UNIVERSIDADE FEDERAL DO MARANHÃO

Programa de Pós-Graduação em Ciência da Computação

Marcus Vinicius Silva Lima de Oliveira

Atenção Multiescala em Redes U-Net: Uma Abordagem para Segmentação de Rins, Tumores e Cistos em Tomografia Computadorizada

São Luís - MA

2026

Marcus Vinicius Silva Lima de Oliveira

Atenção Multiescala em Redes U-Net: Uma Abordagem para Segmentação de Rins, Tumores e Cistos em Tomografia Computadorizada

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Programa de Pós-Graduação em Ciência da Computação

Universidade Federal do Maranhão

Orientador: Prof. Dr. Geraldo Braz Junior

São Luís - MA

2026

Ficha gerada por meio do SIGAA/Biblioteca com dados fornecidos pelo(a) autor(a).
Diretoria Integrada de Bibliotecas/UFMA

Silva Lima de Oliveira, Marcus Vinicius.
Atenção Multiescala em Redes U-Net: Uma Abordagem para Segmentação de Rins, Tumores e Cistos em Tomografia Computadorizada / Marcus Vinicius Silva Lima de Oliveira. - 2026.
92 f.

Orientador(a): Geraldo Braz Junior.
Dissertação (Mestrado) - Programa de Pós-graduação em Ciência da Computação/ccet, Universidade Federal do Maranhão, Auditório Núcleo de Computação Aplicada, 2026.

1. Câncer Renal. 2. Kits23. 3. Atenção Multiescala. 4. Segmentação Semântica. 5. Squeeze-and-excitation. I. Braz Junior, Geraldo. II. Título.

Marcus Vinicius Silva Lima de Oliveira

Atenção Multiescala em Redes U-Net: Uma Abordagem para Segmentação de Rins, Tumores e Cistos em Tomografia Computadorizada

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Ciência da Computação, ao Programa de Pós-Graduação em Ciência da Computação, da Universidade Federal do Maranhão.

Trabalho aprovado São Luís - MA, 23 de Abril de 2026:

Prof. Dr. Geraldo Braz Junior
Orientador
Universidade Federal do Maranhão

Prof. Dr. João Dallyson Sousa de Almeida
Examinador Interno
Universidade Federal do Maranhão

Profa. Dra. Aura Conci
Examinadora Externa
Universidade Federal Fluminense

São Luís - MA
2026

Aos meus pais, futura esposa e amigos

Agradecimentos

A Deus, em primeiro lugar, por todas as bênçãos e graças que guiaram meus passos ao longo desta caminhada.

Aos meus pais, que me apoiaram com amor e dedicação em cada etapa desta jornada, sempre fazendo o possível e o impossível para que eu pudesse estudar, incentivando minhas escolhas e estando ao meu lado nos momentos de dúvidas e de conquistas. Pelos sacrifícios silenciosos e pelo cuidado de quem nunca mediu esforços, toda a minha gratidão.

À minha futura esposa Milena Serejo, pelo amor incondicional, pela paciência e pela companhia que tornaram esses anos mais leves. Pelo suporte essencial nos momentos mais exigentes do mestrado e por ser, silenciosamente, uma das peças fundamentais desta conquista.

Ao meu irmão, pelo apoio constante ao longo deste trajeto, pela presença nos momentos que mais importaram e por sempre me incentivar a dar o melhor de mim e a crescer como pessoa.

Ao meu orientador Prof. Dr. Geraldo Braz Junior, pela confiança depositada desde que ingressei no VipLab em 2020, ainda na graduação, e pelo direcionamento e paciência que me acompanharam em cada etapa desta trajetória. Sou profundamente grato pelos conselhos, pelos ensinamentos e, sobretudo, pela amizade construída ao longo desses anos, presença fundamental na minha formação acadêmica e pessoal.

Aos amigos que fiz ao longo desta jornada, na graduação e no mestrado, tanto no VipLab quanto no NCA, pela acolhida generosa e pela convivência que tornaram o caminho mais leve e enriquecedor. Em especial a Gabriel Costa, Daniel Pinto, Alison Mendes, Weslley Kelson, João Pedro, Vitor Rogério e Alexandre Pessoa, pela amizade, pelos conselhos e pelo apoio que levarei para a vida. A presença de cada um contribuiu, à sua maneira, para que esta etapa fosse vivida com mais alegria e significado.

À memória dos meus avós, que partiram durante esta jornada, mas cuja presença permanece em cada passo que dou. Em especial ao meu avô Salvador, que sempre sonhou ver os filhos e os netos formados, sonho que hoje, ao me formar, tenho a honra de realizar por ele.

Ao Núcleo de Computação Aplicada (NCA) e ao Programa de Pós-Graduação em Ciência da Computação, pelo suporte institucional e pelas condições oferecidas ao longo desta pesquisa.

Aos professores, membros da banca examinadora, pela disponibilidade para avaliação deste trabalho.

"A educação tem raízes amargas, mas os seus frutos são doces."

(Aristóteles)

Resumo

A detecção e o diagnóstico precoce do câncer renal desempenham um papel fundamental no prognóstico e no tratamento, elevando significativamente as chances de cura e de sobrevivência dos pacientes. Nesse contexto, os avanços nos exames radiológicos permitem ao médico especialista realizar, por meio de imagens, a análise e identificação de lesões suspeitas. A Tomografia Computadorizada (TC) destaca-se como uma ferramenta amplamente utilizada no diagnóstico do câncer renal, devido à sua capacidade de gerar um grande volume de imagens detalhadas das estruturas internas do corpo, incluindo os rins, os cistos e os tumores. Entretanto, a análise manual desses exames é um processo trabalhoso e suscetível a erros decorrentes de fadiga e distração, o que evidencia a necessidade de métodos computacionais que auxiliem na segmentação automática dessas estruturas. Nos últimos anos, as Redes Neurais Convolucionais (CNNs) têm se destacado nessa tarefa, fornecendo suporte relevante aos especialistas em imagens médicas. Nesta pesquisa, desenvolvemos um modelo convolucional denominado Dual-Scale SE U-Net, que explora a extração de características em múltiplas escalas por meio de convoluções paralelas (3×3 e 7×7), combinadas com mecanismos de atenção por canal baseados no módulo *Squeeze-and-Excitation* (SE), integrados à arquitetura U-Net para segmentação de rins, cistos e tumores em imagens de tomografia computadorizada. Como etapa de pré-processamento, as imagens foram redimensionadas de 512×512 para 256×256 pixels, visando maior eficiência computacional. A metodologia produziu resultados promissores no conjunto de dados do desafio KiTS23, avaliado por meio de validação cruzada com cinco iterações, apresentando coeficiente de similaridade *Dice* de 90,94% para Rins e Massas, 89,52% para Massas Renais e 86,27% para Tumores Renais. Na análise por estruturas individuais, foram obtidos valores de 93,80% para rins, 92,76% para cistos e 86,27% para tumores, evidenciando a eficácia da abordagem proposta como ferramenta de apoio ao diagnóstico médico.

Palavras-chave: Câncer Renal; Tomografia Computadorizada; Redes Neurais Convolucionais; Segmentação Semântica; U-Net; Atenção Multiescala; Squeeze-and-Excitation; KiTS23.

Abstract

Early detection and diagnosis of renal cancer play a crucial role in patient prognosis and treatment, significantly increasing the chances of survival and cure. In this context, advances in radiological imaging have enabled medical specialists to analyze and identify suspicious lesions more effectively. Computed Tomography (CT) stands out as a widely used tool for renal cancer diagnosis due to its ability to generate high-resolution images of internal body structures, including the kidneys, cysts, and tumors. However, manual analysis of these exams is time-consuming and error-prone, often affected by fatigue and distraction, highlighting the need for computational methods to support the automatic segmentation of these structures. In recent years, Convolutional Neural Networks (CNNs) have shown significant potential in this task, providing valuable support to medical imaging specialists. In this study, we propose a convolutional model named Dual-Scale SE U-Net, which leverages multi-scale feature extraction through parallel convolutions (3×3 and 7×7), combined with channel attention mechanisms based on the *Squeeze-and-Excitation* (SE) module, integrated into the U-Net architecture for the segmentation of kidneys, cysts, and tumors in CT images. As a preprocessing step, the images were resized from 512×512 to 256×256 pixels to improve computational efficiency. The proposed methodology achieved promising results on the KiTS23 dataset, evaluated using five-fold cross-validation, yielding Dice similarity coefficients of 90.94% for Kidneys and Masses, 89.52% for Renal Masses, and 86.27% for Renal Tumors. In the analysis of individual structures, the model achieved 93.80% for kidneys, 92.76% for cysts, and 86.27% for tumors, demonstrating the effectiveness of the proposed approach as a supportive tool for medical diagnosis.

Keywords: Renal Cancer; Computed Tomography; Convolutional Neural Networks; Semantic Segmentation; U-Net; Multi-scale Attention; Squeeze-and-Excitation; KiTS23.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

Adam	*Adaptive Moment Estimation*
ASPP	*Atrous Spatial Pyramid Pooling*
BN	*Batch Normalization*
CAH	*Classes de Avaliação Hierárquica*
CAT	*Computed Axial Tomography (Tomografia Axial Computadorizada)*
CNN	*Convolutional Neural Network (Rede Neural Convolucional)*
CPU	*Central Processing Unit (Unidade Central de Processamento)*
CR	*Câncer de rim)*
DCFL	*Dice Categorical Focal Loss*
DSC	*Dice Similarity Coefficient*
ECA	*Efficient Channel Attention*
ELU	*Exponential Linear Unit*
FC	*Fully Connected*
FN	*Falso Negativo*
FP	*Falso Positivo*
GAP	*Global Average Pooling*
GCA	*Global Channel Attention*
GELU	*Gaussian Error Linear Unit*
GFLOPs	*Giga Floating Point Operations per Second*
GPU	*Graphics Processing Unit (Unidade de Processamento Gráfico)*
Grad-CAM	*Gradient-weighted Class Activation Mapping*
GSA	*Global Spatial Attention*
HD	*Hausdorff Distance*

HD95	*95th percentile Hausdorff Distance*
HU	*Hounsfield Units*
IMC	*Índice de Massa Corporal*
IoU	*Intersection over Union*
JCC	*Jaccard Similarity Coefficient*
KiTS23	*Kidney Tumor Segmentation Challenge 2023*
KNN	*k-Nearest Neighbors*
LeakyReLU	*Leaky Rectified Linear Unit*
MLP	*Multi-Layer Perceptron*
MSD	*Mean Surface Distance*
NIfTI	*Neuroimaging Informatics Technology Initiative*
OBJ	*Object File Format*
PET	*Tomografia por Emissão de Pósitrons*
Pixel	*Picture Element*
PPM	*Pyramid Pooling Module*
PSPNet	*Pyramid Scene Parsing Network*
RAM	*Random Access Memory (Memória de Acesso Aleatório)*
RCC	*Carcinoma de Células Renais*
ReLU	*Rectified Linear Unit*
ResNet	*Residual Network*
RNA	*Rede Neural Artificial*
RNC	*Rede Neural Convolucional*
ROI	*Region of Interest (Região de Interesse)*
SE	*Squeeze-and-Excitation*
TC	*Tomografia Computadorizada*
ViT	*Vision Transformer*

VN	*Verdadeiro Negativo*
VOE	*Volumetric Overlap Error*
Voxel	*Volume Element*
VP	*Verdadeiro Positivo*
VRAM	*Video Random Access Memory*
YOLO	*You Only Look Once*

Sumário

1 Introdução

O câncer representa um conjunto heterogêneo de doenças caracterizadas por transformações celulares sequenciais, nas quais as células progressivamente adquirem propriedades anormais de proliferação, de evasão da morte celular e de capacidade invasiva. Apesar da notável heterogeneidade genômica e fenotípica entre os diferentes tipos tumorais, estas neoplasias convergem em características biológicas essenciais que governam sua progressão(HANAHAN, 2022). O impacto do diagnóstico oncológico estende-se além das manifestações clínicas, impondo uma carga significativa socioeconômica, funcional e psicossocial aos pacientes e às suas famílias (KITAW et al., 2025). Em resposta a esta complexa realidade, avanços substanciais têm sido alcançados nas estratégias de detecção precoce do câncer. Tais progressos são impulsionados, principalmente, pela integração de tecnologias de imagem molecular e anatômica, exemplificada pela tomografia por emissão de pósitrons acoplada à tomografia computadorizada (PET/TC), que possibilita a detecção simultânea de alterações metabólicas incipientes e a correlação espacial precisa.

A vigilância epidemiológica do câncer constitui um pilar fundamental para a formulação de políticas públicas e o direcionamento estratégico de recursos em saúde, permitindo a análise sistemática da magnitude, dos padrões de distribuição e da evolução temporal da doença (SANTOS et al., 2023). Dados da *American Cancer Society* ilustram a relevância desse monitoramento: em 2024, foram projetados aproximadamente 2.001.140 casos de câncer e 611.720 óbitos nos Estados Unidos, com destaque para a redução de 33% na taxa de mortalidade observada entre 1991 e 2021, equivalente a mais de 4 milhões de mortes prevenidas (SIEGEL; GIAQUINTO; JEMAL, 2024). Paralelamente a esses avanços, emergem desafios epidemiológicos significativos, incluindo o crescimento na incidência de seis dos dez tipos de câncer mais frequentes e a ascensão do câncer colorretal como principal causa de morte oncológica entre homens abaixo de 50 anos, uma mudança dramática em relação à quarta posição que ocupava no final dos anos 1990 (SIEGEL; GIAQUINTO; JEMAL, 2024).

No cenário brasileiro, as estatíticas do Instituto Nacional de Câncer para o triênio 2023-2025 apontam aproximadamente 704 mil casos novos anuais, sendo 483 mil quando excluído o câncer de pele não melanoma. A distribuição geográfica desses casos reflete disparidades regionais marcadas: as regiões Sul e Sudeste concentram 65,5% da incidência total e apresentam perfil epidemiológico que se aproxima do observado em países de alta renda, evidenciando a complexa heterogeneidade do panorama oncológico nacional (SANTOS et al., 2023).

O câncer de rim (CR) é o 12º câncer mais comum no mundo. Sua incidência global

está em ascensão, estimada em 400.000 novos casos por ano. O carcinoma de células renais claras é o tipo mais frequente da doença em adultos. Além disso, o câncer de rim apresenta uma taxa de mortalidade mundial expressiva, estimada em 175.000 mortes por ano. Projeções atuais sugerem que a incidência continuará a aumentar na próxima década, com estimativa de que o número global de casos chegue a 475,4 mil até 2030 (CIRILLO; INNOCENTI; BECHERUCCI, 2024).

O carcinoma de células renais (RCC) representa aproximadamente 2-3% de todas as neoplasias malignas em adultos, sendo o tumor sólido mais comum do rim, correspondendo a cerca de 90% das malignidades renais. Trata-se de uma neoplasia predominantemente silenciosa, com aproximadamente 90% dos casos adultos distribuídos entre os subtipos células claras (75%, o mais frequente e agressivo), papilar e cromófobo (PADALA et al., 2020). A patologia demonstra predominância no sexo masculino, com razão de 1,5:1 em relação às mulheres, e o pico de incidência ocorre entre os 60 e 70 anos. Em termos de impacto global, estimou-se a ocorrência de 434.840 novos casos e 155.953 mortes em 2022 (BEX et al., 2025).

Estudos apontam que o crescimento da incidência de câncer decorre principalmente do envelhecimento da população, da expansão demográfica e da exposição crescente a fatores de risco evitáveis (BRAY et al., 2024). Entre os principais determinantes modificáveis da doença, o tabagismo emerge como fator predominante, sendo responsável por 18-30% dos casos, enquanto a obesidade contribui com 13,4% dos casos atribuíveis ao IMC elevado, e a hipertensão amplifica o risco em até 10 vezes. Adicionalmente, a exposição ocupacional a compostos nefrotóxicos, como cádmio e amianto, representa um risco significativo, e todos esses fatores convergem para um mecanismo patogênico comum: a indução de lesão renal aguda e o desenvolvimento de doença renal crônica, condições que elevam substancialmente o risco de progressão para o câncer renal (CIRILLO; INNOCENTI; BECHERUCCI, 2024).

A maioria dos casos de câncer de rim (CR) é descoberta por acaso em exames de imagem realizados por outros motivos. Um estudo demonstrou que o diagnóstico foi incidental em 60% dos pacientes no geral e em 87% dos casos em estágio inicial. A "tríade clássica" de sintomas, composta por dor no flanco, sangue na urina e massa abdominal, é considerada incomum hoje em dia (BEX et al., 2025).

A detecção e o diagnóstico precoce do câncer renal são fundamentais para estabelecer estratégias terapêuticas eficazes e aumentar significativamente as chances de cura. O prognóstico está diretamente relacionado ao estágio da doença no momento do diagnóstico:a sobrevida em 5 anos é de 93% para a doença localizada (estágio I), caindo para 72,5% na presença de envolvimento linfonodal regional (estágios II/III) (PADALA et al., 2020). Já em casos metastáticos (estágio IV), a taxa de sobrevida reduz-se a apenas 12%. Este cenário torna-se ainda mais crítico, considerando que aproximadamente um terço dos pacientes já apresenta metástases no momento do diagnóstico inicial e que entre 20% e

50% dos casos tratados cirurgicamente desenvolvem doença metastática posteriormente (PADALA et al., 2020). Esses dados reforçam a importância crucial da identificação precoce da neoplasia renal, quando as opções terapêuticas são mais eficazes e o prognóstico é mais favorável.

1.1 Definição do Problema

Embora as tecnologias de segmentação automática representem ferramentas valiosas no auxílio diagnóstico, sua aplicação em imagens de TC renal enfrenta limitações significativas. Estruturas como tumores e cistos, caracterizadas por morfologia irregular e baixo contraste, frequentemente resultam em segmentações imprecisas e na detecção inadequada de bordas, gerando falsos positivos que comprometem a confiabilidade diagnóstica.

Na tomografia computadorizada, desafios intrínsecos como a similaridade entre tecidos adjacentes, a variabilidade anatômica renal e a complexidade na identificação de patologias dificultam a segmentação automatizada.

O volume crescente de imagens e a natureza repetitiva da análise expõem o radiologista à fadiga cognitiva e ao estresse visual, com estudos indicando que a frequência de erros tende a atingir seu ápice em torno de 9 horas de turno. Nesse cenário, ferramentas de Inteligência Artificial tornam-se recursos fundamentais por atuarem como uma "rede de segurança" capaz de mitigar falhas perceptivas e vieses cognitivos (ALEXANDER et al., 2022).

Esta dissertação propõe um método de segmentação semântica para a identificação automática de rins, tumores e cistos em exames de TC, visando aprimorar a acurácia diagnóstica e o planejamento terapêutico oncológico.

1.2 Questões de Pesquisa

As questões de pesquisa relacionadas ao problema de segmentação de rins e tumores em tomografia computadorizada são:

- Q1: Qual o impacto de mecanismos de atenção (canal, espacial ou híbridos) na capacidade de modelos de segmentação para distinguir regiões de interesse (tumores, cistos) de tecidos circundantes com características semelhantes em imagens de TC?
- Q2: Como a integração de convoluções em multiescala com mecanismos de atenção impacta a capacidade do modelo de capturar simultaneamente características locais (detalhes anatômicos) e globais (contexto da imagem) em diferentes resoluções?

1.3 Objetivo

Este trabalho tem como objetivo geral a segmentação das regiões de rim, tumor e cisto em imagens de TC através de uma versão modificada da rede U-Net, incorporando blocos Dual-Scale SE, visando melhorar a precisão na segmentação de estruturas anatômicas e patológicas.

1.3.1 Objetivos Específicos

Para alcançar o objetivo geral deste trabalho, alguns objetivos específicos devem ser atingidos:

- Investigar a utilização de mecanismos de atenção multiescala em uma arquitetura U-Net para melhorar a extração de características;
- Otimizar os hiperparâmetros da arquitetura proposta, visando maximizar a eficiência e a acurácia da segmentação;
- Validar os métodos propostos por meio de experimentos aplicados em bases de dados públicas de imagens de tomografia computadorizada, amplamente reconhecidas na literatura;
- Comparar o desempenho do modelo proposto com métodos existentes na literatura, utilizando como base as métricas de *Dice Score*.

1.4 Organização da Dissertação

Os demais capítulos desta dissertação foram estruturados da seguinte forma:

- O Capítulo 2 descreve trabalhos relacionados à problemática da segmentação de rins, cistos e tumores renais em imagens de tomografia computadorizada.
- O Capítulo 3 apresenta a fundamentação teórica necessária à construção da pesquisa proposta. Neste capítulo, são abordados os temas referentes ao câncer renal, às tomografias computadorizadas, ao processamento de imagens, à aprendizagem de máquina, às métricas de avaliação e às arquiteturas de redes neurais utilizadas.
- Os Capítulos 4 e 5 descrevem todas as etapas que compõem a metodologia proposta para esta pesquisa, assim como os resultados obtidos, discussões e comparações com pesquisas correlatas;
- Por fim, o Capítulo 6 apresenta as considerações finais sobre os resultados e propostas de trabalhos futuros.

2 Trabalhos Relacionados

Este capítulo apresenta uma revisão de literatura sobre a segmentação de estruturas renais em imagens de tomografia computadorizada, abordando especificamente a delimitação de rins, cistos e tumores. A análise concentra-se na aplicação de técnicas de aprendizado profundo, examinando tanto a identificação de elementos associados ao carcinoma renal quanto as modificações propostas em redes neurais para extração de informações clínicas relevantes. Essas tecnologias de segmentação automática auxiliam médicos no diagnóstico e no planejamento cirúrgico, impulsionando diversas pesquisas voltadas ao desenvolvimento e à validação desses sistemas.

Os estudos selecionados foram organizados em dois eixos temáticos principais: Modificações Arquiteturais e Mecanismos de Atenção, que englobam alterações na estrutura interna das redes para melhor captura de contexto; e Processamento Multi-estágio e Análise de Incerteza, que agrupam abordagens sequenciais, métodos de *ensemble* e quantificação de confiabilidade diagnóstica. A validação majoritária dos métodos ocorre na base de dados KiTS23, cuja estrutura introduz classes de avaliação hierárquica (CAH). Estas categorias ampliadas agrupam os resultados da segmentação em conjuntos lógicos (Rins e Massas, Massas Renais e Tumores), seguindo conceitos que serão detalhados na Seção 5.2.

2.1 Modificações Arquiteturais e Mecanismos de Atenção

A adaptação de codificadores e a inserção de módulos de atenção em redes convolucionais constituem estratégias centrais para superar limitações na extração de características em exames volumétricos. Qian et al. (2023) desenvolvem uma arquitetura híbrida que combina a nnU-Net com o *Swin Transformer*. A premissa metodológica deste estudo reside na complementaridade, pois as camadas convolucionais processam características locais e estruturais, enquanto o mecanismo de autoatenção do Transformer modela dependências de longo alcance e de contexto global. Essa fusão arquitetural permite que a rede explore simultaneamente informações espaciais finas e relações semânticas distantes, superando a limitação intrínseca das convoluções tradicionais que possuem campos receptivos restritos. A integração de janelas deslizantes hierárquicas do *Swin Transformer* à estrutura de codificador-decodificador da nnU-Net demonstra que a combinação de paradigmas arquiteturais complementares pode elevar a capacidade de generalização do modelo sem exigir um aumento proporcional de parâmetros treináveis.

Em uma linha similar de otimização estrutural, Myronenko et al. (2023) aplicam a técnica de busca por arquitetura neural, especificamente o algoritmo DiNTS, para adaptar as redes SegResNet e SwinUNETR. O objetivo desta abordagem é automatizar

a definição de uma topologia que capture contextos amplos sem incorrer em custos computacionais excessivos, tipicamente associados ao empilhamento de múltiplas camadas convolucionais profundas. A busca automatizada explora um espaço de configurações que equilibra profundidade, largura e tipo de operação, permitindo a descoberta de arquiteturas otimizadas para características específicas do domínio médico, como a heterogeneidade morfológica de tumores renais e a variabilidade de protocolos de aquisição tomográfica.

A exploração de módulos de pooling piramidais como alternativa ao codificador padrão da U-Net é proposta por Matos et al. (2024). A arquitetura CPP-UNet combina o *Pyramid Pooling Module (PPM)* e o *Atrous Spatial Pyramid Pooling (ASPP)* no estágio de codificação, permitindo a captura de informações contextuais em múltiplas escalas por meio de convoluções dilatadas com taxas crescentes. O PPM extrai características aplicando operações de pooling em diferentes resoluções espaciais, enquanto o ASPP emprega convoluções com dilatação progressiva para observar objetos em contextos globais sem aumentar o número de parâmetros. A fusão desses módulos no codificador da U-Net proporciona uma representação mais rica e multirresolução das estruturas renais, facilitando a diferenciação de tecidos com textura e intensidade similares, como cistos e tumores sólidos. Essa estratégia arquitetural demonstra que a combinação de técnicas de pooling piramidal pode superar abordagens baseadas em convoluções padrão, especialmente em cenários onde a variabilidade de tamanho e forma das lesões exige análise em múltiplas escalas.

A comparação entre arquiteturas de codificador-decodificador é explorada por Jariwala et al. (2024), que avaliam o desempenho da DeepLabv3+ em relação à U-Net tradicional para segmentação de rins e tumores renais em imagens de tomografia computadorizada. A DeepLabv3+ se diferencia pela incorporação de convoluções dilatadas no módulo ASPP e pela presença de um decodificador simplificado que refina as bordas dos objetos segmentados por meio de conexões de baixo nível. Os autores argumentam que a arquitetura DeepLabv3+ oferece vantagens em termos de captura de contexto global e de precisão nas fronteiras de segmentação, atributos essenciais para delinear tumores com margens irregulares. A análise comparativa evidencia que a escolha da arquitetura de decodificação impacta diretamente a qualidade da segmentação, especialmente em regiões com transições abruptas de intensidade ou sobreposição anatômica.

Para enfrentar o desafio da variação intra-classe, particularmente na distinção entre tumores e cistos renais, Hu e Peng (2023) propõem um sistema composto onde uma U-Net 3D atua inicialmente filtrando regiões de interesse, seguida pela aplicação da *Global Spatial Channel Attention Network* (GSCA-Net). Esta rede integra mecanismos de atenção espacial global (GSA) e de atenção de canal global (GCA), projetados para refinar a discriminação de tecidos com características radiológicas semelhantes. O módulo GSA calcula pesos de atenção sobre o domínio espacial completo, priorizando regiões

anatomicamente relevantes e suprimindo artefatos de fundo, enquanto o módulo GCA recalibra os canais de características para enfatizar mapas de ativação mais informativos. A integração desses mecanismos permite que a rede ajuste dinamicamente sua atenção com base na complexidade local da imagem, resultando em segmentações mais robustas em cenários de baixo contraste ou presença de heterogeneidade tecidual.

De forma análoga, Li, Peng e Zhang (2023) modificam a estrutura interna da nnU-Net inserindo um mecanismo de atenção residual. Os autores defendem que essa alteração direciona o gradiente durante o treinamento para características mais relevantes, facilitando a identificação progressiva e especializada de rins, cistos e massas tumorais. A atenção residual funciona como um filtro adaptativo que amplifica sinais discriminativos e atenua informações redundantes, melhorando a eficiência do aprendizado em redes profundas. Já Chen e Zhang (2023) estruturam a rede ResUNet substituindo o codificador padrão da U-Net por blocos residuais. Esta modificação arquitetural, combinada a etapas de recorte de regiões de interesse, visa preservar a integridade do sinal em redes profundas e permite a extração de características sensíveis a variações sutis na morfologia renal. A introdução de conexões residuais mitiga o problema de degradação de gradiente, viabilizando o treinamento de arquiteturas mais profundas sem comprometer a estabilidade numérica ou a convergência.

A substituição de codificadores por arquiteturas consolidadas em visão computacional também é explorada para maximizar o detalhamento das segmentações. Vispi (2023) integram a EfficientNet-B5 como codificador na estrutura U-Net, utilizando otimização via *GridSearch* para ajuste de hiperparâmetros. A escolha da EfficientNet-B5 justifica-se por sua competência no processamento de imagens de alta resolução, fator crítico para delinear bordas irregulares de lesões renais. A arquitetura EfficientNet emprega escalamento composto que equilibra profundidade, largura e resolução de entrada, resultando em uma representação hierárquica eficiente de características visuais.

