Análise de Duas Abordagens de Treinamento da U-Net para Segmentação de Vértebras em Radiografias: Máscaras Geradas por Pontos Anotados e pelo *Segment Anything Model*

Yuri Junqueira Tobias
yuri.r.tobias@gmail.com
Departamento de Informática
Universidade Federal do Paraná
Curitiba, Paraná, Brazil

Eduardo Todt
todt@inf.ufpr.br
Departamento de Informática
Universidade Federal do Paraná
Curitiba, Paraná, Brazil

ABSTRACT

The prevalence of vertebral fractures has increased, especially those caused by osteoporotic compression, becoming a sensitive public health issue. The diagnosis of such injuries can be performed in different ways, many of which have limitations regarding cost, availability, and time to analyze the results. Among the possible evaluation options is the use of conventional radiographs (X-rays), a mechanism that has been refined lately. This raises the possibility of applying machine learning techniques as a promising alternative to support and accelerate clinical diagnosis. To address this problem, this study proposes the use of deep learning-based computer vision for vertebral segmentation in radiographic images, considering it an important step for some lesion classification mechanisms. In this case, lateral radiographic images of the lumbar spine were used in conjunction with the respective annotations provided by the BUU-LSPINE dataset. From this, a segmentation scheme was structured based on the U-Net model. Due to the lack of precise annotations, the use of two distinct sets of masks for training was proposed: one based on quadrilaterals generated from the annotated vertices and another using masks generated semi-automatically through the *Segment Anything Model*. Next, the performance of the proposed approaches was evaluated using metrics commonly adopted in image segmentation activities: Dice-Sørensen and Jaccard coefficients. The results obtained reinforce the viability of the U-Net model in the application of the activity in question and indicate that the two strategies adopted for mask creation result in similar performances, with average test performances around 0.974 (Dice) and 0.949 (Jaccard).

PALAVRAS-CHAVE

Segmentação, Imagens Médicas, Lesões Vertebrais.

1 INTRODUÇÃO

A coluna vertebral suporta o esqueleto do corpo humano e o sistema nervoso, proporcionando mobilidade e sensibilidade. Dentre suas principais funções, destacam-se: proteção da medula espinhal e dos nervos espinhais, suporte estrutural, flexibilidade e amortecimento ao corpo. Sendo assim, patologias da coluna podem levar a resultados extenuantes na qualidade de vida [1].

Nesse sentido, a osteoporose — doença associada à coluna vertebral — enfraquece e fragiliza os ossos, aumentando o risco de fraturas ósseas mesmo após quedas leves. Tais episódios são recorrentes, especialmente em adultos de idade mais avançada, e sua prevalência crescente tem consolidado a patologia como um problema expressivo de saúde pública. No Brasil, em consonância com a tendência de envelhecimento populacional, estima-se que aproximadamente 50% das mulheres e 20% dos homens com idade superior a 50 anos sofrerão ao menos uma fratura osteoporótica ao longo da vida, com projeções de incidência crescente até o ano de 2050 [2].

Uma das formas de confirmar o diagnóstico de fraturas vertebrais é por meio de radiografias convencionais, processo que tem sido aperfeiçoado e refinado tanto por métodos semiquantitativos quanto quantitativos ao longo das últimas décadas [3]. Adicionalmente, quando não é possível obter uma conclusão apenas com as radiografias, a tomografia computadorizada e a ressonância magnética da coluna complementam o diagnóstico, auxiliando ainda na distinção entre fraturas agudas (recentes) ou crônicas (antigas). Contudo, tais métodos de avaliação costumam ser demorados, onerosos e, por vezes, indisponíveis em ambientes de atenção primária, em que ocorre a avaliação inicial de pacientes com dores na região lombar [4].

Em resposta a essas limitações, a aplicação de técnicas de aprendizado de máquina (*machine learning* — ML), especialmente no cenário de imagens médicas, tem ganhado destaque como uma alternativa promissora para automatizar e acelerar o diagnóstico de fraturas vertebrais. Modelos de ML, em especial redes de aprendizado profundo (*deep learning* — DL), têm demonstrado capacidade não apenas de prover acesso imediato e remoto [4], mas também de oferecer classificações precisas e consistentes de condições como a degeneração de discos intervertebrais ou a presença de fraturas [5]. Tais avanços são especialmente relevantes no contexto de radiografias lombares, em que uma análise automatizada pode reduzir a variabilidade entre avaliadores, minimizar custos e ampliar o acesso ao diagnóstico em ambientes com recursos especializados escassos.

Em vista disso, o uso de radiografias digitais aliado à aplicação de técnicas de aprendizagem profunda, visando à segmentação precisa e à classificação confiável de discos e corpos vertebrais, demonstra potencial considerável para auxiliar no diagnóstico e na tomada de decisões preventivas contra fraturas, sobretudo as decorrentes de compressão osteoporótica. Todavia, o desenvolvimento de um sistema de diagnóstico auxiliado por computador (*Computer-Aided Diagnosis* — CADx) acurado é uma tarefa desafiadora, dadas as variações topológicas e deformidades ósseas, o baixo contraste das imagens e os distintos campos de visão nos exames radiológicos. Nesse contexto, a busca por algoritmos de segmentação vertebral robustos torna-se de grande valia, sobretudo para o aprimoramento das técnicas de classificação.

