Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Weslley dos Santos Silva

Avaliação do impacto da Forma, Textura e Campo Receptivo na Segmentação de Vasos Sanguíneos

São Carlos
2025

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Weslley dos Santos Silva, realizada em 15/09/2025.

Comissão Julgadora:

Prof. Dr. Cesar Henrique Comin (UFSCar)

Prof. Dr. Alexandre Luis Magalhães Levada (UFSCar)

Prof. Dr. Diego Saqui (UFLA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Weslley dos Santos Silva

Avaliação do impacto da Forma, Textura e Campo Receptivo na Segmentação de Vasos Sanguíneos

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Visão Computacional

Orientador: César Henrique Comin

São Carlos
2025

Este trabalho é dedicado a Araci Pereira dos Santos, Renata dos Santos e Marcos Aurélio dos Santos.

Agradecimentos

Aos meus familiares. Ao meu companheiro, Éberton, por cuidar, motivar e dar suporte à mim todos os dias para que fosse possível chegar até aqui. A Jéssica e Guilherme por sempre me darem apoio e me recordar dos meus feitos por uma perspectiva otimista e de muito carinho. Aos meus amigos Ellen, Giovanni, Semira, Juliano, Matheus, Bruna e Karen por estarem ao meu lado durante os desafios de desenvolver este trabalho. Ao meu orientador, César Henrique Comin, pela oportunidade e confiança em me permitir ter sempre muita autonomia mesmo em um projeto tão desafiador, pelo acompanhamento próximo e por sua paciência, dedicação e disponibilidade em cada etapa deste trabalho. Seu apoio e orientação foram fundamentais para a realização desta pesquisa. A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro, que tornou possível a realização deste trabalho.

"Cê vai atrás desse diploma,
com a fúria da beleza do Sol, entendeu?
Faz isso por nós, faz essa por nós,
te vejo no pódio!"
(Emicida - Amarelo)

Resumo

A segmentação de vasos sanguíneos em imagens médicas é um passo crucial para o diagnóstico de diversas patologias, mas apresenta desafios únicos devido à complexidade e à variabilidade das estruturas presentes nestas. Além disso, a falta de interpretabilidade das Redes Neurais Convolucionais (CNNs) representa um obstáculo significativo à adoção clínica, pois dificulta a depuração dos modelos, limita melhorias no design e pode reduzir a confiança no diagnóstico. Neste estudo, realizamos experimentos sistemáticos para avaliar separadamente a contribuição da forma e da textura de vasos sanguíneos e também do campo receptivo de CNNs para a segmentação de tecidos vasculares. Primeiramente, para avaliar o papel da textura e da intensidade dos pixels, removemos seletivamente essas características em recortes de imagens utilizando embaralhamento e normalização de pixels. Em seguida, para isolar a influência da forma, treinamos modelos de segmentação utilizando apenas os contornos externos ou linhas centrais dos vasos como entrada, eliminando informações internas da textura. Por fim, investigamos a quantidade de contexto necessária para uma segmentação robusta, variando sistematicamente o campo receptivo teórico da rede. Os resultados mostraram que, embora a intensidade dos pixels seja mais relevante que a textura dos vasos, as CNNs conseguem manter alto desempenho mesmo na ausência das duas características. Além disso, as redes não conseguem extrapolar facilmente a forma completa dos vasos utilizando apenas seus contornos ou eixos centrais como entrada. Por fim, verificamos que o campo receptivo efetivo utilizado pelas redes é de aproximadamente 20 pixels nos conjuntos de dados empregados nos experimentos, embora informações globais possam levar a uma pequena melhora na acurácia em imagens de fundo de olho.

Palavras-chave: Segmentação de vasos sanguíneos, interpretabilidade, redes neurais, campo receptivo.

Abstract

Blood vessel segmentation in medical images is a crucial step for the diagnosis of a wide range of pathologies, but it presents unique challenges due to the complexity and variability of the structures involved. Moreover, the lack of interpretability of Convolutional Neural Networks (CNNs) represents a significant barrier to clinical adoption, as it complicates model debugging, limits design improvements, and can reduce diagnostic confidence. In this study, we conducted systematic experiments to separately evaluate the contributions of shape, texture, and receptive field to the performance of CNNs in blood vessel segmentation. First, to assess the role of texture and pixel intensity, these features were selectively removed from image patches using pixel shuffling and normalization. Next, to isolate the influence of shape, segmentation models were trained using only the vessels' outer contours or centerlines as input, thereby removing internal texture information. Finally, we investigated the amount of context required for robust segmentation by systematically varying the network's theoretical receptive field. We found that, although pixel intensity is more relevant than vessel texture, CNNs can still maintain high performance even in the absence of both features. In addition, the networks are not able to easily extrapolate the full vessel shape when provided only with contours or centerlines as input. Finally, we observed that the effective receptive field used by the networks is approximately 20 pixels in the datasets employed, although global information can lead to a slight improvement in accuracy for fundus images.

Keywords: Blood vessel segmentation, interpretability, neural networks, receptive field.

Lista de ilustrações

Lista de tabelas

Lista de siglas

CNN Rede Neural Convolucional

ERF Campo Receptivo Efetivo

FLOP Operações de Ponto Flutuante

HDC Rede Convolucional Dilatada Híbrida

ILSVRC ImageNet Large-Scale Visual Recognition Challenge

ReLU Unidade Linear Retificada

RNA Rede Neural Artificial

Resnet Rede Neural Residual

RF Campo Receptivo

TRF Campo Receptivo Teórico

Sumário

Capítulo 1

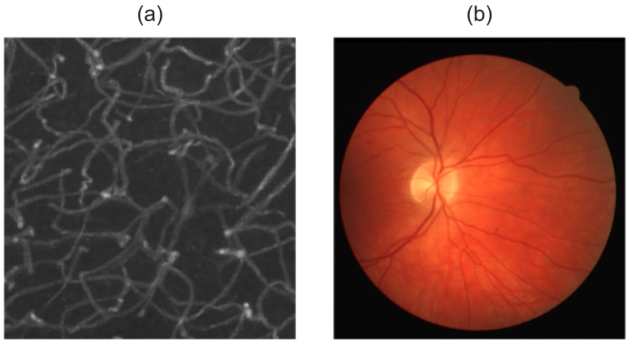
Introdução

A classificação e segmentação automática de vasos sanguíneos têm demonstrado resultados eficientes e promissores no campo de estudo de imagens médicas há alguns anos, principalmente a partir das redes neurais convolucionais (CNNs) (WANG et al., 2015; KASSIM et al., 2017). Na literatura, é possível identificar importantes metodologias de processamento de imagens e aprendizado de máquina produzidas para o estudo de vasos sanguíneos (MOCCIA et al., 2018; KIRST et al., 2020; TODOROV et al., 2020; MUTUA; KASAMANI; REICH, 2025). É notável também que as redes neurais artificiais despontaram como o estado da arte em muitas tarefas de segmentação de imagens biomédicas (ISENSEE et al., 2021; ISENSEE et al., 2024; MA et al., 2024).

Treinar um modelo que seja invariante à mudanças de forma, textura e ao contraste de captura das imagens é imprescindível para um modelo robusto e capaz de ser aplicado em diagnósticos médicos. Por exemplo, modelos de segmentação de vasos sanguíneos utilizados para a análise de patologias e processos fisiológicos (MOOKIAH et al., 2021; CHEN et al., 2023; LI et al., 2022). As características visuais dos vasos sanguíneos podem variar de forma significativa. Eles apresentam uma combinação de atributos distintos, incluindo suas formas alongadas e ramificadas, sua textura interna e o contraste de intensidade em relação ao tecido. Por exemplo, grandes artérias em uma imagem de fundo de olho podem estar claramente delineadas, enquanto capilares finos em uma imagem de microscopia ruidosa podem ser dificilmente distinguíveis do fundo (Figura 1). Um modelo que favoreça excessivamente a textura, por exemplo, pode ser facilmente confundido por artefatos de imagem, ao passo que um que dependa em demasia de uma forma prototípica aprendida pode falhar na segmentação de vasos com morfologias anormais.

Ainda que esta tarefa seja complexa, entender a forma de ponderação e o aprendizado das informações pode minimizar erros ocasionados por pixels espúrios, descontinuidade

Figura 1 – Exemplos de imagens de vasos sanguíneos sob características de captura distintas.



(a) Imagem de microscopia confocal de vasos do córtex de camundongos, na qual pode ser difícil distinguir os vasos do fundo devido ao ruído presente na captura. (b) Imagem de fundo de olho, ao qual os vasos sanguíneos estão mais visíveis e delineadas em relação ao fundo. Fonte: Adaptado de Silva et al. (2025); Staal et al. (2004).

de vasos sanguíneos, artefatos indesejados ou regiões com pouca diferença de contraste. Adicionalmente, pode guiar as escolhas de pré-processamento e os parâmetros de treino conforme os padrões observados nas imagens disponíveis para treinamento das tarefas. Entretanto, apesar de sua impressionante acurácia, esses modelos são frequentemente tratados como caixas-pretas, e muitas vezes falta um entendimento profundo das características visuais em que se baseiam para a tomada de decisão (RUDIN, 2019). Essa lacuna de interpretabilidade possui importantes implicações práticas, dificultando esforços para depurar modelos, aprimorar sua arquitetura e garantir robustez frente a variações nos dados. De forma crítica, para a adoção clínica, a falta de transparência pode reduzir a confiança no diagnóstico e gerar desafios para a aprovação regulatória (NAZIR; DICKSON; AKRAM, 2023).

Uma discussão relevante na comunidade de visão computacional em geral gira em torno do viés forma versus textura. Geirhos et al. (2018) mostraram que CNNs treinadas em conjuntos de imagens naturais em larga escala, como o ImageNet (DENG et al., 2009), apresentam um forte viés para fazer uso das características locais de textura para o reconhecimento de objetos — uma estratégia diferente da visão humana, que prioriza o formato global de objetos.

Além da forma e textura, o campo receptivo de CNNs desempenha papel fundamental na obtenção das informações, seja de forma localizada ou global. Para que uma informação presente no conjunto de dados de entrada seja aprendida pela CNN é necessário que ela seja abrangida pelo seu campo receptivo. Em especial, o campo receptivo efetivo (ERF) das CNNs (LUO et al., 2016) define a região de entrada efetivamente utilizada para segmentar cada pixel. O ERF determina a quantidade de contexto espacial disponível para que o modelo possa aprender e pode variar conforme os parâmetros da rede e/ou as

características do conjunto de dados.

Embora as CNNs sejam capazes de aprender a identificar vasos de maneira eficaz, nem sempre está claro como elas ponderam a importância relativa dos diferentes atributos que permeiam o aprendizado de tarefas. O campo receptivo das CNNs, bem como os parâmetros da sua arquitetura, influencia de forma significativa o desempenho na classificação e segmentação de vasos sanguíneos. Entretanto, como afirmar se uma CNN depende principalmente da textura local, da forma global da vascularização capturada, de um limiar específico de intensidade do pixel avaliado ou de um contexto mínimo que permita que a tarefa tenha um bom desempenho?

Nesse contexto, conduzimos uma investigação em três partes, projetada para avaliar de forma sistemática a influência da textura, da forma e do campo receptivo da rede sobre o desempenho da segmentação. Primeiro, para examinar o papel da textura e da intensidade, avaliamos o impacto de remover das amostras as informações sobre a intensidade dos pixels — normalizando os valores para média zero e variância unitária — bem como a relação espacial entre os pixels, embaralhando aleatoriamente suas posições. Em seguida, para identificar a capacidade de uma CNN de segmentar vasos utilizando apenas características de forma, treinamos um modelo para segmentar vasos usando apenas seu contorno externo ou eixo central dos vasos como entrada, ocultando do modelo todas as características internas de textura. Por fim, investigamos a quantidade de contexto necessária para uma segmentação precisa, variando sistematicamente o campo receptivo teórico da rede por meio da alteração do tamanho e da taxa de dilatação dos filtros do modelo. Seguindo o mesmo objetivo, definimos tamanhos específicos de campo receptivo ao dividir as imagens de entrada em recortes não sobrepostos e treinar um modelo de segmentação nesses recortes, o que permite medir diretamente o impacto do desempenho em um campo de visão limitado.

1.1 Objetivos

Objetivos Gerais:

Este estudo tem como objetivo o desenvolvimento de metodologias para a quantificação da influência da forma, textura e do campo receptivo de CNNs na segmentação de vasos sanguíneos em imagens biomédicas.

Objetivos específicos:

1. Explorar e aplicar técnicas para compreender e interpretar a relação entre as características de modelos de CNNs e a tarefa de segmentação de vasos sanguíneos.

2. Avaliar o impacto de perturbações da informação de textura e intensidades dos pixels na tarefa de segmentação de vasos utilizando CNNs.

3. Avaliar a eficácia de uma CNNs ao segmentar vasos sanguíneos considerando dados estruturais da forma, isto é, sem a presença de textura.

4. Avaliar como a variação de parâmetros da rede, e a consequente alteração do campo receptivo, impactam o desempenho de segmentação de CNNs.

5. Investigar oportunidades de otimização da segmentação por CNNs, buscando ganho de desempenho com diminuição da complexidade dos modelos.

6. Avaliar o impacto de perturbações da informação de textura e intensidades dos pixels na tarefa de segmentação de vasos utilizando CNNs.

Os demais capítulos deste trabalho estão estruturados da seguinte maneira. No Capítulo 2, apresentamos uma contextualização dos principais conceitos presentes neste trabalho. No Capítulo 3, apresentamos uma visão geral do estado da arte relacionado aos estudos aqui conduzidos. No Capítulo 4, descrevemos os conjuntos de dados utilizados nas análises, bem como uma descrição detalhada dos três experimentos mencionados anteriormente. Os resultados de cada experimento são apresentados no Capítulo 5. No Capítulo 6 apresentamos a conclusão do estudo.

Capítulo 2

Conceitos Básicos

2.1 Aprendizado Profundo

O aprendizado profundo utiliza conceitos inspirados no funcionamento de neurônios biológicos para a construção das camadas pertencentes a uma Rede Neural Artificial (RNA) (LEK; PARK, 2008). As camadas de uma RNA são divididas em três tipos: camada de entrada, camadas ocultas e camada de saída. A camada de entrada recebe os dados que são passados às camadas ocultas, compostas por uma ou mais camadas intermediárias responsáveis pelo processamento dos dados e variam em quantidade a depender da arquitetura empregada na rede. A cada camada, as informações são propagadas dos neurônios das camadas anteriores para as camadas seguintes, alimentando a informação camada a camada através de conexões ponderadas pelos pesos da rede, até a camada de saída que produz o resultado da rede.

O processo de aprendizado profundo ocorre pelo ajuste iterativo dos pesos ao longo das camadas ocultas (RUMELHART; HINTON; WILLIAMS, 1986), utilizando algoritmos de otimização, como o gradiente descendente. Nesse método, os pesos da rede são atualizados de forma a minimizar uma função de perda, movendo-se na direção oposta ao gradiente do erro em relação aos parâmetros do modelo. O cálculo desses gradientes é realizado por meio do algoritmo de retropropagação do erro (*backpropagation*), que aplica a regra da cadeia para propagar o erro da saída em direção às camadas anteriores, ajustando os pesos proporcionalmente à sua contribuição para o erro total. Esse processo iterativo permite que a rede aprenda representações hierárquicas dos dados e refine progressivamente seus parâmetros até convergir para uma solução que minimize o erro de predição.