2.2 Processamento Multi-estágio, Eficiência e Adaptação de Domínio

Estratégias que decompõem o problema de segmentação em etapas sequenciais ou que quantificam a confiabilidade do modelo têm demonstrado eficácia na redução de falsos positivos. Alexandru e Popescu (2025) propõem um fluxo de trabalho modular operando predominantemente em 2D, iniciado por um modelo YOLO para detecção da região renal. A segmentação subsequente é bifurcada em duas redes independentes: uma U-Net com codificador EfficientNet-B4 processa rins e massas, e outra U-Net com *Mix Transformer* (MiT-B2) foca especificamente em tumores. O estudo reporta resultados competitivos, argumentando que a especialização de redes 2D aliada a um pós-processamento 3D oferece

uma alternativa computacionalmente eficiente frente a modelos volumétricos densos.

Nessa mesma linha de detecção preliminar, Tanasković et al. (2024) exploram o YOLOv8 como ferramenta de segmentação de instâncias em fatias 2D extraídas de volumes tomográficos. A premissa metodológica reside na capacidade do YOLOv8 de processar imagens em tempo real, mantendo alta precisão na identificação de tumores renais. Os autores argumentam que modelos pré-treinados em tarefas genéricas de detecção de objetos, quando adaptados a dados médicos por meio de ajuste fino, conseguem generalizar adequadamente para segmentação de lesões, evitando o treinamento custoso de arquiteturas volumétricas completas. Essa abordagem facilita a integração clínica ao reduzir requisitos de hardware, permitindo inferência rápida em ambientes com recursos limitados.

A eficiência computacional também é central na proposta de Karunanayake et al. (2025), que estruturam um pipeline de dois estágios combinando *Vision Transformers* (ViTs) e redes convolucionais. No primeiro estágio, um modelo baseado em ViT realiza a segmentação do órgão renal em baixa resolução, aproveitando a capacidade dos Transformers de capturar dependências de longo alcance e contexto global. Subsequentemente, uma U-Net 3D equipada com mecanismos de atenção processa a região de interesse identificada em alta resolução, refinando a delimitação de tumores pequenos e de baixo contraste. A divisão de tarefas entre um modelo especializado em localização global e outro focado em detalhamento fino permite equilíbrio entre precisão diagnóstica e viabilidade computacional, demonstrando que a fusão de paradigmas arquiteturais complementares supera limitações inerentes ao uso isolado de CNNs ou Transformers.

Para mitigar o impacto da complexidade volumétrica sem comprometer a resolução espacial, Alonso-Monsalve et al. (2025) introduzem redes convolucionais esparsas baseadas em subvariedades (*Submanifold Sparse Convolutional Networks*). A metodologia aplica esparsificação de voxels ao filtrar regiões de interesse com base em limiares de Unidades Hounsfield, retendo apenas estruturas anatomicamente relevantes e descartando tecidos como ar e osso. Essa representação esparsa é processada por uma U-Net modificada que opera exclusivamente sobre voxels significativos, reduzindo drasticamente o consumo de memória e tempo de inferência. A estratégia de dois estágios adotada pelos autores inicia com uma rede esparsa em baixa resolução para localização grosseira, seguida por refinamento em alta resolução dentro da região delimitada. Os resultados evidenciam que a esparsificação não apenas viabiliza o processamento de volumes inteiros em GPUs convencionais, mas também preserva a fidelidade da segmentação, representando um avanço significativo em termos de sustentabilidade energética e escalabilidade para aplicações clínicas de grande porte.

No contexto de cascatas baseadas na nnU-Net, Wang et al. (2023) implementam um processo onde a primeira instância da rede extrai uma região de interesse grosseira, que serve de entrada para uma segunda instância dedicada ao refinamento de tumores

e cistos. Salahuddin et al. (2023) sofisticam este conceito ao aplicar uma *bounding box* que engloba ambos os rins após a localização preliminar, recortando a imagem para uma segunda passagem de segmentação multiclasse. O diferencial metodológico deste trabalho reside no uso de uma taxa de aprendizado cíclica, que permite extrair os pesos do modelo em múltiplos máximos locais durante o treinamento. Isso gera um *ensemble* implícito que auxilia na redução de dimensionalidade e na detecção de falsos positivos, sem o custo de treinar múltiplos modelos do zero.

A validação da segurança diagnóstica por meio da análise de incerteza é abordada por Sina Ziaee, Maleki e Ovens (2025) com o método Rel-UNet. A técnica emprega o escalonador *Stochastic Gradient Descent with Warm Restarts* para forçar a rede a convergir para diferentes mínimos locais, coletando *checkpoints* nas épocas específicas do treinamento. A combinação dessas predições gera mapas de incerteza que superam métodos tradicionais, atingindo um *Expected Calibration Error* reduzido comparado ao *Monte Carlo Dropout*, mostrando-se superior na identificação de dados fora da distribuição. Paralelamente, Uhm et al. (2023) exploram o paralelismo ao processar a entrada simultaneamente em duas U-Nets 3D com resoluções distintas, original e reamostrada. A inovação reside no pós-processamento multiescala, que descarta regiões segmentadas que não apresentam interseção nas saídas das duas redes, funcionando como um filtro de consistência espacial para a segmentação final.

A transferência de domínio e a harmonização radiométrica são abordadas por Stoica, Breaban e Barbu (2023), que mantêm a base da U-Net 3D, mas implementam um protocolo de transferência entre a base de dados KiPA22 e o KiTS23. A metodologia envolve correspondência de histograma para equalização radiométrica e um regime intensivo de aumento de dados, incluindo rotação, escalonamento e aplicação de ruído e desfoque gaussiano, demonstrando que o enriquecimento do sinal de entrada pode elevar o desempenho da arquitetura padrão mesmo diante de variações nos protocolos de aquisição.

Finalmente, a reconstrução tridimensional de malhas geométricas a partir de segmentações volumétricas é explorada por Demirtaş, İnner e Kavak (2025), que propõem três pipelines híbridos para a conversão de máscaras NIfTI em representações de superfície no formato OBJ. O primeiro modelo utiliza um fluxo de *Marching Cubes* baseado em limiarização de densidade e filtragem gaussiana para suavizar o efeito de escada dos voxels. O segundo modelo introduz interpolação spline para ganho de resolução e o filtro de difusão anisotrópica de Perona-Malik, que permite reduzir o ruído preservando a integridade das bordas. O terceiro modelo estende essa estrutura ao aplicar uma técnica de suavização de vértices baseada em k-vizinhos mais próximos (KNN) sensível a normais, o que otimiza a topologia da malha ao considerar tanto a proximidade espacial quanto o alinhamento das superfícies. A análise qualitativa e quantitativa demonstrou que esta última abordagem é superior na preservação de interfaces anatômicas críticas e na redução de erros de

contorno, tornando o método ideal para aplicações que exigem alta fidelidade visual, como planejamento cirúrgico e impressão 3D.

A Tabela 1 contém uma síntese das pesquisas relacionadas. Ao analisá-la, observa-se que a maioria dos métodos propostos utiliza variações da arquitetura U-Net (RONNEBERGER; FISCHER; BROX, 2015) como espinha dorsal, evidenciando a consolidação das redes de codificador-decodificador como paradigma dominante na segmentação de estruturas renais em tomografia computadorizada. Contudo, as convoluções padrão apresentam limitações inerentes, como campos receptivos restritos e dificuldade em modelar dependências de longo alcance, o que motiva a incorporação de mecanismos complementares. Nesse sentido, a adaptação dessas redes a módulos de atenção espacial e de canal (HU; PENG, 2023), blocos residuais (CHEN; ZHANG, 2023), pirâmides de *pooling* (MATOS et al., 2024) e arquiteturas híbridas com *Transformers* (QIAN et al., 2023) surge como alternativa promissora para ampliar a eficácia na tarefa de segmentação semântica em imagens médicas. Analisando o estado da arte, percebe-se que essas modificações frequentemente incorporam avanços oriundos de contextos não médicos, integrando-os às redes de segmentação de forma a aprimorar a identificação de estruturas anatômicas em exames e contribuir para o diagnóstico mais eficiente do carcinoma renal.

Tabela 1 – Resumo dos trabalhos relacionados baseados no dataset KiTS23.

Categoria	Trabalho	Técnica(s) Principal(is)	Coeficiente Dice		
			Rins e Massas	Massas	Tumor
Arquitetural	Qian et al. (2023)	nnU-Net + Swin Transformer	-	-	68,70%
	Myronenko et al. (2023)	SegResNet + SwinUNETR	95,60%	79,20%	75,80%
	Li, Peng e Zhang (2023)	nnU-Net + Atenção Residual	93,60%	-	67,00%
	Stoica, Breaban e Barbu (2023)	3D U-Net + Transferência de Domínio	94,70%	76,00%	71,30%
	Tanasković et al. (2024)	YOLOv8n-seg (2D)	-	-	79,00%
	Matos et al. (2024)	CPP-UNet (PPM + ASPP)	92,84%	**92,08%**	88,17%
	Karunanayake et al. (2025)	ViT + 3D UNet (Dual-Stage)	97,00%	-	88,00%
	Demirtaş, İnner e Kavak (2025)	Mesh Reconstruction (Model 3)	-	-	**98,00%**
	Jariwala et al. (2024)	DeepLabv3+	**98,22%**	-	-
	Alonso-Monsalve et al. (2025)	Sparse U-Net	95,80%	85,70%	80,30%
Multi-estágio	Hu e Peng (2023)	3D U-Net + GSCA-Net	93,30%	74,40%	67,90%
	Vispi (2023)	EfficientNet-B5 + U-Net	97,71%	81,39%	73,81%
	Chen e Zhang (2023)	ResUNet + Recorte de ROI	89,44%	85,85%	**85,91%**
	Alexandru e Popescu (2025)	YOLO + EfficientNet-B4/MiT-B2	95,82%	-	62,60%
	Wang et al. (2023)	nnU-Net (Duas Etapas)	86,60%	54,50%	49,00%
	Salahuddin et al. (2023)	nnU-Net + Taxa de Ciclo (Ensemble)	94,00%	**86,50%**	83,50%
	Uhm et al. (2023)	3D U-Net Paralela (Multiescala)	**97,90%**	82,60%	85,70%
	Sina Ziaee, Maleki e Ovens (2025)	Rel-UNet (Múltiplos Mínimos Locais)	80,00%	70,20%	64,10%

A decomposição do problema em múltiplos estágios sequenciais também se destaca como estratégia recorrente, seja por meio de cascatas baseadas na nnU-Net (WANG et al., 2023; SALAHUDDIN et al., 2023), seja pela combinação de detectores preliminares com redes de refinamento (ALEXANDRU; POPESCU, 2025; KARUNANAYAKE et al., 2025). Essa abordagem permite equilibrar a precisão diagnóstica e a viabilidade

computacional, ao especializar cada etapa em uma escala ou granularidade distintas do problema. Complementarmente, métodos de quantificação de incerteza (Sina Ziaee; Maleki; Ovens, 2025) e de filtragem por consistência multiescala (UHM et al., 2023) demonstram que a confiabilidade da segmentação automática pode ser elevada sem necessariamente aumentar a complexidade arquitetural, o que representa avanços relevantes para a validação clínica desses sistemas.

2.3 Considerações finais

Neste capítulo, foi apresentada uma revisão de estudos sobre a identificação de rins, cistos e tumores em imagens de tomografia computadorizada. Além de apresentar os métodos e os resultados alcançados por esses trabalhos, comparamos as diferentes estratégias para evidenciar como a tecnologia nessa área tem evoluído. No próximo capítulo, serão apresentados os conceitos teóricos fundamentais para o desenvolvimento desta pesquisa.

3 Fundamentação Teórica

Este capítulo apresenta os conceitos cuja compreensão é necessária para aprofundar o estudo. Nele, são explorados os conhecimentos essenciais sobre o objeto de análise (rins, cistos e tumores renais) e sobre as técnicas de segmentação de imagens de tomografia computadorizada que sustentam a abordagem metodológica adotada.

3.1 Rins e Doenças Renais

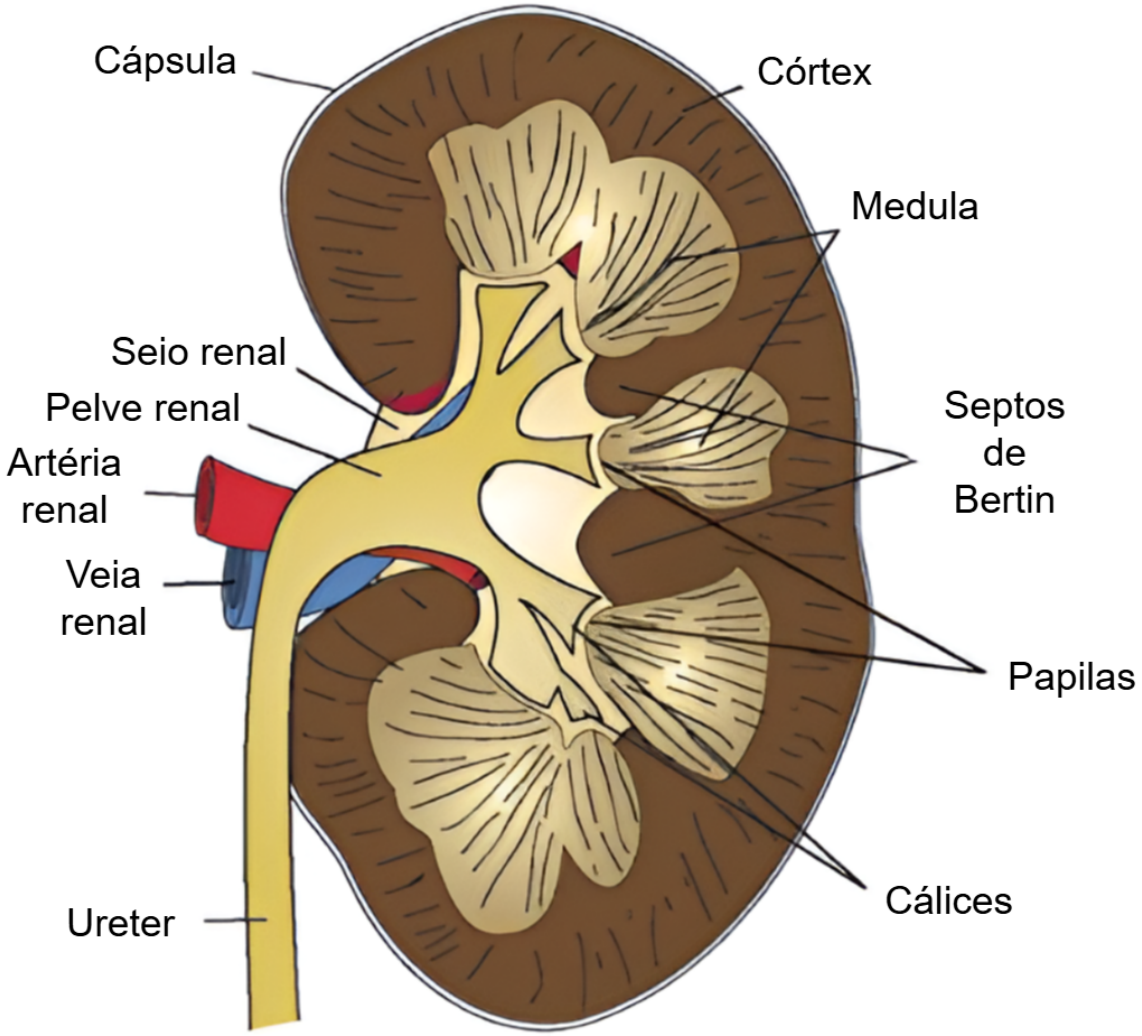
Conhecidos como órgãos em formato de feijão, os rins estão localizados na porção posterior do abdômen, estendendo-se da 12ª vértebra torácica até a 3ª vértebra lombar. Em adultos, cada rim pesa entre 125 e 170 g nos homens e entre 115 e 155 g nas mulheres, e o peso correlaciona-se melhor com a área de superfície corporal. As dimensões médias de cada rim são de 11 a 12 cm de comprimento, 5 a 7,5 cm de largura e 2,5 a 3 cm de espessura. Além disso, os rins apresentam movimentos ascendentes e descendentes de acordo com a respiração. Acima deles, encontram-se as glândulas suprarrenais. Devido à sua posição anatômica, o rim direito costuma ser menor e localizado um pouco abaixo do esquerdo (ZHOU et al., 2017; SBN - Sociedade Brasileira de Nefrologia, 2023).

O rim apresenta uma organização morfofuncional distinta, dividida em duas regiões principais: o córtex, localizado externamente, e a medula, situada na porção interna. Essa disposição estrutural pode ser observada na Figura 1, na qual não ocorre de maneira aleatória, mas está associada à distribuição específica dos segmentos do néfron, unidade funcional do rim. Tal organização fornece a base para processos especializados que vão desde a filtragem de substâncias presentes no plasma até o controle refinado da concentração final da urina.

O córtex renal, camada mais externa do órgão, distingue-se por sua aparência granular e coloração mais escura e representa cerca de 70% da massa total do órgão. Sua principal função está relacionada à filtragem do sangue e às etapas iniciais de processamento do filtrado. Essa região apresenta duas subdivisões estruturais: o labirinto cortical e os raios medulares. O labirinto cortical constitui a maior porção funcional do córtex, abrigando os glomérulos, responsáveis pela filtração, bem como os túbulos contorcidos proximais e distais, onde ocorrem processos de reabsorção e secreção. Já os raios medulares, intercalados entre o labirinto, são formados por feixes paralelos de túbulos retos e de ductos coletores que atravessam o córtex em direção à medula, estabelecendo uma integração funcional entre essas duas regiões.

Em contraste com a aparência granular do córtex, a medula renal apresenta-se como

Figura 1 – Estrutura interna do Rim.



Fonte: Adaptado de (ZHOU et al., 2017).

uma região mais pálida e de aspecto estriado, organizada em estruturas cônicas denominadas pirâmides renais. Sua principal função está associada à geração e à manutenção de um gradiente osmótico, fundamental para a concentração da urina e para a conservação de água no organismo. Estruturalmente, a medula pode ser subdividida em zonas, de acordo com os segmentos tubulares que a compõem. A medula externa, localizada junto ao córtex, divide-se em uma faixa externa, formada pela porção terminal dos túbulos retos proximais e pelos ramos ascendentes espessos da Alça de Henle, e uma faixa interna, constituída por ramos descendentes finos e ascendentes espessos. Já a medula interna corresponde à porção mais profunda, que se afunila em direção à papila renal. Nessa região, predominam ramos finos (ascendentes e descendentes) da Alça de Henle, bem como os ductos coletores terminais, caracterizada histologicamente pela ausência de ramos ascendentes espessos (ZHOU et al., 2017; NETTER; FRANK, 2000).

Os rins desempenham um papel central na manutenção da homeostase por meio de um conjunto de funções excretoras, reguladoras e endócrinas. Como principal sistema excretor, filtram o plasma para eliminar resíduos metabólicos (ureia, creatinina, ácido úrico) e compostos exógenos. Seu papel regulador é vital para o controle do volume hídrico, do equilíbrio de eletrólitos e do pH corporal. Adicionalmente, contribuem para a regulação da pressão arterial e atuam na produção de hormônios essenciais (ZHOU et al., 2017).

O sistema renal está suscetível a uma vasta gama de patologias que podem comprometer sua homeostase e função, abrangendo desde processos inflamatórios, como as nefrites, até distúrbios metabólicos que resultam na formação de cálculos renais. Dentre as condições de maior complexidade e impacto clínico, destacam-se as neoplasias, que decorrem da transformação e da proliferação celulares desreguladas no parênquima renal. Tais tumores são classificados fundamentalmente em duas categorias: benignos, caracterizados por um crescimento lento e comportamento localizado, e malignos, que exibem um fenótipo agressivo e possuem potencial para disseminação metastática (ATKINS; CHOUEIRI, 2022).

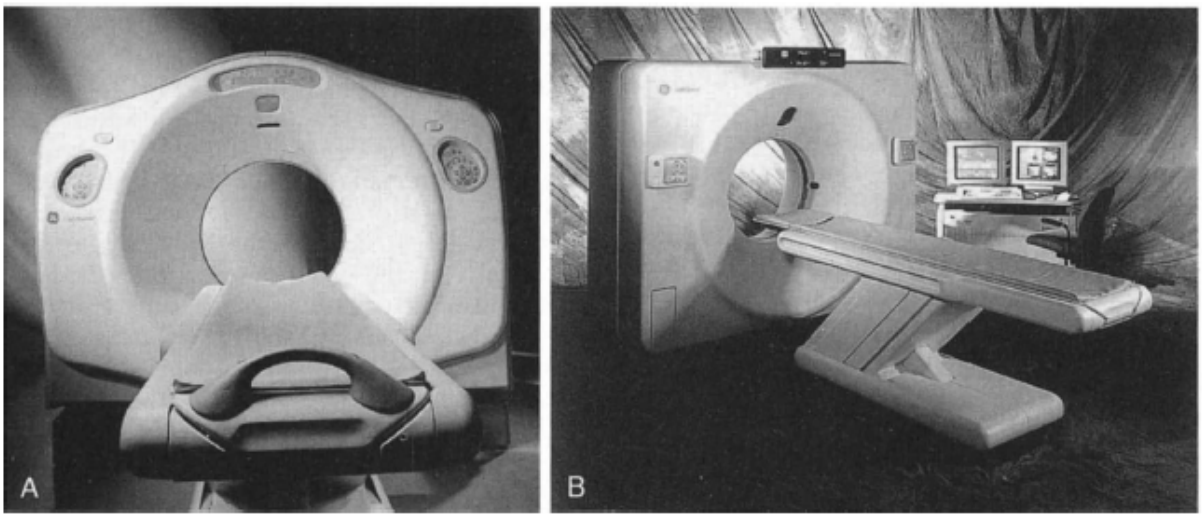
No espectro das neoplasias malignas, o Carcinoma de Células Renais (CCR) figura como a entidade histológica predominante, correspondendo a 80-85% de todos os tumores primários do rim. Originado no córtex renal, o CCR não constitui uma doença singular, mas sim um grupo heterogêneo de subtipos histologicamente distintos. Os subtipos mais prevalentes são o carcinoma de células claras (75-85%), o carcinoma papilífero (10-15%) e o carcinoma cromófobo (5-10%). A histogênese destes subtipos é específica: enquanto os carcinomas de células claras e papilífero derivam do epitélio do túbulo proximal, o subtipo cromófobo origina-se das células intercaladas do sistema ductal coletor. Adicionalmente, é crucial ressaltar que qualquer um desses subtipos pode sofrer um processo de dediferenciação sarcomatoide, uma transformação que confere ao tumor características de um sarcoma de alto grau e associa-se a um comportamento clínico marcadamente mais agressivo e a um prognóstico reservado (DAINA; CORTINOVIS; REMUZZI, 2023).

3.2 Tomografia Computadorizada

A tomografia computadorizada (TC) é um método de imagem que marcou o diagnóstico médico, permitindo a visualização detalhada de estruturas internas. Seu desenvolvimento é creditado ao físico sul-africano Allan Cormack, que publicou sua pesquisa pioneira em 1963 após trabalhos iniciados em 1956-57, e ao engenheiro inglês Godfrey Hounsfield, que desenvolveu um protótipo no final dos anos 1960. Por suas inovações, ambos receberam o Prêmio Nobel de Fisiologia ou Medicina em 1979. O primeiro ensaio clínico da tecnologia foi realizado com sucesso em 1971, em Londres, marcando sua transição para a prática médica. Contudo, a cronologia da invenção pode ser mais complexa, pois

evidências sugerem que pesquisadores de Kiev já haviam solucionado o problema teórico da TC no final da década de 1950, mas seus trabalhos só foram conhecidos no Ocidente no início dos anos 1980, quando os originais já haviam desaparecido (PETRIK et al., 2006; VAUGHAN; MAYOSI, 2007).

Figura 2 – Aparelho Tomográfico. A Gantry LightSpeed QX/i, composto pelo gantry (pórtico) do tomógrafo e pela mesa de suporte ao paciente. B Gantry LightSpeed Plus Multi-Slice, composto pelo gantry (pórtico) do tomógrafo, console do operador e mesa de suporte ao paciente.



Fonte:Parks (2001).

O funcionamento do tomógrafo (Figura 2) baseia-se na atenuação dos raios X à medida que atravessam diferentes tecidos do corpo humano. Essa atenuação segue a Lei de Beer-Lambert, segundo a qual a intensidade do feixe emergente diminui exponencialmente em função do coeficiente de atenuação linear (μ) do material, que depende da densidade e da composição atômica dos tecidos. Durante um exame, a máquina de TC projeta um feixe de fótons de raios X por múltiplos ângulos, completando uma rotação de 360°. As variações nos coeficientes de atenuação resultam em diferenças na quantidade de fótons detectados após atravessarem o corpo. Um processador de computador utiliza essas diferenças para gerar um conjunto de dados brutos que representam a estrutura tridimensional examinada. Para que essa informação possa ser visualizada e analisada em uma tela bidimensional, o computador processa o volume reconstruído e o exibe como uma série de imagens sequenciais, conhecidas como “fatias” (HERMENA; YOUNG, 2023).

Cada fatia é, essencialmente, uma imagem de uma seção transversal muito fina do corpo, como se estivéssemos observando o interior de uma única fatia de um pão. Tecnicamente, cada ponto em uma imagem de TC (um pixel) representa a densidade média de um pequeno volume de tecido do paciente (um voxel). A densidade de cada voxel é calculada e expressa em Unidades Hounsfield (HU), que determinam a intensidade em

tons de cinza desse pixel na imagem final. Ao apresentar essas fatias em sequência, os profissionais de saúde podem "percorrer" virtualmente a anatomia do paciente, camada por camada, permitindo uma avaliação detalhada e a identificação de anomalias (JIANG, 2009).

O hardware de um sistema de TC é composto por vários componentes essenciais. O gerador fornece a energia elétrica de alta e baixa voltagem necessária para gerar raios X. A unidade de varredura, conhecida como gantry, é a estrutura que contém o tubo de raios X, os detectores de fótons e os elementos de blindagem, que giram em uníssono ao redor do paciente. O tubo de raios X é responsável por converter a eletricidade em fótons de raios X, um processo em que apenas 1% da energia elétrica se transforma em fótons, enquanto 99% se dissipa como calor. Os detectores, posicionados em oposição ao tubo, absorvem e contam os fótons que atravessam o paciente, convertendo-os primeiro em luz visível e, depois, em sinais elétricos. A mesa do paciente desliza para dentro do gantry, permitindo a varredura da área de interesse (HERMENA; YOUNG, 2023).

Após a aquisição, os sinais elétricos gerados pelos detectores são amplificados e convertidos de formato analógico para digital por um conversor analógico-digital. Esse conjunto de dados digitais brutos é armazenado e processado por algoritmos matemáticos para gerar a imagem final. A densidade de cada tecido na imagem é representada por um valor na escala de Unidades Hounsfield (HU). Essa escala é uma transformação linear dos coeficientes de atenuação de raios X medidos, padronizada com base em duas substâncias de referência: a água destilada, definida como 0 HU, e o ar, definido como -1000 HU (HERMENA; YOUNG, 2023; JIANG, 2009).

3.3 Processamento de Imagens Digitais

A origem do processamento de imagens remonta a um período anterior ao uso de computadores. Um dos primeiros registros significativos ocorreu na década de 1920, na indústria jornalística, por meio do sistema de cabo submarino Bartlane, que permitia a transmissão de fotografias entre Londres e Nova York em menos de três horas. Apesar de envolver imagens transmitidas discretamente, essa prática não se enquadra no conceito moderno de processamento de imagens digitais, uma vez que não fazia uso de computadores. O desenvolvimento efetivo dessa área está associado ao surgimento dos primeiros computadores digitais e aos avanços do programa espacial. Um marco fundamental foi registrado em 1964, quando o Jet Propulsion Laboratory (JPL) aplicou técnicas computacionais para corrigir distorções nas imagens da Lua transmitidas pela sonda Ranger 7. A partir desse momento, o processamento de imagens expandiu-se rapidamente para diferentes áreas, incluindo a medicina, com a criação da tomografia axial computadorizada (CAT) no início da década de 1970, além de aplicações em geografia e arqueologia (GONZALEZ;

WOODS, 2010).