2 TRABALHOS RELACIONADOS

Nos últimos anos, com os avanços na área da saúde e a busca por maior longevidade e bem-estar, a segmentação eficaz e confiável emergiu como uma etapa preliminar desafiadora para a classificação de vértebras, especialmente no auxílio à prevenção e diagnóstico de fraturas por compressão vertebral osteoporóticas. Nesse sentido, diversos estudos têm explorado novas abordagens ou refinamentos de métodos existentes para melhorar a precisão e a confiabilidade desta etapa. Dessa forma, a presente seção tem por finalidade apresentar as principais descobertas e metodologias recentes no domínio da segmentação vertebral, fornecendo o apoio e a contextualização necessária para o presente trabalho.

A literatura recente demonstra que o uso de Redes Neurais Convolucionais (CNNs) é promissor para automatizar a segmentação e rotulagem das principais estruturas da coluna. O trabalho de Lu et al. [6], em que um dos objetivos é utilizar redes neurais profundas para rotulagem automatizada em nível de coluna, faz uso de uma arquitetura U-Net com um ajuste de curva da coluna para a segmentação e rotulagem dos corpos e discos vertebrais. Essa metodologia é complementada por uma CNN multi-tarefa para a classificação de estenose. A metodologia aplicada inicia-se com a segmentação dos corpos vertebrais utilizando a arquitetura U-Net, com adição de normalização em lote antes de cada ativação da ReLU e uma função sigmoide para gerar o mapa de probabilidade pixel a pixel da segmentação. Nesse caso, as máscaras de *ground-truth* foram feitas gerando caixas delimitadoras a partir de marcações manuais dos quatro cantos de cada corpo vertebral. Quanto aos resultados, o DSC médio para a detecção de vértebras foi de 0,93 com desvio padrão de 0,02.

Outro trabalho que apresenta conclusões relevantes nesse contexto é o Robusto Esquema de Segmentação Vertebral para Diagnóstico Médico [7], que ao levar em consideração patologias e deformidades apresentadas em radiografias e tomografias computadorizadas (TC), apresenta uma nova estrutura de seleção de vértebras capaz de lidar com tais variações de forma eficiente. A estrutura integra uma arquitetura U-Net com um conjunto de níveis paramétricos, anteriormente muito comum em tarefas de segmentação de imagens naturais e médicas, visando aprimorar a extração precisa do formato dos ossos e discos. Nesse caso, em relação ao dados de imagens e detalhes de implementação, foram usados dois conjuntos de dados diferentes para a realização do experimento e o *framework* foi implementado em tensorflow (python). Em comparação com métodos publicados anteriormente ([8]), o método proposto apresentou maior exatidão e precisão.

Adiante, o trabalho proposto por Kim et al. [9] reforça a importância de uma segmentação rápida e precisa para medir a taxa de compressão vertebral (VCR). Os autores propuseram a segmentação de vértebras a partir de imagens 2D usando aprendizado profundo, gerando um algoritmo que mede a VCR com base nos dados obtidos pela segmentação. Nesse caso, foram coletadas radiografias de 339 pacientes com distúrbios da coluna, sendo que todas foram anonimizadas antes da inclusão no estudo. Foi utilizado uma proposta de U-Net residual recorrente multidilatada (MDR2-UNet), que consiste em Bloco Residual Multidilatado (MDRB) e Bloco Residual Recorrente (RRB). Como resultado, o coeficiente de similaridade dos dados foi de 0,929, levando a resultados precisos e produzindo alta confiabilidade em VCRs.

Abordando dificuldades específicas em imagens radiográficas, especialmente a interferência de estruturas sobrepostas, Kim et al. [10] apresentou um método de segmentação hierárquica estruturada. Essa abordagem combina as vantagens de dois métodos de aprendizado profundo: um para a identificação seletiva das vértebras lombares e outro, baseado na arquitetura M-Net, para segmentação fina individual. O desempenho do método foi validado por 160 imagens radiográficas lombares, resultando em um coeficiente de similaridade de Dice médio de 0,916 ± 0,02, indicando que o método proposto alcança identificação precisa de cada vértebra lombar e segmentação fina de vértebras individuais.

Mais recentemente, Altini et al. [11] complementa as descobertas apresentadas pelos trabalhos já mencionados, propondo uma estrutura que combina aprendizado profundo (DL) e metodologias clássicas de aprendizado de máquina. A saber, a proposta compreende duas fases: uma segmentação binária automatizada da coluna de forma integral, explorando a arquitetura V-Net (rede neural convolucional 3D) proposta por Shen et al. [12], e um procedimento semiautomatizado que viabiliza a localização de centroides de vértebras usando algoritmos tradicionais. Para executar os experimentos, foram utilizadas 214 TC da coluna extraídas dos dados dos desafios VerSe'20 - desafio realizado em 2020 cujo objetivo é propor soluções práticas para rotulagem e segmentação vertebral, sendo que são fornecidos conjuntos de dados de TC da coluna vertebral em larga escala, consistindo em 374 exames de 355 pacientes. Quanto as medidas de qualidade consideradas, analogamente às adotadas em trabalhos anteriores, foram consideradas duas classes: medidas baseadas na sobreposição volumétrica, como o DSC - medidas que permitem calcular o grau de similaridade entre a previsão e a verdade fundamental -, e medidas baseadas no conceito de distância de superfície, como a distância máxima simétrica da superfície (MSSD) e a distância média da superfície simétrica (ASSD).

Observa-se que a maioria dos trabalhos apresentados reforça o fato de que a segmentação precisa das vértebras é uma tarefa desafiadora e demorada, frequentemente exigindo múltiplas etapas de processamento dos dados. Além disso, a maioria dos estudos apresentou avanços significativos em relação às soluções previamente propostas. No que diz respeito ao uso de aprendizado de máquina, a maioria dos trabalhos empregou a mesma rede neural convolucional (U-NET) em pelo menos uma de suas etapas, evidenciando seu potencial promissor, embora sejam necessários ajustes para a obtenção de resultados mais expressivos.