O objetivo do treinamento de uma RNA é minimizar uma função de perda que quantifica o erro da saída da rede, e assim obter um processamento mais preciso. Um exemplo

comum é a função de entropia cruzada, amplamente empregada em tarefas de classificação para medir a diferença entre a distribuição prevista pela rede e a distribuição verdadeira das classes (ZHANG; SABUNCU, 2018). Ademais, o treinamento visa garantir que a rede seja capaz de generalizar seu aprendizado para dados diferentes dos utilizados no treinamento, evitando um sobreajuste (*overfitting*) do modelo ao conjunto de dados, possibilitando alto desempenho em um novo conjunto de dados.

O aprendizado profundo permitiu realizar a extração e classificação de distintas características presentes nos dados de diferentes áreas da computação, por exemplo, a detecção de objetos em imagens na visão computacional (HINTON et al., 2012), a tradução automática de textos no processamento de linguagem natural (MIKOLOV et al., 2013) e reconhecimento de fala (GRAVES; MOHAMED; HINTON, 2013).

2.2 CNN

Na visão computacional, as CNNs obtiveram destaque no processo de classificação de imagens a partir da conhecida AlexNet (RUSSAKOVSKY et al., 2015). Em 2012, a rede proposta por Alex Krizhevsky atingiu acurácia de 84,7% ficando em primeiro lugar na competição anual chamada ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) (RUSSAKOVSKY et al., 2015).

CNNs são constituídas de uma série de camadas convolucionais, que representam transformações equivariantes dos dados. Em uma camada equivariante, transformações geométricas aplicadas na entrada dos dados correspondem a uma transformação proporcional nos dados de saída. Esta propriedade é desejável quando se trata de redes aplicadas na segmentação de objetos (GHOSH et al., 2019).

Além das convoluções, são aplicadas amostragens (*pooling*) que introduzem o princípio de invariância às CNNs. A invariância refere-se à capacidade da saída da rede permanecer inalterada diante de transformações, como translações, rotações ou mudanças de escala de um objeto. Uma vez que as CNNs precisam se manter robustas independentemente da transformação geométrica que os pixels possam ser submetidos, ou seja, identificar o objeto de forma independente da posição ou rotação que esse se encontra na imagem (GHOSH et al., 2019). Assim, as CNNs são capazes de aprender a relação entre um pixel e sua vizinhança, extrair características dos dados de forma local e serem invariantes ou equivariantes à escala, translação, rotação e outras transformações (JARRETT et al., 2009).

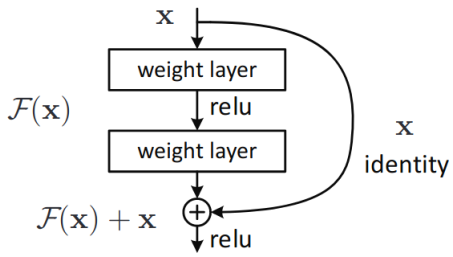
2.2.1 Redes Neurais Residuais

O treinamento por aprendizado profundo tende a apresentar problemas à medida que a profundidade de uma rede aumenta. Os gradientes calculados durante a retropropagação

podem se tornar excessivamente pequenos, causando um desaparecimento de gradiente, ou excessivamente grandes, ocasionando o problema conhecido como gradiente explosivo. Ambos os problemas comprometem o processo de ajuste dos pesos e dificultam o aprendizado e convergência dos modelos. A Rede Neural Residual (Resnet) (HE et al., 2016) representa uma arquitetura com o objetivo de minimizar o problema da degradação de desempenho no aprendizado de redes neurais profundas. Essa abordagem possibilitou o treinamento de redes substancialmente mais profundas, com centenas de camadas como por exemplo as ResNet-101 e ResNet-152.

A metodologia Resnet introduz atalhos no aprendizado a partir dos dados de entrada de camadas iniciais somados à entrada das camadas posteriores da rede. Os blocos residuais, como são chamados, são unidades básicas compostas por camadas convolucionais da rede. A Figura 2 mostra um exemplo de bloco residual. A entrada do bloco x é utilizada em uma camada convolucional seguida da função de ativação Unidade Linear Retificada (ReLU) e outra camada convolucional. A saída desse processo, *F(x)*, é somada à entrada original por meio de uma conexão de "salto". Ao final, uma nova aplicação da ativação ReLU é utilizada para inserção de não linearidade ao processo.

Figura 2 – Ilustração de um Bloco Residual.



Fonte: He et al. (2016)

2.2.2 Redes U-Net

Em geral, CNNs exigem uma grande quantidade de dados anotados para que o treinamento seja bem-sucedido. A demanda por técnicas que sejam mais eficientes frente a esse problema levou à arquitetura de rede neural U-net, uma arquitetura completamente convolucional (RONNEBERGER; FISCHER; BROX, 2015). A arquitetura U-net é comumente utilizada na segmentação de imagens médicas, especialmente pela anotação manual ser um processo caro e demorado e a arquitetura possibilitar bons resultados com uma quantidade de dados reduzida.

A rede U-net consiste em uma fase de contração e uma fase de expansão dos dados de entrada. A contração dos dados ocorre em um modelo comum de CNNs começando a partir da aplicação repetida de duas convoluções 3x3, cada uma seguida por uma aplicação

ReLU, e a aplicação de uma camada de *max-pooling* 2x2 com passo 2 para redução da imagem. A cada redução da resolução, o número de canais é dobrado (RONNEBERGER; FISCHER; BROX, 2015).

A fase expansiva da rede U-Net, principal característica da sua arquitetura, começa com um *upsampling* do mapa de características por meio de uma convolução ascendente 2x2, aumentando a resolução espacial e reduzindo pela metade o número de canais de características. O mapa ampliado é então concatenado com um mapa de características recortado da fase de contratação para alinhar as resoluções das ativações. Em seguida, duas convoluções 3x3 seguidas por funções de ativação ReLU são realizadas para extrair características e reduzir progressivamente os canais. Esse ciclo de *upsampling*, concatenação e convolução dupla é repetido ao longo da fase expansiva para restaurar a resolução original e manter as características extraídas. Na última etapa, uma convolução 1x1 mapeia as ativações, reduzindo o número de canais para o número desejado de classes da segmentação, produzindo estimativas para cada pixel da entrada original (RONNEBERGER; FISCHER; BROX, 2015).

2.3 Campo Receptivo de CNNs

As CNNs são capazes de aprender e representar características complexas a partir dos dados em sua entrada. O aprendizado de características realizado por essas redes começa mapeando características de forma local e expandindo o aprendizado de características à medida que a rede se torna mais profunda.

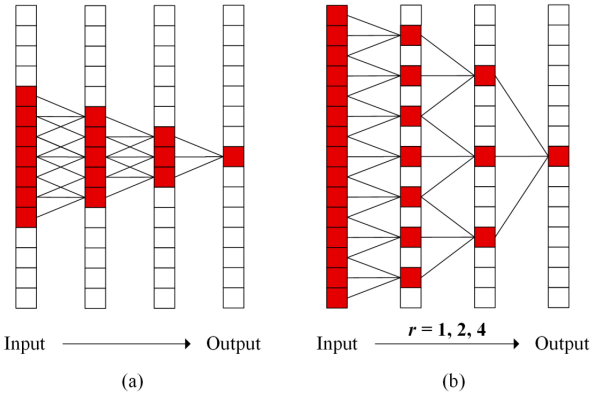
Cada neurônio na primeira camada oculta de uma CNN está conectado com uma pequena parcela dos pixels contidos na camada de entrada. Esse particionamento da análise da imagem de entrada é chamado de campo receptivo local e é processado por meio dos filtros de convoluções em CNNs. O percorrimento dos filtros ao longo da camada de entrada tem o comportamento muito similar ao de se observar uma figura através de uma pequena janela, sendo essa janela o nosso campo receptivo. Cada visualização feita por essa janela dá origem a uma ativação que será processada pela camada posterior. Assim, a camada de entrada é percorrida de janela em janela até que a imagem tenha sua codificação completa na camada oculta à sua frente.

Em uma CNNs os pesos de ponderação conhecidos em RNAs são compartilhados e utilizados para o cálculo da resultante no neurônio da camada seguinte. Esses pesos recebem o nome de pesos compartilhados justamente por serem os mesmos ao longo de uma mesma camada oculta.

Sendo assim, para cada campo receptivo local, existe um neurônio na camada oculta seguinte que recebe o resultado ponderado das informações recebidas das camadas anteriores a ele. Na Figura 3(a) podemos observar que, conforme as redes se tornam mais profundas, os campos receptivos aumentam de tamanho à medida que as informações são

propagadas através das camadas, dando a cada neurônio uma visão cada vez mais global da informação contida na camada de entrada, podendo chegar ao estágio em que cada neurônio tem acesso a toda a informação contida em uma imagem.

Figura 3 – Campo receptivo em CNNs.



(a) Uma CNN de três camadas com operações de convolução convencional 1D utilizando filtros de tamanho 3. (b) Uma CNN de três camadas usando operações de convolução dilatada 1D com taxa de dilatação crescente exponencialmente (r = 1, 2, 4). Fonte: (HAO et al., 2020)

2.3.1 Operações com impacto no campo receptivo: Filtro, Dilatação, Passo

O tamanho do campo receptivo de uma CNN pode ser aumentado de outras formas além do número de camadas da rede. O tamanho do filtro, a dilatação da convolução e o passo utilizados são importantes para o campo receptivo.

Um filtro de convolução maior oferece um campo receptivo proporcionalmente maior, entretanto, o aumento do número de parâmetros treináveis torna a otimização dos pesos mais difícil (DONG; XIE; LI, 2023). Simonyan e Zisserman (2014) demonstraram que o empilhamento de três camadas de filtros 3x3 pode atingir o mesmo campo receptivo de uma camada com filtro de 7x7. Além do empilhamento de múltiplas camadas incorporar maior não linearidade aos dados por meio das funções de ativação, aumentando o poder discriminatório do modelo.

A dilatação da convolução por sua vez possibilita ampliar o campo receptivo sem aumentar o número de parâmetros de convolução ou modificação da resolução da imagem. A vantagem em relação ao aumento do tamanho dos filtros é o menor custo computacional durante o treinamento, visto que não se tem um aumento do número de parâmetros. Convoluções dilatadas podem ser vantajosas no processo de segmentação de imagens por

manterem a resolução da entrada até a saída da rede, não sendo necessário adicionar camadas de *upsampling* à rede.

A razão de dilatação específica o espaçamento entre cada elemento do filtro no cálculo da convolução. À medida que a taxa de dilatação aumenta, o campo receptivo aumenta, conforme demonstrado na Figura 3(b). Assim, enquanto convoluções dilatadas ganham informações no contexto amplo da entrada, podem ser limitadas para o aprendizado de pequenos detalhes de imagens. O aumento da razão de dilatação pode implicar que os pixels sejam muito distantes, diminuindo o aprendizado de características locais da imagem.

É comum nas abordagens de aprendizado profundo a aplicação de técnicas que realizem a sumarização dos dados, no caso de imagens a sumarização é feita através da redução da resolução. A redução da resolução pode ser benéfica tanto por permitir à rede aprender características dos objetos em tamanhos distintos quanto por reduzir o número de Operações de Ponto Flutuante (FLOPs) realizadas pela rede, o que torna o treinamento menos custoso computacionalmente.

Assim, são frequentemente utilizadas camadas de *pooling* para diminuir a resolução das imagens de entrada. Essa técnica realiza o agrupamento do valor de múltiplos pixels da janela observada em um único pixel na camada seguinte, normalmente calculando a média (*avarege-pooling*) ou o máximo (*max-pooling*) entre eles. Semelhantemente, o aumento do passo da convolução, do inglês *stride*, define quantos pixels serão pulados entre os avanços do filtro. É comum em convoluções 3x3 o uso de $passo = 2$, reduzindo pela metade a resolução da imagem de entrada.

Entretanto, a redução da resolução pode levar à perda de informações. Nas camadas de *max-pooling*, por exemplo, apenas características de destaque são transmitidas para camadas seguintes. Ainda que a sumarização contribua para um aumento expressivo do campo receptivo nas CNNs e reduza o custo computacional de FLOPs é preciso cautela em imagens médicas, uma vez que informações podem ser perdidas. A Tabela 1 resume as principais implicações das técnicas de aumento do campo receptivo em termos de custo computacional e capacidade de manutenção de informação ao longo da segmentação.

Tanto as características biológicas quanto as técnicas de captura de imagens precisam ser levadas em conta no estudo de imagens biomédicas, em especial de vasos sanguíneos. Vasos apresentam características distintas de comprimentos, espessuras e porções de aglomeração. Por sua vez, as técnicas de captura dos vasos possuem limitações em relação a contrastes, oclusão de vasos, resolução e altos custos financeiros para obtenção das imagens. Portanto, é necessário o manejo adequado das técnicas para possibilitar o aumento do campo receptivo e evitar a perda de informação e aumento do custo computacional para modelos de segmentação de vasos sanguíneos. Este trabalho visa comparar a eficácia dessas técnicas para a segmentação de tecidos vasculares.

Tabela 1 – Avaliação qualitativa de técnicas que impactam o campo receptivo de CNNs.

Técnica	**Custo computacional**	**Número de parâmetros**	**Informação sobre vasos finos**
Número de camadas	Aumento considerável	Aumento considerável	Manutenção
Tamanho de filtro	Aumento	Aumento considerável	Manutenção
Dilatação	Igual	Igual	Possível perda
Passo	Redução	Igual	Possível perda

Fonte: Do Autor

Capítulo 3

Revisão Bibliográfica

A interpretação de imagens médicas apresenta desafios únicos devido à complexidade e à variabilidade das estruturas presentes nestas. Sendo assim, é um campo em constante evolução com avanços significativos impulsionados pela visão computacional. Nesse sentido, a capacidade de identificar informações contextuais em diferentes escalas é essencial para uma análise precisa e abrangente em imagens médicas.

Redes neurais convolucionais como *ResNet*, *GoogleNet* e *DeepLab*, têm revolucionado a compreensão de imagens naturais desde a classificação até a segmentação. A arquitetura do *GoogleNet*, também conhecida como *Inception*, introduziu módulos de convolução chamados de "inception modules", que combinam filtros de diferentes tamanhos em paralelo, permitindo capturar informações em várias escalas espaciais, o que possibilita a extração de características complexas em diferentes níveis de abstração (SZEGEDY et al., 2014). A *DeepLab* por sua vez, é uma arquitetura projetada para a segmentação semântica de imagens naturais. Sua principal contribuição para o avanço do estado da arte se dá pelo uso de convoluções *atrous*, que são convoluções dilatadas. Tais convoluções permitem aumentar o campo receptivo sem aumentar significativamente o número de parâmetros, melhorando assim a eficiência computacional e preservando a resolução espacial dos dados. Esta abordagem tem sido fundamental para impulsionar o estado da arte em segmentação de imagens, em especial por ter a capacidade de introduzir informação de contexto e diferentes escalas no processo de aprendizado e possibilitar o aumento do campo receptivo das redes sem o aumento do custo computacional.