O processamento digital de imagens pode ser definido como o campo que utiliza computadores para processar imagens digitais. Uma imagem, por sua vez, é compreendida como uma função bidimensional, f(x,y), na qual x e y são coordenadas espaciais (plano), e a amplitude de f em qualquer par de coordenadas (x,y) é chamada de intensidade ou nível de cinza da imagem naquele ponto. Portanto, denomina-se imagem digital quando os valores de x, y e da intensidade são quantidades finitas e discretas (GONZALEZ; WOODS, 2010; FILHO; NETO, 1999).

Uma imagem digital é composta por um número finito de elementos, onde cada um possui localização e valor específicos, esses elementos são denominados de elementos pictóricos ou, mais comumente, pixels. O interesse nesta área provém de duas aplicações principais: primeiro, a melhoria das informações visuais para a interpretação humana; e segundo, o processamento de dados, que visa a percepção por máquinas, além de otimizar o armazenamento e a transmissão.

As técnicas aplicadas variam desde processos de baixo nível, mais primitivos, focados em melhorias, como a remoção de ruídos e o ajuste de contraste. O nível médio avança para a segmentação, descrição e classificação dos objetos presentes na imagem. Por fim, o nível alto interpreta os resultados do nível médio, realizando uma análise cognitiva da cena para extrair seu significado, conforme apontam (GONZALEZ; WOODS, 2010).

Um sistema clássico de processamento digital de imagens é fundamentado em cinco etapas sequenciais e interdependentes. O processo inicia-se com a aquisição, na qual imagens são capturadas e digitalizadas por sensores. Segue-se o pré-processamento, uma fase crucial em que algoritmos são aplicados para aprimorar a qualidade da imagem, visando à eliminação de ruídos e ao realce de características visuais. A terceira etapa, a segmentação, consiste em particionar a imagem, isolando objetos de interesse do fundo com base em atributos como cor, textura ou intensidade. Subsequentemente, na representação e na descrição, extraem-se características quantitativas dos objetos segmentados, transformando informações visuais em um conjunto de dados descritivos. Por fim, na etapa de reconhecimento de padrões, esses dados são utilizados para classificar e identificar os objetos, frequentemente empregando técnicas de aprendizado de máquina, como a aprendizagem profunda, para extrair padrões das imagens (GONZALEZ; WOODS, 2010). Um método específico pode abranger apenas algumas dessas etapas, conforme as necessidades do problema a ser resolvido.

3.4 Redes Neurais Artificiais

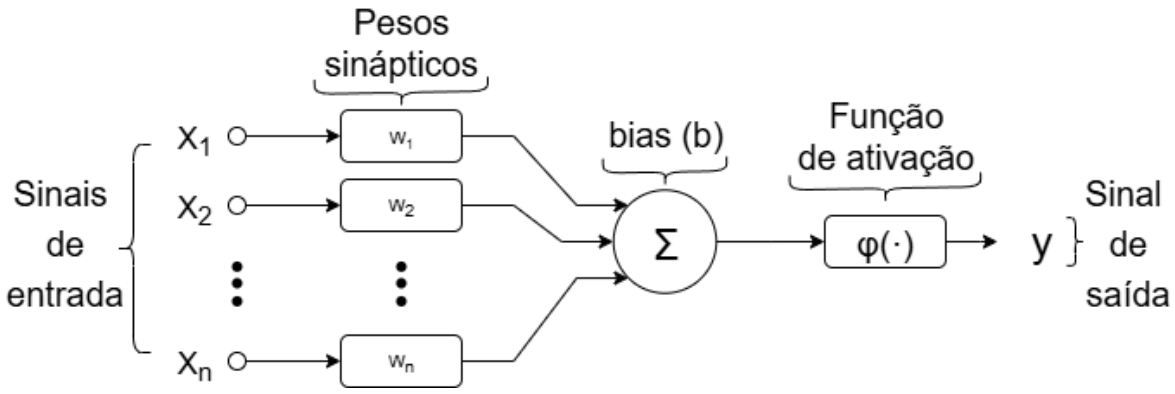
As redes neurais artificiais (RNAs) constituem sistemas computacionais voltados ao aprendizado de máquina, à representação de conhecimento e à otimização de respos-

tas em cenários complexos (CHEN et al., 2019). Inspiradas no sistema nervoso biológico, especialmente no funcionamento do cérebro humano, essas redes utilizam modelos matemáticos que emulam a atividade neuronal para processar informações de entrada e gerar respostas adaptativas (LIVINGSTONE, 2008; GERVEN; BOHTE, 2017). A inspiração no funcionamento dos neurônios biológicos, que recebem, processam e transmitem sinais por meio das sinapses, fundamenta o conceito de aprendizado nas RNAs, permitindo a extração de conhecimento a partir de dados (PACHECO; PEREIRA, 2018; HAYKIN, 2001a). Dessa forma, as RNAs reproduzem, de maneira simplificada, mecanismos de interpretação e adaptação, explicando sua ampla aplicabilidade em diferentes áreas do conhecimento e sua relevância no aprimoramento de sistemas complexos (HAYKIN, 2001b).

3.4.1 Neurônio Artificial

O neurônio artificial é considerado a unidade fundamental para o funcionamento das Redes Neurais Artificiais (RNA) (HAYKIN, 2001a). A proposta inicial desse conceito foi apresentada por McCulloch (1943), em um trabalho pioneiro que estabeleceu as bases teóricas para aplicações práticas das RNA. Nesse estudo, os autores introduziram uma abstração matemática do neurônio biológico, que representa, de forma simplificada, os processos de recepção e transmissão de sinais do cérebro humano. Essa formulação constituiu a base para o desenvolvimento de modelos computacionais mais avançados e contribuiu para a consolidação do campo de estudo das redes neurais.

Figura 3 – Neurônio artificial.



Fonte: Adaptado de (KOVÁCS, 2002).

Graficamente, o funcionamento de um neurônio artificial pode ser representado pela Equação 3.1, ilustrada na Figura 3, que expressa o processo de combinação linear dos sinais de entrada, seguido da aplicação de uma função de ativação. O papel fundamental da função de ativação é introduzir não linearidade no modelo, possibilitando que a rede aprenda representações complexas e não apenas relações lineares entre variáveis.

$$f(x) = \varphi\left(\sum_{i=1}^{n} x_i w_i + b\right), \tag{3.1}$$

onde nesta formulação, os termos $x_1, x_2, ..., x_n$ correspondem às variáveis de entrada do modelo, enquanto $w_1, w_2, ..., w_n$ denotam os pesos sinápticos associados a cada entrada i. O parâmetro b representa o viés (*bias*), responsável por deslocar a função de ativação, e φ designa a função de ativação aplicada ao somatório ponderado. Dessa maneira, cada neurônio agrupa suas entradas e, posteriormente, aplica uma transformação não linear. Entre as funções tradicionalmente utilizadas destacam-se a identidade, a sigmoide e a tangente hiperbólica (PACHECO; PEREIRA, 2018; TAYE, 2023). Mais recentemente, o avanço das redes neurais profundas levou ao emprego de funções de ativação adicionais, como a ReLU (*Rectified Linear Unit*) e suas variantes (LeakyReLU, ELU, GELU, entre outras), que se tornaram amplamente adotadas por sua eficiência no treinamento de modelos complexos e pela mitigação do problema do desaparecimento do gradiente (MAAS et al., 2013; ZHANG; LU; ZHAO, 2024; HENDRYCKS; GIMPEL, 2023).

3.4.2 *Perceptron* de Múltiplas Camadas (MLP)

Entre as diversas arquiteturas de Redes Neurais Artificiais (RNA), o *Perceptron* de Múltiplas Camadas (MLP) é uma das mais utilizadas e estudadas em diferentes áreas. Sua estrutura é composta por três partes fundamentais: a camada de entrada, responsável por receber os dados, uma ou mais camadas intermediárias, conhecidas como camadas ocultas, e a camada de saída, que fornece a resposta final. Os neurônios das camadas ocultas exercem papel essencial, pois permitem que a rede aprenda representações mais complexas a partir das informações recebidas (LI et al., 2012; PACHECO; PEREIRA, 2018; SILVA, 2004; BANSAL, 2006).

O treinamento de uma MLP geralmente é realizado pelo algoritmo de retropropagação do erro (backpropagation). Esse procedimento é dividido em duas etapas. Na primeira, chamada de propagação direta, os dados de entrada percorrem sucessivamente as camadas da rede até que uma resposta seja produzida na saída. Em seguida, essa saída é comparada ao valor esperado, e a diferença entre ambos gera o erro. Na segunda etapa, conhecida como retropropagação, esse erro é transmitido no sentido inverso, da saída para a entrada, ajustando-se os pesos sinápticos para reduzir progressivamente as discrepâncias (VORA; YAGNIK; SCHOLAR, 2014; DINIZ et al., 2021).

O treinamento inicia-se com a definição de pesos aleatórios pequenos, atualizados segundo a regra de aprendizado baseada no gradiente do erro. Essa atualização é realizada de forma iterativa, partindo da camada de saída e retornando até as camadas iniciais. A regra geral de ajuste pode ser expressa pela Equação 3.2,

$$w_{ij}(t+1) = w_{ij}(t) + \eta\delta_j x_i, \tag{3.2}$$

onde w_{ij} representa o peso associado à conexão entre os neurônios i e j, x_ié a entrada recebida, η é a taxa de aprendizagem e δ_j corresponde ao gradiente de erro do neurônio j.

Considerando a função de ativação sigmoide, quando o neurônio em questão pertence à camada de saída, o cálculo de δ_j é realizado conforme a Equação 3.3:

$$\delta_j = y_j(1 - y_j)(d_j - y_j), \tag{3.3}$$

onde d_j é a saída desejada e y_j a saída efetivamente obtida pelo neurônio j. O termo $y_j(1 - y_j)$ corresponde à derivada da função sigmoide. Já para neurônios localizados nas camadas ocultas, o valor de δ_j é obtido considerando a contribuição dos neurônios da camada seguinte, como mostrado na Equação 3.4:

$$\delta_j = x_j(1 - x_j)\sum_k \delta_k w_{jk}, \tag{3.4}$$

onde x_j representa a saída do neurônio j na camada oculta após a aplicação da função de ativação, k percorre todos os neurônios da camada posterior ao neurônio j, e $x_j(1 - x_j)$ corresponde à derivada da função sigmoide.

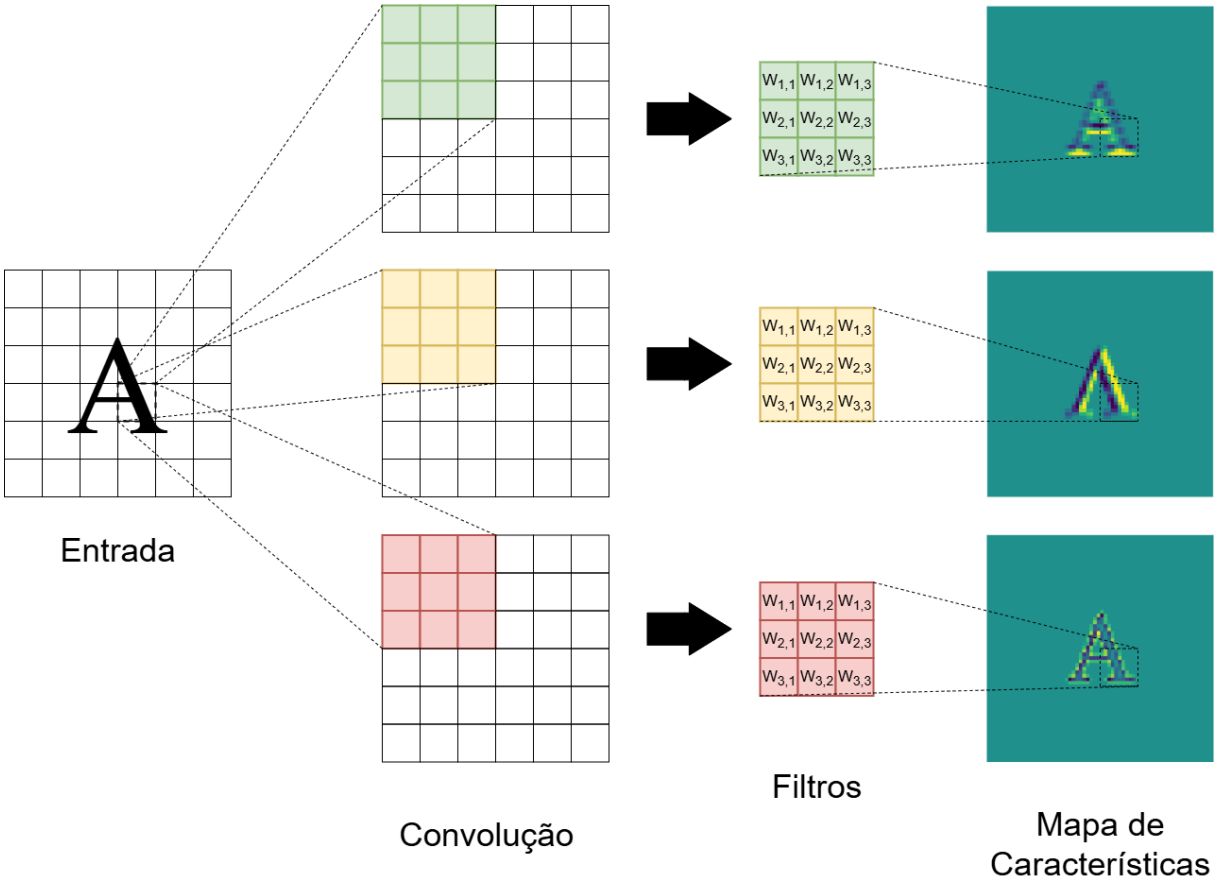
Esse ciclo de propagação e atualização dos pesos é repetido para todos os exemplos do conjunto de treinamento até que um critério de parada seja atendido. Normalmente, utiliza-se como critério o número máximo de épocas de treinamento ou a convergência do erro para um valor satisfatório. Embora a função sigmoide seja utilizada neste trabalho para fins de formulação do algoritmo de retropropagação, é importante ressaltar que outras funções de ativação, como ReLU (*Rectified Linear Unit*) e tangente hiperbólica, podem ser empregadas em MLPs modernas.

3.5 Convolução e Redes Neurais Convolucionais

A convolução é uma operação matemática fundamental no processamento de sinais e imagens digitais. Seu princípio básico consiste na aplicação de um filtro, também chamado de núcleo, sobre uma matriz de entrada. Esse filtro percorre sistematicamente cada posição da matriz e, em cada deslocamento, realiza combinações lineares entre os valores dos pixels da região local e os coeficientes do filtro. O resultado é um novo mapa de saída que pode realçar padrões de interesse, como bordas, texturas ou formas específicas (GONZALEZ; WOODS, 2010), como mostrado na Figura 4.

Do ponto de vista computacional, a convolução permite que informações locais sejam processadas de forma eficiente, pois cada elemento do filtro aprende a reconhecer um

Figura 4 – Representação da operação de convolução.



Fonte: Adaptado de (LI et al., 2021).

aspecto específico da entrada. Essa característica é especialmente relevante em imagens, nas quais a proximidade espacial entre *pixels* contém informações significativas. Assim, filtros menores que a imagem original conseguem capturar variações locais, que podem ser combinadas posteriormente em níveis mais complexos de representação (LI et al., 2021).

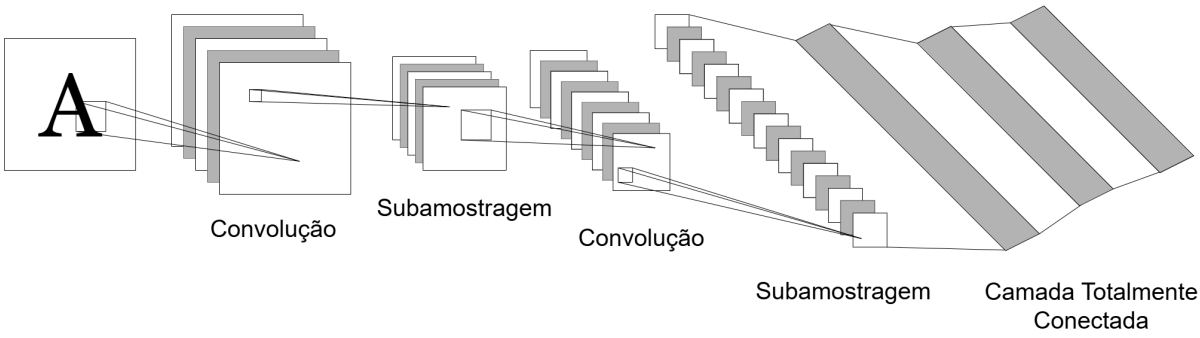
A operação de convolução pode ser descrita pela Equação 3.5. Nela, um filtro w é aplicado à imagem de entrada f para gerar a imagem resultante g. Os índices x e y percorrem toda a matriz de entrada, enquanto os índices s e t percorrem os elementos do filtro. Os valores de a e b correspondem, respectivamente, a $(m-1)/2$ e $(n-1)/2$, sendo m e n as dimensões do filtro.

$$g(x,y) = \sum_{s=-a}^{a} \sum_{t=-b}^{b} w(s,t) f(x-s, y-t) \tag{3.5}$$

A partir desse conceito, surgiram as redes neurais convolucionais (CNNs, do inglês *Convolutional Neural Networks*), que constituem uma extensão das redes neurais artificiais, mas adaptadas para explorar a estrutura espacial dos dados. Diferentemente das redes totalmente conectadas, em que todos os neurônios de uma camada se ligam a todos os

da camada seguinte, as CNNs limitam as conexões a regiões locais da entrada, imitando parcialmente o funcionamento do córtex visual biológico (LECUN et al., 1998; LI et al., 2021).

Figura 5 – Arquitetura padrão de uma Rede Neural Convolucional.



Fonte: Adaptado de (LECUN et al., 1998; LECUN; KAVUKCUOGLU; FARABET, 2010).

A arquitetura padrão de uma CNN é composta por três tipos principais de camadas: convolucionais, de subamostragem e totalmente conectadas (KANG; WANG, 2014) (Figura 5). Nas camadas convolucionais, múltiplos filtros treináveis são aplicados à entrada, gerando mapas de características que capturam padrões de naturezas distintas. Em seguida, as camadas de subamostragem reduzem a resolução espacial dos mapas de características, preservando apenas as informações mais relevantes e reduzindo o custo computacional (LECUN et al., 1998). Para isso, a entrada é dividida em regiões, das quais se extrai um único valor representativo, como a média ou o valor máximo de cada região. Esse procedimento torna a rede menos sensível a pequenas variações morfológicas locais, além de diminuir a complexidade computacional e reduzir a quantidade de neurônios necessários nas camadas subsequentes (LI et al., 2021; ZHAO et al., 2024).

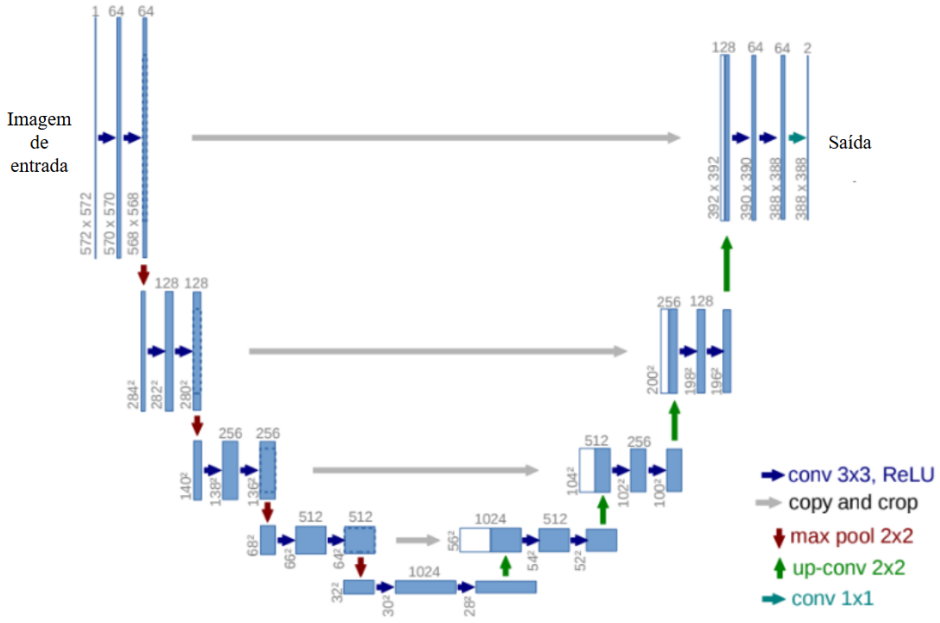
Após sucessivas etapas de convolução e subamostragem, os mapas de características são transformados em vetores unidimensionais e enviados às camadas totalmente conectadas. Nessa fase, a rede atua de forma semelhante a um perceptron multicamadas, integrando as informações extraídas nas etapas anteriores para realizar a classificação ou outra tarefa de aprendizado (FERNANDES, 2013).

3.6 U-Net

A U-Net é uma arquitetura de rede neural convolucional proposta por Ronneberger, Fischer e Brox (2015) para a segmentação semântica de imagens biomédicas, que se tornou notável por seu alto desempenho mesmo com poucos dados de treinamento. Sendo baseada em codificadores e decodificadores, em que a parte de codificação extrai características e reduz a resolução da imagem de entrada, enquanto a parte do decodificação realiza a

reconstrução detalhada da imagem segmentada. Essa estrutura permite que a rede aprenda características em diferentes escalas e capture informações contextuais em vários níveis de abstração.

Figura 6 – Arquitetura da U-Net.



Fonte: Ronneberger, Fischer e Brox (2015).

A codificação consiste na aplicação repetida de blocos operacionais, cada um composto por duas convoluções com filtros (kernel) 3x3 (sem preenchimento), seguidas por uma função de ativação ReLU, e uma operação de max pooling de 2x2 com passo (stride) 2 para realizar o downsampling (redução de amostragem). A cada etapa de downsampling, a resolução espacial do mapa de características é reduzida pela metade, enquanto o número de canais de características é dobrado, o que permite que a rede aprenda representações cada vez mais complexas e contextuais.

A fase de decodificação tem como objetivo aumentar a resolução do mapa de características, reconstruir a máscara de segmentação e permitir uma localização precisa. Em cada etapa, uma convolução transposta de 2x2 dobra a resolução e divide o número de canais pela metade. A principal inovação da U-Net é a concatenação do mapa expandido com o mapa de características de alta resolução correspondente do caminho de contração, conhecido como conexões e salto (skip connections). Esse mapa precisa ser cortado para que as dimensões sejam compatíveis antes da união. Essa conexão crucial combina as informações contextuais com os detalhes de localização, permitindo que duas convoluções de 3x3 com ReLU refinem as informações e gerem uma predição mais precisa. A arquitetura padrão da rede U-Net é ilustrada na Figura6(GONZALEZ; WOODS, 2010).

3.6.1 Redes Convolucionais Deformáveis

As Redes Convolucionais Deformáveis foram propostas por Dai et al. (2017) como uma extensão das CNNs tradicionais, para melhorar a capacidade de modelagem de transformações geométricas. Em arquiteturas convolucionais convencionais, as posições de amostragem do kernel são fixas e definidas por uma grade regular, o que limita a adaptação do campo receptivo a variações como mudanças de escala, rotação, espelhamento e deformações não rígidas.

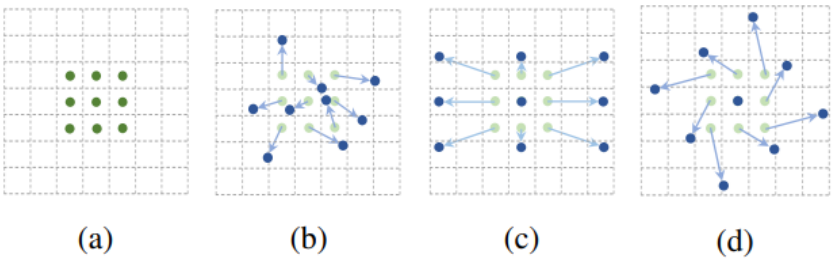
Na convolução tradicional, considerando um mapa de características de entrada x e um mapa de saída y, a resposta em uma posição espacial p_0 é definida por:

$$y(p_0) = \sum_{p_n \in R} w(p_n)\, x(p_0 + p_n) \tag{3.6}$$

onde R representa a grade regular do kernel (por exemplo, os nove deslocamentos de um filtro 3×3) e $w(p_n)$ são os pesos aprendidos associados a cada posição da grade.

Entretanto, como ilustrado na Figura 7, a convolução tradicional realiza a amostragem sempre nas mesmas posições relativas (pontos verdes), independentemente do conteúdo da imagem. Essa limitação pode dificultar a modelagem de objetos com formas e tamanhos variados.

Figura 7 – Ilustração das posições de amostragem em convoluções 3×3. (a) Grade regular de amostragem utilizada na convolução tradicional (pontos verdes). (b)–(d) Exemplos de grades deformadas na convolução deformável, nas quais os pontos azuis representam as novas posições de amostragem e as setas indicam os deslocamentos aprendidos.



Fonte: Dai et al. (2017).

A convolução deformável modifica essa operação ao introduzir deslocamentos aprendíveis Δp_n para cada posição da grade R. Dessa forma, a Equação 3.6 passa a ser definida como:

$$y(p_0) = \sum_{p_n \in R} w(p_n)\, x(p_0 + p_n + \Delta p_n) \tag{3.7}$$

onde o termo $p_n + \Delta p_n$ define posições de amostragem irregulares e adaptativas. Como pode ser observado na Figura 7, os pontos azuis representam as novas posições deformadas, enquanto as setas indicam os deslocamentos aprendidos.

Como os deslocamentos Δp_n geralmente assumem valores reais, a amostragem em posições fracionárias é realizada por interpolação bilinear. Assim, o valor de $x(p)$ em uma posição arbitrária p é calculado como:

$$x(p) = \sum_q G(q, p)\, x(q) \tag{3.8}$$

onde q percorre as posições inteiras do mapa de características e $G(\cdot, \cdot)$ representa o kernel de interpolação bilinear. Esse kernel bidimensional pode ser decomposto em dois kernels unidimensionais:

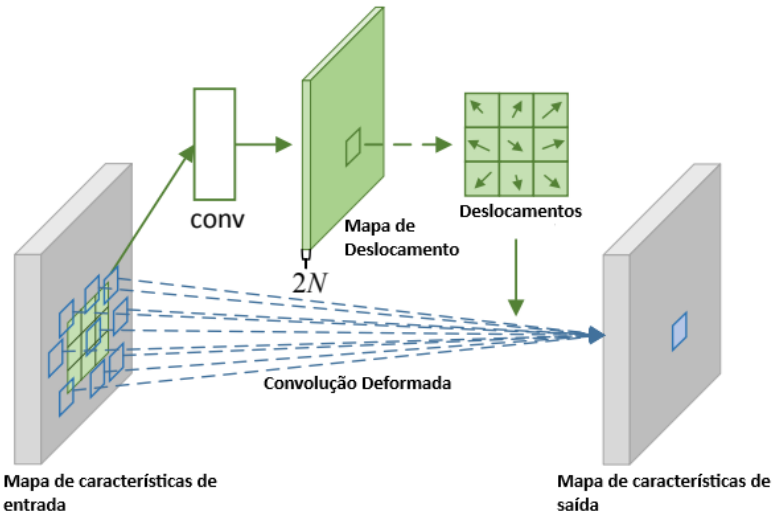
$$G(q, p) = g(q_x, p_x)\, g(q_y, p_y) \tag{3.9}$$

com

$$g(a, b) = \max(0, 1 - |a - b|). \tag{3.10}$$

A implementação apresentada na Equação 3.8 é computacionalmente eficiente, pois o kernel $G(q, p)$ é diferente de zero apenas para um pequeno conjunto de posições vizinhas.