3 METODOLOGIA

Nesta seção, serão apresentadas as escolhas metodológicas adotadas para a realização prática do estudo. A divisão deste segue o fluxograma ilustrado pela figura 1: a subseção 3.1 descreve o conjunto de dados; a 3.2 detalha a arquitetura U-Net; a 3.3 aborda as operações de pré-processamento, incluindo a geração de máscaras; e a subseção 3.4 define as métricas utilizadas para a avaliação do modelo.

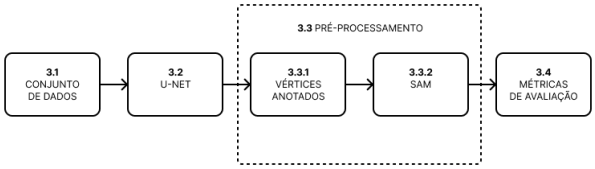


Figura 1: Fluxograma com o esquema de subseções da metodologia.

3.1 Conjunto de Dados

Os experimentos deste trabalho foram realizados utilizando o *dataset* BUU-LSPINE: um conjunto de dados aberto, de origem tailandesa, voltado à detecção de espondilolistese da coluna lombar [13]. O *dataset* é composto por pares de radiografias digitais em diferentes resoluções, sendo que cada par inclui uma incidência anteroposterior (AP) e uma lateral (LA) da coluna lombar. Além disso, todas as imagens são acompanhadas por um arquivo CSV contendo as respectivas anotações realizadas por profissionais da medicina. Tais arquivos fornecem as coordenadas dos quatro cantos (vértices) de cada vértebra identificada, permitindo a geração de máscaras para o treinamento e a avaliação do modelo escolhido. Devido a essas características, o *dataset* foi considerado adequado aos objetivos deste estudo.

Além do conjunto de dados selecionado, foram avaliados outros *datasets*, como o *VerSe* 2019, o *VinDr-SpineXR* e o *UK Biobank*, alguns dos quais indicados pela literatura correlata. Embora apresentassem imagens radiográficas em incidência lateral da coluna lombar com características importantes, como diversidade de amostras e resolução razoável, tais dados mostraram-se limitados para os objetivos desta pesquisa, especialmente devido à ausência de anotações (marcações) em conformidade com o desejado.

3.2 U-Net

A escolha da U-Net como arquitetura de aprendizagem profunda justifica-se, primordialmente, por sua ampla adoção na literatura em tarefas de segmentação de imagens médicas. Adicionalmente, sua eficiência computacional, permitindo um tempo de treinamento viável em uma GPU de médio desempenho (Nvidia RTX A4000), aliada à consistência dos resultados obtidos, reforça a adequação dessa arquitetura aos objetivos deste estudo.

A implementação e a execução do modelo U-Net basearam-se no trabalho disponibilizado publicamente por Vesal [14], sob licença MIT, que permite seu uso, alteração e distribuição. Além da disponibilidade do repositório para a reprodução de estudos científicos, outro fator que contribuiu para a escolha deste é à sua divisão lógica e objetiva, o que facilitou a compreensão e a realização das adaptações necessárias. Tais modificações incluíram a adição da métrica *Intersection over Union* (IoU) para monitoramento durante o treinamento e avaliação nos testes, além de ajustes nos parâmetros e na rotina de leitura dos dados.

Quanto à estrutura do código, destacam-se cinco arquivos principais: trainer.py, predict.py, dataset.py, dilated_unet.py e metric.py. Os dois primeiros contêm as rotinas gerais de treinamento e teste, as quais utilizam a arquitetura definida no penúltimo. O terceiro implementa as funções de partição e leitura de dados, enquanto o último implementa as funções para avaliação do modelo. O código foi desenvolvido em linguagem Python, utilizando as seguintes bibliotecas: Albumentations (v2.0.8), Keras (v3.9.2), MatPlotlib (v3.10.3), NumPy (v2.3.0), OpenCV, Pandas (v2.3.0), sckit-image (v0.25.0), TensorFlow (v2.19.0) e Torch (v2.7.0).

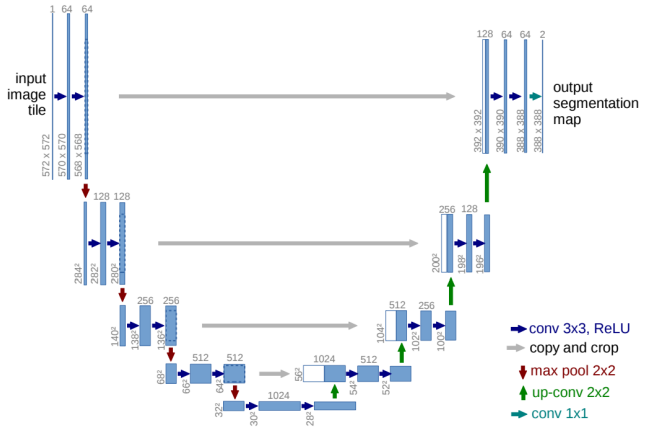


Figura 2: Arquitetura de uma Rede U-Net [15].

O modelo segue a arquitetura proposta por Ronneberger et al. [15], ilustrada na Figura 2. O caminho de contração é composto por quatro blocos, cada um contendo duas camadas de convolução 3x3 seguidas por ativação ReLU e normalização de lote (*batch normalization*), finalizando com uma operação de *max-pooling* 2x2 para redução de amostragem. De forma simétrica, o caminho expansivo apresenta quatro blocos com operações de *upsampling* 2x2 seguidas de duas convoluções. Na camada de saída, uma convolução 1x1 é aplicada para mapear os vetores de característica para o número de classes definido — neste caso, duas. Ao todo, a rede possui 23 camadas convolucionais.