A tomada de decisão de uma CNN é majoritariamente influenciada pelo seu campo receptivo (RF), ou seja, a região de entrada que afeta sua saída. Estudos demonstram que o campo receptivo efetivo (ERF) de uma CNN é consideravelmente menor do que o campo receptivo teórico (TRF) (LUO et al., 2016; ZHANG; MAZUROWSKI, 2024). No contexto

de imagens médicas, há evidências de que o tamanho do RF está fortemente relacionado ao desempenho do modelo (BEHBOODI et al., 2020; SYTWU; GROSCHNER; SCOTT, 2022). Além disso, as investigações sobre o campo receptivo para tarefas de segmentação semântica em imagens naturais são direcionadas para a manipulação dos parâmetros operacionais das CNNs de forma a obter maior performance com menor impacto ao custo computacional. Wang et al. (2018) por exemplo, propõe uma Rede Convolucional Dilatada Híbrida (HDC). A HDC proposta pelo autor tem como característica a capacidade de aumentar o campo receptivo de forma eficiente, agregando assim informações globais do conjunto de entrada sem que um efeito de grade, do inglês *gridding*, ocorra na obtenção de informações pela rede. A HDC utiliza uma série de convoluções com taxas de dilatação distintas, concatenando em série essas convoluções como em blocos de redes residuais (WANG et al., 2018). Os autores destacam que a escolha adequada das taxas de dilatação pode aumentar o campo receptivo eficientemente e melhorar a acurácia para segmentação de objetos relativamente grandes.

Por sua vez, a rede nnU-net, uma das arquiteturas amplamente utilizadas na segmentação de imagens biomédicas, apresenta uma forma de automatizar a escolha dos parâmetros de redes neurais utilizadas para segmentação de imagens 2D e 3D (ISENSEE et al., 2021). A abordagem utiliza uma estrutura de aprendizagem profunda que, de forma autônoma, toma as principais decisões necessárias para transferir uma arquitetura básica para diferentes conjuntos de dados a partir de critérios de inferência de parâmetros. Os parâmetros inferidos codificam as adaptações de acordo com o conjunto de dados e modificam a topologia da rede. Um dos critérios dessa inferência de parâmetros consiste na adaptabilidade da arquitetura ao tamanho espacial do conjunto de treinamento. Os autores definem como um critério para essa adaptação espacial do modelo a necessidade de assegurar que o campo receptivo da rede cubra totalmente a imagem de entrada (ISENSEE et al., 2021).

Notavelmente, os modelos baseados em aprendizagem profunda são muito promissores em uma ampla gama de tarefas em imagens médicas, desde a segmentação de estruturas anatômicas até a identificação de regiões patológicas (ANTONELLI et al., 2022). Ainda assim, os modelos atuais de segmentação de imagens médicas possuem como limitação a natureza específica da tarefa aprendida. Esses modelos podem apresentar um desempenho inferior quando aplicados em um problema de segmentação distinto ou até mesmo em tipos de imagens diferentes das utilizadas no treinamento (MA et al., 2024). Para resolver tal problema Ma et al. (2024) desenvolveu uma abordagem que apresenta alto desempenho em diferentes tarefas de segmentação de imagens biomédicas.

O modelo, chamado *MedSAM*, foi desenvolvido a partir de um grande conjunto de imagens biomédicas composto por aproximadamente 1,5 milhões de pares imagens-máscaras, abrangendo mais de 30 tipos de câncer e 10 distintos tipos de imagens. O modelo é configurado para dar maior prioridade às regiões dentro de caixas delimitadoras. As caixas

delimitadoras são desenhadas ao redor das regiões de interesse nas imagens. O *MedSAM* apresenta um potencial significativo no avanço do estado da arte. Entretanto, ainda que a abordagem projetada tenha tido êxito em dados de características heterogêneas, a segmentação de vasos sanguíneos não está inclusa entre os sucessos da rede. O texto destaca como limitação a dificuldade encontrada na segmentação de vasos sanguíneos a partir de regiões de interesse. A principal justificativa diz respeito à estrutura ramificada de vasos e artérias que compartilhariam a mesma caixa delimitadora em imagens de fundo de olho, por exemplo.

A busca por arquiteturas inovadoras focadas em ganhos de desempenho tem gerado modelos cada vez mais complexos, capazes de localizar vasos com precisão, mas oferecendo pouca compreensão sobre a classificação de cada pixel. Essa falta de interpretabilidade representa uma barreira à aplicação clínica, motivando a priorização de desempenho em detrimento à interpretabilidade dos modelos (CHADDAD et al., 2023; NAZIR; DICKSON; AKRAM, 2023; SINGH; SENGUPTA; LAKSHMINARAYANAN, 2020). O trabalho de Geirhos et al. (2018) demonstrou que as CNNs apresentam um forte viés em função da textura local para o reconhecimento de objetos. Esses resultados incentivaram novas pesquisas focadas na quantificação e manipulação do viés de textura (DAI et al., 2022; HEINERT et al., 2024; TRIPATHI et al., 2023). Frequentemente, esses trabalhos utilizam técnicas de transferência de estilo para criar perturbações das características originais ou treinamento em conjuntos de dados estilizados para forçar os modelos a dependerem da forma. Estudos recentes (CHUNG; PARK, 2022; ISLAM et al., 2021; HERMANN; CHEN; KORNBLITH, 2020) defendem que um modelo ideal não deve focar em uma das características, mas sim encontrar um equilíbrio que faça sentido para a tarefa específica que está sendo realizada. Esse pensamento é particularmente importante em imagens médicas, onde tanto a forma quanto a textura da vasculatura possuem informações importantes para diagnósticos.

Uma abordagem de remoção da textura utilizando um filtro *mean shift* foi proposta por Dai et al. (2022), técnica que realiza suavização de imagens substituindo cada pixel pela média dos valores vizinhos (FUKUNAGA; HOSTETLER, 1975). Eles também utilizaram perturbações de forma que desfocam os contornos dos objetos e uma técnica de remoção de topologia implementada como embaralhamento de recortes da imagem. Procedimentos semelhantes foram adotados por Heinert et al. (2025). No entanto, as técnicas propostas não eliminam completamente as características de interesse.

Mütze et al. (2024) introduziram decomposições de características semelhantes às definidas em nosso trabalho. Na decomposição de textura, emprega-se um diagrama de Voronoi, com células preenchidas por texturas geradas a partir de recortes pertencentes a uma mesma categoria de anotação. Na decomposição das características de forma, foi aplicado o método *Holistically-Nested Edge Detection* (HED)(XIE; TU, 2015) para detectar as bordas dos objetos, que então são usadas como entrada para treinar a rede. No

entanto, os métodos propostos não se aplicam a vasos sanguíneos, devido à sua estrutura fina e bordas suaves. Nosso estudo gera as características por meio de manipulações mais diretas para remover características de intensidade e textura.

Loos, Pardasani e Awasthi (2024) realizaram uma larga investigação que se relaciona bem com nosso trabalho. Eles modificaram a profundidade e os tamanhos de filtro de um modelo U-Net para controlar sistematicamente o campo receptivo e avaliaram o modelo em dois conjuntos de dados médicos sintéticos e seis conjuntos de dados reais. Os autores observaram que o tamanho do RF é especialmente relevante para conjuntos de dados com baixo contraste entre as classes alvo e fundo, e que aumentar o campo receptivo além do tamanho dos objetos segmentados quase não traz ganhos de desempenho. A principal diferença em relação à nossa análise é que consideramos uma tarefa de segmentação semântica, na qual os vasos não apresentam uma estrutura global predefinida, e incorporamos convoluções dilatadas nas variações do modelo — uma estratégia comum para ampliar o contexto sem aumentar o número de parâmetros.

Quando se trata de vasos sanguíneos, é necessário que um modelo possua RF amplo para garantir a continuidade dos vasos, ao mesmo tempo em que se mantém boa precisão local para delinear bordas finas. Características da arquitetura de um modelo, como convoluções dilatadas e subamostragem, podem contribuir para o balanço entre esses dois aspectos.

Até onde se tem conhecimento, nenhum estudo prévio realizou experimentos sistemáticos para avaliar separadamente a contribuição da forma, da textura e do campo receptivo no desempenho de CNNs para a segmentação de vasos sanguíneos. Nosso estudo busca preencher essa lacuna por meio da realização de uma série de experimentos direcionados a este desafio.

Capítulo 4

Metodologia

Este capítulo descreve os conjuntos de dados, ferramentas e processos utilizados no desenvolvimento da pesquisa deste trabalho. O principal objetivo das seções presentes no capítulo é apresentar as justificativas da metodologia empregada e garantir o entendimento dos experimentos propostos de forma que a reprodutibilidade seja garantida. O capítulo está estruturado da seguinte maneira: inicialmente, são apresentados os conjuntos de dados utilizados no desenvolvimento deste trabalho e posteriormente, são descritos os experimentos conduzidos, conforme lista abaixo.

1. Análises referentes à relevância da informação de textura e intensidades na classificação
    - Análise exploratória
    - Classificação sem informação de borda
    - Classificação com informação de borda
    - Perturbação seletiva de informações de textura e intensidade
2. Análise da influência da forma para a segmentação de vasos sanguíneos
    - Abordagem com Contornos e Esqueleto
3. Análises referentes ao impacto do nível de contexto na qualidade da segmentação
    - Variação seletiva dos parâmetros da rede
    - Treinamento para segmentação a partir de recortes

4.1 Descrição de hardware e bibliotecas utilizadas

Os experimentos foram conduzidos em um computador de alto desempenho equipado com uma GPU NVIDIA GeForce RTX 3080, utilizada para acelerar o processamento das operações matriciais envolvidas no treinamento e na inferência das redes neurais. O ambiente de execução foi configurado em sistema operacional baseado em Linux, com suporte às bibliotecas NumPy, Pandas, Scikit-learn, Scikit-image, Pillow e Matplotlib, empregadas para manipulação, análise e visualização dos dados. Todo o material relacionado às publicações e códigos-fonte encontra-se disponível em um repositório público[1].

4.2 Descrição dos dados utilizados

Os experimentos utilizaram os conjuntos de dados *VessMAP* (SILVA et al., 2025) e *DRIVE* (STAAL et al., 2004). As classes consideradas neste estudo correspondem aos pixels de vasos sanguíneos e aos pixels de fundo das imagens. Optou-se por não realizar o balanceamento das classes, mantendo-se a proporção original de pixels presente nos conjuntos de dados, a fim de garantir a comparabilidade com experimentos prévios, como no caso do modelo W-Net (GALDRAN et al., 2022).

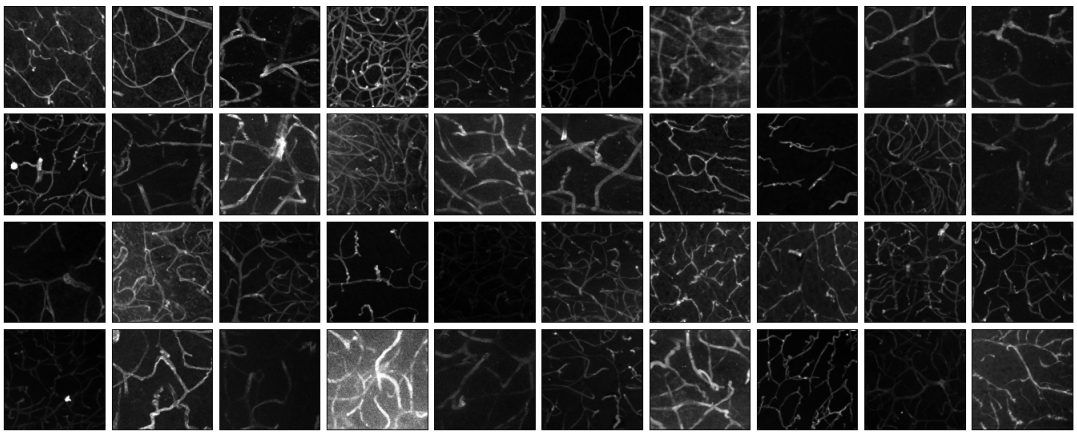
O conjunto de dados *VessMAP* é composto por 100 imagens de microscopia de fluorescência do córtex de camundongos, cada uma com resolução de 256×256 pixels. A Figura 4 mostra algumas imagens contidas no conjunto. O conjunto de dados *VessMAP* nos permite avaliar vasos sanguíneos e imagens com variações de características que são importantes para identificar se a rede neural treinada foi capaz de abstrair eficientemente as imagens. Entre essas características podemos destacar variações nos níveis de ruído, contraste, tamanho dos vasos sanguíneos, além da presença de alguns artefatos de captura das imagens.

O conjunto de dados *DRIVE* por sua vez tem um papel importante para validar se a metodologia é aplicável a imagens de vasos sanguíneos de diferentes domínios. Sendo o *DRIVE* amplamente empregado na avaliação de algoritmos de segmentação de vasos, assim, foi utilizado como um parâmetro de comparação e validação para os resultados obtidos com o *VessMAP* durante o presente estudo. O *DRIVE* é composto de 20 imagens de treinamento e 20 imagens de teste de fundo de olho, cada uma com dimensões de 584×565 pixels.

As imagens de microscopia do conjunto *VessMAP* possuem diferenças significativas quanto às imagens de fundo de olho do conjunto de dados *DRIVE*. No *DRIVE*, vasos mais espessos geralmente apresentam maior contraste em relação ao fundo do que no *VessMAP*. Além disso, as amostras do *DRIVE* possuem uma estrutura de captura muito bem definida, visto que esta inclui o disco óptico e as bordas da retina, uma vez que

[1] Repositório Público no GitHub | weslley-ss/Publications

Figura 4 – Exemplo de imagens do conjunto de dados *VessMAP*.



Fonte: Adaptado de Silva et al. (2025)

abrangem toda a retina, enquanto as amostras do *VessMAP*, adquiridas em regiões muito pequenas do córtex, não apresentam uma estrutura global definida.

4.3 Descrição das métricas utilizadas

As métricas de avaliação consideradas foram acurácia, AUC, índice de Dice, sensibilidade e especificidade. O índice de Dice mede a similaridade entre a segmentação predita e a segmentação de referência, sendo definido como o dobro da interseção entre as regiões preditas e reais dividido pela soma de seus tamanhos, assumindo valores entre 0, quando não há sobreposição, e 1, quando a correspondência é perfeita. A sensibilidade indica a proporção de pixels corretamente identificados como pertencentes à classe de interesse, no atual caso os vasos sanguíneos, enquanto a especificidade mede a proporção de pixels corretamente reconhecidos como pertencentes à classe negativa, fundo. A acurácia representa a proporção total de predições corretas, sendo utilizada para avaliar, nas tarefas de segmentação, o percentual de pixels classificados corretamente, e, nas tarefas de classificação de imagens, a proporção de imagens atribuídas à classe correta. Já a área sob a curva ROC (AUC) avalia o desempenho global do modelo em distinguir entre classes positivas e negativas, considerando diferentes limiares de decisão, sendo que quanto mais próxima de 1 melhor é a capacidade discriminativa do modelo.

4.4 Relevância da informação de textura e intensidades na tarefa de classificação

Nosso primeiro experimento consiste em analisar o potencial de identificação de vasos sanguíneos utilizando apenas informações locais das imagens. O aprendizado das ca-

racterísticas locais ausentes de contextos de vasos sanguíneos e do fundo das imagens foi analisado por testes de classificação de recortes de imagens ocupados por vasos sanguíneos ou pelo fundo da imagem. É esperado que no aprendizado local para essas duas classes observemos maior viés para os padrões de informações mais simples, ou seja, variação na disposição espacial dos pixels de vaso e fundo ou na variação absoluta das intensidades de pixels. Para avaliar qual dessas características é mais relevante para a classificação de vasos sanguíneos por CNNs utilizamos o conjunto de dados *VessMAP*.