Figura 8 – Exemplo de convolução deformável 3×3



Adaptado de Dai et al. (2017).

A Figura 8 ilustra o mecanismo de aprendizado dos deslocamentos. Como mostrado na figura, os offsets são gerados por uma camada convolucional adicional, aplicada ao

mesmo mapa de características de entrada. Essa camada produz $2N$ canais, correspondentes a N deslocamentos bidimensionais (um para cada ponto da grade R).

Durante o treinamento, tanto os pesos da convolução principal quanto os parâmetros responsáveis pela geração dos deslocamentos são aprendidos simultaneamente por retropropagação. Isso permite que o campo receptivo se adapte dinamicamente ao conteúdo da imagem, ajustando-se automaticamente ao formato e à escala dos objetos.

Resultados experimentais obtidos por Dai et al. (2017) demonstram que essa capacidade de adaptação espacial melhora significativamente o desempenho em tarefas que exigem localização precisa, como detecção de objetos e segmentação semântica. Dessa forma, a convolução deformável pode substituir diretamente camadas convolucionais tradicionais em arquiteturas já consolidadas, mantendo a compatibilidade estrutural e adicionando capacidade de modelagem geométrica adaptativa.

3.7 Blocos Estruturais em Redes Neurais Convolucionais

As Redes Neurais Convolucionais (CNNs) organizam-se por meio de blocos estruturais que processam os dados de forma hierárquica, transformando progressivamente representações de baixo nível em características de alto nível mais abstratas (LECUN; BENGIO; HINTON, 2015). Essa organização modular, presente tanto na U-Net quanto em arquiteturas mais recentes, fundamenta-se em componentes básicos que se repetem ao longo da rede, permitindo padronização arquitetural e maior profundidade estrutural.

O bloco convolucional básico, já apresentado na arquitetura U-Net por meio das convoluções 3×3 seguidas por ReLU, constitui a unidade fundamental de processamento. Nesse bloco, filtros aprendidos são aplicados localmente à entrada, explorando a estrutura espacial dos dados. Uma propriedade central desse mecanismo é o compartilhamento de pesos (*weight sharing*), no qual todos os neurônios de um mesmo mapa de características utilizam o mesmo conjunto de filtros. Essa estratégia reduz significativamente o número de parâmetros do modelo e favorece a captura de padrões equivariantes a deslocamentos espaciais, sendo a invariância parcial posteriormente reforçada por operações de pooling (LECUN; BENGIO; HINTON, 2015).

Embora a U-Net apresentada utilize blocos simples (convolução seguida de função de ativação), arquiteturas modernas frequentemente incorporam normalização em lote (*Batch Normalization*) entre a convolução e a ativação (IOFFE; SZEGEDY, 2015). Essa técnica normaliza as ativações para apresentar média zero e variância unitária em cada mini-lote, contribuindo para maior estabilidade do treinamento e permitindo o uso de taxas de aprendizado mais elevadas. Formalmente, a operação de normalização pode ser expressa como:

$$\mathrm{BN}(x) = \gamma \frac{x - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} + \beta \tag{3.11}$$

onde $\mu_{\mathcal{B}}$ e $\sigma_{\mathcal{B}}^2$ representam, respectivamente, a média e a variância do mini-lote, ϵ é uma constante para estabilidade numérica, e γ e β são parâmetros treináveis que permitem à rede recuperar representações adequadas após a normalização.

A organização hierárquica desses blocos, como observado no caminho de contração da U-Net, possibilita a extração de características em múltiplas escalas. Camadas iniciais tendem a detectar padrões locais simples, como bordas e texturas, enquanto camadas mais profundas identificam estruturas complexas e semanticamente mais ricas. As operações de *pooling*, como o max pooling 2×2 utilizado na U-Net, reduzem progressivamente a resolução espacial dos mapas de características, ampliando o campo receptivo efetivo da rede e contribuindo para maior robustez a pequenas variações locais (KRIZHEVSKY; SUTSKEVER; HINTON, 2012).

Para tarefas de predição densa, a organização em arquitetura encoder-decoder exemplifica uma estratégia eficiente de estruturação desses blocos. No encoder, sucessivos blocos convolucionais e operações de redução espacial promovem a extração de representações cada vez mais abstratas e contextuais. No decoder, operações de upsampling restauram progressivamente a resolução espacial, permitindo a reconstrução detalhada da saída. As *skip connections* da U-Net resolvem a tensão entre semântica global e localização precisa ao combinar características provenientes de múltiplas resoluções, possibilitando que informações de granularidade fina orientem a reconstrução espacial detalhada (LONG; SHELHAMER; DARRELL, 2015; RONNEBERGER; FISCHER; BROX, 2015).

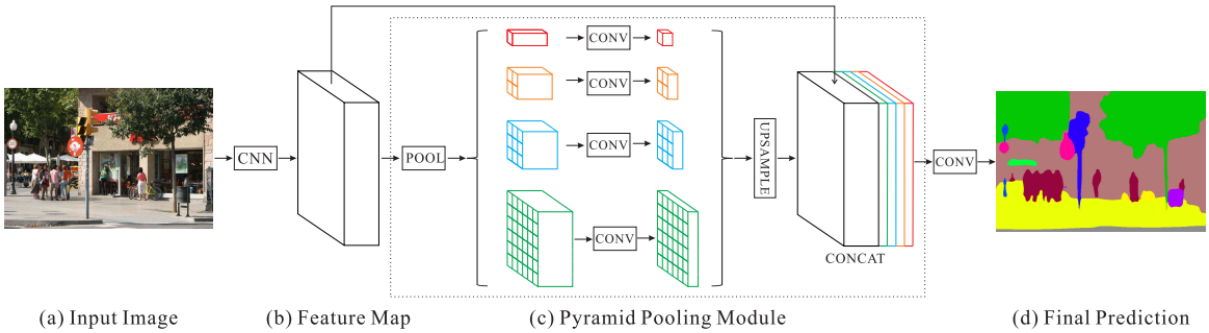
3.8 Módulo de Agrupamento em Pirâmide

O Módulo de Agrupamento em Pirâmide (PPM) é um componente central da Rede de Análise de Cena em Pirâmide (PSPNet), projetado para agregar informações de contexto de diferentes regiões em um modelo de previsão em nível de *pixel* (ZHAO et al., 2017).

A arquitetura base foi inicialmente proposta como componente central da rede PSPNet (ZHAO et al., 2017), com o objetivo de resolver a falta de informações de contexto global em redes de análise de cenas. Nesta arquitetura, utiliza-se uma rede ResNet (HE et al., 2016) pré-treinada para extrair um mapa de características a partir da imagem de entrada. Sobre este mapa, o Módulo de Agrupamento em Pirâmide é aplicado para coletar informações de contexto em múltiplas escalas. A estrutura do módulo funde características de quatro níveis distintos da pirâmide. O nível mais abrangente, destacado em vermelho na arquitetura, realiza um agrupamento global para gerar uma saída única. Os níveis

seguintes dividem o mapa de características em várias sub-regiões, criando representações agrupadas para diferentes áreas da imagem. Como resultado, as saídas de cada nível da pirâmide possuem mapas de características com tamanhos distintos, conforme ilustrado na Figura 9.

Figura 9 – Representação do Módulo de Agrupamento em Pirâmide.



Fonte: Zhao et al. (2017).

Destaca-se que, após cada nível da pirâmide, aplica-se uma convolução 1×1 para reduzir a dimensionalidade, seguida por uma interpolação bilinear que expande essas características reduzidas, resultando em um mapa de características com dimensões equivalentes às do original. Por fim, as características de todos os níveis são concatenadas, formando o recurso global final do agrupamento em pirâmide.

O módulo de agrupamento em pirâmide, conforme descrito por Zhao et al. (2017), permite ajustar o número de níveis e seus respectivos tamanhos de acordo com o mapa de características de entrada. Essa estrutura utiliza filtros convolucionais (*kernels*) com dimensões variadas para capturar informações de diferentes sub-regiões, sendo crucial manter uma diferença adequada nas representações geradas em múltiplos estágios. O modelo proposto contempla quatro níveis, com tamanhos 1×1, 2×2, 3×3 e 6×6, podendo empregar tanto *max pooling* quanto *average pooling*.

3.9 Módulos de Atenção

Mecanismos de atenção permitem que redes neurais foquem seletivamente nas partes mais relevantes de sua entrada, priorizando informações de forma adaptativa durante o processamento (BAHDANAU; CHO; BENGIO, 2016). Em arquiteturas encoder-decoder para tradução neural, a atenção resolve o problema da representação fixa: em vez de comprimir toda a entrada em um único vetor de contexto, o modelo acessa dinamicamente diferentes partes da representação conforme necessário.

Dado um conjunto de vetores de entrada $\{h_1, \ldots, h_n\}$ e um estado de *query* s, o vetor de contexto c é computado como:

$$c = \sum_{i=1}^{n} \alpha_i h_i, \quad \text{onde} \quad \alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)} \tag{3.12}$$

Os scores e_i medem a compatibilidade entre o *query* s e cada elemento de entrada h_i. Os pesos α_i são normalizados via *softmax*, formando uma distribuição de probabilidade sobre as posições de entrada.

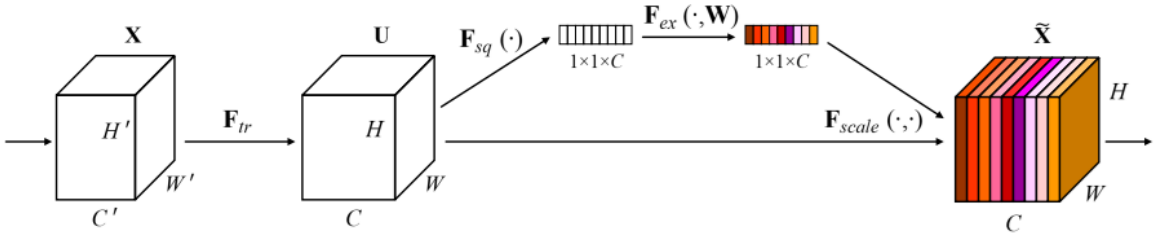
Em tradução neural (BAHDANAU; CHO; BENGIO, 2016; LUONG; PHAM; MANNING, 2015), a atenção opera sobre dimensões temporais, aprendendo alinhamentos entre palavras de sequências de entrada e saída. Em visão computacional (XU et al., 2016), a atenção opera sobre dimensões espaciais de mapas de *features*, focando em regiões específicas da imagem relevantes para a tarefa.

Em arquiteturas convolucionais, não há sequência temporal explícita ou estados recorrentes. A atenção deve operar sobre características espaciais e de canais extraídas pelas convoluções. Diferentemente da atenção espacial, que enfatiza regiões da imagem, a **atenção por canal** modela dependências semânticas globais entre mapas de características, recalibrando a importância de diferentes canais com base no conteúdo da entrada. Esta abordagem dá origem a arquiteturas como *Squeeze-and-Excitation Networks* (HU; SHEN; SUN, 2018), que exploram a interdependência entre canais convolucionais para melhorar a capacidade representacional da rede.

3.9.1 *Squeeze-and-Excitation*

O bloco Squeeze-and-Excitation (SE) tem como objetivo modelar a interdependência entre os canais de características em redes neurais convolucionais (CNNs). Esse mecanismo aprimora o desempenho das redes ao recalibrar dinamicamente a relevância de cada canal em um mapa de características. Conforme descrito por Hu, Shen e Sun (2018), o SE introduz um processo de recalibração adaptativa dos canais em três etapas principais: Squeeze (compressão), Excitation (excitação) e, por fim, combinação e escala. A estrutura desse bloco pode ser visualizada na Figura 10.

Figura 10 – Estrutura do bloco *Squeeze-and-Excitation*



Fonte: Hu, Shen e Sun (2018)

Na imagem, $X \in R^{H' \times W' \times C'}$ representa um tensor de características de entrada com altura H', largura W' e C' canais. A transformação F_{tr} representa uma operação convolucional que mapeia X para um novo tensor de características $U \in R^{H \times W \times C}$, com dimensões espaciais $H \times W$ e C canais.

A operação de *Squeeze* ou compressão, visa agregar as informações espaciais de cada canal em um único descritor. Esta função, denotada por F_{sq} na Figura 10, captura efetivamente o contexto global de cada mapa de características. Para alcançar isso, o bloco SE utiliza o global average pooling para comprimir o tensor U ao longo de suas dimensões espaciais ($H \times W$). O resultado é um vetor de descritores de canal $z \in R^{1 \times 1 \times C}$, em que cada elemento de z corresponde à média global do canal correspondente em U. Este vetor z pode ser interpretado como uma coleção de estatísticas que descrevem o conteúdo de cada canal em toda a imagem, permitindo que a rede obtenha uma percepção global do campo receptivo.

A seguir, a etapa de *Excitation* ou excitação, representada pela função $F_{ex}(\cdot, W)$ na Figura 10, visa modelar explicitamente as interdependências entre os canais. O vetor z, proveniente da etapa de *Squeeze*, é então processado por uma arquitetura análoga a um *bottleneck*. Esta consiste em uma primeira camada totalmente conectada (FC) com uma função de ativação $ReLU$, que reduz a dimensionalidade do vetor, introduzindo não-linearidade e diminuindo a complexidade computacional. Subsequentemente, uma segunda camada totalmente conectada, com uma função de ativação sigmoide, restaura a dimensionalidade do vetor para o número original de canais C. A saída desta etapa é um vetor de ativações, também de tamanho C, em que cada valor escalar, no intervalo [0,1], representa o peso ou a importância relativa de cada canal. A função sigmoide garante que a rede aprenda a enfatizar múltiplos canais simultaneamente, ao contrário de uma função como a *softmax*, que imporia uma competição exclusiva entre eles.

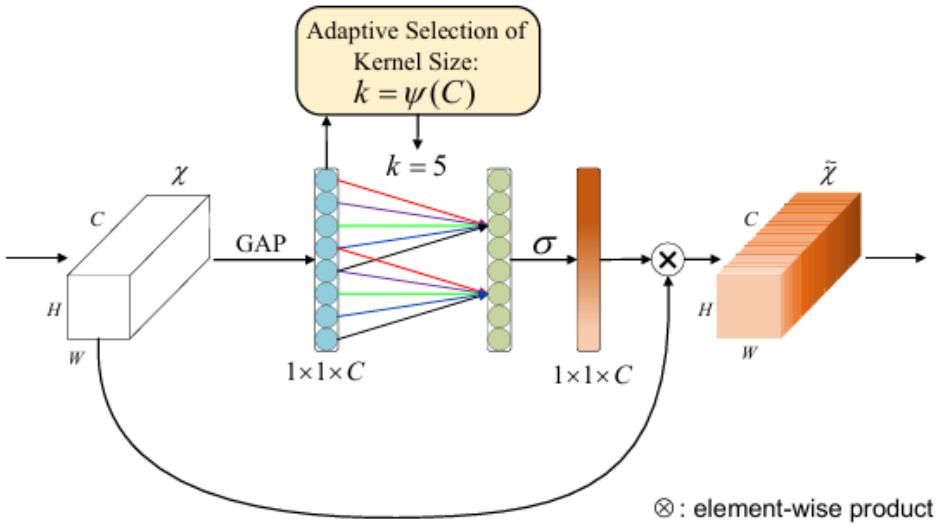
Finalmente, na etapa de Combinação e Escala, representada por $F_{escala}(\cdot, \cdot)$ na Figura 10, os pesos calculados na fase de *Excitation* são utilizados para recalibrar o tensor de características original U. Esta operação é realizada por meio de uma multiplicação canal a canal (*channel-wise multiplication*) entre as ativações da excitação e os mapas de características de U. O resultado é um novo tensor $\widetilde{X} \in R^{H \times W \times C}$, no qual os canais que contêm características mais informativas para a tarefa em questão são acentuados, enquanto aqueles com informações menos relevantes são suprimidos. Desta forma, o bloco SE permite que a rede neural melhore a sua sensibilidade às características mais pertinentes, aprimorando significativamente o poder de representação do modelo.

3.9.2 *Efficient Channel Attention*

O módulo Efficient Channel Attention (ECA) representa uma evolução do mecanismo de atenção de canal proposto no bloco Squeeze-and-Excitation, buscando manter a

eficácia enquanto reduz significativamente a complexidade computacional. Conforme apresentado por Wang et al. (2020), o ECA aborda as limitações do SE ao evitar a redução de dimensionalidade e capturar interações locais entre canais de forma mais eficiente. A estrutura desse módulo pode ser visualizada na Figura 11.

Figura 11 – Estrutura do módulo *Efficient Channel Attention*



Fonte: Wang et al. (2020)

Similarmente ao bloco SE, o módulo ECA recebe como entrada um tensor de características $X \in R^{H \times W \times C}$, onde H, W e C representam altura, largura e número de canais, respectivamente. A primeira etapa consiste na agregação de características por meio de *Global Average Pooling* (GAP), produzindo um vetor de descritores $y \in R^{1 \times 1 \times C}$, onde cada elemento representa a média global de um canal específico.

A principal inovação do ECA reside na geração dos pesos de atenção. Diferentemente do SE, que utiliza duas camadas totalmente conectadas com redução de dimensionalidade, o ECA emprega uma convolução 1D de tamanho de kernel k aplicada diretamente ao vetor y, sem qualquer redução de dimensionalidade. Esta abordagem preserva a relação explícita entre canais e seus pesos de atenção, além de modelar interações locais considerando apenas cada canal e seus k vizinhos mais próximos. A operação pode ser expressa como:

$$\omega_i = \sigma\left(\sum_{j=1}^{k} w^j \cdot y_i^j\right), \quad y_i^j \in \Omega_i^k \tag{3.13}$$

onde Ω_i^k representa o conjunto de k canais adjacentes ao canal i, y_i^j denota os elementos do vetor y dentro da vizinhança local, w^j são os pesos compartilhados da convolução 1D, e σ denota a função sigmoide.

Para evitar ajuste manual do parâmetro k, Wang et al. (2020) propõem sua seleção adaptativa baseada na dimensão do canal C. O mapeamento não-linear estabelece que

canais de alta dimensionalidade tenham interações de longo alcance, enquanto canais de baixa dimensionalidade tenham interações de curto alcance:

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{3.14}$$

onde $|\cdot|_{odd}$ indica o número ímpar mais próximo, com $\gamma = 2$ e $b = 1$.

Finalmente, os pesos de atenção $\omega \in R^{1\times1\times C}$ são aplicados ao tensor original X por meio de multiplicação, elemento a elemento, produzindo o tensor recalibrado $\widetilde{X} \in R^{H\times W\times C}$.

3.10 Avaliação dos Modelos

Para avaliar a eficácia do método de segmentação proposto, foram utilizadas métricas consolidadas na área de segmentação de imagens médicas. A escolha dessas métricas visa quantificar tanto a qualidade da classificação ao nível de *pixel* quanto a exatidão na delimitação dos objetos. A análise é conduzida por meio da comparação entre as máscaras preditas pela rede e as máscaras de referência (*ground truth*), anotadas por especialistas (MÜLLER; SOTO-REY; KRAMER, 2022). As métricas focadas na classificação de *pixels* são calculadas com base na matriz de confusão apresentada na Tabela 2.

Tabela 2 – Matriz de Confusão

		Classe Predita	
		Positivo	**Negativo**
Classe Verdadeira	**Positivo**	**Verdadeiro Positivo** (VP)	**Falso Negativo** (FN)
	Negativo	**Falso Positivo** (FP)	**Verdadeiro Negativo** (VN)

Fonte: Adaptado de (TING; VIJAYAKUMAR; SCHAAL, 2010).

Os componentes principais da matriz de confusão são obtidos pela comparação entre a predição do modelo e a *ground truth*. Nela, os Verdadeiros Positivos (VP) correspondem aos acertos diretos, quando o modelo identifica corretamente a classe de interesse (por exemplo, *pixels* de uma patologia devidamente segmentados). Já os Falsos Positivos (FP) ocorrem quando instâncias negativas são classificadas incorretamente como positivas, como no caso de regiões sadias marcadas como doentes. Os Verdadeiros Negativos (VN) representam as instâncias corretamente identificadas como não pertencentes à classe de

interesse, embora, em segmentação, sua contribuição prática seja menos relevante devido à grande predominância do fundo. Por fim, os Falsos Negativos (FN) representam falhas de detecção, isto é, quando o modelo não consegue identificar instâncias que de fato pertencem à classe positiva, como *pixels* de uma lesão que não foram segmentados (MÜLLER; SOTO-REY; KRAMER, 2022).

Nesta pesquisa, são empregadas métricas comumente usadas em abordagens de segmentação semântica, como o coeficiente de similaridade *Dice* e o índice de *Jaccard* (EELBODE et al., 2020; MÜLLER; SOTO-REY; KRAMER, 2022; XU, 2017), bem como medidas baseadas em distância, como a Distância de Hausdorff (HD).Adicionalmente, são utilizadas métricas volumétricas como o Erro de Sobreposição Volumétrica (VOE) e a Distância Média de Superfície (MSD) (TAHA; HANBURY, 2015).

O coeficiente de *Dice* é uma métrica usada para quantificar a concordância espacial entre a segmentação predita por um modelo e a máscara de referência (*ground truth*), medindo o grau de sobreposição entre as duas. Ele é matematicamente equivalente ao *F1-score* no caso binário, sendo calculado pela razão entre o dobro da interseção das regiões e a soma das áreas de ambas. Matematicamente, o coeficiente pode ser expresso pela Equação 3.15.

$$DSC = \frac{2VP}{2VP + FP + FN}. \tag{3.15}$$

O índice de *Jaccard*, também chamado de coeficiente de *Jaccard* ou *Intersection over Union* (IoU), é uma métrica utilizada para quantificar a similaridade entre dois conjuntos. Seu valor é obtido a partir da razão entre o número de elementos na interseção dos conjuntos e o número de elementos na união deles, conforme representado na Equação 3.16.

$$JCC = \frac{VP}{VP + FP + FN}. \tag{3.16}$$

O Erro de Sobreposição Volumétrica (VOE, do inglês *Volumetric Overlap Error*) é uma métrica complementar ao índice de *Jaccard*, definida como o complemento deste, ou seja, VOE = 1 - JCC. Esta métrica quantifica o erro de sobreposição volumétrica entre a segmentação predita e a segmentação de referência, expressando a fração da união dos volumes que não pertence à interseção entre as segmentações. Valores menores de VOE indicam maior concordância entre as segmentações, com VOE = 0 representando perfeita sobreposição (TAHA; HANBURY, 2015). A formulação do VOE é apresentada na Equação 3.17.

$$VOE = 1 - \frac{VP}{VP + FP + FN} = \frac{FP + FN}{VP + FP + FN}. \tag{3.17}$$

Por outro lado, a Distância de Hausdorff (HD), apresentada na Equação 3.18, quantifica o erro máximo de sobreposição entre dois contornos, sendo amplamente utilizada para avaliar o desalinhamento entre uma segmentação predita e a segmentação de referência. Formalmente, ela é definida como o máximo entre duas medidas: o maior das distâncias

mínimas de cada ponto do conjunto A em relação ao conjunto B, e o maior das distâncias mínimas de cada ponto de B em relação a A (HUTTENLOCHER; KLANDERMAN; RUCKLIDGE, 2002; TAHA; HANBURY, 2015).

$$HD(A,B) = \max\left\{\max_{a\in A}\min_{b\in B}\|a-b\|_2,\ \max_{b\in B}\min_{a\in A}\|a-b\|_2\right\} \tag{3.18}$$

Essa formulação captura o pior caso de divergência espacial entre os contornos. No entanto, devido à sua sensibilidade a outliers, é comum o uso da versão robusta, denominada HD95. Esta variante considera o 95º percentil das distâncias mínimas entre os pontos dos contornos, desconsiderando os 5% de maiores discrepâncias e proporcionando uma avaliação mais estável da qualidade da segmentação em aplicações médicas (TAHA; HANBURY, 2015).

A Distância Média de Superfície (MSD, do inglês Mean Surface Distance) é uma métrica baseada em distância que complementa a análise fornecida pela Distância de Hausdorff ao avaliar o erro médio de delimitação entre as superfícies das segmentações, em vez do erro máximo (TAHA; HANBURY, 2015). Diferentemente da HD, a MSD é menos sensível a *outliers*, pois considera a média das distâncias mínimas entre os pontos das superfícies correspondentes.

Formalmente, a MSD é definida como a média bidirecional das distâncias mínimas entre cada ponto da superfície da segmentação predita e a superfície da segmentação de referência, conforme apresentado na Equação 3.19.

$$MSD(A,B) = \frac{1}{2}\left(\frac{1}{|A|}\sum_{a\in A}\min_{b\in B}\|a-b\|_2 + \frac{1}{|B|}\sum_{b\in B}\min_{a\in A}\|a-b\|_2\right) \tag{3.19}$$

Valores menores de MSD indicam maior concordância espacial entre as superfícies segmentadas, refletindo uma melhor precisão na delimitação das bordas das estruturas de interesse em segmentações médicas (TAHA; HANBURY, 2015).

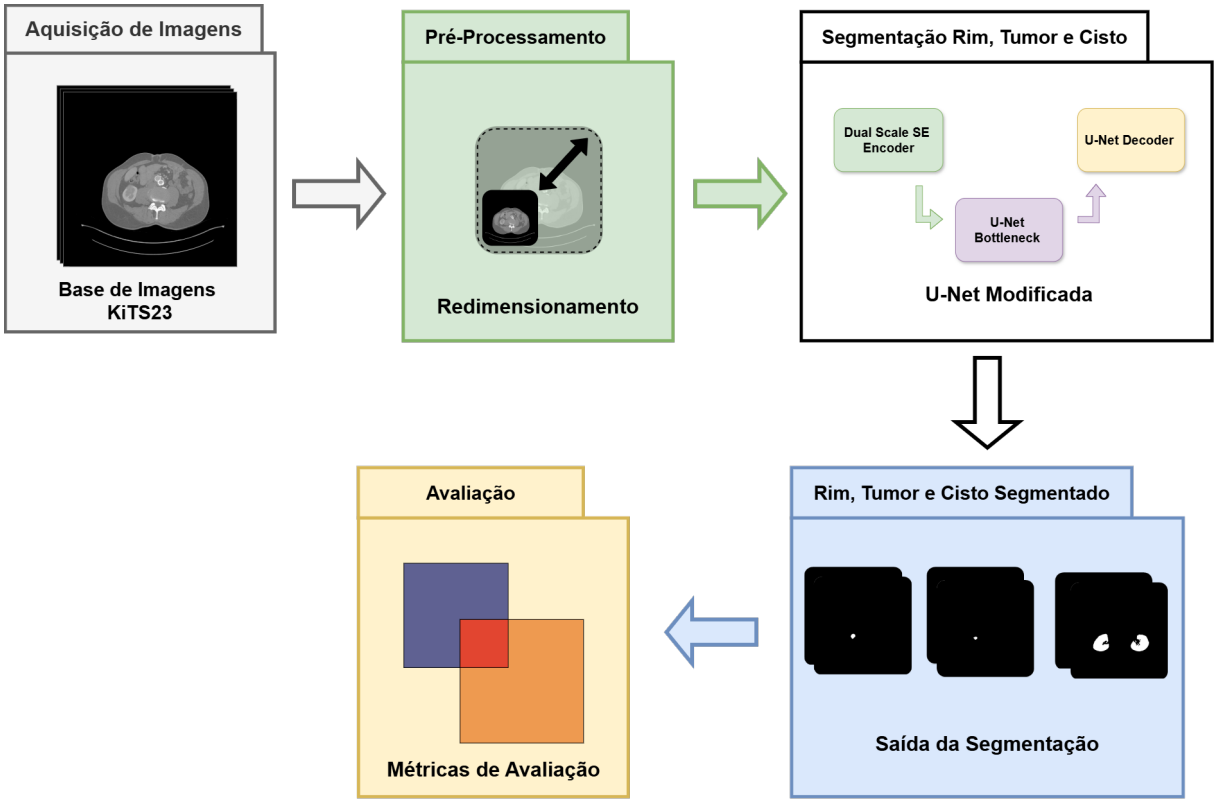
3.11 Considerações Finais

Este capítulo apresentou os fundamentos teóricos que sustentam o desenvolvimento desta pesquisa, abrangendo desde aspectos anatômicos e patológicos dos rins até os conceitos de aprendizado profundo relevantes para a tarefa de segmentação. Foram discutidas as bases das redes neurais convolucionais, a arquitetura U-Net, o Módulo de Agrupamento em Pirâmide (PPM), os mecanismos de atenção por canal e as métricas de avaliação adotadas. O conjunto desses conceitos constitui o arcabouço teórico necessário para a compreensão da metodologia proposta, apresentada no Capítulo 4.