3.3 Pré-processamento

Visto que o conjunto de dados não fornece máscaras detalhadas do contorno vertebral, foi necessário adotar estratégias para a geração desses rótulos (*labels*), fundamentais para o treinamento do modelo. Foram selecionadas duas abordagens principais: (1) a geração de quadriláteros preenchidos a partir dos quatro vértices fornecidos pelo *dataset*; e (2) a geração de máscaras utilizando o *Segment Anything Model* (SAM) — um modelo robusto para segmentação de propósito geral. As subseções a seguir detalham ambas abordagens.

3.3.1 Máscaras geradas a partir dos vértices anotados. Uma vez que as máscaras não estavam disponíveis no formato esperado — fornecendo um contorno aproximado das vértebras — foi necessário gerá-las utilizando as coordenadas presentes no CSV. Para tanto, foi implementado um algoritmo em Python utilizando as bibliotecas NumPy e OpenCV. Esse programa recebe o caminho para um repositório com arquivos CSV e gera um conjunto de imagens de saída representando as máscaras (binárias), feitas a partir da união de quadriláteros formados pelos 4 cantos de cada vértebra. Em

seguida, dado que o algoritmo utilizado para treinamento e teste da U-Net espera a máscara no formato NPY e imagens com tamanhos iguais, tanto as máscaras quanto as imagens foram redimensionadas para 256x256 e as máscaras foram convertidas para arquivos NPY. A figura 3 ilustra esse tipo de máscara, apresentada no formato PNG.

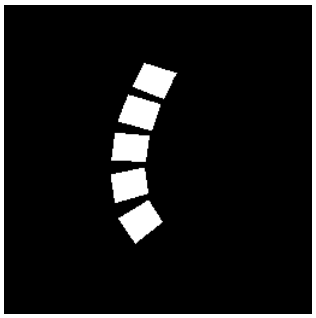


Figura 3: Exemplo de máscara gerada a partir dos vértices anotados.

Por fim, os dados foram divididos em conjuntos de treinamento, validação e teste, em conformidade com as práticas estabelecidas na literatura. Do total disponível, selecionaram-se 60% para o treinamento e 20% para a validação do modelo. Consequentemente, os 20% restantes foram destinados ao conjunto de teste. Em termos quantitativos, essa distribuição resultou em 705 imagens para treinamento, 235 imagens para validação e 235 imagens distintas para o teste final.

3.3.2 Máscaras geradas a partir do SAM. O *Segment Anything Model* é uma arquitetura de segmentação de imagens introduzida recentemente (2023) por Kirillov et al. [16]. O modelo diferencia-se das abordagens tradicionais ao permitir a habilitação da arquitetura por *prompts* — forma de indicar ao modelo qual região da imagem ele deve segmentar —, além de ter sido treinado em uma base de dados massiva, composta por mais de um bilhão de máscaras provenientes de 11 milhões de imagens licenciadas.

Além disso, estudos como o de Huang et al. [17] apontam descobertas consideráveis para o estudo em questão quanto ao uso do SAM, indicando que o modelo (1) demonstra desempenho notável em alguns objetos específicos, apesar de apresentar certa instabilidade em outras situações; (2) tem um desempenho melhor com dicas manuais, especialmente caixas delimitadoras (*bounding boxes*), do que com o modo padrão (segmentar a imagem inteira); e (3) pode ajudar a anotação humana com alta qualidade de rotulagem e menos tempo. Por isso, escolheu-se o mesmo como uma outra possível abordagem para gerar máscaras que viabilizassem o treinamento da U-Net para realização da atividade de segmentação de vértebras em imagens radiográficas.

Ainda, segundo Kirillov et al. [16], a arquitetura do modelo (*SAM*) atende três restrições: um potente codificador de imagem calcula a incorporação de uma imagem, um codificador de *prompts* incorpora *prompts* e, em seguida, ambas as fontes de informação são combinadas em um leve decodificador de máscaras. A figura (4) ilustra a arquitetura do modelo.

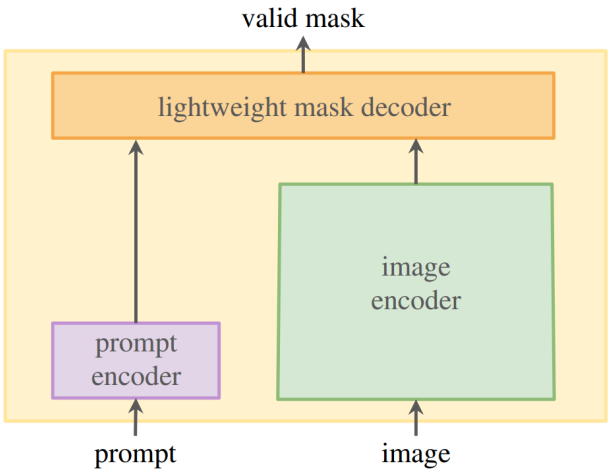


Figura 4: Arquitetura do *Segment Anything Model* [16].

No presente trabalho foi considerada a utilização do modelo com um de seus possíveis *prompts*, seguindo as conclusões de Huang et al. [17], sendo este:

- ***Bounding Boxes***: dado que o *dataset* fornece as coordenadas que indicam os quatro cantos de cada vértebra, é possível obter dois pontos chave — o canto superior esquerdo e o canto inferior direito — que definem uma caixa delimitadora ao redor da região de interesse. Esses dois pontos são obtidos da seguinte forma:
  - Canto superior esquerdo: $(x_{se}, y_{se}) = (\min(x_i), \min(y_i))$
  - Canto inferior direito: $(x_{id}, y_{id}) = (\max(x_i), \max(y_i))$

  Em que x_i e y_i indicam as coordenadas obtidas através dos pontos fornecidos no arquivo csv. A partir desses dois pontos o modelo constrói uma caixa quadrangular em torno da região que deve ser segmentada.