4.4.1 Análise exploratória dos dados

A metodologia empregada iniciou-se com a caracterização dos recortes gerados a partir das imagens de vasos sanguíneos do conjunto de dados *VessMAP*. Foram utilizados recortes de 9x9 pixels, divididos em duas classes: 700 recortes de fundo e 539 recortes de vasos sem informação de borda; a metodologia para obtenção desses recortes é descrita abaixo na subseção 4.4.2. Esses recortes, destinados aos experimentos de classificação sem informação de borda, foram primeiramente analisados de forma exploratória. Utilizando os valores de intensidades dos pixels foram calculadas e avaliadas três medidas principais: a distribuição geral dos valores de intensidade dos pixels, a média de intensidade por imagem de cada classe e, similarmente, o desvio padrão por imagem para cada classe. Para análise, a distribuição das frequências foi avaliada por meio de histogramas. O objetivo foi verificar se a distinção entre vasos e fundo poderia ser tratada apenas pela análise direta da intensidade dos pixels e sua frequência, de forma a fundamentar a necessidade, ou não, de métodos de aprendizado profundo como as CNNs para a classificação e segmentação das imagens médicas.

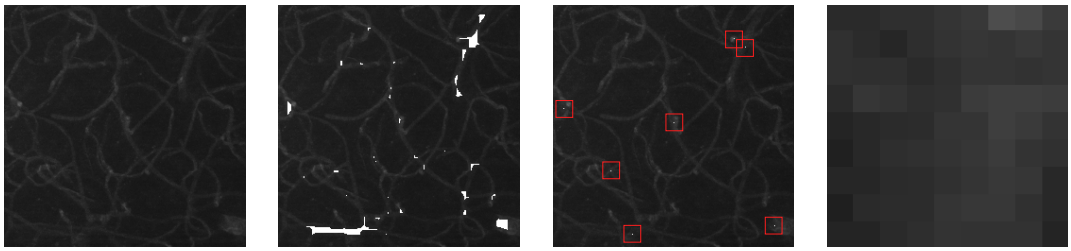
4.4.2 Classificação sem informação de borda

Pelo fato de vasos sanguíneos apresentarem, em sua maioria, espessura diminuta, identificamos a máxima área quadrada que pudesse estar contida dentro de um segmento de vaso, correspondendo a uma janela de 9x9 pixels. Utilizando uma operação de convolução com filtro 9x9 de valor 1 na imagem binária de rótulos, foram identificados os pixels cuja região dentro do filtro estivesse completamente preenchida por pixels de vaso ou de fundo. A regra de seleção dos pixels foi de que a somatória obtivesse valor máximo esperado para o preenchimento dentro da região desejada, uma vez que no rótulo binário as regiões foram anotadas entre 0 e 1, fundo e vaso, respectivamente. A Figura 5(a) exemplifica uma das imagens utilizadas no conjunto e, por sua vez, na Figura 5(b) os pixels elegíveis aos recortes para classificação de vasos sanguíneos. Uma janela 9x9 centrada nos pixels indicados possui apenas pixels pertencentes a vasos.

Para minimizar o efeito de sobreposição dos recortes, algumas seleções foram feitas. Dentre os pixels elegíveis ao recorte, sete foram sorteados mantendo uma distância mínima

entre eles, Figura 5(c). A distância Manhattan foi utilizada como critério para seleção dos pixels. A figura Figura 5(d) exemplifica um dos 7 recortes realizados pelo processo descrito. Em casos nos quais os critérios de tamanho, distância e número de recortes por imagem não foram atendidos, o número mínimo de recortes que possibilitasse número de recortes iguais por classe foi utilizado. Obteve-se 539 recortes de vaso e 700 recortes de fundo. Portanto, o conjunto foi composto de 539 recortes de cada classe, sendo os recortes de fundo sorteados aleatoriamente entre os 700 gerados.

Figura 5 – Processo de seleção de recortes para classificação entre vasos e fundos.



(a) Imagem original. (b) pixels mostrados em branco indicam o centro de regiões 9x9 contendo apenas vasos sanguíneos. (c) Exemplo de seleção de recortes. (d) Visualização de um recorte de vaso. Fonte: Próprio Autor

4.4.3 Classificação com informação de borda

A metodologia apresentada na seção anterior busca quantificar o potencial de identificação de vasos usando apenas o contexto local sem adição de informações de um contexto mais amplo dos vasos sanguíneos. Entretanto, é possível fazer a análise da melhora na classificação quando informações sobre as bordas dos vasos estão presentes nas imagens. Testes complementares foram realizados com variação na proporção de pixels do entorno dos vasos sanguíneos contidos em seus recortes. O objetivo dessa abordagem foi verificar a possível influência da informação de fronteira entre vasos e fundo na acurácia do modelo de classificação. Com essa finalidade, foram criados recortes das imagens contendo uma fração de pixels de vaso e fundo, sendo a classe de vasos composta também por pixels de fundo conforme a proporção estabelecida. Foram considerados recortes contendo 75% e 50% de pixels de vasos sanguíneos, obtendo-se 700 recortes para cada porcentagem. Os conjuntos de dados originais também tiveram seu fracionamento aplicado, gerando conjuntos com proporções de 75% e 50% dos pixels do recorte sendo pixels de vasos sanguíneos, D_{75} e D_{50} respectivamente. As mesmas imagens de fundo utilizadas no experimento anterior foram aplicadas neste experimento.

4.4.4 Perturbação seletiva de informações de textura e intensidade

As variações de contraste entre pixels de vaso e fundo e/ou padrões que possam existir na disposição dos pixels em regiões específicas dos vasos podem gerar artefatos que influenciam no processo de aprendizagem do modelo. Foram elaboradas três variações do conjunto de dados de recortes original D a partir da perturbação das características originais do conjunto de dados. O objetivo é observar o impacto da perda de informação de textura, disposição espacial, e de variação de intensidade dos pixels na distinção pelo modelo entre vasos e fundos, além de identificar quais informações locais são importantes para tal distinção.

As perturbações aplicadas envolvem a normalização da intensidade e a aleatorização das posições dos pixels.

1. Normalização: Uma forma trivial de classificar uma região da imagem como vaso ou fundo é através da comparação da intensidade dos pixels. Pixels de vaso tendem a possuir maior valor de intensidade do que pixels de fundo. O conjunto de dados normalizado objetiva remover informação sobre as intensidades dos pixels contidos nos recortes originais. A normalização realizada utilizou o cálculo por z-score, aplicado conforme a equação:

$$\tilde{x} = \frac{x - \mu}{\sigma} \tag{1}$$

   onde $\tilde{x}$ é o valor normalizado, x é o valor original do pixel, μ é a média da intensidade dos pixels da imagem e σ é o desvio padrão dos pixels da imagem. É importante notar que a normalização de intensidade é relevante para os experimentos envolvendo recortes contendo apenas pixels de vasos ou apenas pixels de fundo, mas também pode ter impacto em recortes contendo sinais de borda. O conjunto de dados de intensidades normalizadas passa a ser denominado D_{-i}.

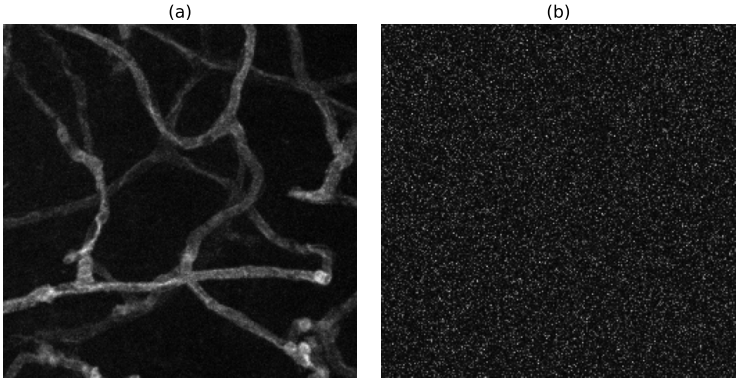
2. Aleatorização: Com a finalidade de investigar a importância da textura aplicou-se uma perturbação ao arranjo espacial original dos pixels. A posição de cada pixel foi embaralhada de forma aleatória entre as posições possíveis dentro de cada recorte, a Figura 6 mostra o impacto visual da aleatorização dos pixels dentro de uma imagem de vasos sanguíneos. Os valores de intensidade dos pixels não foram alterados. O objetivo do conjunto de dados D_{-t} gerado é que os recortes percam a estrutura de textura, de forma a identificar se a rede neural utiliza outras informações além da textura para classificar as imagens. Experimentos similares foram realizados em outros estudos (GEIRHOS et al., 2018).

3. Homogenização: O conjunto de dados homogenizado é formado pela a aplicação conjunta das perturbações de normalização e aleatorização. A aplicação ocorre como descrito nas transformações de forma individual. O objetivo da aplicação

de ambas perturbações é avaliar o desempenho da rede neural na condição mais adversa do teste: quando é removida a informação de textura local e a variação da intensidade dos pixels presentes nos recortes de vaso e fundo, o que torna o recorte homogenizado em relação ao seu recorte original. Este conjunto de dados é representado como $D_{-i,-t}$.

Os três conjuntos de dados resultantes e os dados originais foram utilizados nos experimentos de classificação com e sem informações de borda. Cada conjunto oferece uma perspectiva distinta sobre as informações utilizadas por uma rede neural para diferenciar entre pixels vasculares e não vasculares.

Figura 6 – Exemplo da aplicação da aleatorização dos pixeis da imagem e o impacto visual na características da textura original.

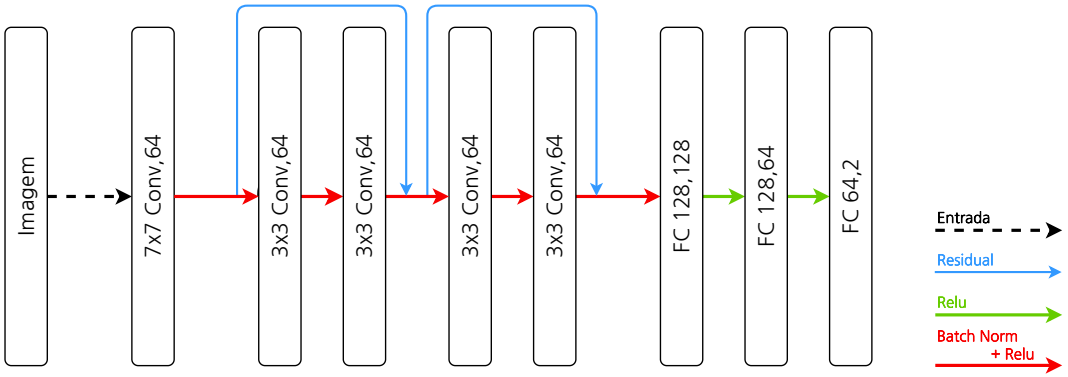


(a) Imagem original identificação as características da textura da imagem e dos vasos sanguíneos de forma completa. (b) Imagem após a aleatorização dos pixels e ausência da textura e organização espacial dos pixels da imagem original. Fonte: Próprio Autor

4.4.5 Treino e avaliação do modelo

O modelo de classificação foi treinado e testado para os conjuntos de dados mencionados anteriormente. Cada conjunto de dados foi dividido aleatoriamente em 80% de recortes para treinamento e 20% para teste. A rede neural utilizada (Figura 7) consiste em dois blocos residuais de 64 canais, treinados por 300 épocas com uma taxa de aprendizado de $0,001$. O *batch size* utilizado teve tamanho 9 e não foram aplicadas técnicas de aumento de dados. As execuções de treinamento foram repetidas cinco vezes com diferentes divisões do conjunto de dados, e a acurácia média da classificação foi calculada e utilizada de forma a garantir maior confiabilidade aos resultados obtidos.

Figura 7 – Rede neural utilizada para classificação local.



Fonte: Próprio Autor

4.5 Análise da influência da forma para a segmentação de vasos sanguíneos

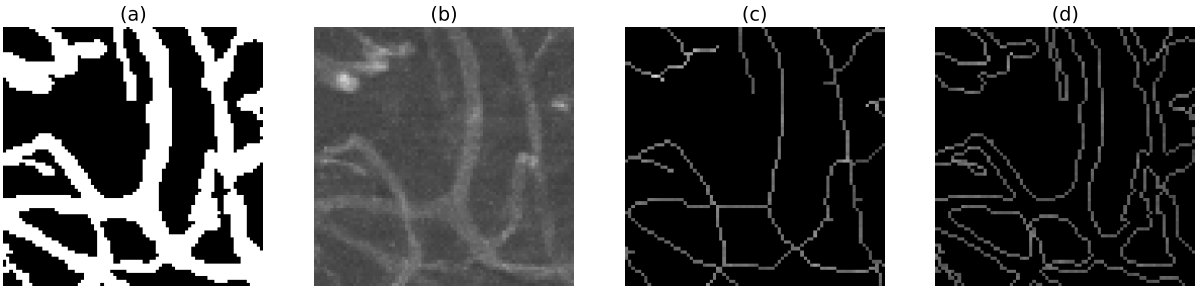
O treino de redes neurais para a segmentação semântica de imagens isolando apenas características de forma dos vasos sanguíneos é uma tarefa difícil. Na literatura, é possível identificar abordagens utilizando transferência de estilo, como a desenvolvida por Geirhos et al. (2018), que substitui a textura de imagens inteiras por texturas selecionadas aleatoriamente. Os autores também experimentaram substituir apenas a textura do objeto principal na imagem, mas essa abordagem levou a resultados inconclusivos. Portanto, não é possível definir tal abordagem como ideal para segmentação semântica.

Realizamos um experimento para avaliar o desempenho das CNNs de segmentar vasos sanguíneos usando apenas informações esparsas da forma, removendo a estrutura interna dos vasos. Dois novos conjuntos de dados foram criados a partir do *VessMAP* para o experimento. O conjunto de dados de contorno, D_c, composto apenas pelos contornos das formas dos vasos sanguíneos, e o conjunto de dados de esqueletos, D_e, composto pela linha central interna aos vasos. Os conjuntos de dados de esqueleto e contorno tiveram o objetivo de avaliar a capacidade da CNN em segmentar os vasos sanguíneos utilizando informações não relacionadas com as intensidades de pixels, mas sim às posições dos pixels das estruturas montadas. Tais conjuntos de dados foram construídos da seguinte maneira:

1. Esqueleto: Foi aplicada a função de esqueletização *skeletonize* da biblioteca *Scikit-image* aos rótulos (Figura 8(a)) do conjunto de dados original com a finalidade de obter os pixels centrais dos vasos sanguíneos. Objetivando tornar o treinamento mais eficiente, as intensidades dos pixels contidos no esqueleto foram mantidos nas imagens, e os demais pixels receberam valor zero (Figura 8(c)).

2. Contornos: Foi aplicada uma função de erosão aos rótulos (Figura 8(a)) do conjunto de dados original. Na sequência, a diferença entre a imagem original e a erodida foi calculada, gerando a imagem de contorno. As intensidades originais da imagem foram então aplicadas aos pixels de contorno (Figura 8(d)).

Figura 8 – Exemplos dos conjuntos de dados utilizados para segmentação focada na forma dos vasos.



(a) Imagem de rótulos. (b) Imagem original. (c) Imagem na qual apenas as intensidades de pixels do esqueleto foram mantidas. (d) Imagem na qual apenas as intensidades dos pixels de contorno foram mantidas. Fonte: Próprio Autor

Amostras dos conjuntos de dados D_c e D_e foram utilizadas como entrada para uma CNN, enquanto os rótulos foram as máscaras de segmentação das amostras originais do conjunto D. No caso do conjunto D_c, a rede deve aprender a preencher o interior dos contornos para prever corretamente os alvos. Assim, dentro dos vasos, ela não pode usar textura ou intensidades de pixels como pistas para a segmentação. Já o conjunto D_e representa um problema com solução indeterminada. A rede não consegue prever de forma única o calibre dos vasos a partir da linha central. Como melhor estimativa, ela pode aprender o calibre médio dos vasos e utilizá-lo para prever a segmentação com base na linha central, ou ainda tentar usar características de intensidade da linha central para estimar o calibre local.