4 Metodologia

Este capítulo aborda a metodologia empregada no desenvolvimento de uma abordagem para a segmentação de rins, cistos e tumores renais em imagens de tomografia computadorizada (TC). Os quatro passos fundamentais da metodologia são apresentados na Figura 12.

Figura 12 – Etapas da metodologia proposta.



Fonte: Acervo do autor

O processo metodológico inicia com a aquisição da versão mais recente da base de dados KiTS23 (HELLER et al., 2023) (Seção 5.2). Em seguida, todos os volumes de TC passaram por uma etapa de pré-processamento, que envolveu o redimensionamento das imagens. Na fase seguinte, é aplicada a arquitetura proposta neste trabalho para segmentação de câncer renal, denominada Dual-Scale SE, baseada na integração de blocos de Pirâmide de *Pooling* (PPM) ao codificador da U-Net. Por fim, os resultados obtidos são avaliados para validar a metodologia.

4.1 Pré-Processamento

Nesta metodologia, o pré-processamento tem como principal objetivo reduzir as dimensões das imagens. Essa etapa busca proporcionar uma execução mais rápida e eficiente do modelo, sobretudo devido às limitações computacionais, que impedem sua execução com os numero de pixels originais dos dados.

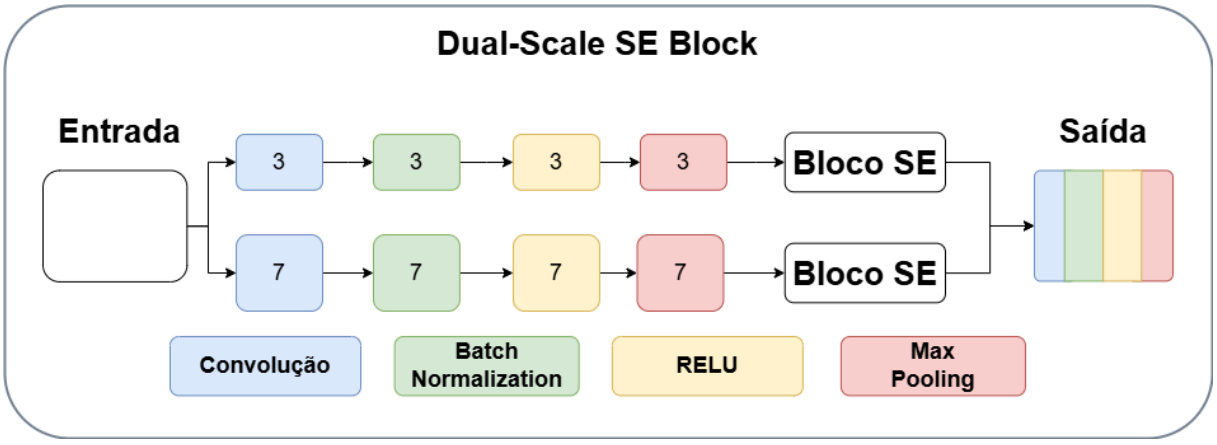
Para contornar as limitações de recursos computacionais durante a etapa de treinamento, adotou-se a estratégia de redimensionar todas as fatias dos exames. Dessa forma, as imagens tiveram suas dimensões reduzidas de 512×512 *pixels* para 256×256 *pixels* utilizando o método de interpolação bilinear, garantindo a viabilidade do treinamento dentro da capacidade computacional disponível.

4.2 Modelo de Segmentação - Dual-Scale SE U-Net

Nesta seção, descreve-se o desenvolvimento de uma arquitetura de rede convolucional para a segmentação de rins, cistos e tumores renais em imagens de TC, baseada em módulos de pirâmide de *pooling*. O modelo proposto, denominado Dual-Scale SE, integra o módulo PPM (ZHAO et al., 2017) e o bloco SE (HU; SHEN; SUN, 2018) à arquitetura U-Net (RONNEBERGER; FISCHER; BROX, 2015). A estrutura e o funcionamento desse modelo são detalhados nas próximas subseções.

Esta pesquisa adotou a estrutura do Bloco *Dual-Scale* SE, conforme detalhado na Figura 13. O processamento inicia-se quando o mapa de características de entrada é direcionado a dois ramos paralelos de processamento. Cada ramo aplica um filtro convolucional de tamanho 3×3 e 7×7, respectivamente.

Figura 13 – Bloco Dual-Scale SE.



Fonte: Acervo do autor

Em cada ramo, o processamento segue uma sequência específica de operações: convolução com função de ativação ReLU, seguida de normalização em lote (*Batch Nor-*

malization). Esta estratégia de processamento paralelo permite a extração simultânea de características complementares em diferentes campos receptivos. O kernel de menor dimensão especializa-se na captura de padrões locais e detalhes finos, enquanto o kernel de maior dimensão é responsável pela extração de informações contextuais de maior abrangência espacial. Tal abordagem multi-escala foi inspirada conceitualmente no *Pyramid Pooling Module* (PPM), que demonstrou a eficácia da agregação de características em diferentes resoluções, preservando assim maior riqueza de detalhes durante o processo de extração.

Imediatamente após a extração de características de cada ramo, e antes da concatenação, cada mapa de características passa por um bloco *Squeeze-and-Excitation* (SE) independente. Esses blocos SE implementam um mecanismo de atenção por canal que recalibra adaptativamente a importância relativa de cada canal. O processo consiste em: primeiro, uma operação de *Global Average Pooling* que comprime a informação espacial; segundo, duas camadas densas com funções de ativação ReLU e Sigmoid que modelam as interdependências entre canais, produzindo pesos de recalibração que são multiplicados, elemento a elemento, pelo mapa de características original.

A aplicação dos blocos SE ocorre de forma independente em cada ramo antes da concatenação, garantindo que a recalibração de canais seja específica para cada escala. Isso é particularmente importante no contexto da segmentação renal, em que estruturas como cistos pequenos podem ser melhor detectadas pelos padrões locais do ramo 3×3, enquanto a delimitação completa do rim requer um contexto espacial amplo, capturado pelo ramo 7×7. Sem essa recalibração independente, a concatenação direta poderia diluir características importantes de escalas específicas, especialmente quando há grande disparidade no número de canais informativos entre os ramos.

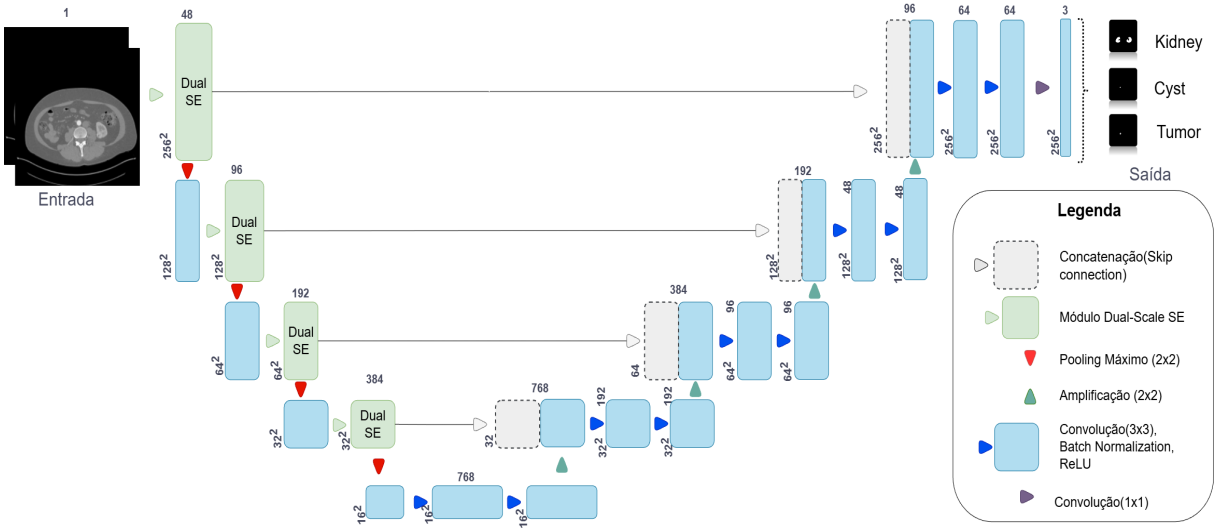
Após a aplicação dos blocos SE, os mapas de características recalibrados de ambos os ramos são concatenados ao longo da dimensão dos canais. Essa concatenação integra as informações multiescala já refinadas pelos mecanismos de atenção. Subsequentemente, aplica-se uma convolução adicional ao tensor concatenado, seguida de *Batch Normalization*, da ativação ReLU e, finalmente, de uma operação de *max pooling* com filtro 2×2 para redução dimensional, como mostra na Figura **??**.

É fundamental notar que a configuração final da arquitetura não foi arbitrária. Foram testadas configurações com dois níveis (3×3 e 7×7) e quatro níveis (3×3,5×5, 7×7 e 11×11). A configuração com dois níveis foi empiricamente determinada como a mais eficaz para a segmentação de rins, cistos e tumores renais, sendo esta a implementada no modelo final.

Seguindo a estrutura fundamental da U-Net tradicional, o modelo de segmentação proposto, denominado Dual-Scale SE UNet, incorpora blocos de atenção com extração multiescala à sua arquitetura. O modelo mantém a configuração característica em duas fases principais: o caminho de contração (encoder), localizado no lado esquerdo, e o caminho

de expansão (decoder), no lado direito. A principal inovação consiste na substituição dos blocos convolucionais convencionais do codificador por blocos Dual-Scale SE, desenvolvidos especificamente para lidar com os desafios da segmentação renal, em que tumores e cistos apresentam tamanhos variados e frequentemente não estão completamente separados do órgão, como ilustrado na Figura 14.

Figura 14 – Modelo arquitetural da Dual-Scale SE UNet.



Fonte: Acervo do autor

A arquitetura Dual-Scale SE UNet representa uma evolução da estrutura padrão da rede U-Net, na qual os blocos de convolução tradicionais do codificador são substituídos por combinações de convoluções multiescala com mecanismos de atenção. Cada bloco Dual-Scale SE processa o mapa de características de entrada por meio de duas convoluções paralelas com kernels de diferentes dimensões (3×3 e 7×7), permitindo a captura simultânea de informações locais e de contextos mais amplos (uma característica essencial para identificar estruturas em regiões de alta similaridade estrutural).

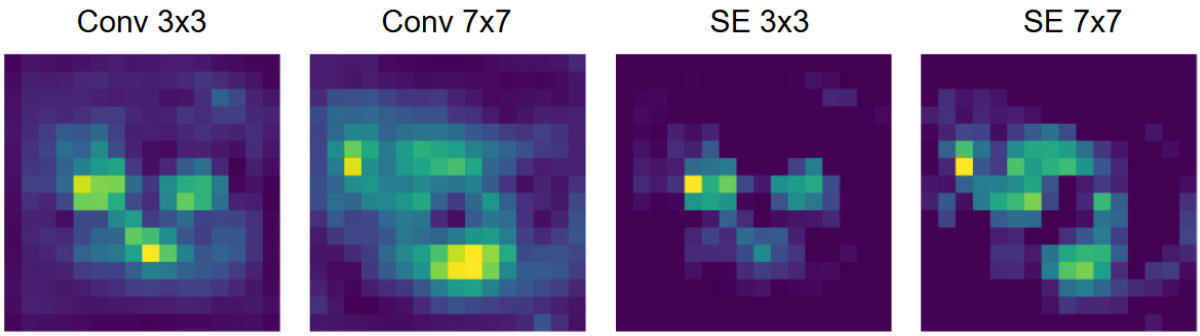
O bloco é composto por dois ramos paralelos de processamento: o primeiro aplica uma convolução 3×3 para extrair características locais detalhadas, enquanto o segundo utiliza uma convolução 7×7 para capturar padrões espaciais mais abrangentes. Após cada operação convolucional, são aplicados mecanismos de *Batch Normalization*, seguidos pela função de ativação ReLU e operação de *Max Pooling*.

A característica peculiar desta arquitetura é a combinação do processamento multinível (dual-scale) com a integração do módulo Squeeze-and-Excitation (SE) ao final de cada ramo convolucional. Esta abordagem híbrida permite que a rede capture informações em diferentes escalas espaciais (através das convoluções 3×3 e 7×7) e, simultaneamente, recalibre a importância de cada canal por meio do mecanismo SE. O módulo SE opera em duas etapas fundamentais: a compressão (squeeze), que aplica Global

Average Pooling para obter uma representação condensada de cada canal, e a recalibração (excitation), que utiliza duas camadas totalmente conectadas, com funções ReLU e Sigmoid, para calcular pesos adaptativos. Esta combinação de extração multiescala com atenção por canal representa o diferencial da arquitetura proposta em relação às abordagens convencionais.

Como demonstrado na Figura 15, o impacto deste mecanismo é visível na comparação entre as ativações convencionais (Conv 3×3 e Conv 7×7) e as ativações com SE (SE 3×3 e SE 7×7), na qual se observa uma redução significativa do ruído e um foco mais seletivo nas regiões relevantes da imagem. Esses pesos recalibram dinamicamente a importância relativa dos canais, permitindo que a rede minimize a dispersão para áreas menos significativas e fortaleça a capacidade de identificar estruturas complexas, mesmo em casos de perda parcial ou total do órgão renal.

Figura 15 – A comparação entre sem atenção (duas primeiras imagens) e com atenção (duas últimas imagens).



Fonte: Acervo do autor

Durante o caminho de contração, os blocos Dual-Scale SE são aplicados sucessivamente em cada nível da rede. Após cada bloco, uma operação de agrupamento máximo com filtro 2×2) e *stride* 2 reduz as dimensões espaciais do mapa de características pela metade, enquanto o número de canais é progressivamente dobrado, seguindo o padrão estabelecido pela U-Net original.

Na fase de decodificação, o modelo realiza operações de upsampling para restaurar progressivamente as dimensões espaciais originais. Cada etapa de expansão é complementada por conexões de salto (*skip connections*) que concatenam os mapas de características correspondentes do caminho de contração, preservando informações espaciais detalhadas. As características concatenadas são então processadas por blocos convolucionais padrão, cada um seguido por *Batch Normalization* e ativação ReLU, garantindo a reconstrução precisa dos mapas de segmentação multiclasse com três canais de saída correspondentes às regiões de rim, cisto e tumor.

4.2.1 Funções de Perda

As funções de perda constituem elementos essenciais na avaliação do desempenho de modelos, mensurando a diferença entre os valores preditos e os rótulos reais do conjunto de dados. Neste trabalho, a tarefa de segmentar múltiplas estruturas, incluindo rins, tumores e cistos, impõe desafios significativos, notadamente o desequilíbrio entre classes e a complexidade em distinguir os diferentes rótulos.

Visando superar tais limitações, empregou-se a função de perda *Dice Categorical Focal Loss*, formada pela integração das funções *Categorical Focal Loss* (LIN et al., 2020) e *Dice Loss* (MILLETARI; NAVAB; AHMADI, 2016), ambas reconhecidas por sua eficácia em aplicações de segmentação (MU et al., 2024; WAKAMATSU; ONO, 2023).

A *Dice Loss*, derivada do coeficiente de Dice, prioriza a otimização da sobreposição espacial entre a predição e o *ground truth*, demonstrando robustez natural ao desequilíbrio de classes ao focar na correta segmentação das estruturas de interesse, independentemente da proporção volumétrica que estas ocupam na imagem. Por sua vez, a *Categorical Focal Loss* estende o conceito da *Focal Loss* para cenários multiclasse, introduzindo um fator de modulação $(1 - p_t)^\gamma$, em que p_t representa a probabilidade predita da classe correta e $\gamma \geq 0$ é o parâmetro de foco. Este fator reduz dinamicamente a contribuição dos exemplos bem classificados (alto p_t), direcionando o processo de aprendizagem para regiões de maior complexidade, como fronteiras entre estruturas e anomalias de pequenas dimensões.

A combinação dessas funções resulta na *Dice Categorical Focal Loss*, expressa como:

$$\mathcal{L}_{DCFL} = \mathcal{L}_{Dice} + \mathcal{L}_{CategoricalFocal}$$

Esta abordagem híbrida permite que o modelo se beneficie simultaneamente da capacidade da *Dice Loss* em otimizar métricas de sobreposição volumétrica e da habilidade da *Categorical Focal Loss* em priorizar o aprendizado de padrões complexos, resultando em segmentações mais precisas e robustas, especialmente em contextos caracterizados por estruturas anatomicamente heterogêneas e volumetricamente desbalanceadas, como observado no conjunto de dados KiTS23.

4.3 Considerações Finais

Este capítulo apresentou, em detalhes, a metodologia desenvolvida para a segmentação de rins, cistos e tumores em imagens de TC. Cada uma das etapas foram descritas, destacando as adaptações realizadas especificamente para a segmentação destas estruturas renais nesta pesquisa. Entre as principais contribuições, destacam-se: o desenvolvimento da arquitetura Dual-Scale SE U-Net, incorporando blocos com convoluções multiescala 3×3 e 7×7 e mecanismo *Squeeze-and-Excitation* ao codificador da U-Net, desenvolvendo

assim uma nova arquitetura profunda para segmentação de rins, cistos e tumores renais em imagens de tomografia computadorizada.

No próximo capítulo, são apresentados os resultados obtidos com a metodologia. Além disso, são apresentados detalhes sobre a configuração experimental dos modelos de segmentação empregados, experimentos realizados na base KiTS23, e uma análise de estudos de casos para validar as etapas da metodologia.

5 Resultados

Este capítulo apresenta os resultados experimentais obtidos com a metodologia proposta para segmentação de rins, tumores e cistos renais na base KiTS23. Inicialmente, são descritos as configurações experimentais e os critérios de avaliação adotados. Em seguida, são apresentados os resultados quantitativos e qualitativos, incluindo estudos de caso e análise por Grad-CAM. Por fim, realiza-se uma comparação crítica com trabalhos relacionados, posicionando o modelo proposto no estado da arte

5.1 Configurações Experimentais

Infraestrututura: A metodologia desta tese foi implementada em Python, com a biblioteca de aprendizado profundo Keras (GULLI; PAL, 2017). Os experimentos foram realizados em um computador equipado com uma CPU Intel Core i5-11400 de 4.20 GHz, 16 GB de RAM e uma placa de vídeo Nvidia GeForce GTX 3060-Ti, 12 GB de VRAM, executando o sistema operacional Ubuntu 20.

Hiperparâmetros: Os hiperparâmetros utilizados nos modelos de segmentação foram: *batch* igual a 4, otimizador estocástico Adam com taxa de aprendizagem inicial de 10^{-4}. Aplica-se um total de 50 épocas utilizando a técnica *Early Stop* (BROWNLEE, 2018). Este mecanismo permite interromper o treinamento da rede quando a métrica monitorada (valor da função de perda nos dados de validação) não apresenta incremento por 3 épocas consecutivas, valor estabelecido empiricamente.

Métricas de Avaliação: As métricas utilizadas para avaliar o desempenho do método proposto e dos modelos reimplementados foram o coeficiente de similaridade de Dice (DSC), a Distância de Hausdorff a 95% (HD95), o Erro de Sobreposição Volumétrica (VOE) e a Distância Média de Superfície (MSD), conforme descrito na Seção de Métricas (Seção 3.10).

Cenário de Avaliação:

O cenário experimental foi concebido com foco na validação da arquitetura Dual-Scale SE UNet, proposta principal deste trabalho. A análise busca verificar de que forma a integração de extração multiescala com mecanismos de atenção por canal influencia o desempenho da segmentação multiclasse de rins, tumores e cistos renais na base KiTS23. A avaliação considera tanto métricas quantitativas quanto análise qualitativa por meio de estudos de caso e mapas de ativação Grad-CAM, permitindo examinar não apenas os resultados estatísticos, mas também a coerência anatômica e o comportamento interno do modelo. Dessa forma, o cenário de avaliação busca oferecer uma análise abrangente da

robustez e da consistência da arquitetura proposta.

Em um segundo momento, são analisadas variações arquiteturais derivadas da proposta principal, incluindo as versões Dual-Scale SE Siamese, Dual-Scale SE com convoluções deformáveis, Dual-Scale SE com Attention Gates, Dual-Scale SE em configuração 2.5D, bem como a variante Dual-Scale ECA. Essas configurações são investigadas com o propósito de avaliar o impacto isolado de modificações complementares sobre o desempenho global. Em um terceiro momento, é realizada uma comparação com outros métodos publicados na literatura.

5.2 Bases de Imagens

Neste estudo, utilizou-se a edição mais recente da base de imagens disponibilizada pelo *Kidney Tumor Segmentation Challenge* (KiTS) (HELLER et al., 2023), lançada em 2023. O desafio tem como propósito incentivar o desenvolvimento de novas abordagens voltadas à segmentação de tumores renais, favorecendo, assim, o aprimoramento dos processos de diagnóstico e tratamento do câncer renal. O conjunto de dados associado inclui informações de pacientes submetidos a diferentes procedimentos terapêuticos, tais como crioablação, nefrectomia parcial ou nefrectomia radical, em situações de suspeita de malignidade. Os dados foram coletados entre 2010 e 2022 no centro médico *M Health Fairview* (M Health Fairview, 2024), resultante da colaboração entre a Universidade de Minnesota (EUA) e a rede de saúde *Fairview Health Services* (Fairview Health Services, 2024).

O KiTS23 (HELLER et al., 2023) introduziu, como inovação, a inclusão de casos obtidos na fase de contraste nefrogênico, caracterizada pelo realce otimizado do parênquima renal após a administração do contraste, enquanto, nas versões anteriores, como o KiTS21, os exames contemplavam apenas a fase arterial tardia, em que o realce se concentra nas estruturas diretamente irrigadas pelo sistema arterial (KULKARNI et al., 2021). Para esta edição, foram disponibilizados 599 exames de tomografia computadorizada (TC), sendo 489 destinados ao treinamento e validação dos modelos, e 110 reservados exclusivamente para o teste e ranqueamento das soluções submetidas pelos participantes. Assim como na edição anterior, as imagens foram disponibilizadas no formato NIFTI (GROUP et al., 2023), que contém os volumes referentes às aquisições de TC juntamente com as anotações fornecidas pelos especialistas.

Por se tratar de uma extensão da base anterior, variações nos protocolos de aquisição e na duração dos exames podem influenciar no número de imagens geradas por paciente. Além disso, o conjunto KiTS23 fornece anotações manuais detalhadas das regiões renais e das patologias associadas, oferecendo suporte à segmentação de rins, tumores e cistos renais. Para a avaliação dos resultados submetidos pelos participantes, o desafio estabelece

quatro macroclasses, denominadas Classes de Avaliação Hierárquica (CAH). Essas classes são definidas a partir do coeficiente Dice aplicado às estruturas segmentadas e organizadas da seguinte forma: Rins e Massas (Rim + Tumor + Cisto), Massas Renais (Tumor + Cisto) e Tumores Renais, como mostra na Figura 3.

Tabela 3 – Classes de Avaliação Hierárquica (CAH)

Categoria	**Rins**	**Tumores**	**Cistos**
Rins e Massas	X	X	X
Massas Renais		X	X
Tumores		X	

5.3 Resultados Quantitativos

Os resultados quantitativos obtidos pelo modelo proposto a partir da validação cruzada com cinco iterações sobre o conjunto de dados KiTS23 são apresentados nas Tabelas 4 e 5. A análise considera métricas de sobreposição volumétrica, representadas pelo coeficiente Dice e pelo Erro Volumétrico de Sobreposição (VOE), além de métricas de precisão de contorno, expressas pela Distância de Hausdorff ao percentil 95 (HD95) e pela Distância de Superfície Média (MSD). Para o coeficiente Dice, valores mais elevados indicam maior concordância com as anotações de referência. Para VOE, HD95 e MSD, valores menores correspondem a melhor desempenho.

5.3.1 Desempenho por Estrutura

Conforme apresentado na Tabela 4, a segmentação do rim obteve o maior coeficiente de Dice entre as estruturas avaliadas, com valor médio de $93,80 \pm 1,16\%$. Esse resultado indica elevada concordância volumétrica entre as segmentações produzidas pelo modelo e as anotações de referência. Observa-se também que o desvio padrão reduzido evidencia a estabilidade do modelo ao longo das cinco iterações da validação cruzada, sugerindo uma adequada capacidade de generalização frente às variações morfológicas dos rins presentes no conjunto de dados.

Tabela 4 – Resultados por estrutura Dice , HD95, MSD e VOE.

Estrutura	**Dice (%)**	**HD95**	**MSD**	**VOE (%)**
Rim	$93,80 \pm 1,16$	$2,75 \pm 0,21$	$0,59 \pm 0,04$	$6,96 \pm 0,40$
Tumor	$86,27 \pm 0,87$	$2,01 \pm 0,09$	$0,72 \pm 0,01$	$12,30 \pm 0,16$
Cisto	$92,76 \pm 1,53$	$0,75 \pm 0,09$	$0,31 \pm 0,04$	$6,05 \pm 0,59$

Fonte: Elaborado pelo autor.

As métricas relacionadas à precisão de contorno corroboram esse comportamento. O valor médio de HD95 de $2,75 \pm 0{,}21$ mm e o MSD de $0{,}59 \pm 0{,}04$ mm indicam que as discrepâncias de borda permanecem relativamente limitadas. De forma complementar, o VOE de $6,96 \pm 0{,}40\%$, cuja interpretação considera valores próximos de zero como indicativos de menor discrepância volumétrica, confirma a elevada correspondência entre as segmentações preditas e as máscaras anotadas manualmente.

A segmentação de cistos renais apresentou desempenho semelhante ao observado para os rins, com coeficiente Dice médio de $92,76 \pm 1,53\%$. Observa-se ainda que essa classe apresentou os menores valores de HD95 e MSD entre todas as estruturas avaliadas, com valores de $0{,}75 \pm 0{,}09$ mm e $0{,}31 \pm 0{,}04$ mm, respectivamente. Esses resultados indicam que as fronteiras dos cistos foram delimitadas com elevada precisão geométrica. Esse comportamento pode estar relacionado às características morfológicas relativamente regulares dos cistos e ao contraste frequentemente mais evidente dessas estruturas em exames de tomografia computadorizada, fatores que tendem a facilitar a identificação de seus contornos. O VOE médio de $6,05 \pm 0,59\%$ reforça a boa correspondência volumétrica obtida nessa classe.

Por outro lado, a segmentação de tumores renais representou a tarefa mais desafiadora entre as estruturas avaliadas, apresentando coeficiente Dice médio de $86,27 \pm 0,87\%$. Embora esse valor seja inferior aos observados para rins e cistos, o baixo desvio padrão indica comportamento consistente entre as diferentes iterações da validação cruzada. Esse resultado sugere que o modelo mantém estabilidade nas predições mesmo diante da elevada heterogeneidade morfológica e textural característica das massas tumorais renais presentes no conjunto KiTS23.

Em relação às métricas de contorno, o HD95 médio de $2{,}01 \pm 0{,}09$ mm representa o segundo melhor resultado entre as estruturas avaliadas, sendo inferior ao observado para o rim, embora superior ao do cisto. Já o MSD médio de $0{,}72 \pm 0{,}01$ mm supera o valor obtido para o rim (0,59 mm), indicando que, embora erros de grande magnitude sejam menos frequentes, há uma distribuição mais ampla de pequenos desvios ao longo da superfície das regiões tumorais. Esse comportamento, aliado ao VOE de $12,30 \pm 0{,}16\%$, reflete a maior complexidade associada à delimitação dessas estruturas, as quais frequentemente apresentam bordas irregulares e variações de intensidade nas imagens de tomografia computadorizada.

