Trabalho de Conclusão de Curso

Segmentação Semântica de Imagens Dermatoscópicas de Lesões de Pele Utilizando Aprendizado Profundo.

Guilherme Volney Mota Amaral
gvma@ic.ufal.br

Orientadores:

Prof. Dr. Thiago Cordeiro
Prof. Dr. Elvys Soares

Maceió, junho de 2023

Guilherme Volney Mota Amaral

Segmentação Semântica de Imagens Dermatoscópicas de Lesões de Pele Utilizando Aprendizado Profundo.

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência de Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientadores:

Prof. Dr. Thiago Cordeiro

Prof. Dr. Elvys Soares

Maceió, junho de 2023

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

A485s Amaral, Guilherme Volney Mota.
Segmentação semântica de imagens dermatoscópicas de lesões de pele utilizando aprendizado profundo / Guilherme Volney Mota Amaral. – 2023.
38 f. : il. color.

Orientador: Thiago Cordeiro.
Coorientador: Elvys Soares.
Monografia (Trabalho de Conclusão de Curso em Ciência da Computação) – Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2023.

Bibliografia: f. 36-38.

1. Câncer de pele. 2. Dermatoscopia. 3. Aprendizagem de máquina. 4. Segmentação semântica. I. Título.

CDU: 004.8 : 616.5-002.621

Agradecimentos

A princípio gostaria de agradecer ao professor Marcelo que aceitou me orientar, por ser bastante atencioso, participativo, solícito e companheiro durante todas as fases do desenvolvimento desse trabalho. Além disso, também pelas diversas manifestações de satisfação com o trabalho que eu vim desempenhando.

Gostaria também de agradecer aos demais professores que me deram oportunidades e me serviram de inspiração durante o curso. Em especial aos professores Márcio Ribeiro, Aydano Machado e Thiago Cordeiro, orientadores de projetos de pesquisa e extensão que participei, por terem acreditado em meu potencial. Também, em especial, ao professor Petrúcio Barros, meu tio, que sempre esteve ao meu lado, me aconselhando e me guiando em todos os âmbitos.

Além destes, gostaria de agradecer também ao professor Elvys Soares (IFAL), que além de ser professor e orientador, é um amigo. Sempre me aconselhando, tentando me mostrar formas amplas de enxergar os problemas e sempre me impulsionando para ir além, fazendo com que os nossos trabalhos juntos sempre fossem muito produtivos. Juntamente, gostaria de agradecer ao Fernando, que me permitiu colaborar no meu primeiro projeto pesquisa, o que me proporcionou muitos aprendizados.

Também, de forma especial, gostaria de agradecer aos meus amigos, monitores e colegas da graduação, que sempre fizeram as manhãs, tardes e noites mais alegres, impulsionando-me nos estudos e sempre me fazendo aprender coisas novas. Em especial, aos amigos Gabriel, Rodrigo, Ricardo, Gustavo, Lucas, Jeferson, Paulo, Thiago, Julios, Pedro, Davi, David.

De forma muito especial, gostaria também de agradecer a minha família, por todo suporte, apoio e incentivo durante minha caminhada na universidade e principalmente nessa etapa final. Minha mãe, em especial, me ajudou muito no TCC, pois, por ser da área da dermatoscopia, me ajudou bastante a entender as necessidades dos médicos e viabilidade da criação do produto.

Por último, mas não menos importante gostaria de agradecer minha namorada Amanda, que esteve presente comigo em todos os momentos, me fazendo feliz e me fortalecendo na caminhada da universidade e da vida.

Resumo

O câncer de pele é uma doença caracterizada pela formação de células malignas a partir dos melanócitos, que são células que dão cor à pele. Apesar de ser o tipo mais recorrente entre todos, o câncer de pele é extremamente tratável nos estágios iniciais da doença. Na dermatoscopia, não há ferramenta que auxilie os médicos a fechar o diagnóstico precoce da doença. Assim, este trabalho propõe uma ferramenta que utilize técnicas de aprendizagem de máquina (segmentação semântica) para otimizar o trabalho dos dermatoscopistas. A validação da ferramenta apontou resultado médio de 0,0971 na perda dice para o conjunto de treino e 0,1724 para o conjunto de validação, permitindo uma análise mais precisa da evolução de lesões de pele conforme o tempo. Apesar da ferramenta apresentar uma boa precisão do ponto de vista das técnicas de aprendizagem de máquina, ainda serão necessários ajustes para atender as expectativas de uma aplicação na área da medicina.

Palavras-chave: Aprendizado Profundo; Visão Computacional; Processamento de Imagem; Segmentação Semântica; Desenvolvimento Web; Dermatologia; Dermatoscopia; Redes Neurais; *Vision Transformer*; Câncer de Pele

Abstract

Skin cancer is a disease characterized by the formation of malignant cells from the melanocytes, which are cells that give color to the skin. Despite being the most recurrent type of all, skin cancer is extremely treatable in the early stages of the disease. In dermoscopy, there is no tool to help physicians to make an early diagnosis of the disease. Thus, this work proposes a tool that uses machine learning techniques (semantic segmentation) to optimize the work of dermatoscopists. The validation of the tool showed an average result of 0.0971 in the loss dice for the training set and 0.1724 for the validation set, allowing a more accurate analysis of the evolution of skin lesions according to time. Although the tool has good accuracy from the point of view of machine learning techniques, adjustments will still be needed to meet the expectations of an application in the medical field.

Key-words: Deep Learning; Computer Vision; Image Processing; Semantic Segmentation; Web Development; Dermatology; Dermoscopy; Neural Networks; Vision Transformer; Skin Cancer

Lista de Figuras

Conteúdo

1

Introdução

O câncer de pele é o tipo de câncer que mais acomete a população brasileira e segundo o Instituto Nacional de Câncer (INCA) estima-se que em 2022 foram diagnosticados no Brasil 229.470 casos de câncer de pele (106.560 em homens e 122.910 mulheres), o que é cerca de 33% de todos os casos de câncer no Brasil[1].

Na área da dermatologia, uma das maiores revoluções tecnológicas para o auxílio do diagnóstico do câncer de pele foi a criação do dermatoscópio [1]. Esse dispositivo permitiu a visualização de lesões cutâneas semelhantemente a um microscópio. Com o auxílio de um dermatoscópio, os dermatologistas conseguem ter bons indicativos de lesões cutâneas malignas ou benignas. Segundo o site do governo brasileiro[2] esses são alguns dos fatores mais importantes para identificar um câncer de pele:

- Manchas pruriginosas (que coçam), lesões que descamam ou sangram
- Mudanças severas na lesão: mudança de tamanho, forma ou cor (conhecido como critérios ABCDE)
- Feridas que não cicatrizam em 4 semanas

É possível colher imagens a partir do dermatoscópio e conduzir estudos a partir desse material. A segmentação semântica desempenha um papel fundamental na dermatoscopia, pois o primeiro passo para a análise de lesões é identificar estruturas dermatoscópicas e classificá-las. Com o avanço da complexidade nas mais diversas tarefas de visão computacional as soluções que utilizam aprendizado profundo foram se aperfeiçoando. Dosovitskiy *et al.* [2] propôs um novo modelo, o *Vision Transformer* (ViT), derivado do modelo *Transformer* da área de Processamento de Linguagem Natural, criado por Vaswani *et al.* [3].

[1] https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/pele-melanoma, https://www.gov.br/inca/pt-br/assuntos/cancer/tipos/pele-nao-melanoma

[2] https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/c/cancer-de-pele

O câncer de pele pode ser dividido em dois subtipos: melanocítico e não melanocítico. Os tumores melanocíticos tem origem nos melanócitos (células produtoras de melanina, substância que determina a cor da pele) e é mais frequente em adultos brancos e entre os dois tipos é o mais grave, devido à sua alta possibilidade de provocar metástase (disseminação do câncer para outros órgãos). Já os tumores de pele não melanocíticos são os mais frequentes e de menor mortalidade, porém, se não tratados adequadamente podem deixar mutilações bastante expressivas. Conforme a Sociedade Brasileira de Dermatologia (SBD), o melanoma tem o pior prognóstico dentre os tumores de pele assim como o maior índice de mortalidade. Porém, tem 90% de chance de cura quando há detecção precoce da doença[3]. A dermatoscopia é um método diagnóstico não invasivo que auxilia na avaliação das lesões pigmentadas da pele e é realizada mediante o emprego do dermatoscópio. Ela consiste na utilização de uma lente de aumento associada a uma luz potente que permite ao dermatologista visualizar estruturas dentro da pele.

O exame dermatoscópico é uma técnica de investigação simples e não invasiva que revela detalhes morfológicos adicionais ao exame dermatológico (estruturas ou achados dermatoscópicos) facilitando o diagnóstico de lesões dermatológicas, uma vez que aumenta a acurácia diagnóstica em comparação ao exame dermatológico efetuado sem o uso desse recurso.