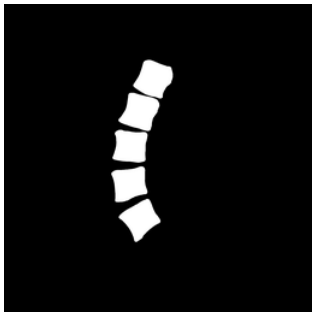


Figura 5: Exemplo de máscara gerada a partir do *Segment Anything Model.*

Isso feito, basta chamar o modelo passando a imagem e as coordenadas da *bounding box* que ele vai segmentar a região indicada, fornecendo assim uma possível segmentação mais suave e ajustada das vértebras. A figura 5 ilustra esse tipo de máscara, apresentada no formato PNG.

3.4 Métricas

Esta subseção descreve as métricas utilizadas para avaliar o desempenho da U-Net. No caso, foram selecionados o **coeficiente de Dice-Sørensen** e o **coeficiente de Jaccard** — comumente referido como *Intersection over Union* (IoU). Ambas são métricas estatísticas que mensuram a semelhança entre dois conjuntos de dados, indicando a proporção de elementos compartilhados, sendo amplamente aplicadas na validação de algoritmos de segmentação de imagens.

- **Coeficiente de Dice-Sørensen**: avalia a similaridade entre dois conjuntos de dados. É definido pelo dobro da área da interseção dos dois conjuntos dividido pela soma das áreas de ambos, conforme fórmula a seguir:

$$\text{Dice}(GT, PR) = \frac{2 \cdot |GT \cap PR|}{|GT \cup PR|}$$

- **Coeficiente de Jaccard (IoU)**: avalia a semelhança entre dois conjuntos de dados. É definido como a razão entre o tamanho da interseção dos conjuntos e o tamanho da sua união, conforme fórmula abaixo:

$$\text{Jaccard}(GT, PR) = \frac{|GT \cap PR|}{|GT \cup PR|}$$

Em ambos os casos, GT (*Ground Truth*) representa a máscara de referência e PR (*Prediction Result*) corresponde ao resultado predito pelo modelo. Os resultados (*scores*) variam entre 0 e um 1, de modo que valores próximos à unidade (1) indicam maior precisão na segmentação.

3.5 Considerações Finais

Nesta seção foram apresentados os materiais e métodos adotados visando uma análise de desempenho da U-Net na atividade de segmentação de vértebras em imagens radiográficas, isso considerando diferentes conjuntos de dados (máscaras) para treinamento. O desempenho final será considerado em termos da pontuação média obtida através das métricas estatísticas descritas acima: coeficiente de Dice-Sørensen e Jaccard.

4 EXPERIMENTOS E RESULTADOS

A seguir serão detalhados os resultados obtidos a partir dos experimentos realizados no estudo em questão. Busca-se avaliar o desempenho da rede neural convolucional U-Net considerando métricas comumente adotadas na tarefa de segmentação semântica de imagens. Para isso, considerando questões de reprodutibilidade do material, a subseção 4.1 descreve o hardware utilizado e demais informações correlatas. As subseções 4.2 e 4.3 apresentam os resultados das predições do modelo de acordo com as respectivas abordagens de treinamento e a subseção **??** traz a análise e conclusão dos resultados.

4.1 Hardware

Os experimentos foram realizados utilizando um servidor Dell, arquitetura x86_64, processador Intel Xeon w3-2435, 64GB de memória RAM, 4TB de disco e, principalmente, GPU Nvidia RTX A4000, com 16GB de memória. Com relação a U-Net, foi feito um *fork* do repositório mencionado na seção 3.2 e, em seguida, um *clone* desse novo repositório no servidor em questão. Quanto ao SAM, também foi feito *clone*, dessa vez do repositório original do modelo, disponível em: https://github.com/facebookresearch/segment-anything/tree/main. Nesse caso, foi feito download e uso do *checkpoint default* ViT-H, variação padrão do modelo [18].

A implementação do código, para gerar as máscaras com as duas abordagens distintas, está presente e disponível para utilização através do seguinte repositório: https://github.com/YuriTobias/TCC.

4.2 Uso de máscaras geradas a partir dos vértices anotados

A seguir serão apresentados os dados obtidos durante as fases de treinamento e teste para o modelo U-Net considerando máscaras geradas a partir dos vértices anotados.

Conforme mencionado na subseção 3.3.1, a fase de treinamento da U-Net contou com 940 imagens; ajustes de parâmetros foram realizados via análise de variação dos resultados. O gráfico da figura 6 ilustra o desempenho do modelo durante uma das 3 execuções desse processo, sendo que no eixo y temos a pontuação média e no eixo x as respectivas épocas — número de iterações em todo o conjunto de dados para treinamento de um modelo de aprendizado de máquina.

Uma vez que a U-Net foi projetada para ser utilizada mesmo com poucos dados de entrada [15], é esperado que sua curva de aprendizagem seja logarítmica. Isso indica, conforme o esperado, que inicialmente o modelo se adapta de modo mais expressivo à realização da tarefa esperada, ao passo que esse aprendizado vai se tornando cada vez mais constante conforme o número de épocas vai aumentando. Tal comportamento pode ser observado pelo que é mostrado na figura 6.

