167/14.3.14>. Citado na p. 24.

KRAHENBUHL, P.; KOLTUN, V. Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In: NEURAL Information Processing Systems. 2011. Citado nas pp. 30, 31, 40.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In: PEREIRA, F.; BURGES, C.; BOTTOU, L.; WEINBERGER, K. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2012. v. 25. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>. Citado na p. 44.

KUANG, X.; CHEUNG, J. P.; WU, H.; DOKOS, S.; ZHANG, T. MRI-SegFlow: A novel unsupervised deep learning pipeline enabling accurate vertebral segmentation of MRI images. **Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS**, 2020-July, p. 1633–1636, 2020. DOI: 10.1109/EMBC44109.2020.9175987. Citado na p. 36.

LARRAZABAL, A. J.; MARTÍNEZ, C.; GLOCKER, B.; FERRANTE, E. Post-DAE: Anatomically Plausible Segmentation via Post-Processing with Denoising Autoencoders. **IEEE Transactions on Medical Imaging**, Institute of Electrical e Electronics Engineers Inc., v. 39, n. 12, p. 3813–3820, dez. 2020. DOI: 10.1109/TMI.2020.3005297. Citado na p. 47.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. **Proceedings of the IEEE**, v. 86, n. 11, p. 2278–2324, 1998. DOI: 10.1109/5.726791. Citado na p. 27.

LIANG, M.; HU, X. Recurrent convolutional neural network for object recognition. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. P. 3367–3375. DOI: 10.1109/CVPR.2015.7298958. Citado nas pp. 36, 43, 44.

LIU, T.; YUAN, Z.; SUN, J.; WANG, J.; ZHENG, N.; TANG, X.; SHUM, H. Y. Learning to detect a salient object. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 33, n. 2, p. 353–367, 2011. DOI: 10.1109/TPAMI.2010.70. Citado na p. 23.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun. 2015. P. 3431–3440. DOI: 10.1109/CVPR.2015.7298965. Disponível em: <https://doi.ieeecomputersociety.org/10.1109/CVPR.2015.7298965>. Citado na p. 27.

LUO, Z.; MISHRA, A.; ACHKAR, A.; EICHEL, J.; LI, S.; JODOIN, P.-M. Non-local Deep Features for Salient Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017. P. 6593–6601. DOI: 10.1109/CVPR.2017.698. Citado na p. 34.

MAHOMED, N.; GINNEKEN, B. van; PHILIPSEN, R. H.; MELENDEZ, J.; MOORE, D. P.; MOODLEY, H.; SEWCHURAN, T.; MATHEW, D.; MADHI, S. A. Computer-aided diagnosis for World Health Organization-defined chest radiograph primary-endpoint pneumonia in children. **Pediatric radiology**, Springer, v. 50, p. 482–491, 2020. Citado na p. 58.

MAITY, A.; NAIR, T. R.; MEHTA, S.; PRAKASAM, P. Automatic lung parenchyma segmentation using a deep convolutional neural network from chest X-rays. **Biomedical**

**Signal Processing and Control**, v. 73, p. 103398, 2022. DOI: 10.1016/j.bspc.2021.103398. Citado nas pp. 34, 47, 54, 55, 58.

MITCHELL, T. M. **Machine Learning**. McGraw-Hill Science/Engineering/Math, 1997. Citado na p. 25.

MUMFORD, D.; SHAH, J. Optimal approximations by piecewise smooth functions and associated variational problems. **Communications on Pure and Applied Mathematics**, v. 42, n. 5, p. 577–685, 1989. DOI: 10.1002/cpa.3160420503. Citado nas pp. 36, 58.

NISHIO, M.; FUJIMOTO, K.; TOGASHI, K. Lung segmentation on chest X-ray images in patients with severe abnormal findings using deep learning. **International Journal of Imaging Systems and Technology**, v. 31, n. 2, p. 1002–1008, 2021. DOI: https://doi.org/10.1002/ima.22528. Citado nas pp. 34, 45, 47, 54, 55, 58.

OH, Y.; PARK, S.; YE, J. C. Deep Learning COVID-19 Features on CXR Using Limited Training Data Sets. **IEEE Transactions on Medical Imaging**, Institute of Electrical e Electronics Engineers Inc., v. 39, n. 8, p. 2688–2700, ago. 2020. DOI: 10.1109/TMI.2020.2993291. Citado nas pp. 34, 38, 47, 54, 55, 58.

PHAM, D. L. Spatial models for fuzzy clustering. **Computer Vision and Image Understanding**, v. 84, n. 2, p. 285–297, 2001. DOI: 10.1006/cviu.2001.0951. Citado nas pp. 36, 58.

QIN, C.; YAO, D.; SHI, Y.; SONG, Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. **Biomedical engineering online**, BioMed Central, v. 17, n. 1, p. 1–23, 2018. Citado na p. 15.

RANZATO, M.; HUANG, F. J.; BOUREAU, Y.-L.; LECUN, Y. Unsupervised Learning of Invariant Feature Hierarchies with Applications to Object Recognition. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. 2007. P. 1–8. DOI: 10.1109/CVPR.2007.383157. Citado na p. 28.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: NAVAB, N.; HORNEGGER, J.; WELLS, W. M.; FRANGI, A. F. (Ed.). **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015**. Cham: Springer International Publishing, 2015. P. 234–241. ISBN 978-3-319-24574-4. Citado nas pp. 34–36, 55.