Com a intenção de aumentar a transparência da pele, a dermatoscopia requer o uso de um líquido de imersão (normalmente um gel de ultrassom), para suavizar a superfície da pele, evitar as interfaces ar/vidro ar/pele (diferentes graus de refração) e reduzir a reflexão da dispersão da luz utilizada no momento do exame.

Uma vez alcançando-se a transparência da pele, camadas mais profundas podem ser avaliadas, baseando-se nos achados dermatoscópicos encontrados.

1.0.1 Estruturas dermatoscópicas

Por regra, o que primeiro se considera no exame é a definição da origem das lesões cutâneas, melanocítica ou não melanocítica, sendo utilizado o algoritmo dos dois passos criado por Soyer et al [4]. Este é aplicado pelo médico na primeira consulta, no momento em que é utilizado o dermatoscópio para mapear todas as lesões de pele do corpo do paciente.

O primeiro passo define a natureza da lesão, considerando-se melanocítica a lesão que apresentar rede pigmentar, glóbulos agregados, área homogênea azulada, estrias e pseudópodes, dentre outros.

[3] https://www.sbd.org.br/doencas/cancer-da-pele/

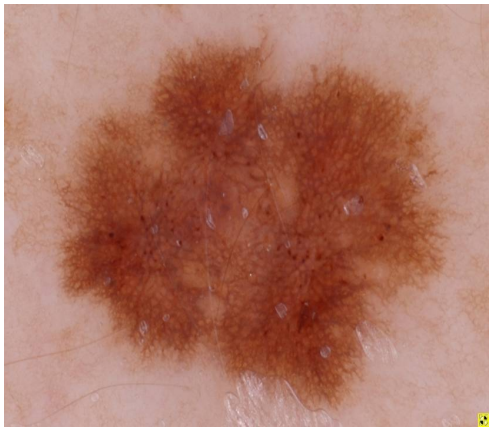

Figura 1.1: Lesão com rede pigmentar

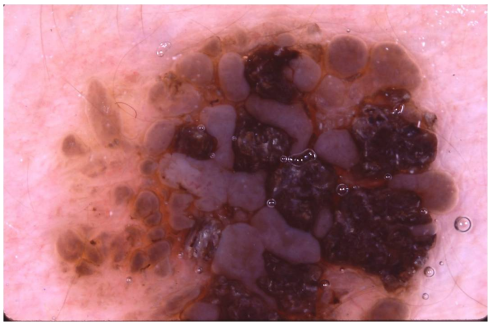

Figura 1.2: Lesão com glóbulos agregados

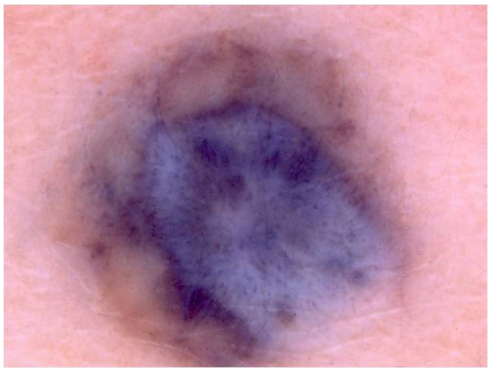

Figura 1.3: Lesão com área homogênea azulada

Figura 1.4: Lesão com estrias e pseudópodes

Portanto, as demais lesões seriam classificadas como não melanocíticas, desde que apresentem estruturas dermatoscópicas clássicas para seus diagnósticos. Exemplos: ninhos ovóides encontrados em um carcinoma basocelular, lagunas vasculares entremeadas por estruturas brancas sugestivas de hemangiomas.

Contudo, algumas lesões apresentam um padrão inespecífico à dermatoscopia, não sendo possível classificá-las em um dos dois grupos acima descritos e com isso, não se podendo afastar a possibilidade diagnóstica de um melanoma. Dessa forma, o paciente é aconselhado a comparecer a uma consulta de retorno, que geralmente acontece num período cerca de 2 meses e meio a 4 meses e meio após a primeira, para acompanhar a lesão e, esperançosamente, ser fechada uma hipótese diagnóstica ou encaminhamento para um patologista.

Considerando a descrição do exame, a identificação dos fatores que determinam o tipo do câncer de pele é prejudicada em virtude do formato e das cores das lesões, como demonstrado nas figuras 1.5 e 1.6 abaixo.

Figura 1.5: Lesão de difícil percepção da sua dimensão

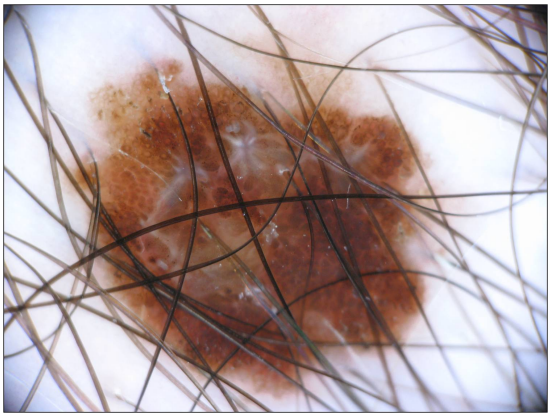

Figura 1.6: Lesão parcialmente coberta por pelos

A análise médica para categorizar o tipo de câncer de pele depende do compreendimento de estruturas dermatoscópicas. Imagens como essas acima dificultam o trabalho dos profissionais na classificação do câncer, uma vez que eles necessitam avaliar a lesão como um todo, o que não é o caso nas figuras acima.

1.1 Motivação

Várias aplicações e ferramentas com foco na medicina vêm sendo implementadas, como demonstrado por Dac-Nhuong Le *et al.* [5], que escreveu um livro listando uma série de tecnologias emergentes para a saúde e medicina que incluem aplicações em realidade virtual, simulações 3D e modelos de inteligência artificial. No entanto, várias dessas inovações terminam demonstrando-se obsoletas ou inválidas, em um contexto que a utilidade real delas termina não sendo aproveitada, pois ainda não foram feitos testes para comprovação da eficácia clínica em muitas delas, conforme apresentado no trabalho de Phillips *et al.* [6].

Ainda assim, uma parte das tecnologias desenvolvidas não foram direcionadas ao público médico, mas ao público leigo. O que pode gerar uma falsa sensação de segurança no diagnóstico ou suspeita médica, mesmo que os próprios programas afirmem não substituir uma conduta profissional com um especialista. Alguns exemplos de ferramentas direcionadas ao público em geral são DermAssist[4], que é um aplicativo de pesquisa de pele guiada do Google Health que visa ajudar o usuário a encontrar informações personalizadas sobre suas preocupações com a pele e DermIA desenvolvido por Shoen et al [7], que é um aplicativo de celular que permite que os usuários captem imagens de lesões de pele suspeitas e sejam avaliadas usando inteligência artificial.

[4] https://health.google/consumers/dermassist/

1.2 Justificativa

A dermatoscopia lança mão de critérios que definem a estrutura de lesões cutâneas, podendo estas serem sugestivas ou não de lesões malignas. O estudo dermatoscópico comparativo, portanto, pode identificar surgimento ou regressão de alguns desses critérios, além de levar em conta o crescimento extensivo da lesão. Sendo o crescimento das lesões lento em muitas das vezes (*Slow Growing Melanoma* (SGM)), o exame dermatoscópico meramente comparativo sem a real definição da variação extensiva de casos como estes pode ser mais respaldado por meio de uma ferramenta que mensure detalhadamente a área lesional.

1.3 Objetivo Geral

O objetivo geral deste trabalho é desenvolver uma ferramenta que visa auxiliar os profissionais na identificação de melanomas precoces, particularmente porque a precocidade diagnóstica determina menor risco de metástase e, consequentemente, uma taxa de cura mais elevada.

1.3.1 Objetivos específicos

Para alcançar o objetivos geral deste trabalho, os seguintes objetivos específicos foram definidos:

1. Estudar técnicas de aprendizagem de máquina para analisar lesões de pele;
2. Escolher um algoritmo para implementar a inteligência artificial;
3. Encontrar uma base de dados para treinar o modelo de aprendizagem de máquina;
4. Realizar testes com o modelo para decidir quais os melhores hiperparâmetros de configuração do algoritmo escolhido;
5. Realizar uma revisão bibliográfica a respeito de trabalhos relacionados;
6. Desenvolver uma ferramenta que implemente soluções para o diagnóstico precoce do câncer de pele;
7. Validar os resultados da ferramenta implementada;
8. Mapear trabalhos futuros;

1.4 Estrutura do trabalho

No capítulo 2, é dado o referencial teórico que fundamenta a motivação desta monografia.