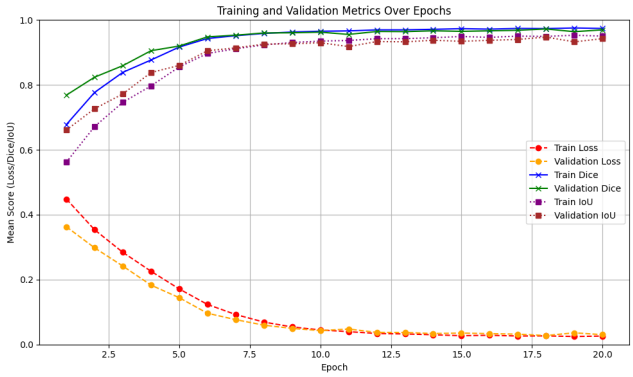


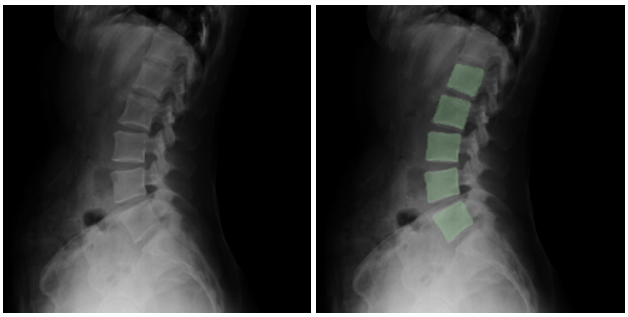
Figura 6: Desempenho do modelo durante a execução 2/3 para a abordagem de vértices anotados

Ainda, a partir dos gráfico, é possível observar que durante o treinamento a U-Net apresentou bons resultados. Em relação a pontuação média das 3 execuções, por exemplo, o modelo obteve coeficientes de Dice de 0,934 e 0,939, e coeficientes de Jaccard de 0,891 e 0,896 para treino e validação, respectivamente. Isso indica que, durante essa primeira fase, o modelo apresentou predições consideravelmente semelhantes e similares as máscaras. Nesse fase, para 20 épocas e 940 imagens, o tempo médio de execução foi de 5 minutos e 51 segundos.

Tabela 1: Pontuação média das métricas para a execução de teste.

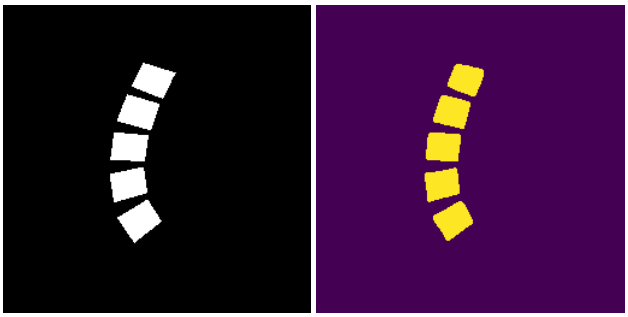
	IoU médio	Dice médio	Tempo de execução (ms)
Dados de teste	0,946	0,972	3408

Com relação as imagens de teste, foram feitas predições para 235 imagens em lotes de 5 imagens, ou seja, 5 predições simultâneas. A cada lote eram calculados os coeficientes médio das predições. A pontuação média final pode ser observada a partir da tabela 1. Quanto aos aspectos visuais, um exemplo de saída pode ser observado na figura 7 e 8.



(a) Imagem original **(b) Imagem com sobreposição**

Figura 7: Exemplo de radiografia usada como entrada para teste dos modelos (a) e radiografia sobreposta com a máscara de predição do modelo (b).



(a) Máscara original **(b) Máscara resultante**

Figura 8: Máscara de *ground-truth* gerada somente a partir dos pontos fornecidos (a) e máscara resultante da predição do modelo (b).

Nesse caso, a figura 7(a) representa a imagem original que foi utilizada como entrada de teste para o modelo. Em seguida, a figura 7(b) apresenta o resultado da segmentação, realizada pelo modelo, a partir da sobreposição da máscara de saída com a imagem original. Ainda, a figura 8(a) apresenta a máscaras de *ground-truth* e a 8(b) a máscara de saída do modelo de forma isolada.

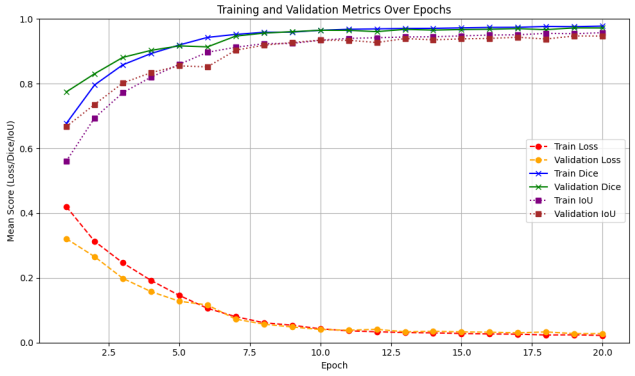


Figura 9: Desempenho do modelo durante a execução 2/3 para a abordagem de uso do SAM

Tabela 2: Pontuação média das métricas para a execução de teste (com SAM).

	IoU médio	Dice médio	Tempo de execução (ms)
Dados de teste	0,953	0,976	3383

4.3 Uso de máscaras resultantes do SAM

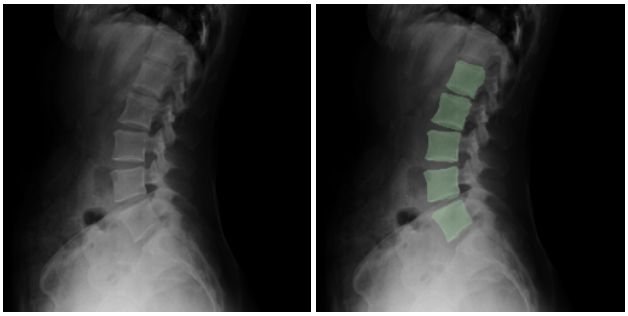
A seguir serão apresentados os dados obtidos durante as fases de treinamento e teste para o modelo U-Net considerando máscaras geradas com uso do modelo *SAM*.