5.3.2 Avaliação no Formato Hierárquico do KiTS23

A Tabela 5 apresenta os resultados obtidos segundo o formato de avaliação hierárquica adotado pelo desafio KiTS23, cujas Classes de Avaliação Hierárquica (CAH) foram descritas na Seção 5.2.

Tabela 5 – Desempenho por iteração segundo as Classes de Avaliação Hierárquica (CAH) do KiTS23 (%)

Iteração	**Rins e Massas (%)**	**Massas Renais (%)**	**Tumores Renais (%)**
1	90,49	88,36	86,08
2	90,32	89,48	86,03
3	91,24	89,19	86,33
4	90,53	88,96	85,11
5	92,14	91,58	87,81
Média ± DP	$90,94 \pm 0,67$	$89,52 \pm 1,09$	$86,27 \pm 0,87$

Fonte: Elaborado pelo autor.

Os resultados médios obtidos foram de $90,94 \pm 0,67\%$ para a categoria Rins e Massas, $89,52 \pm 1,09\%$ para Massas Renais e $86,27 \pm 0,87\%$ para Tumores Renais. Observa-se uma redução gradual do desempenho ao longo das categorias hierárquicas, comportamento esperado em virtude do aumento da especificidade e da complexidade da tarefa de segmentação.

Ao analisar individualmente os resultados por iteração, verifica-se que a quinta iteração apresentou os maiores valores em todas as categorias avaliadas, alcançando valores de $92,14\%$, $91,58\%$ e $87,81\%$ para Rins e Massas, Massas Renais e Tumores Renais, respectivamente. Em contraste, a quarta iteração apresentou o menor desempenho para a categoria Tumores Renais, com valor de $85,11\%$. Essa variação possivelmente está associada à composição específica dos exames presentes nessa partição da validação cruzada.

Os desvios padrão reduzidos observados nas três categorias hierárquicas evidenciam a robustez da abordagem proposta, demonstrando a capacidade do modelo de manter desempenho consistente ao longo das cinco partições da validação cruzada, independentemente da composição dos exames em cada iteração.

5.4 Resultados Qualitativos

Nesta etapa, a investigação concentra-se na análise de casos práticos para a segmentação de estruturas renais, especificamente rins, tumores e cistos. O objetivo é realizar uma avaliação qualitativa que valide o comportamento dos modelos propostos nesta dissertação, indo além das métricas estatísticas.

A análise qualitativa de casos representativos é importante para avaliar a coerência clínica das predições, uma vez que métricas quantitativas, embora essenciais para medir o desempenho global do modelo, nem sempre refletem comportamentos específicos em situações anatômicas mais complexas.

Por meio da inspeção visual de amostras significativas e do emprego de técnicas de

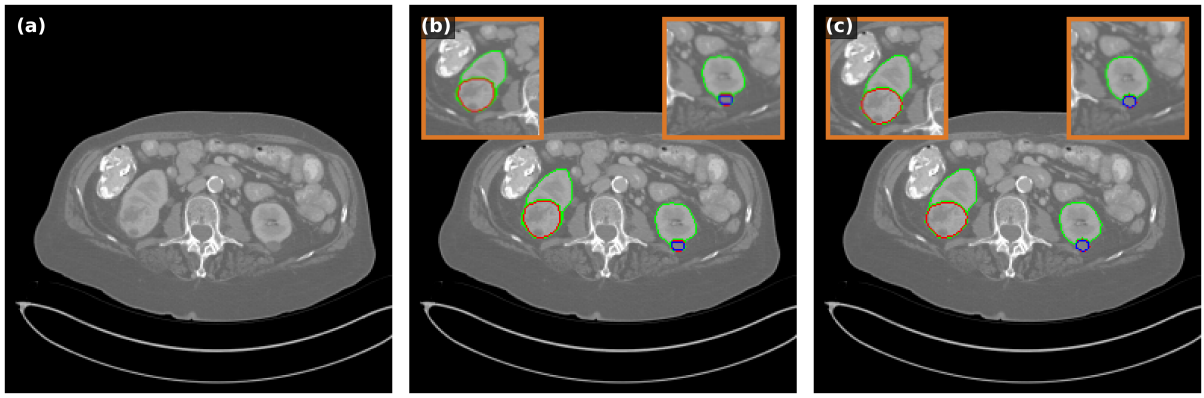
explicabilidade, notadamente o Grad-CAM (SELVARAJU et al., 2017), examinam-se as particularidades do processo de segmentação. Essa metodologia permite identificar padrões de acerto e eventuais vulnerabilidades do modelo frente a diferentes contextos clínicos e anatômicos.

Tal análise funciona como um suporte crítico aos dados quantitativos expostos anteriormente. Ela possibilita uma percepção detalhada sobre a robustez dos algoritmos ao processarem imagens de tomografia computadorizada, que frequentemente apresentam alta complexidade e variações morfológicas acentuadas.

5.4.1 Casos de Sucesso

Exame 00060 - Neste exame de tomografia computadorizada, disponibilizado na base KiTS23, a Figura 16 apresenta, da esquerda para a direita, (a) a fatia original do exame; (b) as anotações do especialista, nas quais os rins estão delimitados em verde, o tumor renal em vermelho e o cisto renal em azul; e (c) a segmentação produzida pelo modelo Dual-Scale SE. As inserções ampliadas nos painéis (b) e (c) destacam as regiões de maior interesse clínico, especificamente o tumor e o cisto renal, possibilitando a comparação direta entre a segmentação de referência e a predição do modelo.

Figura 16 – Exemplo de segmentação dos rins, tumor renal e cisto renal no exame 00060 utilizando o modelo Dual-Scale SE: (a) fatia original da tomografia computadorizada; (b) segmentação de referência (*ground truth*), com rins em verde, tumor em vermelho e cisto em azul; (c) segmentação produzida pelo modelo, seguindo a mesma convenção de cores.

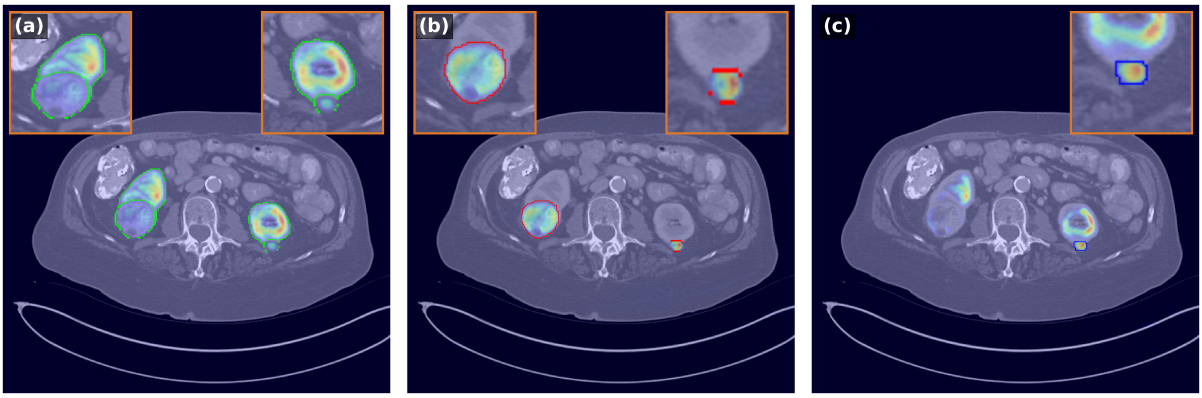


Fonte:Desenvolvido pelo Autor.

Nota-se que o modelo apresentou segmentação consistente das três estruturas de interesse, com contornos preditos próximos às anotações de referência. Destaca-se, particularmente, a identificação do cisto renal, estrutura de dimensões reduzidas e maior grau de dificuldade de segmentação, evidenciando a capacidade do modelo em preservar detalhes anatômicos mesmo em regiões de menor extensão espacial.

Considerando ainda o Exame 00060, a Figura 17 apresenta a aplicação da técnica Grad-CAM ao modelo Dual-Scale SE, exibindo os mapas de ativação sobrepostos à fatia de TC para cada classe segmentada, sendo (a) rim, (b) tumor renal e (c) cisto renal. No painel (a), as ativações de maior intensidade, representadas por tonalidades quentes como amarelo e vermelho, concentram-se predominantemente sobre as regiões renais, enquanto as áreas periféricas apresentam ativação reduzida, associada a tonalidades frias, sugerindo direcionamento seletivo do modelo às estruturas-alvo.

Figura 17 – Mapas de ativação Grad-CAM gerados pelo modelo Dual-Scale SE para o exame 00060, sobrepostos à fatia de tomografia computadorizada: (a) ativação para a classe rim (contornos em verde); (b) ativação para a classe tumor renal (contornos em vermelho); (c) ativação para a classe cisto renal (contorno em azul).



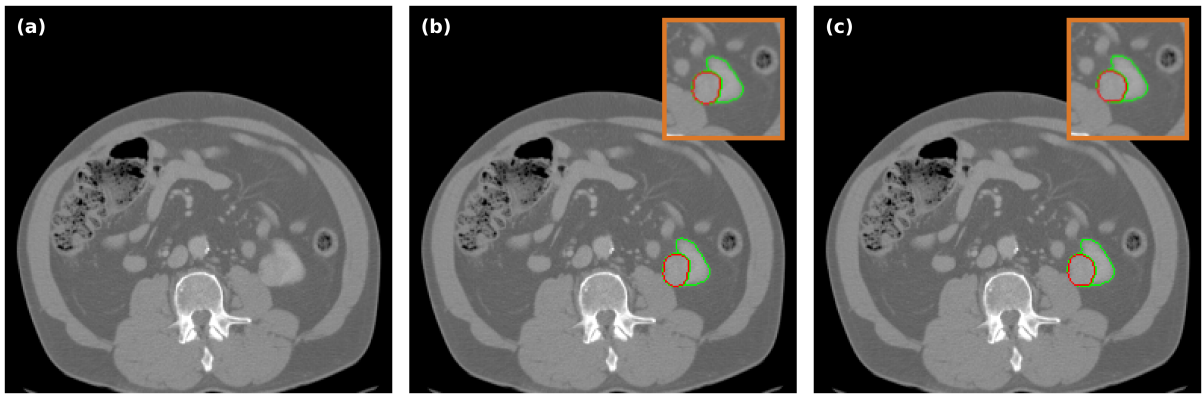
Fonte: Desenvolvido pelo Autor.

No painel (b), o mapa correspondente ao tumor renal evidencia ativação intensa e espacialmente bem delimitada sobre a lesão, conforme observado na inserção ampliada. De maneira análoga, no painel (c), o mapa referente ao cisto renal apresenta ativações concentradas na região correspondente, ainda que com menor intensidade quando comparado às demais classes. Tal comportamento é compatível com o reduzido volume da lesão e com a menor diferenciação de contraste em relação ao tecido renal adjacente.

Em conjunto, os mapas Grad-CAM sugerem que o modelo orienta sua atenção de maneira seletiva e dependente da classe, reforçando a eficácia da abordagem proposta na extração de características relevantes, mesmo diante da coexistência de múltiplas estruturas patológicas em uma mesma fatia.

Exame 00175 - Neste exame de tomografia computadorizada, também pertencente à base KiTS23, a Figura 18 apresenta, da esquerda para a direita, (a) a fatia original do exame; (b) as anotações do especialista, com o rim delimitado em verde e o tumor renal em vermelho; e (c) a segmentação produzida pelo modelo Dual-Scale SE. Diferentemente do caso anterior, observa-se que apenas um rim está visível na fatia analisada, não havendo presença de cisto renal.

Figura 18 – Exemplo de segmentação do rim e tumor renal no exame 00175 utilizando o modelo Dual-Scale SE: (a) fatia original da tomografia computadorizada; (b) segmentação de referência (*ground truth*), com rim em verde e tumor em vermelho; (c) segmentação produzida pelo modelo, seguindo a mesma convenção de cores.



Fonte: Desenvolvido pelo Autor.

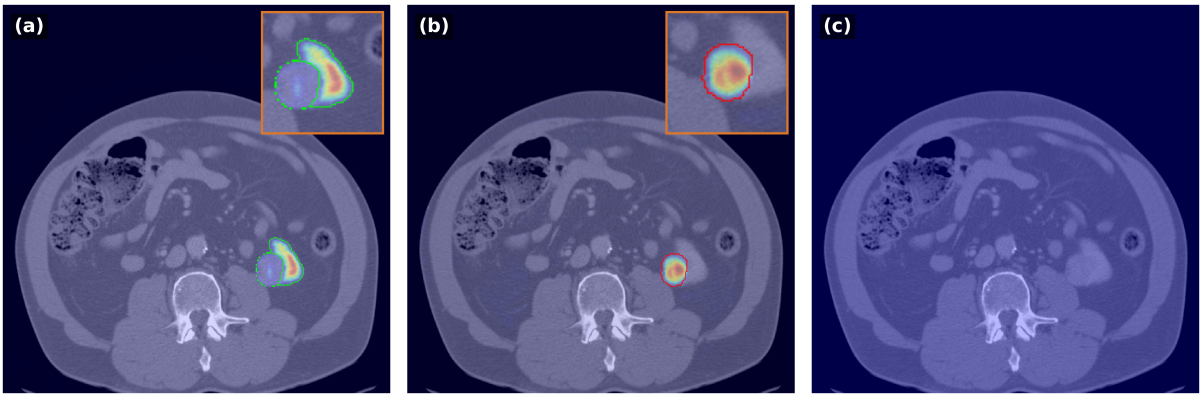
As inserções ampliadas nos painéis (b) e (c) permitem comparação direta entre a segmentação de referência e a predição do modelo. Nota-se que o modelo segmentou adequadamente tanto o rim quanto o tumor renal, apresentando contornos preditos próximos às anotações do especialista. A correta delimitação do rim único presente na imagem evidencia a capacidade do modelo em adaptar sua predição ao contexto anatômico específico da fatia, sem induzir segmentações incorretas na região contralateral.

A Figura 19 apresenta os mapas de ativação obtidos por meio da técnica Grad-CAM, sobrepostos à fatia de TC para cada classe considerada pelo modelo: (a) rim, (b) tumor renal e (c) cisto renal. No painel (a), as ativações de maior intensidade concentram-se na região correspondente ao rim visível, indicando que o modelo direciona sua atenção de forma coerente à estrutura anatômica presente. No painel (b), observa-se ativação intensa e bem delimitada sobre a área tumoral, conforme evidenciado na inserção ampliada.

No painel (c), referente à classe cisto renal, não se observam ativações relevantes na região abdominal, o que é consistente com a ausência dessa estrutura na fatia analisada. Esse comportamento sugere que o modelo não apenas identifica corretamente as classes presentes, mas também demonstra capacidade de suprimir ativações indevidas quando determinada classe não está representada na imagem.

Em conjunto, os resultados qualitativos e os mapas de ativação indicam que o modelo mantém comportamento estável mesmo em cenários com estruturas ausentes ou visualização parcial da anatomia, reforçando sua robustez e capacidade de generalização.

Figura 19 – Mapas de ativação Grad-CAM gerados pelo modelo Dual-Scale SE para o exame 00175, sobrepostos à fatia de tomografia computadorizada: (a) ativação para a classe rim (contornos em verde); (b) ativação para a classe tumor renal (contornos em vermelho); (c) ausência de ativação para a classe cisto renal.

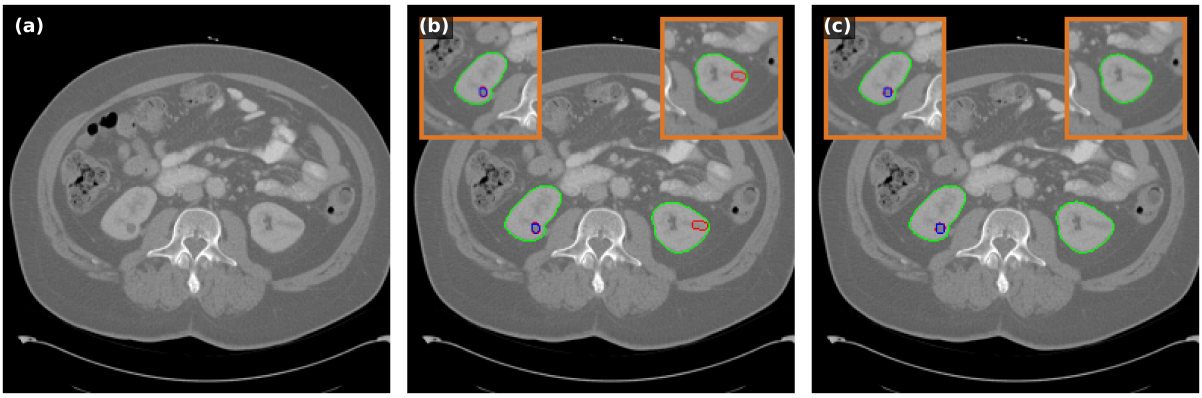


Fonte: Desenvolvido pelo Autor.

5.4.2 Casos de Erro

Exame 00289 - Neste exame de tomografia computadorizada, também pertencente à base KiTS23, a Figura 20 apresenta, da esquerda para a direita, (a) a fatia original do exame; (b) as anotações do especialista, com os rins delimitados em verde, o tumor renal em vermelho e o cisto renal em azul; e (c) a segmentação produzida pelo modelo Dual-Scale SE. Nesta fatia, observam-se ambos os rins, além da presença de duas lesões distintas: um pequeno cisto em um dos rins e uma lesão tumoral no rim contralateral.

Figura 20 – Exemplo de segmentação do rim, tumor renal e cisto renal no exame 00289 utilizando o modelo Dual-Scale SE: (a) fatia original da tomografia computadorizada; (b) segmentação de referência (*ground truth*), com rim em verde, tumor em vermelho e cisto em azul; (c) segmentação produzida pelo modelo, seguindo a mesma convenção de cores.



Fonte: Desenvolvido pelo Autor.

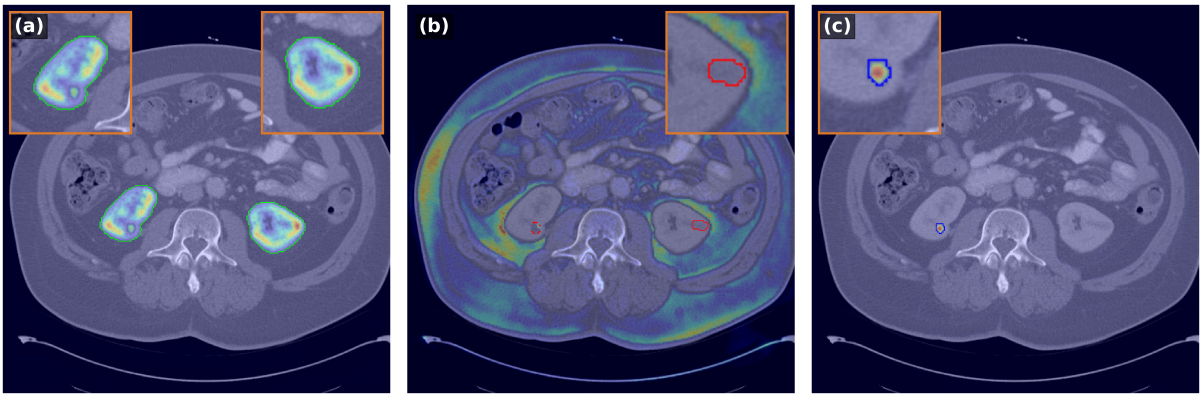
A comparação entre os painéis (b) e (c) evidencia um comportamento assimétrico

do modelo. O contorno renal foi adequadamente segmentado em ambos os lados, com boa correspondência espacial em relação às anotações do especialista, indicando que a rede mantém desempenho consistente para a classe rim. No caso do cisto renal, a predição também apresenta alinhamento satisfatório com a marcação de referência, ainda que com pequenas diferenças no contorno, o que é esperado devido ao reduzido tamanho da estrutura.

Entretanto, é possível observar que o tumor renal anotado no painel (b) não foi segmentado pelo modelo no painel (c). A ausência completa de predição para essa lesão caracteriza um falso negativo, evidenciando uma limitação do modelo na detecção dessa estrutura específica nesta fatia. Considerando que se trata de uma lesão de pequenas dimensões, é possível que fatores como baixo contraste em relação ao tecido renal não acometido ou variabilidade morfológica tenham dificultado sua identificação pela rede.

A Figura 21 apresenta os mapas de ativação obtidos por meio da técnica Grad-CAM para as classes rim, tumor renal e cisto renal. No painel (a), referente à classe rim, as ativações concentram-se adequadamente nas regiões correspondentes aos rins, demonstrando coerência entre a atenção do modelo e a anatomia presente. No painel (c), associado à classe cisto renal, observa-se ativação localizada na região do cisto, em concordância com a segmentação produzida.

Figura 21 – Mapas de ativação Grad-CAM gerados pelo modelo Dual-Scale SE para o exame 00289, sobrepostos à fatia de tomografia computadorizada: (a) ativação para a classe rim (contornos em verde); (b) ausência de ativação significativa na região correspondente ao tumor renal anotado (contornos em vermelho), evidenciando falha de detecção; (c) ativação localizada para a classe cisto renal (contornos em azul).



Fonte: Desenvolvido pelo Autor.

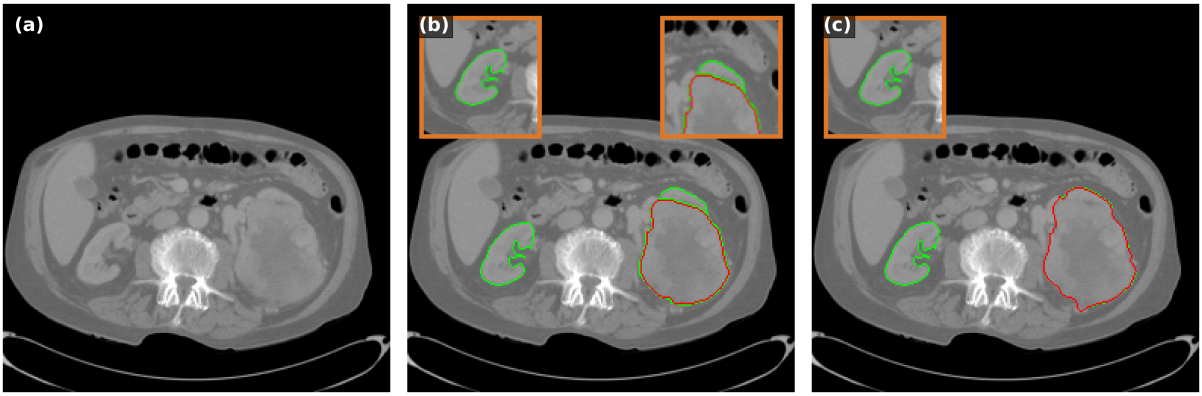
Por outro lado, no painel (b), referente à classe tumor renal, não se observam ativações relevantes na região correspondente ao tumor anotado pelo especialista. A ausência de resposta significativa do Grad-CAM nessa área indica que o modelo não direcionou atenção à lesão tumoral, o que é consistente com a falha de segmentação

observada. Esse comportamento sugere que o erro não está restrito apenas à etapa de decisão final da máscara, mas já se manifesta no padrão de ativação interna da rede.

Em conjunto, os resultados qualitativos indicam que, embora o modelo apresente desempenho robusto para a segmentação renal e para a identificação do cisto nesta fatia, ele demonstra limitação na detecção de tumores pequenos ou de menor contraste, resultando em falso negativo. A análise dos mapas de ativação reforça essa interpretação, ao evidenciar a ausência de foco atencional na região tumoral, contribuindo para uma compreensão mais aprofundada das falhas do modelo neste caso específico.

Exame 00221 - Neste exame de tomografia computadorizada, também pertencente à base KiTS23, a Figura 22 apresenta, da esquerda para a direita, (a) a fatia original do exame; (b) as anotações do especialista, com o rim delimitado em verde e o tumor renal em vermelho; e (c) a segmentação produzida pelo modelo Dual-Scale SE. Observa-se a presença de uma volumosa massa tumoral ocupando grande parte da topografia do rim direito, com importante distorção da anatomia habitual, enquanto o rim contralateral mantém morfologia preservada.

Figura 22 – Exemplo de segmentação do rim e tumor renal no exame 00221 utilizando o modelo Dual-Scale SE: (a) fatia original da tomografia computadorizada; (b) segmentação de referência (*ground truth*), com rim em verde e tumor em vermelho; (c) segmentação produzida pelo modelo, seguindo a mesma convenção de cores.



Fonte: Desenvolvido pelo Autor.

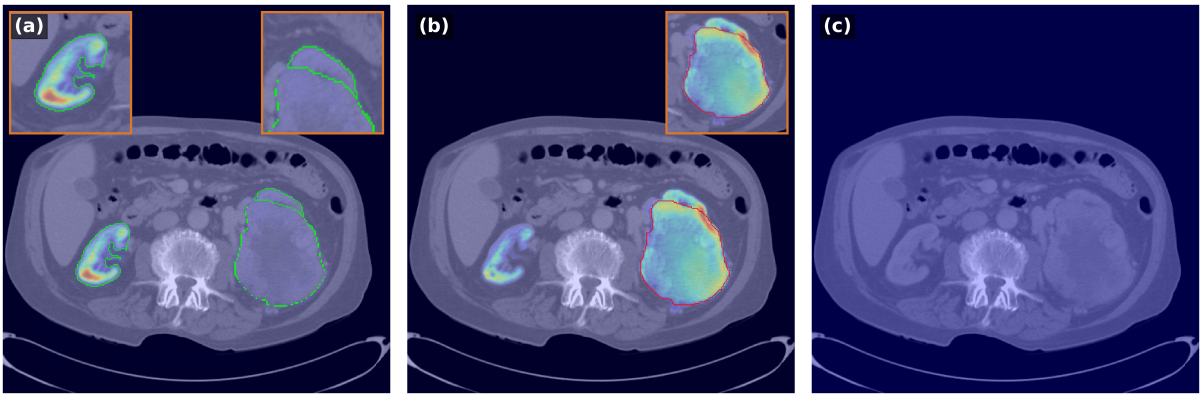
A comparação entre os painéis (b) e (c) evidencia um comportamento distinto para as duas estruturas. O modelo foi capaz de segmentar adequadamente o rim contralateral, apresentando contorno compatível com a anotação de referência. Entretanto, em relação ao rim direito, observa-se que boa parte do parênquima renal não foi corretamente identificada como classe rim, possivelmente em decorrência da extensa substituição tumoral e da perda dos limites anatômicos habituais. Assim, há subsegmentação do rim direito, caracterizando limitação do modelo na delimitação da porção residual do órgão.

Por outro lado, a massa tumoral extensa foi identificada de forma consistente, com a segmentação em vermelho acompanhando de maneira geral os limites da lesão descritos no *ground truth*. Embora existam pequenas discrepâncias nas bordas, especialmente em áreas de transição com tecidos adjacentes, o modelo conseguiu capturar a maior parte do volume tumoral, sem gerar segmentações espúrias relevantes em regiões distantes.

A Figura 23 apresenta os mapas de ativação Grad-CAM para cada classe considerada pelo modelo. No painel (a), referente à classe rim, as ativações concentram-se predominantemente no rim preservado, com menor intensidade na região correspondente ao rim direito acometido, corroborando a dificuldade observada na segmentação dessa estrutura. No painel (b), observa-se ativação intensa e difusa na região da massa tumoral, indicando que a rede direciona sua atenção de forma coerente à área patológica predominante na imagem. No painel (c), não se observam ativações relevantes para a classe cisto renal, comportamento consistente com a ausência dessa estrutura na fatia analisada.

De forma geral, este caso demonstra que, embora o modelo apresente desempenho reduzido na delimitação do rim amplamente infiltrado pela lesão, ele mantém capacidade de identificar e segmentar lesões tumorais volumosas, direcionando adequadamente sua atenção à região de maior relevância patológica.