No capítulo 3 é apresentada a metodologia do trabalho desenvolvido.

No capítulo 4 são mostrados os resultados obtidos, juntamento de uma breve discussão acerca deles.

E, por fim, no capítulo 5, são apresentadas as considerações finais, concluindo este trabalho. Além disso, são apresentados trabalhos futuros, sendo que neste último é mostrado um direcionamento e próximos passos para dar continuidade na ferramenta desenvolvida nesta monografia.

2

Fundamentação Teórica

2.1 Câncer de Pele

Segundo o *International Classification of Diseases 11*[1] e a Sociedade Brasileira de Dermatologia[2] (SBD), o câncer de pele é um tumor primário ou metastático envolvendo a pele. Os tumores de pele malignos primários são mais frequentemente carcinomas ou melanomas que surgem de células da camada epiderme da pele, sejam eles:

- Carcinoma Basocelular (CBC): esse é o tipo mais comum dentre os cânceres de pele. Tem baixa letalidade e pode ser curado em caso de detecção precoce;
- Carcinoma Espinocelular (CEC): segundo mais prevalente dentre todos os tipos de câncer que, assim como o CBC, tem expressa relação com a exposição ao sol;
- Melanoma: tipo menos frequente dentre todos os cânceres de pele, originário especificamente dos melanócitos (células produtoras de melanina), de pior prognóstico e o maior índice de letalidade.

Além desses, a pele pode apresentar outros tipos de neoplasias malignas, primariamente cutâneas ou metastáticas como linfomas e sarcomas.

Pessoas com pele clara e que se queimam com facilidade quando se expõem ao sol têm mais risco de desenvolver a doença[2], enquanto as de pele mais escura, embora mais raramente, também podem sofrer dessa mesma doença.

Além disso, a hereditariedade desempenha um papel central no desenvolvimento do melanoma[2]. Por isso, familiares de pacientes diagnosticados com a doença devem se submeter a exames preventivos regularmente. O risco aumenta quando há casos registrados em familiares de primeiro grau.

[1] https://shorturl.at/btJW8
[2] https://rb.gy/xqrhl

2.1.1 Características Clínicas

O câncer de pele pode se assemelhar a pintas, eczemas ou outras lesões benignas. Dessa forma, é muito importante que cada pessoa faça seu autoexame e se familiarize com as pintas que possui. Às pessoas que apresentam histórico familiar de câncer de pele, bem como aquelas portadoras de múltiplas pintas, pode ser executado o mapeamento de suas lesões cutâneas regularmente, para ser possível acompanhar a evolução de cada uma das pintas.

Os principais sintomas da doença são:

- Uma lesão na pele de aparência elevada e brilhante, translúcida, avermelhada, castanha, rósea ou multicolorida, com crosta central e que sangra facilmente;
- Uma pinta preta ou castanha que muda sua cor, textura, torna-se irregular nas bordas e cresce de tamanho;
- Uma mancha ou ferida que não cicatriza, que continua a crescer apresentando coceira, crostas, erosões ou sangramento.

As figuras 2.1, 2.2, 2.3 e 2.4 são exemplos de câncer de pele considerando as características clínicas apresentadas anteriormente.

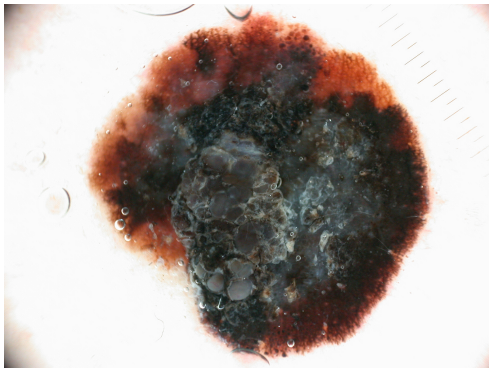

Figura 2.1: Lesão com crosta central

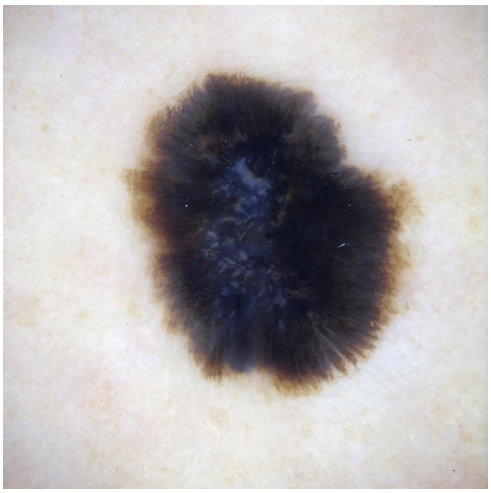

Figura 2.2: Lesão com mudança de cor

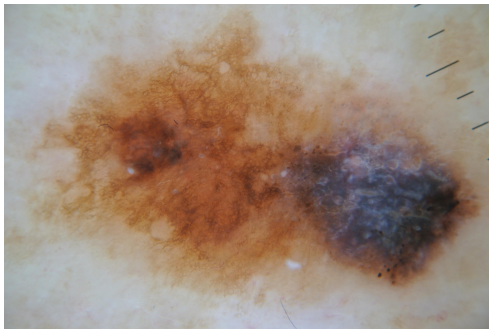

Figura 2.3: Lesão com bordas irregulares

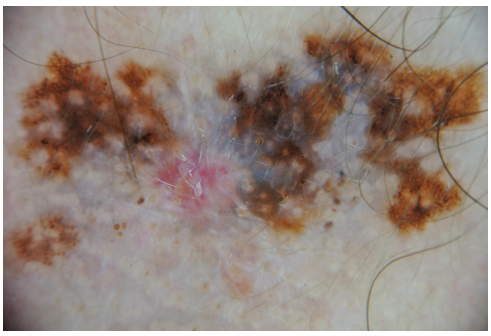

Figura 2.4: Lesão com ferida

Além de todos esses sinais e sintomas, melanomas avançados (espessos) podem apresentar outros, que variam de acordo com a área para onde o câncer avançou, ou seja, para onde o tumor progrediu por meio de metástases. Isso pode incluir nódulos na pele, inchaço nos gânglios linfáticos, falta de ar ou tosse, dores abominais e de cabeça, por exemplo.

2.1.2 Classificação de lesões

A suspeição de uma lesão como cancerígena pode ser dada a partir do exame médico ou autoexame, daí, recorrem-se a alguns critérios para ajudar na classificação dela. Clinicamente, estes são conhecidos como os critérios ABCDE:

- A (assimetria): é caracterizada uma lesão assimétrica aquela que tem o seu centro de interesse divido em quadrantes, sendo estes assimétricos;
- B (bordas): é levado em conta a igualdade entre as bordas da lesão, caso elas sejam desiguais (em sua maioria) maior a probabilidade de classificação em lesão cancerígena;
- C (cor): as cores das lesão podem ser indicativos de lesões cancerígenas na pele;
- D (diâmetro): por norma, lesões com diâmetro maior do que 6mm tem maior chance de se tornarem um câncer;
- E (evolução): com o decorrer do tempo, as lesões se modificam, em tamanho, em cor e em aparência. Mudanças significativas nesses pontos categorizam um câncer de pele.

Os outros critérios são de igual importância no auxílio do diagnóstico do câncer, porém, a evolução dos sinais de pele são muito mais perceptíveis para o profissional após ser aplicada a segmentação semântica nas imagens.

2.2 *Convolutional Neural Network* (CNN)

No aprendizado profundo, CNN (rede neural convolucional ou ConvNet) é uma classe de rede neural artificial (ANN), proposta por Le Cun *et al* [8]. As utilizações das CNNs se mostraram, desde a sua invenção, serem uma das soluções mais eficazes para resolver problemas de classificação, tornando-se uma das mais viáveis alternativas aos métodos tradicionais para classificação dos mais diversos tipos.

Por ser um algoritmo supervisionado, automaticamente pode-se afirmar que um dos pontos mais problemáticos de se trabalhar com uma CNN é o fato de ser necessário existir uma grande quantidade de dados rotulados para a extração de *features*. Para serem extraídas, normalmente são necessários três componentes básicos em uma CNN: camada de convolução, de *pooling* e a rede totalmente conectada. Qualquer arquitetura básica de CNN é composta por esses blocos. A

arquitetura de CNN utilizada nesse trabalho foi baseada em uma proposta por Ronneberger *et al* [9], chamada UNet, que é uma arquitetura de CNN projetada originalmente para a segmentação de imagens biomédicas, mas que também tem sido amplamente utilizada em outras aplicações de segmentação de imagens. A figura 2.5 representa a arquitetura de Ronneberger.