4.5.1 Treinamento e avaliação dos modelos

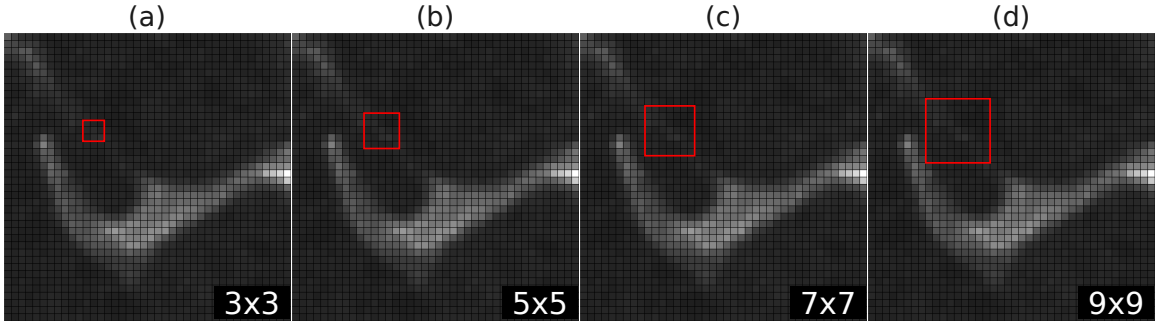
Os conjuntos de dados D_c e D_s foram divididos aleatoriamente em 68, 12 e 20 amostras para treinamento, validação e teste, respectivamente. Os modelos U-Net e W-Net de Galdran et al. (2022) foram utilizados para o aprendizado dessa tarefa. O treinamento seguiu o mesmo código e procedimentos descritos em (GALDRAN et al., 2022), entretanto, fez-se necessário um número maior de épocas para alcançar a convergência. Especificamente, a função de perda de entropia cruzada da segmentação foi otimizada por 2500 épocas, e o modelo com o maior valor de Dice de validação registrado durante o treinamento foi selecionado para o teste. O processo de treinamento foi realizado com

batch size de tamanho 4, taxa de aprendizado de $0,01$ e adotando o uso de estratégias básicas de aumento de dados, incluindo uma combinação de redimensionamento, rotação, translação e espelhamento vertical e horizontal.

4.6 Impacto do nível de contexto na qualidade da segmentação

A quantidade de contexto disponível para uma CNN possui papel crucial no aprendizado e no desempenho da tarefa de segmentação de vasos sanguíneos. A expansão do campo receptivo permite a ampliação da região da imagem considerada pela rede durante o treinamento, permitindo que ela utilize uma gama maior de informações para lidar com situações desafiadoras presentes nos conjuntos de dados. Fatores como oclusões, descontinuidades, presença de ruídos ou baixo contraste entre vaso e fundo são frequentemente associados às limitações das técnicas de captura de imagens, e podem dificultar ou comprometer a precisão da segmentação. Para que a rede neural seja capaz de contextualizar adequadamente a diferença entre pixels regulares e aqueles com artefatos de captura de imagem, é necessário que o campo receptivo tenha uma dimensão suficiente. A Figura 9 ilustra como o aumento do campo receptivo pode melhorar a capacidade discriminatória das CNNs. É possível observar que na região central do segmento do vaso sanguíneo há um contraste reduzido entre os pixels dos vasos e os do fundo. Observa-se que um campo receptivo 3x3 (Figura 9(a)) pode apresentar maior dificuldade para a obtenção de contexto e, por sua vez, na distinção dos pixels do que um campo 9x9 (Figura 9(d)). Ao obter informações do contexto, a rede pode aprender características presentes no contexto da imagem e expandir o campo receptivo efetivo, tendo assim a possibilidade de melhores desempenhos.

Figura 9 – Aumento de contexto a partir da expansão do campo receptivo



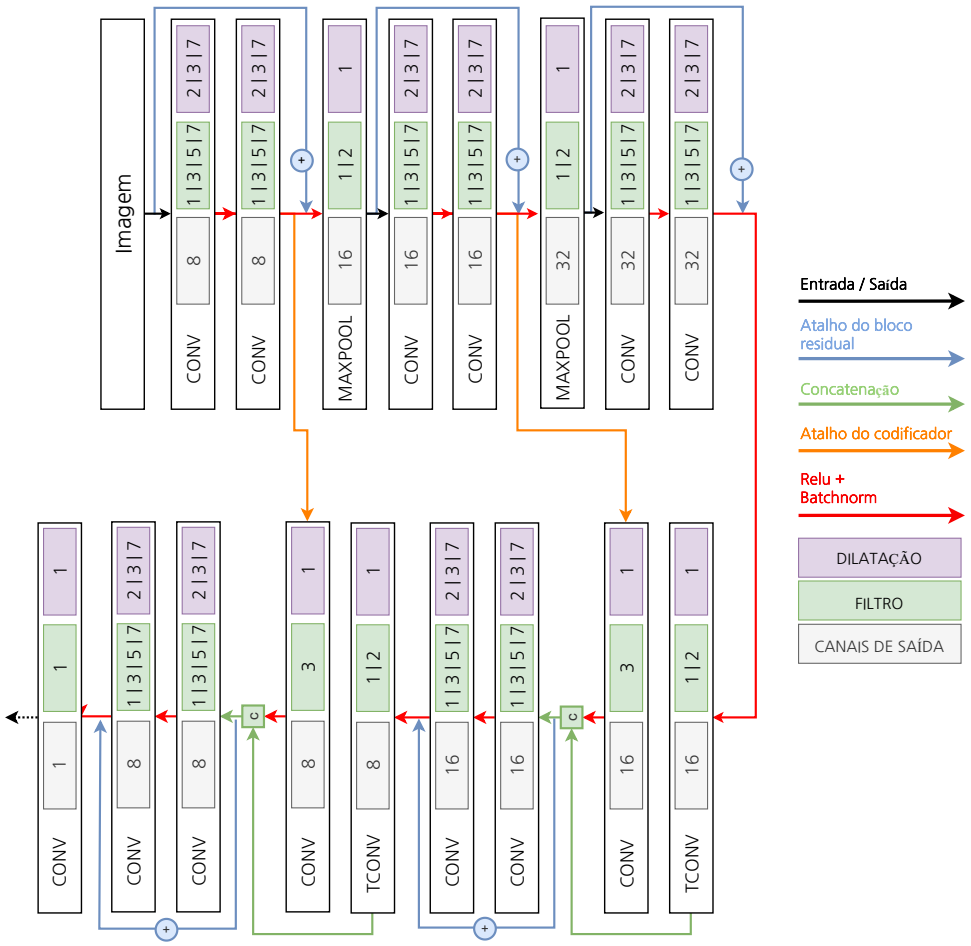
Segmento de vaso sanguíneo com baixo contraste. (a) Campo 3x3 preenchido apenas por pixels de baixo contraste e facilmente interpretáveis como fundo de imagens. (b) Campo 5x5 com maior distinção entre pixels de vaso e fundo. (c) Campo 7x7 capaz de identificar a silhueta do segmento de vaso. (d) Campo 9x9 capaz de abranger regiões do segmento de vaso que apresentam contraste maior e assim acompanhar a redução gradual do contraste do segmento na sua porção central. Fonte: Próprio autor.

Com base nesse cenário, foram desenvolvidas duas abordagens complementares para avaliar a influência do contexto na segmentação. A primeira consistiu em variar seletivamente os parâmetros da rede associados ao tamanho do campo receptivo, analisando a relação entre o campo receptivo e o desempenho do modelo. A segunda abordagem utilizou a divisão das imagens em recortes de diferentes dimensões, de forma a restringir o contexto acessível à rede. Esse procedimento permitiu quantificar o tamanho mínimo de um recorte da imagem necessário antes que o desempenho da segmentação fosse prejudicado pela insuficiência de contexto.

No primeiro experimento, 160 configurações diferentes de modelos foram treinadas utilizando os conjuntos de dados *VessMAP* e *DRIVE*. Foram utilizadas diferentes combinações de tamanho de filtro, taxa de dilatação e quantidade de *downsampling*. A Figura 10 apresenta a arquitetura base utilizada e os parâmetros alterados em cada camada, bem como os valores possíveis dos parâmetros para combinação. Por exemplo, um dos modelos utilizou todos os filtros com tamanho 1, enquanto outro teve o primeiro filtro com tamanho 1 e os demais com tamanho 3, e assim por diante. O modelo com filtros de tamanho 3 em todas as camadas convolucionais, *max pooling* de tamanho 2 e taxa de dilatação igual a 1 corresponde ao mesmo modelo descrito em Galdran et al. (2022). Esse modelo é considerado o *modelo base*, por representar um modelo com boa performance para a segmentação de vasos.

Para cada configuração, o modelo foi treinado por 1000 épocas, utilizando taxa de aprendizado de 0,01, *batch size* igual a 4 e função de perda de entropia cruzada. O modelo que obteve o maior valor de *dice* no conjunto de validação durante o treinamento foi selecionado para a avaliação de desempenho. Cada execução de treinamento foi repetida três vezes para garantir significância estatística.

Figura 10 – Rede U-Net utilizada nos experimentos de campo receptivo.



As caixas verdes e roxas mostram, respectivamente, os tamanhos de filtro e as taxas de dilatação utilizadas. Um tamanho de filtro igual a 1 para a camada de *max pooling* significa que, na prática, nenhum *max pooling* foi aplicado. Fonte: Próprio Autor

Para o conjunto de dados *VessMAP*, as imagens foram divididas aleatoriamente em conjuntos de treinamento, validação e teste contendo, respectivamente, 68, 12 e 20 imagens. Para o conjunto de dados *DRIVE*, a divisão oficial de treinamento foi subdividida aleatoriamente em 16 imagens para treinamento e 4 para validação, enquanto o conjunto oficial de teste contendo 20 imagens foi utilizado para teste. Os dados foram aumentados por meio de redimensionamentos, rotações, translações e espelhamentos horizontais e verticais aleatórios.

O tamanho do campo receptivo foi medido após o treinamento dos modelos. Foram considerados dois tipos de campos receptivos: teórico e efetivo. O campo receptivo teórico (TRF) de um pixel de saída corresponde ao conjunto de pixels de entrada capazes de influenciar o valor desse pixel de saída. Seu cálculo é feito obtendo-se o gradiente de um pixel de saída em relação a todos os pixels de entrada e considerando apenas os valores

não nulos. Como o TRF é um quadrado, o comprimento de um dos lados foi utilizado como medida do seu tamanho. O ERF foi medido utilizando a metodologia de Luo et al. (2016). Como o tamanho pode variar dependendo do pixel de saída, foi feita a média dos valores calculados para 100 pixels selecionados aleatoriamente no conjunto de teste de cada conjunto de dados. A razão ERF é definida como a razão entre os tamanhos do ERF e do TRF. Ela mede a fração do TRF que é de fato utilizada por uma CNN.

O segundo experimento consistiu em dividir cada amostra dos conjuntos de dados *VessMAP* e *DRIVE* em recortes não sobrepostos de tamanho $W \times W$ e treinar o modelo base para segmentar os vasos. O modelo foi treinado utilizando recortes de tamanhos $[4, 8, 16, 32, 64, 128, 256, 512]$. O tamanho 512 não foi utilizado no conjunto *VessMAP* por ser maior do que o tamanho das imagens. Foram empregados os mesmos parâmetros de treinamento, divisões dos dados e aumentos utilizados no experimento anterior. A única exceção foi o *batch size*. Como a quantidade de informação contida em um recorte de imagem varia de acordo com W, um *batch size* fixo resultaria em diferentes quantidades de informação para cada valor de W. Assim, o *batch size* foi calculado para cada tamanho de recorte como:

$$B = 4\left(\frac{H}{W}\right)^2, \tag{2}$$

onde H é o tamanho das imagens do conjunto de dados, todas com proporção quadrada. Para o conjunto *VessMAP*, $H = 256$, e para o *DRIVE*, $H = 512$. Isso garantiu que os *batches* utilizados para cada tamanho de recorte de imagem contivessem sempre o mesmo número de pixels.

Capítulo 5

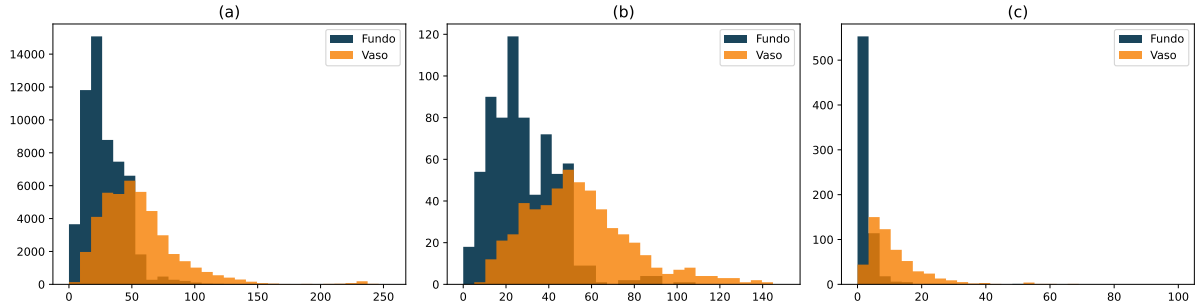
Resultados

Nesta seção, apresentamos os resultados da investigação sobre como diferentes características das imagens de vasos sanguíneos influenciam o desempenho das CNNs nas tarefas de classificação e segmentação de vasos sanguíneos. Inicialmente, realizamos uma análise exploratória da distribuição dos pixels, a fim de compreender a diferença entre vasos e fundo. Em seguida, avaliamos a performance dos modelos em cenários controlados, nos quais manipulamos características específicas das imagens — como intensidade, textura e forma — para quantificar sua relevância na segmentação. Por fim, analisamos o impacto da complexidade dos modelos e do tamanho do campo receptivo na capacidade das redes em extrair informações e utilizá-las para a segmentação dos vasos sanguíneos.

5.1 Análise exploratória

A análise das características presentes nos recortes de vasos sanguíneos e fundo das imagens foi iniciada a partir da análise exploratória dos valores dos pixels presentes nos recortes sem informação de borda feitos a partir do conjunto de dados *Vessmap*. Tal abordagem analisou a possibilidade da classificação dos pixels ser um problema trivial que pudesse ser resolvido apenas observando o valor da intensidade de cada pixel. Na Figura 11 observamos a distribuição geral dos valores de intensidade dos pixels nos recortes, a média dos valores por imagem e o desvio padrão por imagem para cada classe de recortes, vasos e fundo. O histograma dos valores de intensidades contidos na Figura 11(a) demonstra que, apesar dos recortes de fundo possuírem uma concentração marcante nos valores menores que 50, existe uma grande porção de pixels de vasos sanguíneos com suas intensidades sobrepostas às de fundos. Da mesma forma, os valores das médias dos pixels (Figura 11(b)) confirmam a similaridade entre as intensidades de pixels de boa parte dos recortes das

Figura 11 – Distribuição dos pixels para cada classe de recortes.



(a) distribuição geral das intensidades dos pixels, distribuição das (b) médias das intensidades por imagem e dos (c) desvios padrão das intensidades por imagem. Observa-se a sobreposição entre vasos e fundo em (a) e (b), sendo apenas em (c) possível notar maior variabilidade nos recortes de vasos, embora ainda insuficiente para uma separação completa das classes. Fonte: Do autor

duas classes. Apenas a partir do desvio padrão dos pixels calculado para cada imagem (Figura 11(c)) podemos identificar que os valores contidos nos vasos sanguíneos tendem a ter uma variabilidade maior entre eles. Ainda assim, não constitui uma variabilidade suficiente pra distingui-los de forma simples. Sendo assim, a distinção dos pixels entre vasos e fundos não se enquadra como um problema trivial de classificação, e por isso a necessidade do aprendizado de características das imagens a partir das CNNs para a classificação e segmentação de imagens médicas, em especial de vasos sanguíneos.