Conforme indicado na seção 4.1, foi utilizado o *checkpoint default* ViT-H, cuja performance geral demonstrou-se superior à do modelo ViT-B [18], conforme apontado por Huang et al. [17]. Neste trabalho, o mesmo programa responsável por invocar o modelo também realiza a leitura das coordenadas anotadas no arquivo csv e, a partir de tais informações, gera os dois pontos necessários para definir a caixa delimitadora, posteriormente interpretada pelo modelo por meio do parâmetro *box*.

Assim como na seção anterior, os gráficos da figura 9 ilustra o desempenho da U-Net durante a fase de treinamento. Nesse caso, é possível observar que o modelo apresentou resultados próximos aos que foram obtidos pela abordagem anterior. Em relação a pontuação média das 3 execuções, o modelo obteve coeficientes de Dice de 0,938 e 0,940, e coeficientes de Jaccard de 0,894 e 0,898 para treino e validação, respectivamente. Também foram consideradas 20 épocas e 940 imagens, resultando em um tempo médio de execução de 5 minutos e 46 segundos.

Uma vez que a escolha das imagens para os lotes de treinamento são feitas de forma aleatória, o processo de executar a etapa de treinamento mais de uma vez busca garantir que o modelo apresente resultados consistentes independentemente da ordem em que as imagens são escolhidas.

Quanto ao conjunto de teste, foram feitas predições para as mesmas 235 imagens que foram testadas pela abordagem anterior, também em lotes de 5 imagens. A tabela 2 a seguir traz a pontuação média final das predições do modelo e as figuras 10 e 11 ilustram os resultados.



(a) Imagem original **(b) Imagem com sobreposição**

Figura 10: Exemplo de radiografia usada como entrada para teste dos modelos (a) e radiografia sobreposta com a máscara de predição do modelo (b).

(a) Máscara original **(b) Máscara resultante**

Figura 11: Máscara de *ground-truth* gerada a partir do SAM (a) e máscara resultante da predição do modelo (b).

A figura 10(a) apresenta a imagem original e a 10(b) o resultado da predição do modelo, após ter sido treinado considerando as máscaras geradas pelo SAM. Por fim, a figura 11(a) apresenta a máscara de *ground-truth* e a 11(b) a máscara de predição de forma isolada.

Considerando a mesma radiografia apresentada na seção anterior. Na figura 11, é possível observar que tanto a máscara de *ground truth* quanto a máscara resultante apresentam algumas curvaturas a mais, o que é interessante quando se deseja considerar ao máximo a estrutura anatômica das vértebras.

5 CONCLUSÃO

O estudo realizado se propôs a avaliar o desempenho de duas abordagens de treinamento considerando a rede neural convolucional U-Net para a segmentação de vértebras em radiografias digitais da coluna lombar: uma com máscaras geradas a partir dos vértices anotados e outra utilizando máscaras geradas pelo SAM (*Segment Anything Model*). Os resultados indicam que ambas abordagens podem ser utilizadas para treinar o modelo e que este é capaz de gerar predições consistentes que se assemelham as máscaras de *ground-truth*, viabilizando sua utilização tanto como ferramenta auxiliar na análise clínica de vértebras lombares quanto como etapa preliminar para sistemas de classificação de fraturas/lesões.

As duas estratégias adotadas para a geração de máscaras de treinamento resultaram em desempenhos similares, com médias de predição próximas de 0,974 (Dice) e 0,949 (IoU). Quanto aos aspectos visuais, a abordagem baseada em quadriláteros produziu segmentações mais regulares, enquanto que a segunda resultou em predições com contornos mais detalhados, possivelmente mais ajustados à anatomia das vértebras, conforme esperado. A abordagem baseada em vértices anotados pode ser considerada mais indicada quando se tem um especialista para realizar a marcação dos vértices, sendo esta uma anotação simplificada (vértice em vez de contorno). Por outro lado, a abordagem baseada no SAM é indicada quando a anotação não é facilmente disponível, sem contar que seu uso como auxiliar na etapa de pré-processamento demonstra o potencial que pode ser explorado ao se utilizar uma combinação de modelos para se atingir algum objetivo específico, nesse caso, de treinar a U-Net com máscaras geradas de forma semiautomática — passando poucas informações e obtendo a máscara completa.

Em possíveis trabalhos futuros, sugere-se a validação das segmentações por especialistas da área, visando a garantia da precisão anatômica das máscaras e predições. Recomenda-se ainda a utilização do modelo como etapa preliminar em pipelines que incluam modelos de classificação, aproveitando as regiões segmentadas como insumo para a detecção e classificação de lesões. Ademais, podem ser considerados estudos incluindo novas técnicas de geração de máscaras ou mesmo novos modelos de segmentação, além de refinamentos nas abordagens e modelo adotados.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, e da Fundação Araucária de Apoio ao Desenvolvimento Científico e Tecnológico do Estado do Paraná (FA).

REFERÊNCIAS

[1] C. DeSai, V. Reddy, and A. Agarwal. Anatomy, back, vertebral column. In *StatPearls [Internet]*. StatPearls Publishing, Treasure Island (FL), 2023. URL https://www.ncbi.nlm.nih.gov/books/NBK525969/. [Updated 2023 Aug 8].