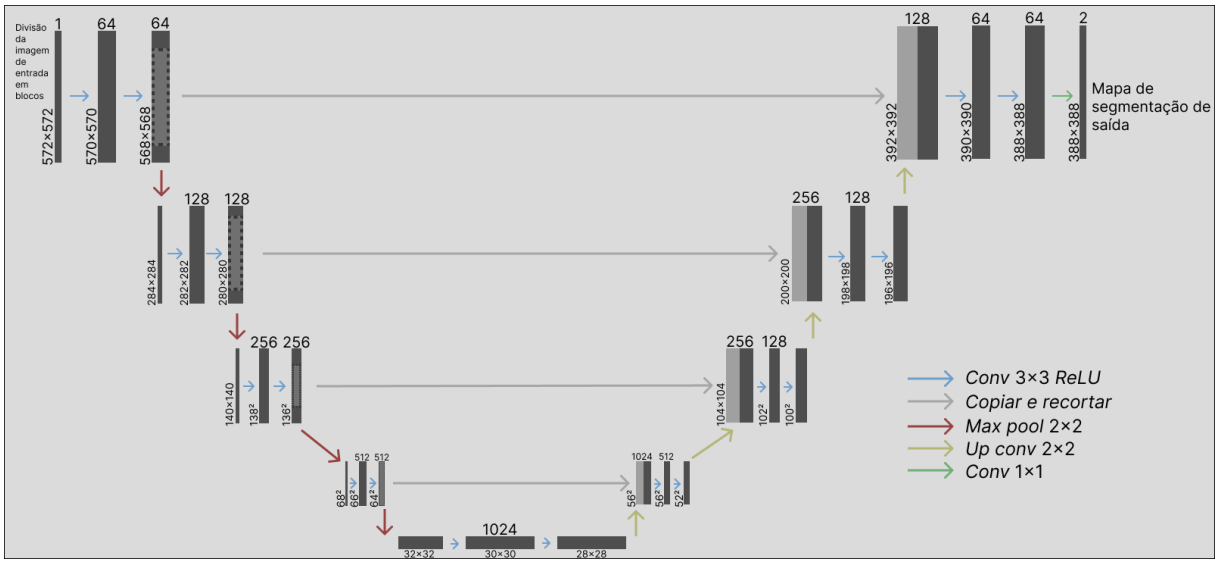

Figura 2.5: Imagem adaptada de Ronneberger da arquitetura UNet (exemplo para 32x32 pixels na resolução mais baixa)

2.2.1 Camada Convolucional

A camada convolucional pode ser considerada a principal de uma CNN e é responsável por extrair as chamadas *features* da entrada. Essa extração se dá por meio de filtros convolucionais de tamanhos reduzidos, de forma que os filtros percorrem todas as dimensões dos dados de entrada realizando uma convolução sobre eles.

A cada processamento de entrada no período de treinamento da rede, os filtros são ajustados em um formato que facilita a identificação de padrões nos dados. No início do treinamento, os filtros são configurados para aprender características mais simples e com o decorrer das épocas, esses filtros, em um processo contínuo, moldam-se para absorver os dados mais profundamente.

2.2.2 Camada de *Pooling*

Essa camada tem como responsabilidade reduzir o tamanho dos dados de entrada. Normalmente, essa camada é ativada após a camada convolucional, com isso, as camadas seguintes a de *pooling* recebem outra representação dos dados de entrada, dessa forma, essa camada permite que a rede aprenda diversas representações de um mesmo dado, evitando o *overfitting*.

A figura 2.6 representa um exemplo de *pooling*, que utiliza o processo de *Max Pooling*.

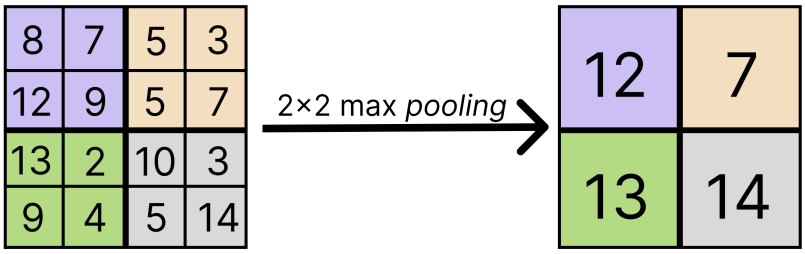

Figura 2.6: Processo de *max pooling*

2.2.3 Camada Totalmente Conectada

A camada totalmente conectada é normalmente formam as últimas camadas da rede. A entrada para a camada totalmente conectada é a saída da camada *pooling* ou convolucional, que é nivelada e, em seguida, alimentada na camada totalmente conectada.

Essa camada se situa no final das redes. Nesse tipo de camada as *features* extraídas são utilizadas para se ter a saída de classificação da rede.

Neste trabalho é utilizada uma arquitetura adaptada da demonstrada em 2.5 como um dos componentes do modelo chamado UNETR, que foi proposta por Hatamizadeh *et al* [10].

2.3 *Vision Transformer*

Arquiteturas baseadas em *self-attention*, em particular os *Transformers* (Vaswani *et al* [3]), tornaram-se o modelo de escolha no processamento de linguagem natural (PLN). A abordagem dominante é pré-treinar em grande corpo de texto e, em seguida, ajustar um conjunto de dados específicos de uma tarefa menor (Bozinovski *et al* [11]).

Na visão computacional, no entanto, as arquiteturas convolucionais permanecem dominantes (Le Cun *et al* [12]) em quantidade. Inspirado nos sucessos das arquiteturas provenientes da área de processamento de linguagem natural, vários projetos tentam mesclar arquiteturas *CNN* com arquiteturas *self-attention* (Ramachandran *et al* [13] e Wang *et al* [14]).

Inspirado no escalamento e no sucesso dos *Transformers* na área de PLN, foi criado um modelo chamado *Vision Transformer* [2] por Dosovitskiy *et al*, que se tornou o novo estado da arte, seguindo a mesma ideia do modelo anteriormente criado [3].

2.3.1 Processamento de dados

O ViT divide uma imagem em *patches* e fornece uma sequência de incorporações lineares como entrada para o modelo. Os *patches* de entrada são tratados da mesma forma que são tratadas os *tokens* (palavras) numa aplicação PLN.

A arquitetura proposta para o ViT é a mais próxima possível da original proposta por Vaswani *et al* [3]. A figura 2.7 representa a arquitetura proposta por Dosovitskiy *et al* [2].

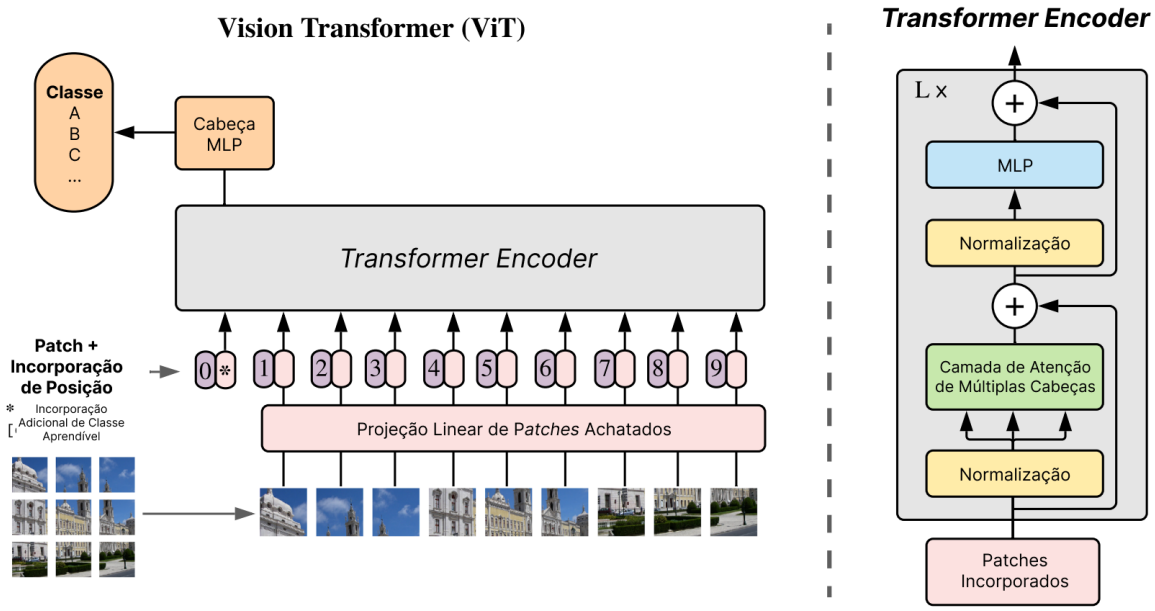

Figura 2.7: Arquitetura proposta do *Vision Transformer* (imagem adaptada do trabalho de Dosovitskiy *et al* [2])

Nessa incorporação linear é adicionada incorporações de posição, que, em seguida alimentarão a sequência resultante de vetores para um *encoder* padrão de *Transformer*. Para realizar a classificação, é utilizada a abordagem padrão de adicionar um *classification token* para esses vetores, que num contexto de imagens, se refere adicionar uma classificação a um pixel.