Figura 23 – Mapas de ativação Grad-CAM gerados pelo modelo Dual-Scale SE para o exame 00221, sobrepostos à fatia de tomografia computadorizada: (a) ativação para a classe rim (contornos em verde), com foco predominante no rim contralateral; (b) ativação para a classe tumor renal (contornos em vermelho), concentrada na volumosa massa tumoral; (c) ausência de ativação para a classe cisto renal, consistente com a não presença dessa estrutura na fatia analisada.



Fonte: Desenvolvido pelo Autor.

5.5 Comparativo de Arquiteturas Avaliadas

Foi realizada uma avaliação sistemática entre todas as variações arquiteturais desenvolvidas neste trabalho, com o objetivo de identificar quais modificações da U-Net

base contribuíram de forma mais efetiva para a segmentação de rins, tumores e cistos renais na base KiTS23. Os resultados consolidados, obtidos via validação cruzada com cinco iterações, são apresentados na Tabela 7 e discutidos a seguir.

Para facilitar a identificação das variações arquiteturais analisadas, a Tabela 6 apresenta um resumo das modificações estruturais introduzidas em cada modelo. Todas as arquiteturas derivam do módulo de extração multiescala Dual-Scale proposto neste trabalho, diferindo apenas pelos mecanismos complementares incorporados em cada configuração.

Tabela 6 – Descrição das variações arquiteturais baseadas no módulo Dual-Scale avaliadas neste trabalho.

Rede	**Descrição das modificações**
U-Net (base)	Arquitetura U-Net padrão sem módulos adicionais.
Dual-Scale SE (128)	Módulo Dual-Scale com atenção SE e entrada redimensionada para 128×128.
Dual-Scale SE + Deformable (128)	Dual-Scale SE com convoluções deformáveis e entrada 128×128.
Dual-Scale ECA	Substituição do mecanismo SE pelo Efficient Channel Attention (ECA).
Dual-Scale SE Siamese	Compartilhamento de pesos entre ramos paralelos do encoder.
Dual-Scale SE + Attention Gates	Inserção de Attention Gates nas conexões de salto da U-Net.
Dual-Scale ECA + Attention Gates	Combinação do mecanismo ECA com Attention Gates.
Dual-Scale SE (2.5D)	Empilhamento de fatias adjacentes $(i-1, i, i+1)$ para incorporar contexto inter-fatia.
Dual-Scale SE UNet	Arquitetura proposta final com extração multiescala Dual-Scale e atenção SE.

Fonte: Elaborado pelo autor.

A rede base U-Net, sem qualquer módulo adicional, obteve coeficientes Dice de 85,10% para tumores, 93,03% para rins e 93,74% para cistos. Esse resultado serve como linha de base para a análise das contribuições de cada modificação. Notavelmente, o Dice para cistos da U-Net base foi o mais elevado entre todas as redes avaliadas, o que sugere que, para estruturas com maior homogeneidade morfológica, modificações arquiteturais adicionais nem sempre resultam em ganho e podem introduzir instabilidade durante o treinamento.

A primeira família de modificações introduzida baseou-se na incorporação do módulo Dual-Scale, combinado com mecanismos de atenção por canal. Para permitir uma comparação direta com a variante que utiliza convoluções deformáveis, cujo custo computacional é significativamente maior, foi definida a configuração Dual-Scale SE (128),

Tabela 7 – Comparação dos resultados de segmentação (Dice) entre as arquiteturas avaliadas na base KiTS23.

Rede	**Dice Tumor (%)**	**Dice Rim (%)**	**Dice Cisto (%)**
U-Net (base)	85,10	93,03	**93,74**
Dual-Scale SE (128)	79,91	90,32	90,65
Dual-Scale SE + Deformable (128)	82,05	91,26	89,79
Dual-Scale ECA	83,80	91,55	91,43
Dual-Scale SE Siamese	84,13	91,82	91,53
Dual-Scale SE + Attention Gates	85,39	93,37	92,58
Dual-Scale ECA + Attention Gates	85,45	93,17	92,73
Dual-Scale SE (2.5D)	85,60	93,29	91,78
Dual-Scale SE UNet	**86,27**	**93,80**	92,76

Fonte: Elaborado pelo autor.

correspondente à arquitetura Dual-Scale SE proposta neste trabalho com redimensionamento das imagens de entrada para 128×128. Essa estratégia foi necessária para viabilizar o treinamento da versão Dual-Scale SE + Deformable (128) sem exceder os limites de memória da GPU, garantindo que ambas as arquiteturas fossem avaliadas sob a mesma resolução espacial.

Nessa configuração reduzida, a Dual-Scale SE (128) obteve Dice de 79,91% para tumores, 90,32% para rins e 90,65% para cistos. Como essa versão utiliza uma resolução espacial menor, ocorre perda de detalhes estruturais finos, particularmente relevantes para tumores e pequenas formações císticas, o que explica a redução do desempenho em relação às arquiteturas treinadas na resolução padrão. Dessa forma, essa configuração deve ser interpretada como uma variante controlada por resolução para análise comparativa, e não como a configuração finalmente otimizada da arquitetura proposta.

A variante Dual-Scale SE + Deformable (128), que substitui as convoluções padrão por convoluções deformáveis mantendo a mesma resolução de entrada (128×128), obteve melhoria em relação à configuração Dual-Scale SE (128) (82,05% / 91,26% / 89,79%). Esse resultado indica que a adaptação geométrica do campo receptivo, proporcionada pelas convoluções deformáveis, contribui para a modelagem das estruturas renais, mesmo em baixa resolução. Entretanto, o desempenho permaneceu inferior ao da U-Net base treinada na resolução original, sugerindo que a limitação espacial impacta mais significativamente o resultado do que o ganho representacional proporcionado pelas deformações.

Além do impacto em desempenho, observou-se aumento significativo no custo computacional da versão com convoluções deformáveis. O tempo médio de treinamento da arquitetura Dual-Scale SE + Deformable (128) foi de aproximadamente 5 horas por execução, enquanto a configuração Dual-Scale SE (128) demandou menos de 1 hora nas mesmas condições experimentais. Esse aumento está associado à maior complexidade das operações deformáveis, que introduzem parâmetros adicionais e etapas extras de

interpolação espacial, elevando o custo de processamento e o consumo de memória.

A variante Dual-Scale SE Siamese introduziu uma estratégia de compartilhamento de pesos entre ramos paralelos do codificador, obtendo Dice de 84,13% para tumores, 91,82% para rins e 91,53% para cistos. Já a arquitetura Dual-Scale ECA substituiu o mecanismo SE pelo bloco Efficient Channel Attention (ECA), reduzindo o custo paramétrico. Os resultados obtidos foram 83,80% para tumores, 91,55% para rins e 91,43% para cistos, apresentando comportamento semelhante à variante Siamese.

A incorporação de portões de atenção (Attention Gates) nas conexões de salto da U-Net resultou em ganhos consistentes em relação às versões sem atenção. A Dual-Scale SE + Attention Gates alcançou 85,39% para tumores, 93,37% para rins e 92,58% para cistos, enquanto a Dual-Scale ECA + Attention Gates obteve 85,45%, 93,17% e 92,73%, respectivamente. Esses resultados indicam que os portões de atenção contribuem para a seleção de características mais discriminativas nas conexões de salto.

Diferentemente das abordagens anteriores, que focam na recalibração de características espaciais e de canal, a próxima variante incorpora contexto volumétrico, mantendo a estrutura Dual-Scale. A arquitetura Dual-Scale SE (2.5D) utiliza o empilhamento de fatias adjacentes como canais de entrada, considerando a fatia central juntamente com as fatias imediatamente anterior e posterior ($i-1$, i e $i+1$). Essa estratégia permite inserir contexto inter-fatia de forma limitada, preservando um custo computacional próximo ao de abordagens 2D. Os resultados obtidos foram Dice de 85,60% para tumores, 93,29% para rins e 91,78% para cistos.

A arquitetura Dual-Scale SE UNet, proposta neste trabalho, obteve os melhores resultados para tumores (86,27%) e rins (93,80%) dentre todas as redes avaliadas, com Dice para cistos de 92,76%. Em comparação com a U-Net base, observou-se superioridade para as classes tumor e rim, enquanto, para a classe cisto, o valor obtido permaneceu inferior ao do modelo base.

A melhoria observada em relação à U-Net base e às demais modificações confirma a eficácia da estratégia multiescala adotada, evidenciando que a combinação de diferentes níveis de contexto espacial contribui para a segmentação das estruturas renais mais complexas.

Em síntese, a comparação entre as variações avaliadas demonstra que: (i) modificações isoladas apresentam ganhos limitados; (ii) a combinação de mecanismos complementares resulta em melhorias progressivas; e (iii) a arquitetura Dual-Scale SE UNet apresentou o melhor desempenho geral, sendo adotada como modelo final deste trabalho.

5.6 Comparação com Trabalhos Relacionados

A base KiTS23 representa um dos desafios mais exigentes na segmentação de estruturas renais em imagens de TC, devido à grande variação morfológica de cistos e tumores, à presença de fronteiras difusas entre estruturas patológicas e parênquima saudável, e ao severo desequilíbrio volumétrico entre classes. A comparação apresentada na Tabela 8 posiciona a Dual-Scale SE UNet proposta em relação às abordagens arquiteturais e multiestágio mais recentes, considerando o desempenho simultâneo nas três hierarquias de avaliação do desafio: *Rins e Massas*, *Massas* e *Tumor*.

Tabela 8 – Comparação dos métodos na base KiTS23.

			Base KiTS23		
Categoria	**Trabalho**	**Método**	**Rins e Massas**	**Massas**	**Tumor**
Arquitetural	Qian et al. (2023)	nnU-Net + Swin Transformer	-	-	68,70%
	Myronenko et al. (2023)	SegResNet + SwinUNETR	95,60%	79,20%	75,80%
	Li, Peng e Zhang (2023)	nnU-Net + Atenção Residual	93,60%	-	67,00%
	Stoica, Breaban e Barbu (2023)	3D U-Net + DA	94,70%	76,00%	71,30%
	Tanasković et al. (2024)	YOLOv8n-seg (2D)	-	-	79,00%
	Matos et al. (2024)	CPP-UNet (PPM + ASPP)	92,84%	92,08%	88,17%
	Karunanayake et al. (2025)	ViT + 3D UNet (Dual-Stage)	97,00%	-	88,00%
	Demirtaş, İnner e Kavak (2025)	Mesh Reconstruction (Model 3)	-	-	98,00%
	Jariwala et al. (2024)	DeepLabv3+	98,22%	-	-
	Alonso-Monsalve et al. (2025)	Sparse U-Net	95,80%	85,70%	80,30%
Multi-estágio	Hu e Peng (2023)	3D U-Net + GSCA-Net	93,30%	74,40%	67,90%
	Vispi (2023)	EfficientNet-B5 + U-Net	97,71%	81,39%	73,81%
	Chen e Zhang (2023)	ResUNet + Recorte de ROI	89,44%	85,85%	85,91%
	Alexandru e Popescu (2025)	YOLO + EfficientNet-B4/MiT-B2	95,82%	-	62,60%
	Wang et al. (2023)	nnU-Net (2 etapas)	86,60%	54,50%	49,00%
	Salahuddin et al. (2023)	nnU-Net + Taxa de Ciclo (Ensemble)	94,00%	86,50%	83,50%
	Uhm et al. (2023)	3D U-Net Paralela (Multiescala)	97,90%	82,60%	85,70%
	Sina Ziaee, Maleki e Ovens (2025)	Rel-UNet (Múltiplos Mínimos Locais)	80,00%	70,20%	64,10%
Modelo Proposto	**Dual-Scale SE UNet**	U-Net Modificada	**90,94%**	**89,52%**	**86,27%**

5.6.1 Comparação com Abordagens Arquiteturais

Entre os métodos de modificação arquitetural, observa-se uma tendência recorrente de incorporar mecanismos baseados em Transformers para modelar dependências de longo alcance. Qian et al. (2023) combinam a nnU-Net com o Swin Transformer, porém reportam apenas o coeficiente Dice para a classe *Tumor* (68,70%), o que limita uma avaliação abrangente das três hierarquias consideradas neste trabalho. Já Myronenko et al. (2023) utilizam as arquiteturas SegResNet e SwinUNETR com busca automatizada de arquitetura, obtendo 95,60% em *Rins e Massas*, 79,20% em *Massas* e 75,80% em *Tumor*. Embora o desempenho na hierarquia mais ampla seja superior ao do modelo proposto, a Dual-Scale SE UNet apresenta valores mais elevados nas classes *Massas* e *Tumor*. Esses resultados sugerem que diferentes estratégias de modelagem de contexto (como autoatenção global e convoluções multiescala) podem influenciar de forma distinta a segmentação de estruturas

com alta variabilidade morfológica. Ressalta-se, contudo, que diferenças de protocolo experimental também podem impactar essa comparação.

A CPP-UNet de Matos et al. (2024), que combina módulos PPM e ASPP no codificador, é conceitualmente a abordagem mais próxima da proposta deste trabalho, pois também explora mecanismos multiescala para captura de contexto. Os valores reportados (92,84% / 92,08% / 88,17%) superam os da Dual-Scale SE UNet nas três hierarquias. Esse comportamento é consistente com o uso de convoluções dilatadas no ASPP, que ampliam o campo receptivo efetivo ao amostrar o contexto em múltiplas taxas de dilatação simultaneamente. Por outro lado, a Dual-Scale SE UNet incorpora blocos Squeeze-and-Excitation, responsáveis pela recalibração adaptativa de canais, o que pode contribuir para a seletividade das representações em estruturas menores ou com limites pouco definidos. A análise isolada desses fatores exigiria estudos ablatórios específicos, representando uma possível direção para trabalhos futuros.

O trabalho de Demirtaş, İnner e Kavak (2025) reporta o maior valor para a classe *Tumor* (98,00%), porém o método inclui uma etapa adicional de reconstrução geométrica com suavização de superfície após a segmentação. Esse pós-processamento tende a regularizar as bordas preditas e pode elevar o coeficiente Dice, o que limita a comparação direta com abordagens baseadas exclusivamente em predição voxel-wise.

Em contraste com métodos que reportam resultados apenas para uma das hierarquias, como Jariwala et al. (2024) (98,22% exclusivamente em *Rins e Massas*) e Tanasković et al. (2024) (79,00% exclusivamente em *Tumor*), a Dual-Scale SE UNet apresenta desempenho mensurável nas três hierarquias simultaneamente (90,94% / 89,52% / 86,27%). Essa característica é relevante no contexto prático, uma vez que a segmentação conjunta de rins, cistos e tumores constitui o cenário típico de aplicação clínica. Além disso, a variação relativamente reduzida entre os resultados das três hierarquias sugere comportamento estável entre diferentes níveis estruturais.

A U-Net Esparsa de Alonso-Monsalve et al. (2025) obtém 95,80% / 85,70% / 80,30%. Embora apresente elevada eficiência computacional, o modelo proposto alcança valores superiores nas classes *Massas* e *Tumor*. Essa diferença pode estar associada à combinação de convoluções multiescala e recalibração intercanal introduzida pelos blocos Squeeze-and-Excitation, que favorecem a discriminação de regiões com transições sutis entre tecidos.

5.6.2 Comparação com Abordagens Multi-estágio

Entre as abordagens multi-estágio, Wang et al. (2023) implementam uma cascata composta por duas instâncias da nnU-Net, obtendo 86,60% / 54,50% / 49,00% nas três hierarquias avaliadas. O modelo proposto apresenta valores superiores, com diferenças

mais expressivas nas classes *Massas* e *Tumor*. Esses resultados indicam que, nas condições experimentais consideradas, o uso de um modelo em estágio único com aprimoramentos no codificador por meio de convoluções multiescala e mecanismos de atenção por canal mostrou desempenho competitivo em relação à estratégia de decomposição sequencial da segmentação.

O método proposto por Salahuddin et al. (2023) utiliza uma estratégia de *ensemble* implícito baseada em taxa de aprendizado cíclica, alcançando 94,00% / 86,50% / 83,50%. Observa-se vantagem desse método na hierarquia *Rins e Massas* (94,00% contra 90,94%), enquanto o modelo proposto apresenta resultado superior na classe *Tumor* (86,27% contra 83,50%). Considerando que a abordagem emprega uma estratégia de *ensemble*, os resultados indicam desempenho competitivo da Dual-Scale SE UNet mesmo sem a adoção de múltiplos modelos ou etapas adicionais de inferência.

Em Uhm et al. (2023), volumes são processados em paralelo por duas U-Nets 3D em diferentes resoluções, resultando em 97,90% / 82,60% / 85,70%. O modelo proposto apresenta valor superior na hierarquia *Massas* (89,52% contra 82,60%), classe diretamente relacionada à segmentação de lesões renais em relação ao parênquima saudável. Esse resultado pode estar associado à combinação de informações locais e contextuais promovida pelo bloco Dual-Scale, que busca integrar padrões multiescala durante o processo de extração de características. Por fim, o *Rel-UNet* de Sina Ziaee, Maleki e Ovens (2025), voltado à quantificação de incerteza, reporta 80,00% / 70,20% / 64,10%, sendo superado pelo modelo proposto nas três hierarquias avaliadas.

5.6.3 Análise Global do Modelo Proposto

A análise conjunta dos resultados obtidos na base KiTS23 evidencia que a Dual-Scale SE UNet apresenta desempenho equilibrado e consistente nas três hierarquias de avaliação do desafio (90,94% para *Rins e Massas*, 89,52% para *Massas* e 86,27% para *Tumor*). Diferentemente de diversas abordagens comparadas, que exibem elevado desempenho na hierarquia mais abrangente, mas reduções mais acentuadas nas classes diretamente associadas às lesões, o modelo proposto mantém variação relativamente reduzida entre os níveis estruturais avaliados (intervalo de 4,67 pontos percentuais). Esse comportamento sugere maior estabilidade na segmentação de estruturas com diferentes graus de complexidade morfológica.

O desempenho observado na classe *Massas* (89,52%) é particularmente relevante, uma vez que essa hierarquia concentra um dos principais desafios clínicos, envolvendo a delimitação de regiões patológicas no parênquima renal saudável, frequentemente caracterizadas por fronteiras imprecisas e por heterogeneidade de contraste. A integração de informações multiescala no codificador amplia o campo receptivo efetivo, sem comprometer significativamente a resolução espacial, favorecendo a captura simultânea de

contexto global e de detalhes locais. Paralelamente, os blocos Squeeze-and-Excitation promovem recalibração adaptativa de canais, contribuindo para a seleção de características discriminativas em cenários com baixo contraste ou elevada variabilidade estrutural.

Ressalta-se, contudo, que comparações diretas entre trabalhos devem ser interpretadas com cautela, visto que diferenças nos protocolos experimentais, estratégias de particionamento dos dados, procedimentos de pré-processamento e critérios de avaliação podem impactar os valores reportados. Ainda assim, os resultados indicam que a estratégia proposta constitui uma alternativa robusta para a segmentação simultânea de rins, cistos e tumores renais, apresentando equilíbrio entre desempenho quantitativo, estabilidade entre classes e coerência qualitativa, evidenciada pelos estudos de caso e pelos mapas de ativação.

Em síntese, a Dual-Scale SE UNet demonstra que a combinação estruturada de extração multiescala e atenção por canal é capaz de aprimorar a representação de estruturas renais complexas sem recorrer a arquiteturas excessivamente profundas ou a pipelines multiestágio. Esse achado reforça a relevância de abordagens arquiteturais que priorizem a integração contextual eficiente e a seletividade representacional como caminho promissor para a segmentação médica em imagens de tomografia computadorizada.

5.7 Limitações da Pesquisa e Metodologia

Embora os resultados obtidos pelo modelo Dual-Scale SE UNet indiquem que a integração de convoluções multiescala com mecanismos de atenção por canal contribui para a segmentação de estruturas renais em tomografia computadorizada, é importante destacar as limitações inerentes à metodologia adotada, a fim de permitir uma análise crítica e contextualizada dos resultados apresentados.

Uma das principais limitações deste trabalho decorre da natureza bidimensional da abordagem proposta. O modelo opera sobre fatias individuais dos volumes de tomografia computadorizada, sem incorporar plenamente a continuidade volumétrica entre cortes adjacentes. Embora a variante Dual-Scale SE (2.5D) tenha sido investigada por meio do empilhamento de fatias vizinhas, essa estratégia fornece apenas uma aproximação do contexto tridimensional, sendo insuficiente para capturar dependências espaciais de longo alcance ao longo do eixo axial. Abordagens volumétricas completas, como as redes U-Net 3D, apresentam maior capacidade para modelar a continuidade morfológica de estruturas como tumores renais, cujos contornos e extensões se distribuem ao longo de múltiplos cortes.

A etapa de pré-processamento também constitui um fator limitante. O redimensionamento das imagens de 512×512 para 256×256 pixels, necessário para viabilizar o treinamento dentro das restrições de memória disponíveis, acarreta perda de detalhes es-

truturais finos. Esse efeito é especialmente relevante para lesões de pequenas dimensões, como cistos e tumores em estágio inicial, nas quais a resolução espacial é determinante para a delimitação adequada dos contornos. Os casos de erro analisados nos estudos de caso corroboram essa hipótese, evidenciando que tumores de menor volume e contraste reduzido constituem os principais desafios para o modelo.

A configuração dos filtros convolucionais do bloco Dual-Scale constitui outro aspecto passível de aprimoramento. A escolha dos kernels de tamanhos 3×3 e 7×7 foi determinada empiricamente, sem a realização de uma busca sistemática de hiperparâmetros. Kernels maiores ampliam o campo receptivo e favorecem a captura de contexto global, mas podem atenuar padrões locais relevantes. Por outro lado, kernels menores priorizam detalhes anatômicos finos, porém com campo receptivo mais restrito. A definição ideal dessas dimensões depende diretamente das características morfológicas das estruturas-alvo, e ajustes inadequados podem comprometer a fase de extração de características, impactando negativamente a diferenciação entre rins, cistos e tumores renais. A aplicação de métodos de otimização de hiperparâmetros, como *Grid Search* ou técnicas de busca *bayesiana*, poderia contribuir para a identificação de configurações mais adequadas.

O processo de recalibração adaptativa de canais, implementado pelos blocos Squeeze-and-Excitation, também apresenta limitações. A razão de redução de dimensionalidade utilizada na camada de *bottleneck* do bloco SE, definida como parâmetro fixo durante o treinamento, pode não ser igualmente eficaz em todos os níveis de profundidade da rede, em que o número de canais e a natureza das representações diferem substancialmente. A otimização individual desse parâmetro para cada nível do codificador representa uma direção promissora para investigações futuras.

As funções de perda adotadas, especificamente a combinação entre Dice Loss e Categorical Focal Loss, embora eficazes para lidar com o desequilíbrio volumétrico entre classes, também são sensíveis aos pesos associados a cada componente. A otimização desses pesos para cada estrutura específica, incluindo fundo, rim, cisto e tumor, poderia contribuir para um treinamento mais discriminativo. Esse ajuste pode ser particularmente relevante para a classe tumor, que apresentou os menores valores de Dice entre as estruturas avaliadas.

Por fim, a validação foi conduzida exclusivamente sobre os 489 volumes públicos do conjunto KiTS23, sem avaliação em bases de dados externas ou em protocolos de aquisição distintos. Esse aspecto limita a análise da capacidade de generalização do modelo diante de variações de equipamento, de protocolo de contraste e de características demográficas da população. Além disso, a ausência de etapas de pós-processamento volumétrico, como a filtragem de regiões desconexas ou a aplicação de restrições anatômicas, representa uma oportunidade adicional de melhoria. Algumas segmentações incorretas observadas nos estudos de caso poderiam ser mitigadas por meio dessas estratégias.

5.8 Considerações Finais

Este capítulo apresentou os resultados obtidos com a arquitetura Dual-Scale SE UNet para a segmentação de rins, tumores e cistos renais na base KiTS23. A avaliação evidenciou desempenho consistente e competitivo, tanto em termos quantitativos quanto qualitativos, com destaque para a estabilidade observada ao longo das iterações de validação cruzada e para o desempenho equilibrado entre as Classes de Avaliação Hierárquica do desafio.

A análise qualitativa, complementada pelos mapas de ativação Grad-CAM, reforçou a coerência anatômica das predições e evidenciou que o modelo direciona sua atenção de forma adequada às regiões de interesse, inclusive em cenários desafiadores. A comparação com trabalhos da literatura confirmou que a Dual-Scale SE UNet alcança desempenho competitivo sem recorrer a arquiteturas excessivamente complexas ou estratégias multi-estágio.

Por outro lado, as limitações identificadas, especialmente relacionadas à natureza bidimensional da abordagem e à ausência de pós-processamento volumétrico, indicam oportunidades claras de aprimoramento a serem exploradas em trabalhos futuros. No próximo capítulo, são apresentadas as conclusões desta dissertação, retomando os objetivos propostos e sintetizando as contribuições alcançadas.

6 Conclusão

O desenvolvimento de metodologias e arquiteturas para a segmentação semântica de estruturas renais e suas patologias associadas em exames de tomografia computadorizada constitui uma tarefa de elevada complexidade, especialmente diante da variabilidade morfológica de tumores e cistos, da similaridade entre tecidos adjacentes e do severo desequilíbrio volumétrico entre classes. A segmentação precisa dessas estruturas é de grande relevância para auxiliar o diagnóstico precoce do câncer renal e subsidiar o planejamento terapêutico oncológico. Neste trabalho, foi desenvolvida uma metodologia que propõe a integração de blocos Dual-Scale SE ao codificador da arquitetura U-Net, com convoluções paralelas de *kernels* 3×3 e 7×7 integradas ao módulo Squeeze-and-Excitation (HU; SHEN; SUN, 2018), com o objetivo de construir um modelo capaz de capturar simultaneamente características locais e contextuais em diferentes resoluções, aplicado ao contexto da segmentação renal.

A metodologia desenvolvida consistiu em quatro etapas principais. Primeiro, foi adquirido o conjunto de dados da edição de 2023 do desafio KiTS (HELLER et al., 2023). Em seguida, foram aplicadas técnicas de pré-processamento, incluindo o redimensionamento das imagens de 512×512 para 256×256 *pixels*. Na terceira etapa, foi empregado o modelo de segmentação proposto, denominado Dual-Scale SE UNet, para a identificação das estruturas renais. Esse modelo incorpora blocos Dual-Scale SE ao codificador da U-Net, com convoluções paralelas de *kernels* 3×3 e 7×7 integradas ao módulo Squeeze-and-Excitation (HU; SHEN; SUN, 2018). Por fim, a precisão das segmentações obtidas foi avaliada por meio de métricas quantitativas de desempenho (coeficiente Dice, HD95, MSD e VOE), além de análise qualitativa baseada em estudos de caso e mapas de ativação Grad-CAM (SELVARAJU et al., 2017).

Os resultados experimentais demonstraram desempenho promissor na segmentação de rins e patologias renais na base KiTS23. A segmentação do rim apresentou o maior coeficiente Dice entre as estruturas avaliadas, com valor médio de $93,80 \pm 1,16\%$, acompanhado de HD95 de $2{,}75 \pm 0{,}21$ mm e MSD de $0{,}59 \pm 0{,}04$ mm. Para a segmentação de cistos renais, o modelo alcançou coeficiente Dice de $92,76 \pm 1,53\%$, com os menores valores de HD95 ($0{,}75 \pm 0{,}09$ mm) e MSD ($0{,}31 \pm 0{,}04$ mm) entre todas as estruturas avaliadas, evidenciando elevada precisão na delimitação geométrica dessas lesões. A segmentação de tumores renais, considerada a tarefa de maior complexidade, obteve Dice de $86,27 \pm 0,87\%$, com comportamento consistente ao longo das cinco iterações da validação cruzada. No formato hierárquico do KiTS23, os resultados médios foram de $90,94 \pm 0{,}67\%$ para a categoria Rins e Massas, $89,52 \pm 1,09\%$ para Massas Renais e $86,27 \pm 0,87\%$ para Tumores Renais, com variação de apenas 4,67 pontos percentuais entre as hierarquias, in-

dicando estabilidade do modelo frente a estruturas de diferentes graus de complexidade morfológica.