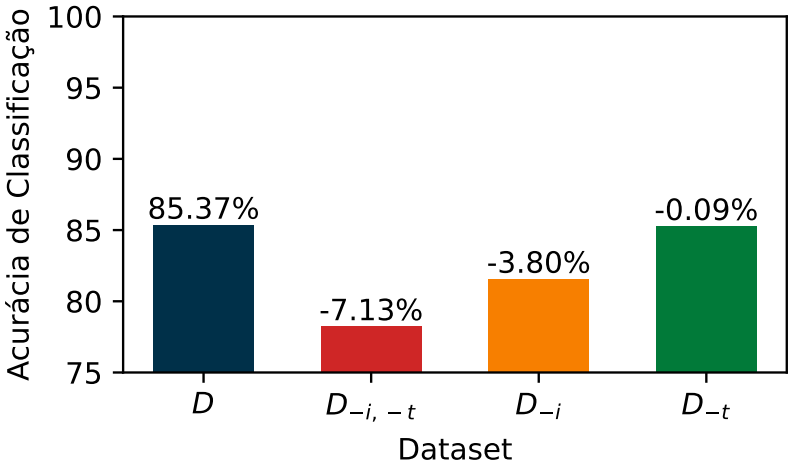
5.2 Manipulação de textura e intensidade

Os testes de classificação realizados com recortes sem informação de borda possibilitam o entendimento sobre a informação contida na textura e intensidade dos pixels para a distinção entre vasos e fundo de imagem. A Figura 12 demonstra a acurácia de classificação dos modelos treinados para o conjunto de dados D e suas modificações da quantidade de informação recebida pela CNN durante o treinamento. Dos resultados observados, temos que a acurácia dos conjuntos de dados original (D), Homogenizado ($D_{-i,-t}$), Normalizado (D_{-i}) e Aleatorizado (D_{-t}) foram 85.37%, 78, 24%, 81.57% e 85.28%, respectivamente.

O conjunto de dados D será utilizado como referência, tendo este a máxima informação quanto às características presentes na textura e intensidade dos pixels das imagens originais. De igual maneira, a perda de informação após a aplicação das perturbações será entendida como a relação direta e proporcional à queda de acurácia observada nos modelos treinados em tais conjuntos de dados.

Em um comparativo com D, o conjunto de dados $D_{-i,-t}$ apresenta uma perda de 7.13% de acurácia. Tal resultado põe à vista que em imagens de vasos sanguíneos obtidas por

Figura 12 – Acurácia de classificação de recortes de vasos sem informação de borda.



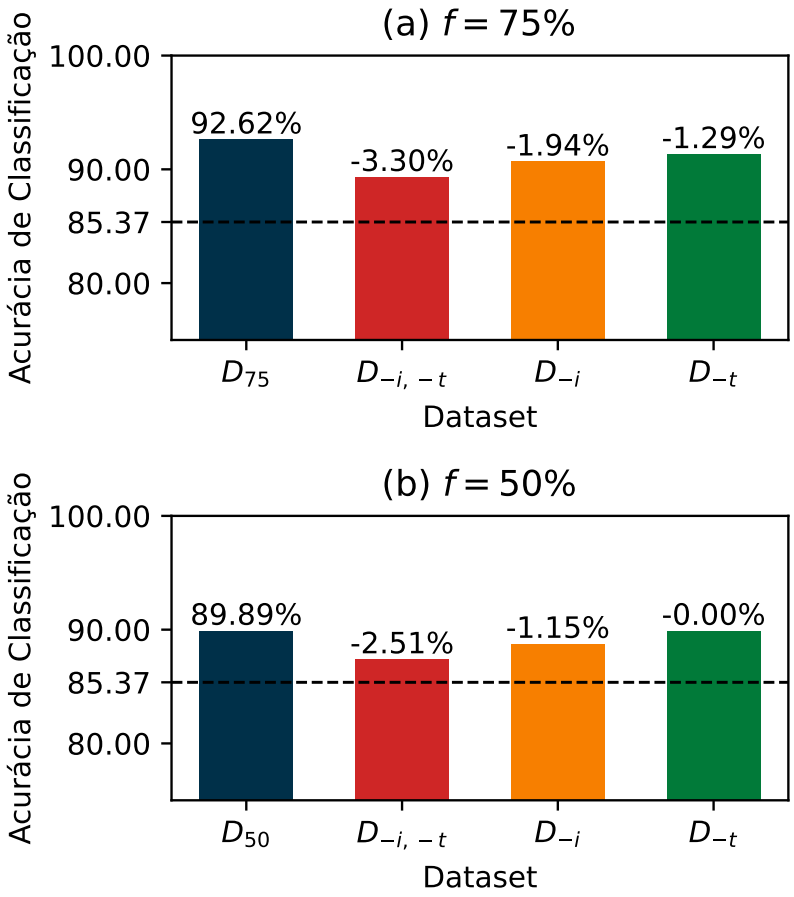
Avaliação de recortes em informação de borda. Os valores negativos indicam a redução na acurácia em comparação com o conjunto de dados D. Fonte: Do autor

microscopia confocal a disposição espacial e a intensidade dos pixels originais são relevantes para o aprendizado das redes e, por sua vez, contribuem para a correta classificação dos pixels, ainda que sem informação do contexto global da imagem. Apesar de sofrer maior perturbação das características originais, $D_{-i,-t}$ ainda é capaz de oferecer informação suficiente para que a rede neural consiga distinguir e classificar os recortes a partir de outras características presentes nos recortes e apresentar mais de 75% de acurácia. O desenvolvimento de técnicas para identificar outras características utilizadas pela rede além das texturas e intensidades dos pixels é um problema em aberto para ser analisado em estudos futuros.

Analisando individualmente cada uma das características da imagem obtém-se que a perda de informação com D_{-t} é menor que em D_{-i}. Portanto, a modificação da disposição espacial dos pixels de vaso e fundo teve menos impacto na classificação dos recortes, tendo uma perda de informação de $-0,09\%$, valor muito próximo ao máximo observado em D. Assim, entendemos que a perturbação das intensidades de pixels é mais impactante ao processo de classificação entre vasos e fundos para as CNNs. Embora para um humano possa ser uma tarefa simples inferir entre vasos e fundos a partir das tonalidades presentes nas imagens, as CNNs funcionam como caixas-pretas no que diz respeito ao seu aprendizado. Por este motivo, compreender quais informações interferem no desempenho do modelo pode contribuir para a aplicação de pré-processamentos e normalizações adequados aos conjuntos de dados de imagens médicas e, especialmente, no contexto de vasos sanguíneos imageados por microscopia confocal.

Ao incluirmos informações da borda dos vasos sanguíneos nos recortes, observamos

Figura 13 – Acurácia de classificação de recortes de vasos contendo frações de 75% e 50% de informação de borda.



Avaliação de recortes de vasos contendo (a) 75% de pixels de vaso e (b) 50% de pixels de vaso. Valores negativos indicam a redução na acurácia em comparação com os conjuntos de dados D_{75} e D_{50}. A linha tracejada indica a acurácia obtida para o conjunto de dados D. Fonte: Do autor

uma melhoria na acurácia de classificação dos modelos de todos os conjuntos de dados conforme ilustrado na Figura 13. Foi observado um ganho de acurácia de $7,25\%$ em $D_{75\%}$ e $4,52\%$ em $D_{50\%}$ comparado ao conjunto D sem informação de borda dos vasos, sugerindo que a geometria ou as variações de contraste presentes nos contornos dos vasos também compõem o aprendizado das CNNs na classificação. Nas análises realizadas neste estudo, a adição da informação de borda manteve a tendência de indicar as intensidades dos pixels como a característica mais relevante para a classificação entre vasos e fundos, visto que os experimentos feitos com D_{-i} têm uma perda maior de acurácia quando comparados a D_{-t}.

Embora a intensidade dos pixels continue sendo o fator determinante para o ganho de informação e aprendizado das CNNs na classificação, ao adicionarmos novas informações,

que neste caso foi o contexto da borda dos vasos sanguíneos, o impacto que as perturbações geram na classificação foi menor. Por exemplo, quando as perturbações de texturas em D_{-t} são realizadas a partir de D_{50}, a aleatorização dos pixels não surte impacto na classificação dos recortes. Da mesma forma, a normalização das intensidades dos pixels também apresenta um impacto menor no aprendizado das CNNs ao compararmos com os experimentos com uma quantidade mais restrita de informação. Esse mesmo efeito também é observado nos demais conjuntos de dados, Figura 13. Isso sugere que, à medida que se enriquece o conteúdo informativo dos recortes, o efeito negativo das perturbações das características da imagem diminui, impactando menos o desempenho da rede. Assim, mesmo em cenários com maior degradação da imagem, à medida que a rede obtém acesso a informações do contexto no qual o vaso está localizado, é possível que ela consiga compensar parte da perda ocasionada pelas perturbações, contribuindo para uma classificação mais robusta.

É importante mencionar que a adição da informação de borda neste experimento apresentou um limite quanto ao ganho da informação. Tal limitação é observável nos resultados da Figura 13(b) marcada pela queda de acurácia dos modelos quando diminui-se os pixels de vaso para D_{50} em comparação a D_{75}, levando a uma queda na acurácia de $2,63\%$. Entendemos que ao adicionarmos uma quantidade maior de pixels que compõem o fundo da imagem nos recortes de vasos sanguíneos, a rede tenha maior dificuldade neste experimento para o aprendizado das características dos vasos sanguíneos, justamente por ter menor exposição aos pixels que são desta classe.

5.3 Segmentação a partir de aspectos de forma

Os experimentos conduzidos com o objetivo de avaliar a capacidade de segmentação de vasos sanguíneos utilizando apenas informações relacionadas à forma destes, isolando tais características das imagens — ou seja, por meio do contorno dos vasos D_c ou de sua linha central D_e — apresentaram resultados limitados. Como mostra a Tabela 2, os modelos não conseguiram atingir um desempenho satisfatório no conjunto de dados *Vessmap*, demonstrando dificuldades em segmentar os vasos a partir das formas.

O melhor desempenho foi observado no experimento com a U-Net aplicada sobre o conjunto D_e, alcançando uma acurácia de 67,55% e *dice* de 57,03%. Entre os modelos treinados com D_c, o modelo W-Net obteve uma acurácia de 62,46% e *dice* de 54,37%, reforçando que ambas as abordagens — contorno e esqueleto — fornecem informações estruturais úteis para a segmentação, entretanto, insuficientes para uma segmentação refinada dos vasos sanguíneos. Além disso, observa-se um padrão consistente entre os modelos quanto à alta sensibilidade e baixa especificidade, sobretudo no caso da U-Net aplicada no conjunto D_c, que apresentou sensibilidade de $93,36\%$, mas especificidade extremamente baixa ($12,03\%$) o que indica uma forte tendência à supersegmentação —

Conjunto de dados	CNN	AUC	ACC	Dice	Spec	Sens
D_c	U-net	52.59	34.99	44.80	12.03	93.36
D_e	U-net	73.44	67.55	57.03	65.80	72.01
D_c	W-net	68.61	62.46	54.37	55.91	79.09
D_e	W-net	69.02	59.12	55.01	50.67	80.63

Tabela 2 – Desempenho dos modelos U-Net e W-Net nos conjuntos de dados D_c e D_e.

ou seja, muitos falsos positivos gerados, o que compromete a confiabilidade do modelo.

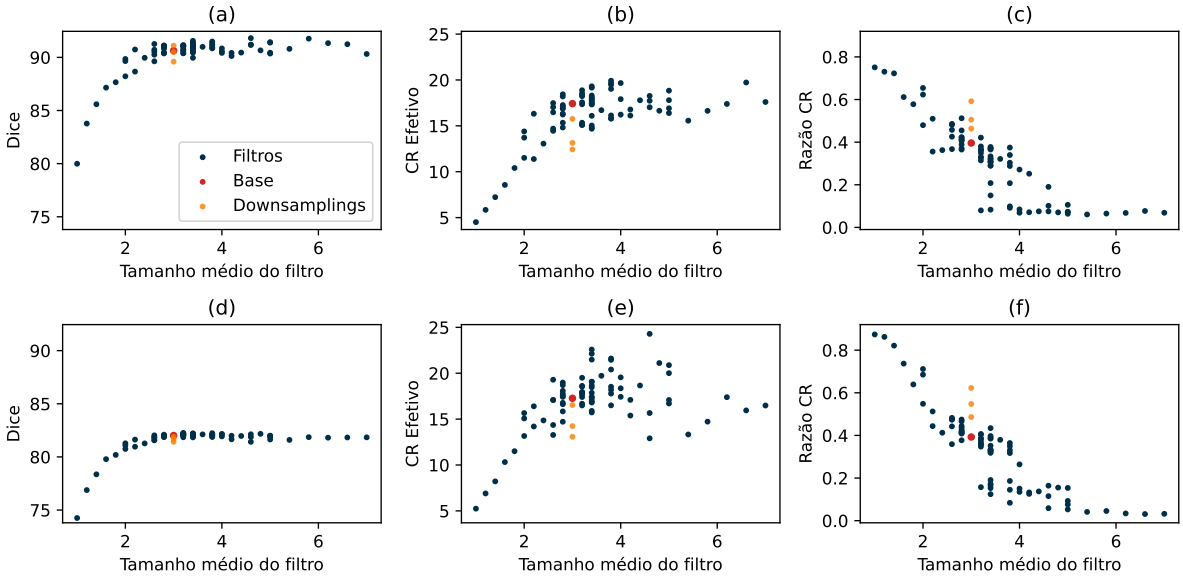
Esses resultados sugerem que, embora a forma dos vasos contenha informações espaciais úteis, ela não é suficiente para permitir uma segmentação precisa e confiável. A remoção da estrutura interna dos vasos limita o desempenho das CNNs. Desta forma, a importância de combinar características espaciais e da captura da imagem é reforçada para o desenvolvimento de abordagens mais robustas para segmentação de vasos sanguíneos.

5.4 Relação entre o desempenho de segmentação e o tamanho do campo receptivo

Ao variarmos os parâmetros da rede neural que estão associados ao campo receptivo, tínhamos em vista analisar como o tamanho deste campo poderia contribuir na performance dos modelos de segmentação de vasos sanguíneos. Como o mesmo número de filtros foi mantido em todo o experimento e apenas seus tamanhos sofreram variações, o tamanho médio dos filtros reflete o campo receptivo da CNN e também o número de parâmetros treináveis da rede. O campo receptivo tende a crescer com o aumento dos filtros, o que pode beneficiar tarefas de segmentação. Aumentar o tamanho dos filtros também adiciona mais parâmetros treináveis e, por sua vez, incorpora maior complexidade ao modelo sem alterar a estrutura geral. A configuração experimental permitiu observarmos a relação direta entre a complexidade dos modelos e o tamanho do campo receptivo. A Figura 14 apresenta os resultados obtidos após as variações de filtro aplicadas nos treinos realizados utilizando os conjuntos de dados *Vessmap* (Figuras 14(a), (b) e (c)) e *DRIVE* (Figuras 14(d), (e) e (f)).