Após esses dados serem passados para a camada do *encoder*, ele passa pela *Multi Layer Perceptron*, que é uma rede de classificação de duas camadas com GeLU (*Gaussian Error Linear Unit* [15]) no final, usada como uma saída do *Transformer*.

A função GeLU é uma generalização da função ReLU (Rectified Linear Unit), que é amplamente usada em redes neurais. A GeLU é uma função suave e diferenciável, definida pela seguinte equação:

$$\mathrm{GeLU}(x) = \frac{1}{2}x\left(1+\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \tag{2.1}$$

Onde *erf* é a função de erro gaussiano. A GeLU é semelhante à ReLU para valores positivos, mas suaviza a transição para valores negativos, o que pode ajudar a melhorar o desempenho do modelo.

2.3.2 *Patching*

O *patching* é o processo de quebra das imagens em partes menores, conhecida como pixel. Cada *patch* irá representar um pixel que será convertido num vetor de pixels, que por sua vez serão

passados adiante para o *Transformer encoder*.

2.3.3 *Transformer encoder*

O *Transformer encoder* é composto por dois componentes principais: os mecanismos de *self-attention* e uma rede neural *feed-forward*.

- *Self-attention* (*Multi-Head Attention*): aceita *encodings* de entrada do codificador anterior e pesa sua relevância entre si para gerar *encodings* de saída. No contexto de visão computacional, o componente *self-attention* pesa a relevância de um pixel dentro do contexto da imagem inteira, processo que torna mais fácil a tarefa de segmentação semântica;
- *Feed-forward* (MLP): processa cada codificação de saída individualmente. Esses *encodings* de saída são então passados para o próximo codificador como sua entrada.

2.3.4 MLP *Head*

É nessa camada que é gerada a classificação pixel a pixel das imagens. No contexto deste trabalho, para a tarefa de segmentação semântica, é a camada na qual serão classificados os pixels dentro da área de interesse de uma lesão de pele.

Sendo considerado um dos modelos mais promissores na área de visão computacional, por combinar uma arquitetura de rede neural com os princípios de *Transformers* e estar obtendo ótimos resultados em várias aplicações, como em [10] e [16], o ViT foi escolhido para realizar a tarefa de segmentação semântica das lesões de pele.

2.4 Métrica *dice* e função de perda

A métrica *dice* foi escolhida para ser utilizada no modelo, pois é uma das métricas mais utilizadas para qualificar modelos de segmentação semântica e é dada pela seguinte fórmula:

$$\text{Métrica } dice = \frac{2 \times (A \bigcap B)}{A + B} \tag{2.2}$$

Onde A e B são, respectivamente, os dados de saída do ViT e o *ground-truth*. Esse cálculo é feito para termos o valor que será utilizado no cálculo da função de perda:

$$\text{Perda } dice = 1 - \text{Métrica } dice \tag{2.3}$$

O resultado essa equação diz respeito a performance do modelo na tarefa proposta. Quanto mais perto de 0 a perda *dice* quer dizer que o modelo está performando bem.

3

Metodologia

Neste capítulo será detalhado o conjunto de dados usado (3.1) e o pré-processamento aplicado nele (3.1.1), do algoritmo de aprendizado profundo escolhido para a resolução do problema (3.2), as descrições das ferramentas utilizadas no desenvolvimento da aplicação (3.3), e, por fim, a aplicação criada (3.4).

3.1 *Dataset ISIC 2016*

O *dataset* escolhido para treinamento e validação do modelo de aprendizado profundo está disponível no site da companhia *The International Skin Imaging Collaboration* (ISIC[1]). A partir de 2016, o ISIC patrocinou desafios anuais para a comunidade de ciência da computação em associação com as principais conferências de visão computacional. Dentre esses desafios anuais, foi selecionado o desafio do ano de 2016 juntamente de seu *dataset*[2]. O objetivo geral do desafio é desenvolver ferramentas de análise de imagem para permitir o diagnóstico automatizado de melanoma a partir de imagens dermatoscópicas.

A estrutura do *dataset* é composta por 2 pastas principais: *test* e *train* e dentro de cada uma delas os *inputs* e *ground-truth*, onde ficam armazenadas as imagens e as máscaras das lesões de pele, que são usadas para o treinamento e validação das respostas do modelo. Com o objetivo de atingir os melhores resultados no aprendizado do modelo, é necessário pré-processar os dados, nesse caso, aplicando as técnicas de redução de numerosidade e *data augmentation*

3.1.1 Pré-Processamento dos Dados

O pré-processamento de dados é um processo fundamental para o aprendizado profundo, uma vez que pode tanto reduzir o tempo de execução quanto aumentar a qualidade das predições

[1] https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main
[2] https://challenge.isic-archive.com/landing/2016/

dos algoritmos usados, quando aplicado corretamente. Devido às características únicas das imagens de lesões e a dificuldade de obtenção das mesmas, foram utilizados 2 conceitos de pré-processamento de dados: Redução de Numerosidade e *Data Augmentation*.

Redução de Numerosidade

O *dataset* ISIC 2016 possui 2558 imagens, distribuídas entre conjuntos de treino e validação.

- Número inicial de imagens do conjunto de treino: 1800
- Número inicial de imagens do conjunto de validação: 758

No entanto, muitas das imagens contidas no conjunto de treino poderiam representar dificuldades no aprendizado do modelo. Dessa forma, foram retiradas do conjunto de treino 484 imagens, divididas igualitariamente entre imagens de *input* e do *ground-truth*. Essas imagens do *ground-truth* necessitam fazer parte do conjunto de treino pois se trata de um algoritmo supervisionado, que é uma técnica de aprendizado em que os dados de treinamento são rotulados em pares, a entrada e o *ground-truth*. Dessa forma, foram removidas 242 imagens de cada um desses subconjuntos do conjunto de treino. O critério utilizado para remoção das imagens foi baseado em imagens que fugiam do "padrão" de imagens obtidas a partir de um dermatoscópio. As figuras abaixo mostram um exemplo de uma imagem "padrão" e outras imagens que fogem dele.

Figura 3.1: Imagem "padrão" retirada de um Dermatoscópio

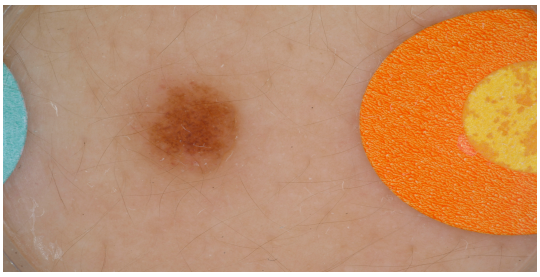

Figura 3.2: Imagem fora do "padrão" retirada de um Dermatoscópio

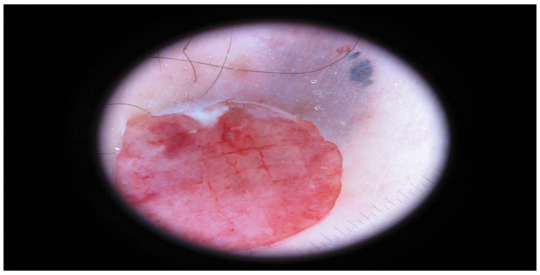

Figura 3.3: Imagem fora do “padrão“ retirada de um Dermatoscópio

As figuras 3.2 e 3.3 são representantes de um conjunto de imagens que podem fazer com que o modelo de aprendizado profundo aprenda de forma errônea, assim dificultando as chances de sucesso na implementação final do modelo. Dessa forma, essa imagem e outras que possuem mesmas características foram retiradas do conjunto final de imagens de treino. A figura 3.2 por conta do adesivo laranja, que é comumente usado como marcação para lesões de risco e a figura 3.3 por conta de suas bordas pretas, pois a maioria dos dermatoscópios produzem uma imagem quadrada, sem as bordas pretas arredondadas. Como podemos ver na figura 3.1, não existem marcações adesivas nem bordas pretas arredondadas, sendo uma ótima candidata a imagem que irá compor o conjunto de treino dos dados de entrada para o modelo.

Após a remoção de todas as imagens, o conjunto de treino no *dataset* permaneceu com 1316 imagens, divididas entre imagens de *input* e *ground-truth*, resultando em 658 imagens em cada um desses subconjuntos. Já o conjunto de validação não sofreu nenhum tipo de remoção, uma vez que o intuito da aplicação é ser de uso para todos os médicos, assim quanto maior a variedade de imagens presentes nesse conjunto, melhor. Dessa maneira, permaneceram as 658 imagens iniciais propostas pelo *dataset* ISIC 2016.