Retomando os objetivos propostos, o primeiro consistia em investigar a utilização de mecanismos de atenção multiescala em uma arquitetura U-Net para melhorar a extração de características. Esse objetivo foi atendido por meio do desenvolvimento dos blocos Dual-Scale SE, que integram ramos convolucionais paralelos com *kernels* de dimensões distintas e módulos Squeeze-and-Excitation (HU; SHEN; SUN, 2018), permitindo a recalibração adaptativa dos canais em diferentes escalas de representação. O segundo objetivo, relativo à otimização dos hiperparâmetros da arquitetura, foi conduzido por meio de busca Bayesiana com o framework Optuna, resultando na seleção de dois ramos convolucionais paralelos com *kernels* 3×3 e 7×7, *batch* igual a 4 e taxa de aprendizado otimizada para o treinamento do modelo. O terceiro objetivo, de validação experimental em bases de dados públicas reconhecidas na literatura, foi cumprido com a aplicação do modelo à base KiTS23 (HELLER et al., 2023), avaliada por meio de validação cruzada estratificada em cinco dobras. Por fim, o quarto objetivo, de comparação com métodos existentes com base no coeficiente Dice, foi realizado a partir dos resultados obtidos nas categorias hierárquicas do KiTS23, confrontados com abordagens da literatura especializada.

Em relação às questões de pesquisa, Q1 investigou o impacto de mecanismos de atenção na capacidade de distinguir regiões de interesse de tecidos circundantes com características semelhantes. Os resultados indicam que a atenção por canal, implementada via módulos Squeeze-and-Excitation, contribuiu para a diferenciação entre tumores, cistos e rim, conforme evidenciado pelos coeficientes Dice de $86,27\%$ para tumores e $92,76\%$ para cistos, estruturas que apresentam maior dificuldade de delimitação em razão da similaridade com tecidos adjacentes. A análise dos mapas Grad-CAM (SELVARAJU et al., 2017) reforça essa interpretação, revelando regiões de ativação concentradas nas estruturas-alvo. Q2 investigou como a integração de convoluções multiescala com mecanismos de atenção impacta a captura simultânea de características locais e globais em diferentes resoluções. A arquitetura proposta respondeu a essa questão por meio dos ramos paralelos com *kernels* 3×3 e 7×7: o primeiro orientado à captura de detalhes anatômicos finos, como bordas irregulares de tumores e morfologia de cistos, e o segundo voltado à extração de informações contextuais mais abrangentes, contribuindo para a delimitação do parênquima renal como um todo.

A abordagem proposta possui potencial para auxiliar tarefas de apoio ao diagnóstico médico, contribuindo para a identificação automatizada de estruturas renais e de suas principais patologias em exames de tomografia computadorizada. A segmentação automatizada dessas estruturas pode favorecer análises quantitativas mais consistentes e apoiar especialistas no processo de avaliação clínica e planejamento terapêutico.

Como perspectivas futuras, recomenda-se a extensão da arquitetura proposta para

abordagens volumétricas tridimensionais. A estratégia 2.5D adotada neste trabalho incorpora contexto inter-fatia de forma limitada, ao considerar apenas as fatias imediatamente adjacentes como canais de entrada. A adaptação dos blocos Dual-Scale SE para operar diretamente sobre volumes 3D permitiria explorar de maneira mais completa as dependências espaciais entre fatias consecutivas, aspecto particularmente relevante para a detecção de tumores de pequenas dimensões, cuja continuidade volumétrica pode fornecer informações discriminativas não disponíveis na análise fatiada. Adicionalmente, recomenda-se avaliar a capacidade de generalização da metodologia proposta para além do domínio renal, por meio de sua aplicação em outras estruturas anatômicas abdominais, como fígado, baço e pâncreas, e em bases de dados de múltiplos centros, como o KiPA22. Essa avaliação permitiria verificar em que medida a combinação de extração multiescala e recalibração adaptativa de canais, elementos centrais da arquitetura Dual-Scale SE, é transferível a contextos com diferentes características morfológicas, protocolos de aquisição e distribuições de classes, contribuindo para o desenvolvimento de modelos de segmentação mais versáteis e clinicamente aplicáveis.

Dessa forma, os resultados obtidos indicam que a arquitetura Dual-Scale SE UNet constitui uma abordagem promissora para a segmentação automática de estruturas renais em imagens de tomografia computadorizada, contribuindo para o avanço de métodos computacionais aplicados ao suporte ao diagnóstico médico. Parte dos resultados apresentados nesta dissertação foi publicada no artigo *Dual-Scale SE UNet: Uma Rede Neural Convolucional Aprimorada com SE Blocks para Segmentação Renal em Imagens de TC*, aceito no XXV Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2025).

Referências

ALEXANDER, R.; WAITE, S.; BRUNO, M. A.; KRUPINSKI, E. A.; BERLIN, L.; MACKNIK, S.; MARTINEZ-CONDE, S. Mandating limits on workload, duty, and speed in radiology. *Radiology*, Radiological Society of North America, v. 304, n. 2, p. 274–282, 2022. Citado na página 19.

ALEXANDRU, B.-M.; POPESCU, D. Two-stage neural network pipeline for kidney and tumor segmentation. *IEEE Access*, v. 13, p. 146326–146339, 2025. Citado 3 vezes nas páginas 23, 26 e 73.

ALONSO-MONSALVE, S.; WHITEHEAD, L. H.; AURISANO, A.; SANCHEZ, L. E. Submanifold sparse convolutional networks for automated 3d segmentation of kidneys and kidney tumours in computed tomography. *arXiv preprint arXiv:2511.04334*, 2025. Citado 4 vezes nas páginas 24, 26, 73 e 74.

ATKINS, M. B.; CHOUEIRI, T. K. Epidemiology, pathology, and pathogenesis of renal cell carcinoma. *UpToDate Retrieved June*, v. 9, p. 1–3, 2022. Citado na página 30.

BAHDANAU, D.; CHO, K.; BENGIO, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. Disponível em: <https://arxiv.org/abs/1409.0473>. Citado 2 vezes nas páginas 44 e 45.

BANSAL, R. C. Overview and literature survey of artificial neural networks applications to power systems (1992-2004). *Journal of Institution of Engineers (India)–Electrical Engineering*, v. 86, n. 1, p. 282–296, 2006. Citado na página 35.

BEX, A.; GHANEM, Y. A.; ALBIGES, L.; BONN, S.; CAMPI, R.; CAPITANIO, U.; DABESTANI, S.; HORA, M.; KLATTE, T.; KUUSK, T. et al. *European Association of Urology guidelines on renal cell carcinoma: the 2025 update*. [S.l.]: Elsevier, 2025. Citado na página 18.

BRAY, F.; LAVERSANNE, M.; SUNG, H.; FERLAY, J.; SIEGEL, R. L.; SOERJOMATARAM, I.; JEMAL, A. Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 74, n. 3, p. 229–263, 2024. Citado na página 18.

BROWNLEE, J. *Better deep learning: train faster, reduce overfitting, and make better predictions*. 151 Calle de San Francisco: Machine Learning Mastery, 2018. Citado na página 58.

CHEN, C.; ZHANG, R. An ensemble of 2.5 d resunet based models for segmentation of kidney and masses. In: *International Challenge on Kidney and Kidney Tumor Segmentation*. Vancouver, Canada: Springer, 2023. p. 47–53. Citado 3 vezes nas páginas 23, 26 e 73.

CHEN, M.; CHALLITA, U.; SAAD, W.; YIN, C.; DEBBAH, M. Artificial neural networks-based machine learning for wireless networks: A tutorial. *IEEE Communications Surveys & Tutorials*, IEEE, v. 21, n. 4, p. 3039–3071, 2019. Citado na página 34.

CIRILLO, L.; INNOCENTI, S.; BECHERUCCI, F. Global epidemiology of kidney cancer. *Nephrology Dialysis Transplantation*, v. 39, n. 6, p. 920–928, 02 2024. ISSN 0931-0509. Disponível em: <https://doi.org/10.1093/ndt/gfae036>. Citado na página 18.

DAI, J.; QI, H.; XIONG, Y.; LI, Y.; ZHANG, G.; HU, H.; WEI, Y. *Deformable Convolutional Networks*. 2017. Disponível em: <https://arxiv.org/abs/1703.06211>. Citado 3 vezes nas páginas 40, 41 e 42.

DAINA, E.; CORTINOVIS, M.; REMUZZI, G. Kidney diseases. *Immunological Reviews*, Wiley Online Library, v. 313, n. 1, p. 239–261, 2023. Citado na página 30.

DEMIRTAŞ, M. A.; İNNER, A. B.; KAVAK, A. Hybrid 3d mesh reconstruction models of ct images for deep learning based classification of kidney tumors. *Engineering Proceedings*, MDPI, v. 104, n. 1, p. 79, 2025. Citado 4 vezes nas páginas 25, 26, 73 e 74.

DINIZ, J. O.; QUINTANILHA, D. B.; NETO, A. C. S.; SILVA, G. L. da; FERREIRA, J. L.; NETTO, S. M.; ARAUJO, J. D.; CRUZ, L. B. D.; SILVA, T. F.; MARTINS, C. M. da S. et al. Segmentation and quantification of covid-19 infections in ct using pulmonary vessels extraction and deep learning. *Multimedia Tools and Applications*, Springer, v. 80, n. 19, p. 29367–29399, 2021. Citado na página 35.

EELBODE, T.; BERTELS, J.; BERMAN, M.; VANDERMEULEN, D.; MAES, F.; BISSCHOPS, R.; BLASCHKO, M. B. Optimization for medical image segmentation: theory and practice when evaluating with dice score or jaccard index. *IEEE Transactions on Medical Imaging*, IEEE, v. 39, n. 11, p. 3679–3690, 2020. Citado na página 49.

Fairview Health Services. *Leading The Way To Better Healthcare*. 2024. Acesso em: 11 de setembro 2025. Disponível em: <https://www.fairview.org/>. Citado na página 59.

FERNANDES, B. J. T. Redes neurais com extração implícita de características para reconhecimento de padrões visuais. Universidade Federal de Pernambuco, 2013. Citado na página 38.

FILHO, O. M.; NETO, H. V. *Processamento digital de imagens*. [S.l.]: Brasport, 1999. Citado na página 33.

GERVEN, M. van; BOHTE, S. M. Artificial neural networks as models of neural information processing. 2017. Citado na página 34.

GONZALEZ, R. C.; WOODS, R. E. *Processamento digital de imagens. 3. ed.* [S.l.]: São Paulo: Pearson Prentice Hall, 2010. Citado 3 vezes nas páginas 33, 36 e 39.

GROUP, D. F. W. et al. *NIfTI:—neuroimaging informatics technology initiative*. 2023. Disponível em: <https://nifti.nimh.nih.gov/>. Acesso em: 11 de setembro 2025. Citado na página 59.

GULLI, A.; PAL, S. *Deep learning with Keras*. Birmingham, UK: Packt Publishing Ltd, 2017. Citado na página 58.

HANAHAN, D. Hallmarks of cancer: new dimensions. *Cancer discovery*, American Association for Cancer Research, v. 12, n. 1, p. 31–46, 2022. Citado na página 17.

HAYKIN, S. Kalman filters. *Kalman filtering and neural networks*, Wiley Online Library, p. 1–21, 2001. Citado na página 34.

HAYKIN, S. *Redes neurais: princípios e prática*. Porto Alegre, RS, Brasil: Bookman Editora, 2001. Citado na página 34.

HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. San Juan, PR, USA: IEEE, 2016. p. 770–778. Citado na página 43.

HELLER, N.; ISENSEE, F.; MAIER-HEIN, K. H.; HOU, X.; XIE, C.; LI, F.; NAN, Y.; MU, G.; LIN, Z.; HAN, M. et al. *The kits23 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes*. 2023. Disponível em: <https://kits-challenge.org/kits23/>. Acesso em: 11 de setembro 2025. Citado 4 vezes nas páginas 51, 59, 79 e 80.

HENDRYCKS, D.; GIMPEL, K. *Gaussian Error Linear Units (GELUs)*. 2023. Disponível em: <https://arxiv.org/abs/1606.08415>. Citado na página 35.

HERMENA, S.; YOUNG, M. *CT-scan Image Production Procedures*. 2023. StatPearls Publishing. In: StatPearls [Internet]. Treasure Island (FL). Updated 2023 Aug 8. Disponível em: <https://www.ncbi.nlm.nih.gov/books/NBK574548/>. Citado 2 vezes nas páginas 31 e 32.

HU, J.; SHEN, L.; SUN, G. Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2018. p. 7132–7141. Citado 4 vezes nas páginas 45, 52, 79 e 80.

HU, X.; PENG, Y. Gsca-net: A global spatial channel attention network for kidney, tumor and cyst segmentation. In: *International Challenge on Kidney and Kidney Tumor Segmentation*. Vancouver, BC, Canada: Springer, 2023. p. 67–76. Citado 3 vezes nas páginas 22, 26 e 73.

HUTTENLOCHER, D. P.; KLANDERMAN, G. A.; RUCKLIDGE, W. J. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, IEEE, v. 15, n. 9, p. 850–863, 2002. Citado na página 50.

IOFFE, S.; SZEGEDY, C. *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*. 2015. Disponível em: <https://arxiv.org/abs/1502.03167>. Citado na página 42.

JARIWALA, T. A.; MEHTA, P. C.; MEHTA, M. A.; JOSHI, V. C. Kidney and kidney tumour segmentation from 3d ct scan using deeplabv3+. In: IEEE. *2024 IEEE Region 10 Symposium (TENSYMP)*. [S.l.], 2024. p. 1–6. Citado 4 vezes nas páginas 22, 26, 73 e 74.

JIANG, H. Computed tomography: principles, design, artifacts, and recent advances. *Bellingham, Washington USA (Published by SPIE and John Wiley & Sons, Inc.): SPIE*, 2009. Citado na página 32.

KANG, K.; WANG, X. Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*, 2014. Citado na página 38.

KARUNANAYAKE, N.; LU, L.; YANG, H.; GENG, P.; AKIN, O.; FURBERG, H.; SCHWARTZ, L. H.; ZHAO, B. Dual-stage ai model for enhanced ct imaging: Precision segmentation of kidney and tumors. *Tomography*, MDPI, v. 11, n. 1, p. 3, 2025. Citado 3 vezes nas páginas 24, 26 e 73.

KITAW, T. A.; TILAHUN, B. D.; ZEMARIAM, A. B.; GETIE, A.; BIZUAYEHU, M. A.; HAILE, R. N. The financial toxicity of cancer: unveiling global burden and risk factors–a systematic review and meta-analysis. *BMJ Global Health*, BMJ Publishing Group Ltd, v. 10, n. 2, 2025. Citado na página 17.

KOVÁCS, Z. L. *Redes neurais artificiais*. [S.l.]: Editora Livraria da Fisica, 2002. Citado na página 34.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, v. 25, 2012. Citado na página 43.

KULKARNI, N. M.; FUNG, A.; KAMBADAKONE, A. R.; YEH, B. M. Computed tomography techniques, protocols, advancements, and future directions in liver diseases. *Magnetic Resonance Imaging Clinics*, Elsevier, v. 29, n. 3, p. 305–320, 2021. Citado na página 59.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *nature*, Nature Publishing Group UK London, v. 521, n. 7553, p. 436–444, 2015. Citado na página 42.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998. Citado na página 38.

LECUN, Y.; KAVUKCUOGLU, K.; FARABET, C. Convolutional networks and applications in vision. In: IEEE. *Proceedings of 2010 IEEE international symposium on circuits and systems*. Paris, France: IEEE, 2010. p. 253–256. Citado na página 38.

LI, J.; CHENG, J.-h.; SHI, J.-y.; HUANG, F. Brief introduction of back propagation (bp) neural network algorithm and its improvement. In: SPRINGER. *Advances in Computer Science and Information Engineering: Volume 2*. Zhengzhou, China, 2012. p. 553–558. Citado na página 35.

LI, Z.; LIU, F.; YANG, W.; PENG, S.; ZHOU, J. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, IEEE, v. 33, n. 12, p. 6999–7019, 2021. Citado 2 vezes nas páginas 37 e 38.

LI, Z.; PENG, Y.; ZHANG, Z. Two-stage segmentation framework with parallel decoders for the kidney and kidney tumor segmentation. In: *International Challenge on Kidney and Kidney Tumor Segmentation*. Vancouver, BC, Canada: Springer, 2023. p. 83–92. Citado 3 vezes nas páginas 23, 26 e 73.

LIN, T.-Y.; GOYAL, P.; GIRSHICK, R.; HE, K.; DOLLáR, P. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 42, n. 2, p. 318–327, 2020. Citado na página 56.

LIVINGSTONE, D. J. *Artificial neural networks: methods and applications*. United Kingdom, UK: Springer, 2008. v. 458. Citado na página 34.

LONG, J.; SHELHAMER, E.; DARRELL, T. Fully convolutional networks for semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 3431–3440. Citado na página 43.

LUONG, M.-T.; PHAM, H.; MANNING, C. D. *Effective Approaches to Attention-based Neural Machine Translation*. 2015. Disponível em: <https://arxiv.org/abs/1508.04025>. Citado na página 45.

M Health Fairview. *Healthcare founded in academics*. 2024. Acesso em: 11 de setembro 2025. Disponível em: <https://mhealthfairview.org/About-Us>. Citado na página 59.

MAAS, A. L.; HANNUN, A. Y.; NG, A. Y. et al. Rectifier nonlinearities improve neural network acoustic models. In: ATLANTA, GEORGIA, USA. *Proc. icml*. California, USA, 2013. v. 30, n. 1, p. 3. Citado na página 35.

MATOS, C. E. F.; JUNIOR, G. B.; ALMEIDA, J. D. S. d.; PAIVA, A. C. d. Cpp-unet: Combined pyramid pooling modules in the u-net network for kidney, tumor and cyst segmentation. *IEEE Latin America Transactions*, v. 22, n. 8, p. 642–650, 2024. Citado 4 vezes nas páginas 22, 26, 73 e 74.

MCCULLOCH, W. S. A logical calculus of the ideas imminent in nervous activity. *Bull. Math. Biophys.*, v. 5, p. 115–133, 1943. Citado na página 34.

MILLETARI, F.; NAVAB, N.; AHMADI, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: IEEE. *2016 fourth international conference on 3D vision (3DV)*. California, USA: IEEE, 2016. p. 565–571. Citado na página 56.

MU, Y.; NGUYEN, T.; HAWICKHORST, B.; WRIGGERS, W.; SUN, J.; HE, J. The combined focal loss and dice loss function improves the segmentation of beta-sheets in medium-resolution cryo-electron-microscopy density maps. *Bioinformatics Advances*, Oxford University Press, v. 4, n. 1, p. vbae169, 2024. Citado na página 56.

MÜLLER, D.; SOTO-REY, I.; KRAMER, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, Springer, v. 15, n. 1, p. 210, 2022. Citado 2 vezes nas páginas 48 e 49.

MYRONENKO, A.; YANG, D.; HE, Y.; XU, D. Automated 3d segmentation of kidneys and tumors in miccai kits 2023 challenge. *arXiv preprint arXiv:2310.04110*, 2023. DOI:<https://doi.org/10.48550/arXiv.2310.04110>. Citado 3 vezes nas páginas 21, 26 e 73.

NETTER, F. H.; FRANK, M. *Atlas de anatomia humana*. 2. ed. New Jersey: Artmed, 2000. 401–416 p. Citado na página 29.

PACHECO, C. A. R.; PEREIRA, N. S. Deep learning conceitos e utilização nas diversas áreas do conhecimento. *Revista Ada Lovelace*, 2018. Citado 2 vezes nas páginas 34 e 35.

PADALA, S. A.; BARSOUK, A.; THANDRA, K. C.; SAGINALA, K.; MOHAMMED, A.; VAKITI, A.; RAWLA, P.; BARSOUK, A. Epidemiology of renal cell carcinoma. *World journal of oncology*, v. 11, n. 3, p. 79, 2020. Citado 2 vezes nas páginas 18 e 19.

PARKS, E. T. Basic principles of computed tomography. *Oral and Maxillofacial Surgery Clinics of North America*, Elsevier, v. 13, n. 4, p. 547–567, 2001. Citado na página 31.

PETRIK, V.; APOK, V.; BRITTON, J. A.; BELL, B. A.; PAPADOPOULOS, M. C. Godfrey hounsfield and the dawn of computed tomography. *Neurosurgery*, LWW, v. 58, n. 4, p. 780–787, 2006. Citado na página 31.

QIAN, L.; LUO, L.; ZHONG, Y.; ZHONG, D. A hybrid network based on nnu-net and swin transformer for kidney tumor segmentation. In: *International Challenge on Kidney and Kidney Tumor Segmentation*. Vancouver, BC, Canada: Springer, 2023. p. 30–39. Citado 3 vezes nas páginas 21, 26 e 73.

RONNEBERGER, O.; FISCHER, P.; BROX, T. U-net: Convolutional networks for biomedical image segmentation. In: NAVAB, N.; HORNEGGER, J.; WELLS, W. M.; FRANGI, A. F. (Ed.). *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, 2015. p. 234–241. ISBN 978-3-319-24574-4. Citado 5 vezes nas páginas 26, 38, 39, 43 e 52.

SALAHUDDIN, Z.; KUANG, S.; LAMBIN, P.; WOODRUFF, H. C. Leveraging uncertainty estimation for segmentation of kidney, kidney tumor and kidney cysts. In: *International Challenge on Kidney and Kidney Tumor Segmentation*. Vancouver, BC, Canada: Springer, 2023. p. 40–46. Citado 4 vezes nas páginas 25, 26, 73 e 75.

SANTOS, M. d. O.; LIMA, F. C. d. S. d.; MARTINS, L. F. L.; OLIVEIRA, J. F. P.; ALMEIDA, L. M. d.; CANCELA, M. d. C. Estimativa de incidência de câncer no brasil, 2023-2025. *Revista Brasileira de Cancerologia*, v. 69, n. 1, p. e–213700, fev. 2023. Disponível em: <https://rbc.inca.gov.br/index.php/revista/article/view/3700>. Citado na página 17.

SBN - Sociedade Brasileira de Nefrologia. *Compreendendo os rins*. 2023. Disponível em: <https://www.sbn.org.br/o-que-e-nefrologia/compreendendo-os-rins/>. Acesso em: 16 de Setembro 2025. Citado na página 28.

SELVARAJU, R. R.; COGSWELL, M.; DAS, A.; VEDANTAM, R.; PARIKH, D.; BATRA, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE international conference on computer vision*. Venice, Italy: IEEE, 2017. p. 618–626. Citado 3 vezes nas páginas 63, 79 e 80.

SIEGEL, R. L.; GIAQUINTO, A. N.; JEMAL, A. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, Wiley Online Library, v. 74, n. 1, p. 12–49, 2024. Citado na página 17.

SILVA, A. *Algoritmos para diagnóstico assistido de nódulos pulmonares solitários em imagens de tomografia computadorizada. 2004. 140f.* Tese (Doutorado) — Tese (Doutorado)-Departamento de Informática, Pontifícia Universidade . . . , 2004. Citado na página 35.

Sina Ziaee, S.; Maleki, F.; Ovens, K. Rel-UNet: Reliable Tumor Segmentation via Uncertainty Quantification in nnU-Net. *arXiv e-prints*, p. arXiv:2503.09633, mar. 2025. Citado 5 vezes nas páginas 25, 26, 27, 73 e 75.

STOICA, G.; BREABAN, M.; BARBU, V. Analyzing domain shift when using additional data for the miccai kits23 challenge. *arXiv preprint arXiv:2309.02001*, 2023. DOI:<https://doi.org/10.48550/arXiv.2309.02001>. Citado 3 vezes nas páginas 25, 26 e 73.

TAHA, A. A.; HANBURY, A. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, Springer, v. 15, n. 1, p. 29, 2015. Citado 2 vezes nas páginas 49 e 50.

TANASKOVIĆ, I.; IČAGIĆ, S.; ŠOLIĆ, I.; RAKIĆ, B. Adapting yolov8 for kidney tumor segmentation in computed tomography. In: IEEE. *2024 9th International Conference on Smart and Sustainable Technologies (SpliTech)*. [S.l.], 2024. p. 1–5. Citado 4 vezes nas páginas 24, 26, 73 e 74.

TAYE, M. M. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, MDPI, v. 12, n. 5, p. 91, 2023. Citado na página 35.

TING, J.-A.; VIJAYAKUMAR, S.; SCHAAL, S. Encyclopedia of machine learning. In: *Encyclopedia of Machine Learning*. New York, USA: Springer USA, 2010. p. 613–624. Citado na página 48.

UHM, K.-H.; CHO, H.; XU, Z.; LIM, S.; JUNG, S.-W.; HONG, S.-H.; KO, S.-J. Exploring 3d u-net training configurations and post-processing strategies for the miccai 2023 kidney and tumor segmentation challenge. *arXiv preprint arXiv:2312.05528*, 2023. DOI:<https://doi.org/10.48550/arXiv.2312.05528>. Citado 5 vezes nas páginas 25, 26, 27, 73 e 75.

VAUGHAN, C. L.; MAYOSI, B. M. Origins of computed tomography. *The Lancet*, Elsevier, v. 369, n. 9568, p. 1168, 2007. Citado na página 31.

VISPI, A. Recursive learning reinforced by redefining the train and validation volumes of an encoder-decoder segmentation model. In: *International Challenge on Kidney and Kidney Tumor Segmentation*. Vancouver, BC, Canada: Springer, 2023. p. 126–138. Citado 3 vezes nas páginas 23, 26 e 73.

VORA, K.; YAGNIK, S.; SCHOLAR, M. A survey on backpropagation algorithms for feedforward neural networks. *Int. J. Eng. Dev. Res*, v. 1, n. 3, p. 193–197, 2014. Citado na página 35.

WAKAMATSU, K.; ONO, S. Transunet with unified focal loss for class-imbalanced semantic segmentation. *Artificial Life and Robotics*, Springer Nature, v. 29, n. 1, p. 101–106, 2023. Citado na página 56.

WANG, Q.; WU, B.; ZHU, P.; LI, P.; ZUO, W.; HU, Q. *ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks*. 2020. Disponível em: <https://arxiv.org/abs/1910.03151>. Citado na página 47.

WANG, Y.; DAI, Y.; ZHANG, J.; YIN, J. Cascaded nnu-net for kidney and kidney tumor segmentation. In: *International Challenge on Kidney and Kidney Tumor Segmentation*. Vancouver, BC, Canada: Springer, 2023. p. 114–119. Citado 4 vezes nas páginas 24, 26, 73 e 74.

XU, K.; BA, J.; KIROS, R.; CHO, K.; COURVILLE, A.; SALAKHUTDINOV, R.; ZEMEL, R.; BENGIO, Y. *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*. 2016. Disponível em: <https://arxiv.org/abs/1502.03044>. Citado na página 45.

XU, Z. Martin bland (2015): An introduction to medical statistics. *Statistical Papers*, v. 58, n. 3, p. 953–954, set. 2017. Citado na página 49.

ZHANG, S.; LU, J.; ZHAO, H. *Deep Network Approximation: Beyond ReLU to Diverse Activation Functions*. 2024. Disponível em: <https://arxiv.org/abs/2307.06555>. Citado na página 35.

ZHAO, H.; SHI, J.; QI, X.; WANG, X.; JIA, J. Pyramid scene parsing network. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. Honolulu, HI, USA: IEEE, 2017. p. 2881–2890. Citado 3 vezes nas páginas 43, 44 e 52.

ZHAO, X.; WANG, L.; ZHANG, Y.; HAN, X.; DEVECI, M.; PARMAR, M. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, Springer, v. 57, n. 4, p. 99, 2024. Citado na página 38.

ZHOU, X.; LASZIK, Z.; NADASDY, T.; D'AGATI, V.; CLAPP, W.; PARIKH, S.; DIEZ, A.; AYOUB, I.; ALVARADO, A.; ROVIN, B.; LASZIK, Z.; D'AGATI, V.; STOKES, M. B.; MARKOWITZ, G.; D'AGATI, V.; SAXENA, R.; KUPERMAN, M.; RAJORA, N.; SATOSKAR, A.; JEN, K.-Y. Silva's diagnostic renal pathology. 03 2017. Citado 3 vezes nas páginas 28, 29 e 30.