[2] Sociedade Brasileira de Endocrinologia e Metabologia. Dia mundial da osteoporose 2023 - 20 de outubro, 2023. URL https://www.endocrino.org.br/noticias/dia-mundial-da-osteoporose-2023-20-de-outubro/.

[3] Harry K. Genant, Chun Y. Wu, Cornelis van Kuijk, and Michael C. Nevitt. Vertebral fracture assessment using a semiquantitative technique. *Journal of Bone and Mineral Research*, 8(9):1137–1148, September 1993. ISSN 1523-4681. doi: 10.1002/jbmr.5650080915.

[4] Kazuma Murata, Kenji Endo, Takato Aihara, Hidekazu Suzuki, Yasunobu Sawaji, Yuji Matsuoka, Hirosuke Nishimura, Taichiro Takamatsu, Takamitsu Konishi, Asato Maekawa, Hideya Yamauchi, Kei Kanazawa, Hiroo Endo, Hanako Tsuji, Shigeru Inoue, Noritoshi Fukushima, Hiroyuki Kikuchi, Hiroki Sato, and Kengo Yamamoto. Artificial intelligence for the detection of vertebral fractures on plain spinal radiography. *Scientific Reports*, 10(1), November 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-76866-w.

[5] Fabio Galbusera, Gloria Casaroli, and Tito Bassani. Artificial intelligence and machine learning in spine research. *JOR SPINE*, 2(1), March 2019. ISSN 2572-1143. doi: 10.1002/jsp2.1044.

[6] Jen-Tang Lu, Stefano Pedemonte, Bernardo Bizzo, Sean Doyle, Katherine P. Andriole, Mark H. Michalski, R. Gilberto Gonzalez, and Stuart R. Pomerantz. Deepspine: Automated lumbar vertebral segmentation, disc-level designation, and spinal stenosis grading using deep learning, 2018. URL https://arxiv.org/abs/1807.10215.

[7] Faisal Rehman, Syed Irtiza Ali Shah, Naveed Riaz, and Syed Omer Gilani. A robust scheme of vertebrae segmentation for medical diagnosis. *IEEE Access*, 7: 120387–120398, 2019. doi: 10.1109/ACCESS.2019.2936492.

[8] Sewon Kim, Won C. Bae, Koichi Masuda, Christine B. Chung, and Dosik Hwang. Fine-grain segmentation of the intervertebral discs from mr spine images using

deep convolutional neural networks: Bsu-net. *Applied Sciences*, 8(9), 2018. ISSN 2076-3417. doi: 10.3390/app8091656. URL https://www.mdpi.com/2076-3417/8/9/1656.

[9] Dong Hyun Kim, Jin Gyo Jeong, Young Jae Kim, Kwang Gi Kim, and Ji Young Jeon. Automated vertebral segmentation and measurement of vertebral compression ratio based on deep learning in x-ray images. *Journal of Digital Imaging*, 34(4): 853–861, July 2021. ISSN 1618-727X. doi: 10.1007/s10278-021-00471-0.

[10] Kang Cheol Kim, Hyun Cheol Cho, Tae Jun Jang, Jong Mun Choi, and Jin Keun Seo. Automatic detection and segmentation of lumbar vertebrae from x-ray images for compression fracture evaluation. *Computer Methods and Programs in Biomedicine*, 200:105833, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2020.105833. URL https://www.sciencedirect.com/science/article/pii/S0169260720316667.

[11] Nicola Altini, Giuseppe De Giosa, Nicola Fragasso, Claudia Coscia, Elena Sibilano, Berardino Prencipe, Sardar Mehboob Hussain, Antonio Brunetti, Domenico Buongiorno, Andrea Guerriero, Ilaria Sabina Tatò, Gioacchino Brunetti, Vito Triggiani, and Vitoantonio Bevilacqua. Segmentation and identification of vertebrae in ct scans using cnn, k-means clustering and k-nn. *Informatics*, 8(2), 2021. ISSN 2227-9709. doi: 10.3390/informatics8020040. URL https://www.mdpi.com/2227-9709/8/2/40.

[12] Chen Shen, Fausto Milletari, Holger R. Roth, Hirohisa Oda, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. Improving V-Nets for multi-class abdominal organ segmentation. In Elsa D. Angelini and Bennett A. Landman, editors, *Medical Imaging 2019: Image Processing*, volume 10949, page 109490B. International Society for Optics and Photonics, SPIE, 2019. doi: 10.1117/12.2512790. URL https://doi.org/10.1117/12.2512790.

[13] Podchara Klinwichit, Watcharaphong Yookwan, Sornsupha Limchareon, Krisana Chinnasarn, Jun-Su Jang, and Athita Onuean. Buu-lspine: A thai open lumbar spine dataset for spondylolisthesis detection. *Applied Sciences*, 13(15), 2023. ISSN 2076-3417. doi: 10.3390/app13158646. URL https://www.mdpi.com/2076-3417/13/15/8646.

[14] Sulaiman Vesal. Vertebrae segmentation. https://github.com/sulaimanvesal/vertebraeSegementation, 2020. Accessed: 2025-06-12.

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[17] Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Jiongquan Chen, Chaoyu Chen, Sijing Liu, Haozhe Chi, Xindi Hu, Kejuan Yue, Lei Li, Vicente Grau, Deng-Ping Fan, Fajin Dong, and Dong Ni. Segment anything model for medical images? *Medical Image Analysis*, 92: 103061, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2023.103061. URL https://www.sciencedirect.com/science/article/pii/S1361841523003213.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. 2021. URL https://arxiv.org/abs/2010.11929.