Data Augmentation

Partindo do princípio em que o número de imagens foi reduzido do *dataset* original, o número final é considerado insuficiente para um modelo de aprendizado profundo. Dessa forma, foi utilizado o conceito de Data Augmentation, ao serem utilizadas funções transformadoras em cada uma das imagens, para serem geradas mais amostras.

As técnicas de modificação de imagens para a criação de novos dados usadas foram os *flips* verticais e horizontais. Assim, ao invés de 658 ser o número de imagens de entrada para o conjunto de treino, agora o número de imagens é 658x3, já que possuímos 3 subconjuntos de imagens: as originais, as com *flip* vertical e as com *flip* horizontal, totalizando em um número de 1974 de imagens de treino. As imagens 3.5 e 3.6 representam a imagem 3.4 transformada com os *flips* vertical e horizontal respectivamente.

Figura 3.4: Lesão de pele sem transformações

Figura 3.5: Lesão de pele com *flip* vertical

Figura 3.6: Lesão de pele com *flip* horizontal

Além dessas modificações, as imagens também sofreram uma normalização pixel a pixel. Essa transformação foi aplicada considerando as diversas lentes das câmeras e alterações na iluminação dos diversos ambientes em que as fotos foram tiradas. Ao normalizarmos os pixels de cada uma das imagens, colocamos elas em intervalos de valores que não destoam muito, então mantendo um certo padrão de coloração de cada uma das imagens.

O conjunto de dados de validação não sofreu nenhum tipo de aumento no número de imagens, mas diferentemente das imagens retiradas do conjunto de treino, que foram retiradas, o conjunto de validação permaneceu íntegro, com 379 imagens de entrada.

Considerando os números de imagens dispostas no *dataset*, a proporção de imagens de treino e validação se deu em: 83,9% para o conjunto de teste e 16,10% para o de validação.

3.2 Implementação do Modelo

Para a implementação do modelo, foram utilizados 2 algoritmos principais: Rede Neural e Vision Transformer (ViT). A implementação da arquitetura do modelo se deu por conta da biblioteca MONAI, com o modelo chamado de UNETR[3] que permite uma série de configurações de hiperparâmetros, dentre eles, foram escolhidos os seguintes:

Hiperparâmetro	**Valor**
in_channels	3
out_channels	1
img_size	224
spatial_dims	2
num_heads	8
dropout_rate	0.3

Cada um desses valores escolhidos possui uma razão específica para aplicação dos mesmos na implementação do modelo:

- *in_channels*: para maximizar as *features* absorvidas pelo modelo, optou-se por trabalhar com imagens RGB, que possuem 3 canais;
- *out_channels*: como a tarefa de segmentação semântica aplicado a esse problema resulta numa imagem com somente duas cores, optou-se por ter o número de canais de saída em 1;
- *img_size*: por conta da limitação do poder computacional para execução das etapas de treinamento do modelo, foi escolhido um valor arbitrário de 224 pixels para a imagem;
- *spatial_dims*: é indicado o número de dimensões que as imagens de entrada para o modelo terão, nesse caso, são imagens 2D;
- *num_heads*: assim como para o tamanho da imagem, o *num_heads* também implica em sérios problemas computacionais, dessa maneira, foi escolhido um valor de 8 cabeças;
- *dropout_rate*: para evitar o *overfitting* no treinamento dos dados usa-se uma taxa de inativação dos neurônios, para que o modelo não se vicie em usar somente os melhores neurônios para conseguir os melhores resultados.

Ao combinarmos os algoritmos em uma única arquitetura, como visto por Hatamizadeh et al [10] que criou o UNETR, pode ser feito um modelo que realize segmentação de imagens médicas 3D. Nesse caso, essa mesma arquitetura proposta também serve para imagens 2D, que é a forma em que foi utilizada neste trabalho. O modelo importado não é pré-treinado, então é preciso ter um conjunto de dados específico para treinar o ViT com a arquitetura proposta.

[3] https://docs.monai.io/en/stable/_modules/monai/networks/nets/unetr.html

3.3 Frentes de Desenvolvimento

Para o desenvolvimento da aplicação, foi necessário o desenvolvimento em 2 frentes: o *Frontend*, responsável por disponibilizar uma interface gráfica para utilização da aplicação e o *Backend*, responsável por implementar as regras de negócio, o servidor e a conexão com o banco de dados indexado à aplicação.

Para o desenvolvimento do Frontend foi utilizada a linguagem de programação *JavaScript* juntamente com *Node.js*. A biblioteca utilizada para a montagem da interface gráfica chama-se *React*. Além dessa biblioteca, uma outra permitiu a criação do aplicativo desktop, *Electron*. Essas duas bibliotecas trabalharam em conjunto para permitir a portabilidade da interface web criada para uma aplicação desktop.

Para a segunda frente de desenvolvimento foram necessárias 2 linguagens diferentes com suas respectivas bibliotecas: *JavaScript* com *Node.js* para o desenvolvimento da aplicação desktop e *Python* para a criação do modelo de aprendizado profundo. As bibliotecas de *Python* utilizadas para desenvolvimento foram: *Torchvision*, *MONAI*, *Pillow* e *Numpy*. A construção do ViT foi feita utilizando uma implementação de arquitetura presente na biblioteca *MONAI*, que tem sua implementação baseada em uma outra biblioteca chamada *Pytorch*. O sistema de gerenciamento de banco de dados usado na aplicação foi PostgreSQL. Foram utilizados 2 ambientes de execução para criação do modelo: Google Colaboratory[4] e uma máquina cedida pela USP. Já as bibliotecas para construção da estrutura de suporte de dados da aplicação desktop teve a utilização da biblioteca *Express*.

[4] https://colab.research.google.com/

3.4 Serpens

Considerando a necessidade supracitada de uma ferramenta para auxiliar os médicos o aplicativo Serpens foi criado. As funcionalidades que estarão disponíveis no aplicativo são: login, cadastro de pacientes, cadastro de imagens contendo lesões de pele e histórico dos pacientes contendo a evolução das lesões.

3.4.1 Cenários

Para a exemplificação do aplicativo, foram desenvolvidos protótipos de telas, assim como parte do código-fonte da aplicação. Abaixo, serão descritas cada uma das telas, a importância delas e as funcionalidades implementadas.

Login

O login foi implementado pensando na utilização de computadores compartilhados entre vários médicos, como é o caso de consultórios médicos compartilhados, hospitais, ambulatórios, unidades de pronto atendimento, unidades básicas de saúde, dentre outros. A Figura 3.7 representa o design da tela.

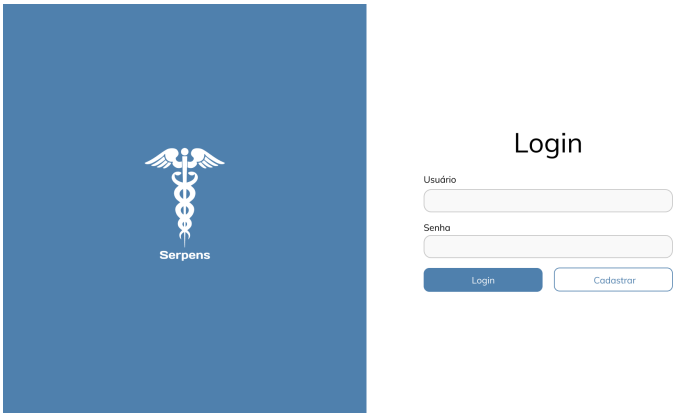

Figura 3.7: Tela de login

Cadastro de Pacientes

Será possível cadastrar pacientes na aplicação, permitindo com que os profissionais acompanhem a evolução das lesões cadastradas no sistema. Essas lesões estão vinculadas diretamente ao paciente, que, por sua vez, tem um histórico único para cada uma das lesões mapeadas em suas consultas dermatológicas. As figuras 3.8 e 3.9 são os layouts das telas de cadastro de pacientes.