SHIRAISHI, J.; KATSURAGAWA, S.; IKEZOE, J.; MATSUMOTO, T.; KOBAYASHI, T.; KOMATSU, K. I.; MATSUI, M.; FUJITA, H.; KODERA, Y.; DOI, K. Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. **American Journal of Roentgenology**, v. 174, n. 1, p. 71–74, 2000. DOI: 10.2214/ajr.174.1.1740071. Citado nas pp. 19–21, 37, 39, 45, 50, 52, 54, 58.

SIMONYAN, K.; ZISSERMAN, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: 3RD International Conference on Learning Representations (ICLR 2015). 2015. P. 1–14. Citado na p. 30.

SINGH, A.; LALL, B.; PANIGRAHI, B. K.; AGRAWAL, A.; AGRAWAL, A.; THANGAKUNAM, B.; CHRISTOPHER, D. J. Deep LF-Net: Semantic lung segmentation from Indian chest radiographs including severely unhealthy images. **Biomedical Signal Processing and Control**, Elsevier Ltd, v. 68, jul. 2021. DOI: 10.1016/j.bspc.2021.102666. Citado nas pp. 34, 54, 55, 58.

THOMAS, G. B.; WEIR, M. D.; HASS, J. **Cálculo**. 12ª edição: Pearson EDucation do Brasil, 2012a. v. 1. Citado na p. 31.

THOMAS, G. B.; WEIR, M. D.; HASS, J. **Cálculo**. 12ª edição: Pearson EDucation do Brasil, 2012b. v. 2. Citado na p. 31.

VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. Attention is All you Need. In: GUYON, I.; LUXBURG, U. V.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (Ed.). **Advances in Neural Information Processing Systems**. Curran Associates, Inc., 2017. v. 30. Disponível em: <https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>. Citado nas pp. 30, 35.

VEKSLER, O.; BOYKOV, Y.; MEHRANI, P. Superpixels and Supervoxels in an Energy Optimization Framework. In: DANIILIDIS, K.; MARAGOS, P.; PARAGIOS, N. (Ed.). **Computer Vision – ECCV 2010**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. P. 211–224. ISBN 978-3-642-15555-0. Citado na p. 23.

VILLATE, J. E. **Eletricidade, Magnetismo e Circuitos**. FEUP, set. 2019. DOI: 10.24840/978-972-99396-6-2. Disponível em: <https://doi.org/10.24840/978-972-99396-6-2>. Citado na p. 18.

WANG, P.; LIU, Y.; CAO, Y.; YANG, X.; LUO, Y.; LU, H.; LIANG, Z.; LAU, R. W. Salient object detection with image-level binary supervision. **Pattern Recognition**, v. 129, p. 108782, 2022. ISSN 0031-3203. DOI: https://doi.org/10.1016/j.patcog.2022.108782. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320322002631>. Citado na p. 34.

WEI, Y.; WEN, F.; ZHU, W.; SUN, J. Geodesic Saliency Using Background Priors. In: FITZGIBBON, A.; LAZEBNIK, S.; PERONA, P.; SATO, Y.; SCHMID, C. (Ed.). **Computer Vision – ECCV 2012**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. P. 29–42. ISBN 978-3-642-33712-3. Citado nas pp. 15, 23, 41, 42, 46.

WOLFE, J. M.; HOROWITZ, T. S. Five factors that guide attention in visual search. **Nature Human Behaviour**, Macmillan Publishers Limited, v. 1, p. 1–8, 3 2017. ISSN 23973374. DOI: 10.1038/s41562-017-0058. Disponível em: <http://dx.doi.org/10.1038/s41562-017-0058>. Citado nas pp. 16, 22.

ZHANG, J.; DAI, Y.; ZHANG, T.; HARANDI, M.; BARNES, N.; HARTLEY, R. Learning Saliency from Single Noisy Labelling: A Robust Model Fitting Perspective. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 43, n. 8, p. 2866–2873, 2021. DOI: 10.1109/TPAMI.2020.3046486. Citado nas pp. 35, 54, 55.

ZHANG, J.; ZHANG, T.; DAI, Y.; HARANDI, M.; HARTLEY, R. Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, p. 9029–9038, 2018. DOI: 10.1109/CVPR.2018.00941. Citado nas pp. 16, 35, 36, 40, 42, 43, 50, 58.

ZHOU, Z.; RAHMAN SIDDIQUEE, M. M.; TAJBAKHSH, N.; LIANG, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In: STOYANOV, D.; TAYLOR, Z.; CARNEIRO, G.; SYEDA-MAHMOOD, T.; MARTEL, A.; MAIER-HEIN, L.; TAVARES, J. M. R.; BRADLEY, A.; PAPA, J. P.; BELAGIANNIS, V.; NASCIMENTO, J. C.; LU, Z.; CONJETI, S.; MORADI, M.; GREENSPAN, H.; MADABHUSHI, A. (Ed.). **Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support**. Cham: Springer International Publishing, 2018. P. 3–11. ISBN 978-3-030-00889-5. Citado na p. 35.

ZHU, W.; LIANG, S.; WEI, Y.; SUN, J. Saliency optimization from robust background detection. **Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition**, IEEE, p. 2814–2821, 2014. DOI: 10.1109/CVPR.2014.360. Citado nas pp. 23, 41, 42, 46.