Ao medirmos e avaliarmos a relação entre o Dice e o tamanho médio do filtro, observamos nas Figuras 14(a) e (d) que a performance dos modelos tende a melhorar com o aumento do tamanho médio dos filtros. De forma complementar, o gráfico que relaciona o ERF ao tamanho médio dos filtros mostra que o campo efetivo também se expande com o aumento do filtro até atingir um ponto de saturação, conforme Figuras 14(b) e (e). Esses comportamentos são consistentes com a ideia de que filtros maiores contribuem para campos receptivos mais amplos, possibilitando uma melhor captação das estruturas nas imagens. Entretanto, as melhorias das performances de segmentação dos modelos apresentaram um platô em torno do tamanho dos filtros em 3.

Figura 14 – Desempenho de segmentação e tamanhos de campo receptivo medidos para arquiteturas U-Net com diferentes tamanhos de filtro.



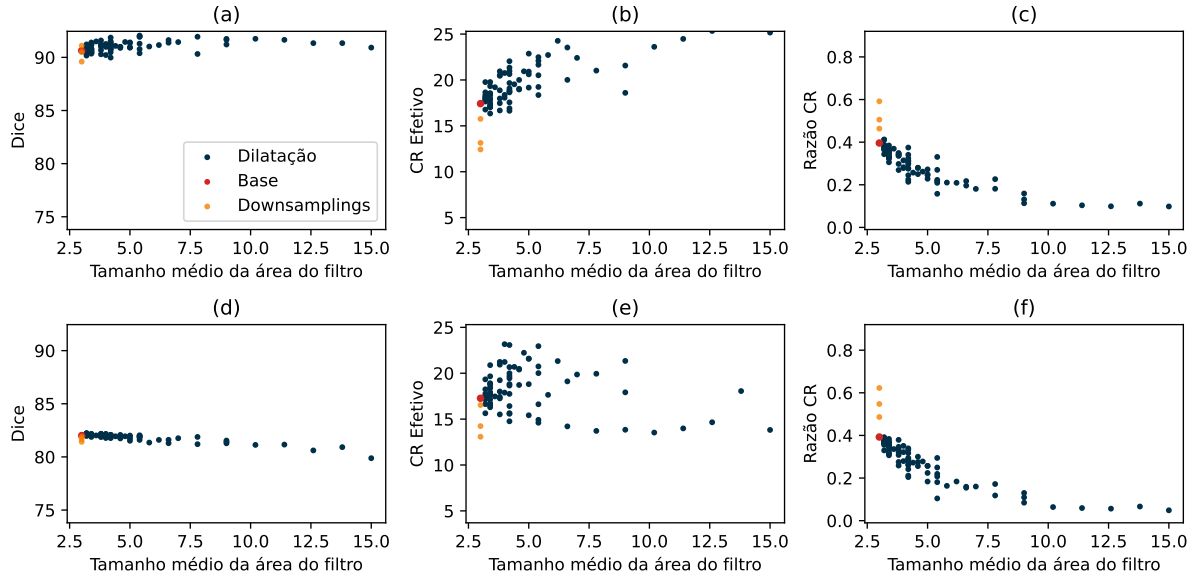
Os gráficos (a), (b) e (c) referem-se ao conjunto de dados *Vessmap*, enquanto os gráficos (d), (e) e (f) referem-se ao conjunto de dados *DRIVE*. O ponto vermelho indica o valor do modelo base, enquanto os pontos laranjas mostram os resultados obtidos ao alterar as camadas de subamostragem (downsampling) e por fim os pontos azuis representam as variações obtidas ao modificar o tamanho dos filtros ao longo das camadas convolucionais. Fonte: Do autor

Esse comportamento pode ser explicado pelo fato de que a expansão do ERF está vinculada à utilidade prática das informações captadas para reduzir a perda e melhorar a performance. Uma vez que o aumento do tamanho dos filtros deixa de gerar ganhos significativos na segmentação, os gradientes associados a regiões mais distantes da área central tornam-se menos relevantes, e a rede deixa de ampliar seu campo efetivo de forma substancial, mesmo que o campo teórico permita.

O fato de a saturação ocorrer em ambas as análises quando o tamanho médio dos filtros é próximo de 3 reforça uma tendência comum na literatura envolvendo CNNs que é a aplicação do filtro 3×3 como um ponto de equilíbrio ideal entre complexidade e eficiência para modelos de segmentação de imagens.

A razão entre o campo efetivo (ERF) e o campo teórico (TRF), Figuras 14(c) e (f), indica o quanto do campo receptivo é utilizado de fato pela CNN na segmentação. Associando essa medida ao tamanho médio dos filtros, foi observado que, embora o ERF aumente com filtros maiores, a saturação de desempenho não é resultado da limitação física do campo teórico. Isso porque a razão ERF/TRF diminui à medida que os filtros se tornam maiores, indicando uma parcela menor do campo teórico sendo utilizada pela rede, mesmo esta tendo a possibilidade de acessar regiões mais amplas. Quanto à suba-

Figura 15 – Desempenho de segmentação e tamanhos de campo receptivo medidos para arquiteturas U-Net com diferentes taxas de dilatação.



Os gráficos (a), (b) e (c) referem-se ao conjunto de dados *Vessmap*, enquanto os gráficos (d), (e) e (f) referem-se ao conjunto de dados *DRIVE*. O ponto vermelho indica o valor do modelo base, enquanto os pontos laranjas mostram os resultados obtidos ao alterar as camadas de subamostragem (downsampling) e por fim os pontos azuis representam as variações obtidas ao modificar o tamanho dos filtros e das taxas de dilatação ao longo das camadas convolucionais. Fonte: Do autor

mostragem, a remoção do *max pooling* não reduziu o Dice significativamente, apesar de diminuir o ERF, sugerindo que a rede foi capaz de compensar a diminuição da informação espacial.

De forma análoga à análise do desempenho dos modelos a partir da variação do tamanho dos filtros, também investigamos o impacto da aplicação de diferentes dilatações, cujo resultado pode ser observado na Figura 15. Como métrica, utilizou-se o tamanho total do filtro após sua dilatação para analisar o quanto a expansão dos filtros através da dilatação poderia impactar na segmentação. No entanto, diferentemente da variação de filtros sem dilatação, não foi possível identificar uma correlação clara entre essa métrica e as variações de Dice ou no ERF dos modelos. Por mais que a dilatação altere o alcance teórico da convolução e potencialmente expanda o campo receptivo, os resultados sugerem que esse efeito não se traduziu de forma consistente em ganhos de performance. Essa inconsistência pode estar associada ao fato de que a dilatação modifica a distribuição espacial dos pixels capturados pelo filtro, mas não necessariamente aumenta a quantidade de informação útil para a segmentação de vasos sanguíneos. Uma vez que as estruturas encontradas em vasos sanguíneos são diminutas, é possível que o aumento das lacunas entre os pontos de amostragem tenha levado à perda de detalhes relevantes das estruturas

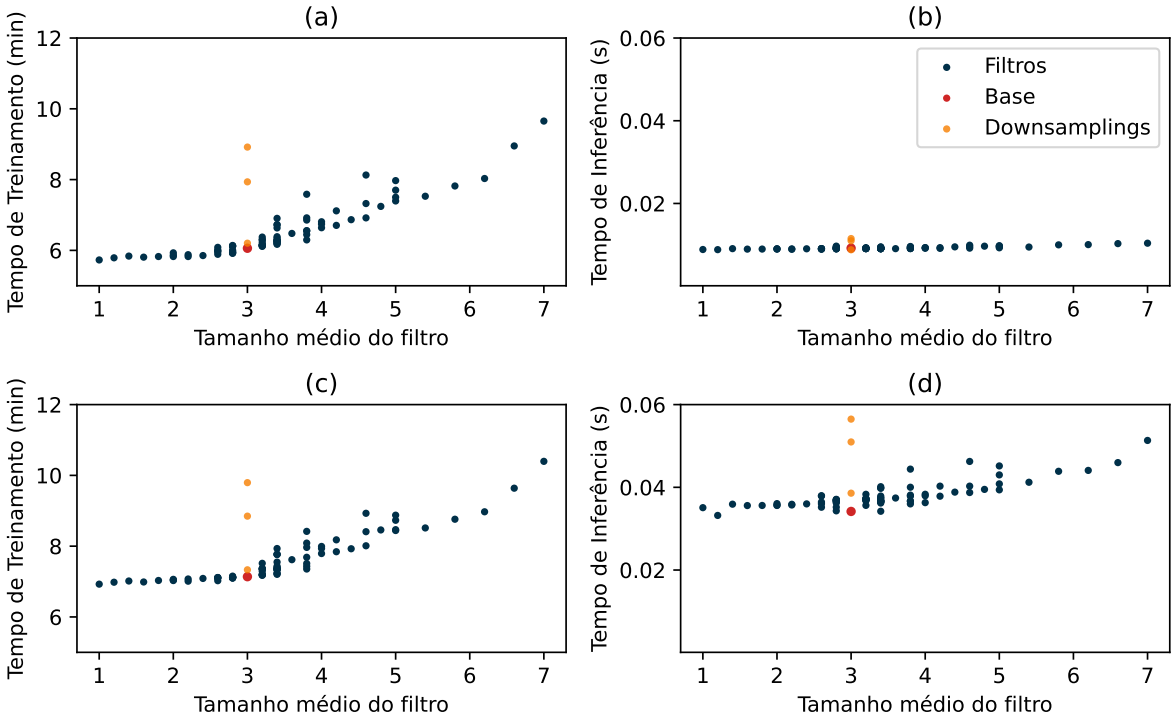
vasculares. Tal efeito pode traduzir a sutil queda do Dice na Figura 15(d) dado o fato de as estruturas vasculares contidas no conjunto de dados *DRIVE* serem pequenas. Estudos futuros incorporando dilatação em arquiteturas mais especializadas ou aplicando-a a diferentes tipos de imagens médicas podem fornecer evidências adicionais para compreender melhor sua aplicação ideal em tarefas de segmentação.

Para avaliar o impacto das variações de filtros e dilatação no desempenho computacional, analisamos o tempo de treinamento e o tempo de inferência dos modelos. O tempo de treinamento representa o tempo necessário para que cada configuração de arquitetura consiga convergir. Já o tempo de inferência corresponde ao tempo necessário para o modelo segmentar uma imagem dos conjuntos de dados de teste. As análises de desempenho dos modelos em função dos tempos de treino e inferência para as variações de filtros e de taxas de dilatação podem ser observadas nas Figuras 16 e 17, respectivamente. Na análise dos tempos de treinamento a partir da variação dos tamanhos dos filtros observamos um aumento gradual do custo computacional a medida que o tamanho médio dos filtros aumentam, Figuras 16(a) e (c), enquanto o tempo de inferência das imagens mantém maior constância independente da complexidade inserida ao modelo com o aumento do filtro, Figuras 16(b) e (d). Similarmente, a análise a partir da aplicação de diferentes dilatações demonstrou um comportamento crescente no tempo de treinamento a medida que aumenta-se a área média do filtro, Figuras 17(a) e (c), entretanto de forma menos correlacionada que a análise do tamanho de filtros. Já no tempo de inferência por imagem observamos maior constância no tempo necessário, tendo assim um comportamento similar ao descrito anteriormente. Destaca-se ainda que os modelos treinados no conjunto de dados *DRIVE* apresentaram um tempo de treinamento e inferência ligeiramente maiores do que no conjunto de dados *Vessmap*.

Para além destas análises quanto ao custo computacional, na tarefa de segmentação de vasos sanguíneos, observou-se que o ERF pode servir em conjunto com a complexidade do modelo para avaliar a eficiência dos parâmetros da rede neural. Mesmo com variações nos parâmetros e aumento da complexidade das arquiteturas, o ERF tendeu a um valor máximo próximo de 20. Isso indica que o campo receptivo máximo alcançado para um determinado conjunto de dados pode orientar a redução da arquitetura do modelo, de forma que a eficiência do campo receptivo seja mantida, mas a complexidade reduzida. Essa abordagem permite identificar até que ponto o aumento da complexidade contribui para a captura de informações relevantes, evitando expansões desnecessárias que não proporcionem ganhos adicionais de desempenho.

Para medir diretamente o campo receptivo necessário para alcançar um bom desempenho de segmentação, o método de extração de recortes de imagem descrito na metodologia foi aplicado aos conjuntos de dados *Vessmap* e *DRIVE*. A Figura 18 apresenta o valor de Dice em função do tamanho dos recortes. O desempenho se estabiliza em 32×32 para o conjunto *Vessmap*, ou seja, um contexto de 32×32 é suficiente para atingir o

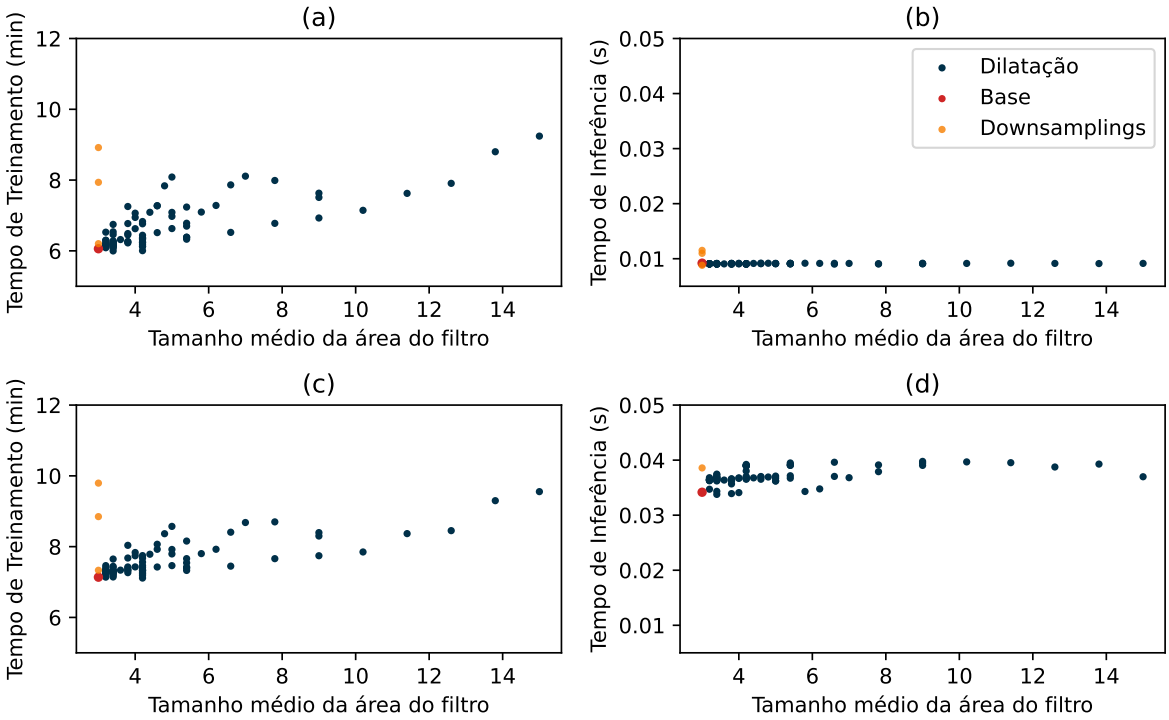
Figura 16 – Desempenho computacional em função do tempo e da variação do tamanho dos filtros em arquiteturas U-Net.



Os gráficos (a), (b) referem-se ao conjunto de dados *Vessmap*, enquanto os gráficos (c), (d) referem-se ao conjunto de dados *DRIVE*. O ponto vermelho indica o valor do modelo base, enquanto os pontos laranjas mostram os resultados obtidos ao alterar as camadas de subamostragem (downsampling) e por fim os pontos azuis representam as variações obtidas ao modificar o tamanho dos filtros ao longo das camadas convolucionais Fonte: Do autor.