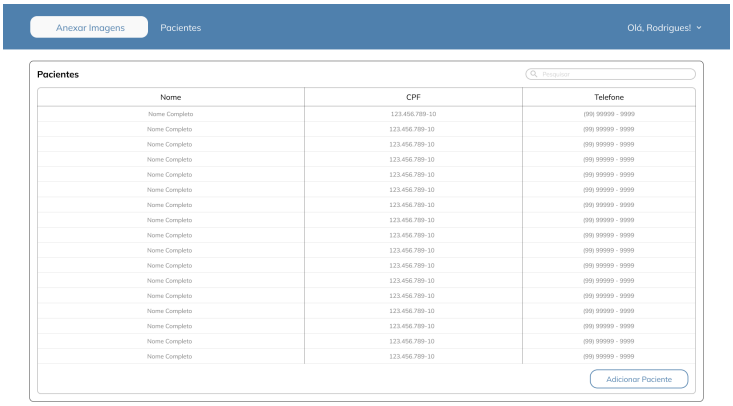

Figura 3.8: Tela de cadastro de pacientes

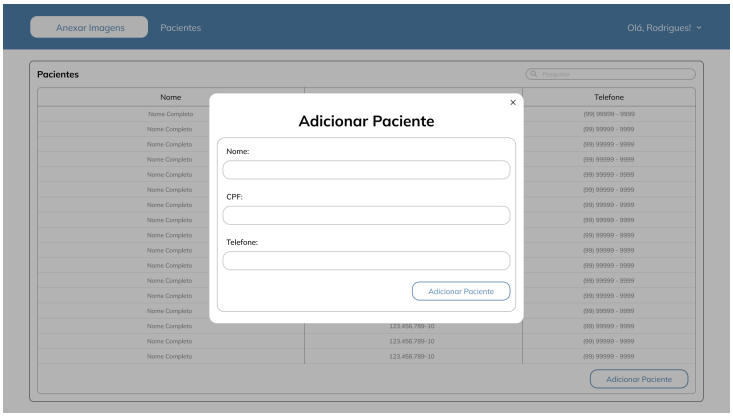

Figura 3.9: Modal de cadastro de pacientes

Cadastro de Imagens

O aplicativo possui uma tela específica para o cadastro de imagens, que serão salvas a fim de gerar um histórico para o acompanhamento das lesões de pele dos pacientes. A figura 3.10 é a tela principal da aplicação, onde o usuário poderá anexar as imagens vinculadas a um paciente já cadastrado no sistema.

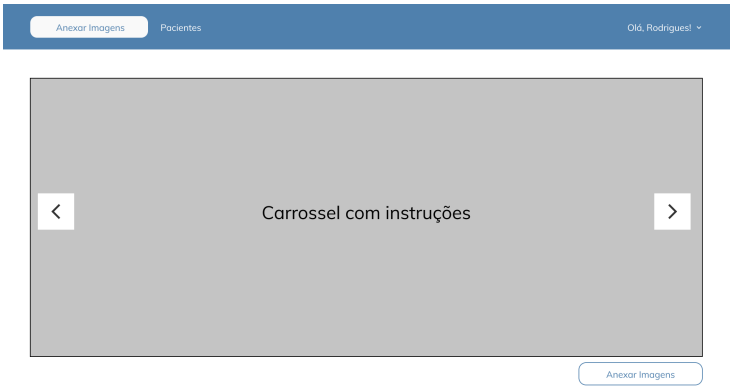

Figura 3.10: Tela de cadastro de imagens

Histórico dos Pacientes

Essa tela contém as informações mais importantes disponibilizadas na aplicação, pois é nela que, de fato, vemos a aplicabilidade do ViT nas imagens de lesões de pele. Nesta tela é possível acompanhar o histórico dos pacientes, vendo as imagens geradas pelo Transformer em cada uma das consultas de rotina para demarcação de lesões de pele, que geralmente são feitas a cada 3 meses. Após as consultas de retorno, o usuário do sistema deverá anexar novas imagens referentes as lesões anteriormente mapeadas, dessa forma, permitindo a construção do histórico de evolução, como especificado no critério E.

Figura 3.11: Tela do histórico do paciente

4

Resultados e Discussões

Neste capítulo serão descritos e discutidos os resultados obtidos a partir do treinamento do ViT para a construção da aplicação.

4.1 Resultados do conjunto de treinamento

Como citado anteriormente, o número de imagens do conjunto de treinamento é de 1316 imagens de lesões cutâneas, considerando o processo de *data augmentation*. O treinamento do ViT durou apenas 50 épocas, devido ao elevado tempo de execução de cada uma delas e das limitações computacionais e os seguintes resultados foram alcançados:

Época	**Perda *dice* Média**
1	0,3374
5	0,1593
10	0,1283
15	0,1203
20	0,1157
25	0,1111
30	0,1057
35	0,1047
40	0,1011
45	0,0986
50	**0,0971**

Tabela 4.1: Perda *dice* média do ViT e suas épocas

Como pode-se observar na tabela 4.1, conforme a passagem das épocas, torna-se notória a aprendizagem do modelo, que conseguiu extrair mais *features* e assim obter melhores resulta-

dos. As últimas épocas já apresentam resultados bastante positivos, considerando a limitação de imagens e do número de épocas do treinamento.

Como pode-se observar na tabela 4.1 os resultados obtidos foram bastante positivos, chegando a uma perda *dice* média de **0,0971** na última época.

4.2 Resultados do conjunto de validação

No conjunto de validação, 379 imagens foram avaliadas, gerando uma perda *dice* média de **0,1724** para as imagens segmentadas semanticamente pelo ViT.

As figuras 4.1 e 4.2 são um comparativo entre, respectivamente, a lesão segmentada semanticamente e a original. Como podemos observar na lesão original, não é possível identificarmos com tamanha precisão a olho nu a dimensão da lesão, pois grande parte dela tem uma cor muito parecida com a pele do paciente. Em contraponto, para a lesão segmentada semanticamente, torna-se mais perceptível a real dimensão da lesão, o que facilita a análise clínica e a tomada de decisão do profissional no diagnóstico precoce das lesões.

Figura 4.1: Lesão segmentada semanticamente

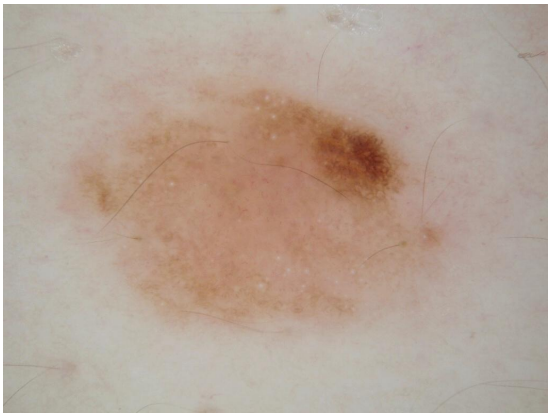

Figura 4.2: Lesão original

Como exposto na seção 3.1 imagens como a 4.3 foram somente retiradas do conjunto de treino. Dessa forma, ao tentar segmentar semanticamente as lesões para imagens desse mesmo tipo, o modelo entregou resultados nem um pouco próximos do ideal, pois não foi exposto a nenhuma amostra semelhante.

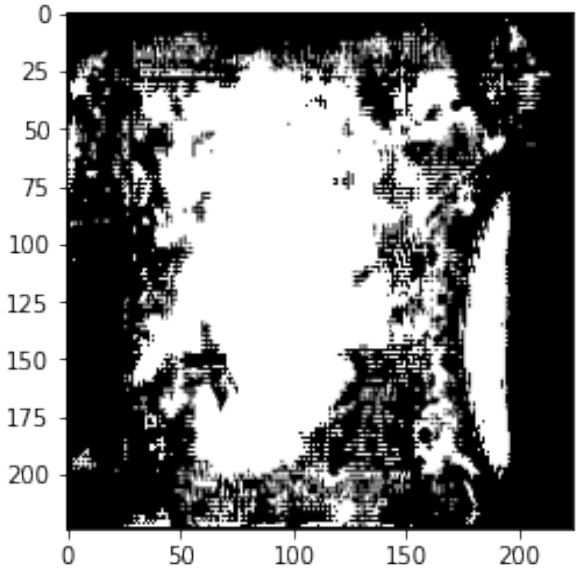

Figura 4.3: Lesão segmentada retirada do conjunto de validação

As figuras 4.5 e 4.7 são uma aplicação em um caso real de uma lesão melanocítica que cresceu no intervalo de 3 meses. As figuras 4.4 e 4.6 representam o resultado da segmentação semântica aplicada pelo ViT.

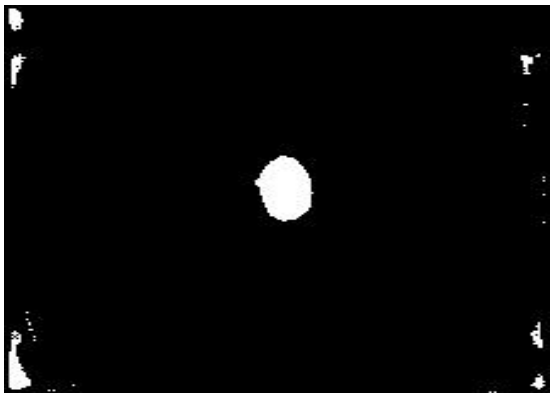

Figura 4.4: Imagem segmentada de nevus melanocítico na primeira consulta

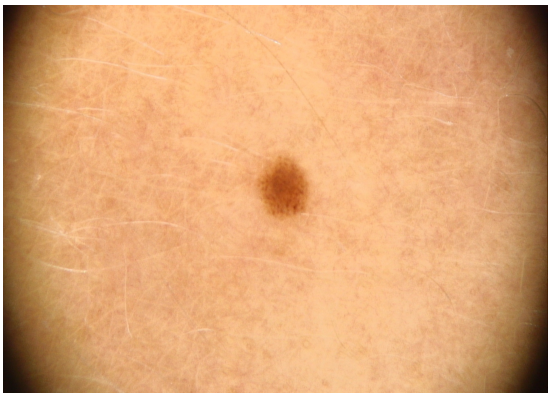

Figura 4.5: Imagem original de nevus melanocítico na primeira consulta

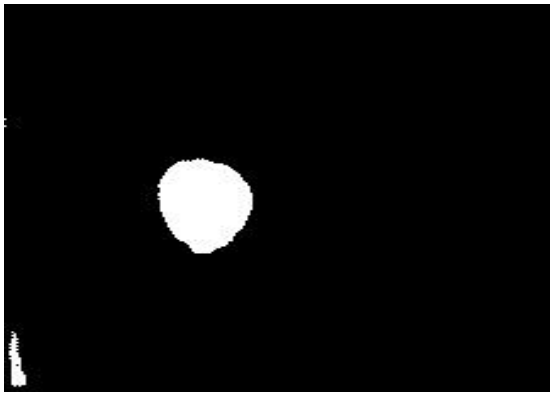

Figura 4.6: Imagem segmentada de nevus melanocítico na consulta de retorno

Figura 4.7: Imagem original de nevus melanocítico na consulta de retorno

Como pode ser observado nas imagens 4.5 e 4.7, são notáveis as semelhanças entre elas e a imagem 3.3 que também foi retirada do conjunto de treino, como detalhado na seção 3.1. Logo, não foi condicionado o aprendizado do modelo para esses tipos de imagem, consequentemente, na borda das imagens 4.4 e 4.6 são apresentados alguns pixels brancos, que para a segmentação semântica representaria uma lesão de pele, o que não é o caso. Isso também pode acontecer devido à configuração dos hiperparâmetros do modelo, como por exemplo o redimensionamento das imagens para uma resolução menor. Dessa forma, considerando a limitação computacional mencionada em 3.2, o modelo pode não ter tido tempo nem informações suficientes para adquirir o conhecimento e segmentar a imagem corretamente. Entretanto, ao analisarmos a imagem descartando as bordas, ainda é possível constatar que a lesão teve um crescimento considerável em 3 meses e a segmentação delas deve facilitar a análise clínica dos médicos.

4.3 Discussão

No conjunto de treinamento a perda *dice* começou bastante alta e isso se dá porque os pesos do modelo são inicializados de forma aleatória. Isso significa que, inicialmente, o modelo não tem informações suficientes para fazer previsões precisas. Portanto, a sobreposição entre as máscaras segmentadas previstas e verdadeiras é baixa, resultando em uma perda *dice* inicialmente baixa. Além disso, nas primeiras épocas, o modelo está aprendendo as características básicas das imagens, como bordas, texturas e cores. Nessa fase inicial, o modelo pode não estar capturando detalhes finos ou informações contextuais mais complexas, levando a uma segmentação menos precisa e, consequentemente, a uma perda *dice* mais baixa.

O resultado da perda *dice* média de **0,0971** poderia ainda ser mais elevado caso o número de épocas fosse maior, no entanto, isso não seria uma garantia, existem algumas ressalvas, como é o caso do *overfitting* que pode ocorrer com um número excessivo de épocas. O *overfitting* ocorre quando o modelo se torna muito especializado nos dados de treinamento específicos e não generaliza bem para novos exemplos e isso se torna mais fácil de acontecer caso o modelo seja submetido a um treinamento que dure muito mais épocas do que o necessário.

Ao observar-se o resultado do conjunto de validação, pode-se afirmar que as imagens são semelhantes o suficiente para comprovar a eficácia do programa e auxiliar no diagnóstico das lesões de pele. O aplicativo deve permitir que os médicos tenham mais facilidade em distinguir imagens de lesões cutâneas melanocíticas e primárias das não melanocíticas e comuns.

5

Conclusão

No início deste projeto, a idealização do resultado era de ser implementada uma ferramenta que pudesse auxiliar dermatologistas numa detecção mais eficiente do câncer de pele.

Através da implementação do aplicativo, ele demonstrou uma eficiência no papel proposto, considerando os mecanismos de inteligência artificial, uma vez que os resultados atingidos foram mais do que satisfatórios para um problema de complexidade alta, que é o caso da tarefa de segmentação semântica para modelos de aprendizado profundo. No entanto, para uma solução que tem atuação na área da medicina é considerado baixo. Podemos comparar os resultados de [17] e [9] que atingiram **0,013** e **0,0797** de perda média nos seus trabalhos, ambos focados em imagens médicas e biomédicas.

Durante o desenvolvimento da aplicação, foi observado a necessidade de incrementação do produto gerado, especialmente em relação aos critérios ABCD e as lesões de pele primárias, que tem um grau de importância igual a implementada. No entanto, essas funcionalidades não estavam no escopo definido para este trabalho, que é de prova de conceito, apesar da sua importância num produto final de *software*. Dito isto, futuramente, serão exploradas oportunidades para implementar funcionalidades relacionadas aos critérios ABCD, a fim de aprimorar ainda mais o desempenho e a usabilidade do sistema proposto.

Como citado em 1.1, um estudo para mensurar a eficácia do programa ainda é necessário. Portanto, visualiza-se a importância de confirmar a aplicabilidade do programa com os usuários de fato, para que não se torne uma aplicação obsoleta. Por fim, deve ser implementado o restante do código do *frontend* e do *backend*, pois a comprovação da eficácia da ferramenta diz respeito somente aos resultados obtidos pelo ViT.

Referências bibliográficas

[1] LEON GOLDMAN. A Simple Portable Skin Microscope for Surface Microscopy. *A.M.A. Archives of Dermatology*, 78(2):246–247, 08 1958. ISSN 0096-5359. **DOI** 10.1001/archderm.1958.01560080106017. URL https://doi.org/10.1001/archderm.1958.01560080106017.

[2] 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. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[4] H P Soyer, G Argenziano, S Chimenti, and V Ruocco. Dermoscopy of pigmented skin lesions. *European journal of dermatology : EJD*, 11(3):270–276; quiz 277, May 2001. URL https://doi.org/10.1046/j.0906-6705.2001.390401.x.

[5] Dac-Nhuong Le, Chung Le, Jolanda Tromp, and Nguyen Nhu. Emerging technologies for health and medicine. 10 2018. **DOI** 10.1002/9781119509875.

[6] Michael Phillips, Jack Greenhalgh, Helen Marsden, and Ioulios Palamaras. Detection of malignant melanoma using artificial intelligence: An observational study of diagnostic accuracy. *Dermatol Pract Concept*, 10(1):e2020011, 2020. ISSN 2160-9381. **DOI** 10.5826/dpc.1001a11. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6936633/.

[7] Ezra Shoen. Dermia: Machine learning to improve skin cancer screening. *Journal of Digital Imaging*, 34(6):1430–1434, 2021. ISSN 1618-727X. **DOI** 10.1007/s10278-020-00395-1. URL https://doi.org/10.1007/s10278-020-00395-1.

[8] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. **DOI** 10.1109/5.726791.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. volume 9351, pages 234–241, 10 2015. ISBN 978-3-319-24573-7. **DOI** 10.1007/978-3-319-24574-4$_2$8.

[10] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021.

[11] Stevo Bozinovski. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44, 09 2020. **DOI** 10.31449/inf.v44i3.2828.

[12] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989. **DOI** 10.1162/neco.1989.1.4.541.

[13] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. **DOI** 10.1109/CVPR.2018.00813.

[14] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 108–126, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58548-8.

[15] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[16] Omid Nejati Manzari, Hamid Ahmadabadi, Hossein Kashiani, Shahriar B. Shokouhi, and Ahmad Ayatollahi. Medvit: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791, 2023. ISSN 0010-4825. **DOI** https://doi.org/10.1016/j.compbiomed.2023.106791. URL https://www.sciencedirect.com/science/article/pii/S0010482523002561.

[17] Hritam Basak, Rohit Kundu, and Ram Sarkar. Mfsnet: A multi focus segmentation network for skin lesion segmentation. *Pattern Recognition*, 128:108673, 2022. ISSN 0031-3203. **DOI** https://doi.org/10.1016/j.patcog.2022.108673. URL https://www.sciencedirect.com/science/article/pii/S0031320322001546.