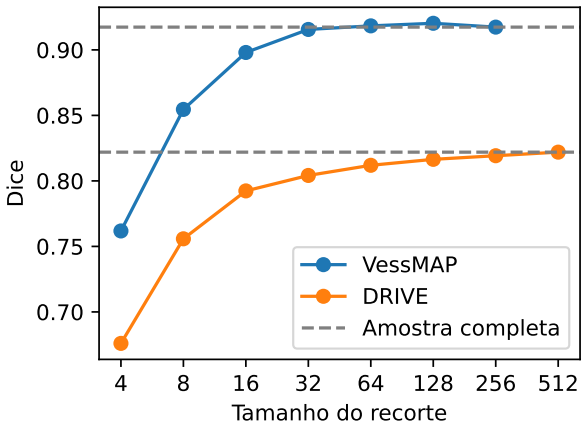
desempenho máximo de segmentação. Esse resultado é consistente com o valor máximo do ERF observado no experimento anterior. No entanto, o desempenho não se estabiliza para o conjunto *DRIVE*. A razão dessa diferença não é clara. Temos por hipótese que este comportamento se deve ao fato de que as amostras do *DRIVE* possuem uma estrutura global, incluindo o disco óptico, vasos grandes e pequenos em posições específicas e uma região de interesse bem definida (a retina).

Figura 17 – Desempenho computacional em função do tempo e da variação do tamanho dos filtros em arquiteturas U-Net com diferentes taxas de dilatação.



Os gráficos (a), (b) referem-se ao conjunto de dados *Vessmap*, enquanto os gráficos (c), (d) referem-se ao conjunto de dados *DRIVE*. O ponto vermelho indica o valor do modelo base, enquanto os pontos laranjas mostram os resultados obtidos ao alterar as camadas de subamostragem (downsampling) e por fim os pontos azuis representam as variações obtidas ao modificar o tamanho dos filtros e das taxas de dilatação ao longo das camadas convolucionais. Fonte: Do autor

Figura 18 – Valor de Dice em função do tamanho dos recortes para os conjuntos de dados *Vessmap* e *DRIVE*.



Fonte: Do autor

Capítulo 6

Conclusão

Conduzimos uma análise sistemática para investigar a influência da textura, da forma e do campo receptivo no desempenho de CNNs para a segmentação de vasos sanguíneos. Nosso objetivo foi mensurar de forma quantitativa como as características visuais contribuem e direcionam os processos de decisão em amostras de microscopia e fotografias de fundo de olho. Por meio de experimentos direcionados ao isolamento de cada uma das características, avaliamos a sua importância relativa para o desempenho das CNNs.

O primeiro experimento, que envolveu a manipulação de textura e intensidade a partir de recortes especializados na informação local das imagens, revelou que a intensidade dos pixels tende a ser uma característica mais crítica do que a informação local da textura quando se trata da classificação entre regiões de vasos e fundo. Os modelos treinados em recortes com perturbação da textura, mas com intensidades preservadas, obtiveram desempenho similar aos modelos treinados nas imagens originais. Mesmo com a remoção simultânea da textura e da intensidade, o desempenho da rede se manteve significativamente acima do nível do acaso, indicando sua capacidade de explorar características estatísticas ainda mais sutis do que a percepção humana é capaz. Além disso, a presença das bordas dos vasos aumentou de forma estável a acurácia da classificação, evidenciando a relevância das características contextuais locais.

A análise do papel das características de forma evidenciou as limitações do uso de informações puramente estruturais para a segmentação dos vasos sanguíneos. Ao desafiarmos os modelos a reconstruir vasos completos apenas a partir de seus contornos ou eixos centrais, estes apresentaram um desempenho insatisfatório, além de uma forte tendência à supersegmentação ocasionada pela não especificidade. Esse resultado evidencia que, embora a forma seja uma característica intuitiva para a visão humana, as CNNs dependem fortemente de estatísticas não triviais dos vasos para uma delimitação precisa

e não conseguem extrapolar facilmente o calibre do vaso apenas a partir das informações parciais da forma.

A análise do campo receptivo da rede mostrou que, para os conjuntos de dados estudados, o ERF utilizado na segmentação tende a saturar em um tamanho relativamente pequeno, de aproximadamente 20 pixels, mesmo quando alterações na arquitetura permitem um campo receptivo teórico muito maior. Para as imagens de microscopia do conjunto VessMAP, que não possuem uma estrutura global consistente, o contexto local foi suficiente para atingir o desempenho máximo. Já nas fotografias de fundo de olho do conjunto DRIVE — nas quais existem marcos anatômicos de abrangência global — observou-se uma pequena melhora no desempenho quando o contexto considerado era maior. Estes resultados sugerem que o tamanho de um campo receptivo ótimo apresenta-se dependente da tarefa, mas que a maior parte das informações utilizadas pelas CNNs para a segmentação de vasos está localizada muito próxima aos pixels segmentados.

Ao contrário das CNNs, a arquitetura *Transformer* (VASWANI et al., 2017) possui um campo receptivo global, o que geralmente é considerado uma vantagem em relação às CNNs. O fato de o contexto local parecer suficiente para a segmentação de vasos sanguíneos corrobora estudos recentes que utilizaram CNNs em vez de *Transformers* para esta e outras tarefas relacionadas à imagens médicas (WITTMANN et al., 2025; ISENSEE et al., 2024).

Uma limitação do nosso estudo é que todos os resultados foram avaliados de forma empírica, uma vez que desenvolver modelos matemáticos sobre o funcionamento interno das CNNs não é uma tarefa de fácil execução. Além disso, apenas dois conjuntos de dados foram utilizados nos experimentos. Embora acreditemos que essas descobertas se generalizem para outros conjuntos de imagens de microscopia e fotografias de fundo de olho, pesquisas futuras poderiam verificar a validade desses resultados em diferentes condições de aquisição de imagem.

As percepções obtidas no presente estudo contribuem para um entendimento mais profundo de como as CNNs operam, especialmente no contexto de imagens médicas, e podem orientar o desenvolvimento de modelos mais robustos, eficientes e interpretáveis para a análise de vasos sanguíneos. Pesquisas futuras poderiam investigar características não percebidas por humanos que são utilizadas pelas redes na ausência de textura e intensidade, bem como estender a análise para outras modalidades de imageamento.

Referências

ANTONELLI, M. et al. **The medical segmentation decathlon. Nat Commun 13: 4128**. 2022.

BEHBOODI, B. et al. Receptive field size as a key design parameter for ultrasound image segmentation with u-net. In: IEEE. **2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)**. [S.l.], 2020. p. 2117–2120.

CHADDAD, A. et al. Survey of explainable ai techniques in healthcare. **Sensors**, MDPI, v. 23, n. 2, p. 634, 2023.

CHEN, C. et al. All answers are in the images: A review of deep learning for cerebrovascular segmentation. **Computerized Medical Imaging and Graphics**, Elsevier, v. 107, p. 102229, 2023.

CHUNG, H.; PARK, K. H. Shape prior is not all you need: Discovering balance between texture and shape bias in cnn. In: **Proceedings of the Asian Conference on Computer Vision**. [S.l.: s.n.], 2022. p. 4160–4175.

DAI, D. et al. Rethinking the image feature biases exhibited by deep convolutional neural network models in image recognition. **CAAI Transactions on Intelligence Technology**, Wiley Online Library, v. 7, n. 4, p. 721–731, 2022.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: IEEE. **2009 IEEE conference on computer vision and pattern recognition**. [S.l.], 2009. p. 248–255.

DONG, Z.; XIE, M.; LI, X. Multi-scale receptive fields convolutional network for action recognition. **Applied Sciences**, MDPI, v. 13, n. 6, p. 3403, 2023.

FUKUNAGA, K.; HOSTETLER, L. The estimation of the gradient of a density function, with applications in pattern recognition. **IEEE Transactions on information theory**, IEEE, v. 21, n. 1, p. 32–40, 1975.

GALDRAN, A. et al. State-of-the-art retinal vessel segmentation with minimalistic models. **Scientific Reports**, Nature Publishing Group UK London, v. 12, n. 1, p. 6174, 2022.

GEIRHOS, R. et al. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. **arXiv preprint arXiv:1811.12231**, 2018.

GHOSH, S. et al. Understanding deep learning techniques for image segmentation. **ACM computing surveys (CSUR)**, ACM New York, NY, USA, v. 52, n. 4, p. 1–35, 2019.

GRAVES, A.; MOHAMED, A.-r.; HINTON, G. Speech recognition with deep recurrent neural networks. In: IEEE. **2013 IEEE international conference on acoustics, speech and signal processing**. [S.l.], 2013. p. 6645–6649.

HAO, X. et al. Unetgan: A robust speech enhancement approach in time domain for extremely low signal-to-noise ratio condition. **arXiv preprint arXiv:2010.15521**, 2020.

HE, K. et al. Deep residual learning for image recognition. In: **Proceedings of the IEEE conference on computer vision and pattern recognition**. [S.l.: s.n.], 2016. p. 770–778.

HEINERT, E. et al. Shape bias and robustness evaluation via cue decomposition for image classification and segmentation. **arXiv preprint arXiv:2503.12453**, 2025.

_____. Reducing texture bias of deep neural networks via edge enhancing diffusion. **arXiv preprint arXiv:2402.09530**, 2024.

HERMANN, K.; CHEN, T.; KORNBLITH, S. The origins and prevalence of texture bias in convolutional neural networks. In: LAROCHELLE, H. et al. (Ed.). **Advances in Neural Information Processing Systems**. [S.l.]: Curran Associates, Inc., 2020. v. 33, p. 19000–19015.

HINTON, G. et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. **IEEE Signal processing magazine**, IEEE, v. 29, n. 6, p. 82–97, 2012.

ISENSEE, F. et al. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. **Nature methods**, Nature Publishing Group, v. 18, n. 2, p. 203–211, 2021.

_____. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: SPRINGER. **International Conference on Medical Image Computing and Computer-Assisted Intervention**. [S.l.], 2024. p. 488–498.

ISLAM, M. A. et al. Shape or texture: Understanding discriminative features in cnns. In: **International Conference on Learning Representations**. [S.l.: s.n.], 2021.

JARRETT, K. et al. What is the best multi-stage architecture for object recognition? In: IEEE. **2009 IEEE 12th international conference on computer vision**. [S.l.], 2009. p. 2146–2153.

KASSIM, Y. M. et al. Microvasculature segmentation of arterioles using deep cnn. In: IEEE. **2017 IEEE International conference on image processing (ICIP)**. [S.l.], 2017. p. 580–584.

KIRST, C. et al. Mapping the fine-scale organization and plasticity of the brain vasculature. **Cell**, Elsevier, v. 180, n. 4, p. 780–795, 2020.

LEK, S.; PARK, Y. Artificial neural networks. In: **Encyclopedia of Ecology, Five-Volume Set**. [S.l.]: Elsevier Inc., 2008. p. 237–245.

LI, H. et al. Human treelike tubular structure segmentation: A comprehensive review and future perspectives. **Computers in Biology and Medicine**, Elsevier, v. 151, p. 106241, 2022.

LOOS, V.; PARDASANI, R.; AWASTHI, N. Demystifying the effect of receptive field size in u-net models for medical image segmentation. **Journal of Medical Imaging**, Society of Photo-Optical Instrumentation Engineers, v. 11, n. 5, p. 054004–054004, 2024.

LUO, W. et al. Understanding the effective receptive field in deep convolutional neural networks. **Advances in neural information processing systems**, v. 29, 2016.

MA, J. et al. Segment anything in medical images. **Nature Communications**, Nature Publishing Group UK London, v. 15, n. 1, p. 654, 2024.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MOCCIA, S. et al. Blood vessel segmentation algorithms—review of methods, datasets and evaluation metrics. **Computer methods and programs in biomedicine**, Elsevier, v. 158, p. 71–91, 2018.

MOOKIAH, M. R. K. et al. A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. **Medical Image Analysis**, Elsevier, v. 68, p. 101905, 2021.

MUTUA, E. N.; KASAMANI, B. S.; REICH, C. Deep learning applications for diabetic retinopathy and retinopathy of prematurity diseases diagnosis: a systematic review. **International Journal of Ophthalmology**, v. 18, n. 8, p. 1594, 2025.

MÜTZE, A. et al. On the influence of shape, texture and color for learning semantic segmentation. **arXiv preprint arXiv:2410.14878**, 2024.

NAZIR, S.; DICKSON, D. M.; AKRAM, M. U. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. **Computers in Biology and Medicine**, Elsevier, v. 156, p. 106668, 2023.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: SPRINGER. **Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18**. [S.l.], 2015. p. 234–241.

RUDIN, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. **Nature machine intelligence**, Nature Publishing Group UK London, v. 1, n. 5, p. 206–215, 2019.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **nature**, Nature Publishing Group UK London, v. 323, n. 6088, p. 533–536, 1986.

RUSSAKOVSKY, O. et al. Imagenet large scale visual recognition challenge. **International journal of computer vision**, Springer, v. 115, p. 211–252, 2015.

SILVA, M. Viana da et al. A new dataset for measuring the performance of blood vessel segmentation methods under distribution shifts. **PloS one**, Public Library of Science San Francisco, CA USA, v. 20, n. 5, p. e0322048, 2025.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. **arXiv preprint arXiv:1409.1556**, 2014.

SINGH, A.; SENGUPTA, S.; LAKSHMINARAYANAN, V. Explainable deep learning models in medical image analysis. **Journal of imaging**, MDPI, v. 6, n. 6, p. 52, 2020.

STAAL, J. et al. Ridge-based vessel segmentation in color images of the retina. **IEEE Transactions on Medical Imaging**, v. 23, n. 4, p. 501–509, 2004.

SYTWU, K.; GROSCHNER, C.; SCOTT, M. C. Understanding the influence of receptive field and network complexity in neural network-guided tem image analysis. **Microscopy and Microanalysis**, Cambridge University Press, v. 28, n. 6, p. 1896–1904, 2022.

SZEGEDY, C. et al. **Going deeper with convolutions. CoRR abs/1409.4842 (2014)**. 2014.

TODOROV, M. I. et al. Machine learning analysis of whole mouse brain vasculature. **Nature methods**, Nature Publishing Group, v. 17, n. 4, p. 442–449, 2020.

TRIPATHI, A. et al. Edges to shapes to concepts: Adversarial augmentation for robust vision. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2023. p. 24470–24479.

VASWANI, A. et al. Attention is all you need. **Advances in neural information processing systems**, v. 30, 2017.

WANG, P. et al. Understanding convolution for semantic segmentation. In: IEEE. **2018 IEEE winter conference on applications of computer vision (WACV)**. [S.l.], 2018. p. 1451–1460.

WANG, S. et al. Hierarchical retinal blood vessel segmentation based on feature and ensemble learning. **Neurocomputing**, Elsevier, v. 149, p. 708–717, 2015.

WITTMANN, B. et al. vesselfm: A foundation model for universal 3d blood vessel segmentation. In: **Proceedings of the Computer Vision and Pattern Recognition Conference**. [S.l.: s.n.], 2025. p. 20874–20884.

XIE, S.; TU, Z. Holistically-nested edge detection. In: **Proceedings of the IEEE international conference on computer vision**. [S.l.: s.n.], 2015. p. 1395–1403.

ZHANG, Y.; MAZUROWSKI, M. A. Convolutional neural networks rarely learn shape for semantic segmentation. **Pattern Recognition**, Elsevier, v. 146, p. 110018, 2024.

ZHANG, Z.; SABUNCU, M. Generalized cross entropy loss for training deep neural networks with noisy labels. **Advances in neural information processing systems**, v. 31, 2018.