Aplicação de Patches Dinâmicos em Vision Transformers em Exames de Papanicolau

Vinícius Henrique Giovanini [1]
Alexei Manso Correa Machado [1]

[1]Instituto de Ciências Exatas e Informática
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – Minas Gerais – Brasil

vgiovanini@sga.pucminas.br, alexeimcmachado@gmail.com

Abstract. *This work presents an approach to improve Vision Transformers (ViT) by implementing dynamic patch capture. Experiments were conducted with different types of models, performing fine-tuning and exploring multiple strategies to identify the most relevant areas in patch capture. The proposed modifications were compared to the traditional ViT model, applying these approaches to the Cell Recognition and Inspection Center (CRIC) dataset, composed of Pap smear images. The results demonstrated that the fine-tuning of the ViT-Small model with Grid patch extraction achieved an accuracy of 81%. In contrast, the best dynamic approach obtained 77%, due to the excessive overlap of the patches.*

Resumo. *Este trabalho apresenta uma abordagem para aprimorar os Vision Transformers (ViT) por meio da implementação de captura de patches de maneira dinâmica. Foram conduzidos experimentos com diferentes tipos de modelos, realizando ajuste fino (fine-tuning) e explorando múltiplas estratégias para identificar as áreas mais relevantes na captura de patches. As modificações propostas foram comparadas ao modelo tradicional do ViT, aplicando-se essas abordagens ao conjunto de dados do Centro de Reconhecimento e Inspeção de Células (CRIC), composto por imagens de exames de Papanicolau. Os resultados demonstraram que o fine-tuning do modelo ViT-Small com extração de patches em Grid alcançou uma acurácia de 81%. Em contrapartida, a melhor abordagem dinâmica obteve 77%, devido à excessiva sobreposição dos patches.*

1. Introdução

O câncer do colo do útero representa um desafio de saúde pública no Brasil, com uma incidência anual de 16.000 casos e uma taxa de mortalidade de 4,86 casos por 100.000 mulheres [Barcelos et al. 2017]. O exame de Papanicolau é usado para detectar precocemente o câncer de colo de útero. Inventado em 1928 por George Papanicolau [Michalas 2000], ele é utilizado para detectar alterações nas células do colo do útero que possam indicar lesões pré-cancerígenas. O processo envolve a coleta de células do colo do útero por meio de uma espátula e de uma escova, que depois são transferidas para uma lâmina, e a análise é geralmente realizada por um profissional de saúde que examina as células através de um microscópio, em busca de anomalias morfológica, o que pode tornar o processo demorado.

Visando aprimorar a precisão na identificação de células anômalas e otimizar a eficiência do processo, soluções baseadas em algoritmos de *deep learning* mostram-se altamente eficazes. Esses algoritmos permitem uma análise automatizada em grande escala, extraindo e interpretando padrões que, em uma abordagem tradicional, exigiriam mais tempo para realizar a classificação. Dessa forma, contribuem para agilizar o processo de triagem e diminuir o tempo necessário para a interpretação dos exames, tornando o procedimento mais eficiente.

Este trabalho adota a abordagem dos *Vision Transformers* (ViT) [Dosovitskiy et al. 2021], uma técnica de *deep learning* recente na literatura, que tem demonstrado grande potencial em comparação com as Redes Neurais Convolucionais (CNNs). O ViT adapta a arquitetura dos *Transformers*, originalmente desenvolvidos para Processamento de Linguagem Natural (NLP), para processar imagens, oferecendo uma nova perspectiva para problemas de classificação. O diagrama da Figura 1 ilustra o fluxo básico de entrada e saída do *Vision Transformer*. A imagem de entrada é dividida em *patches* fixos, transformados em *embeddings* lineares. Esses *embeddings* são processados pelo codificador do *Transformer*, resultando na saída correspondente à classe prevista para a imagem.

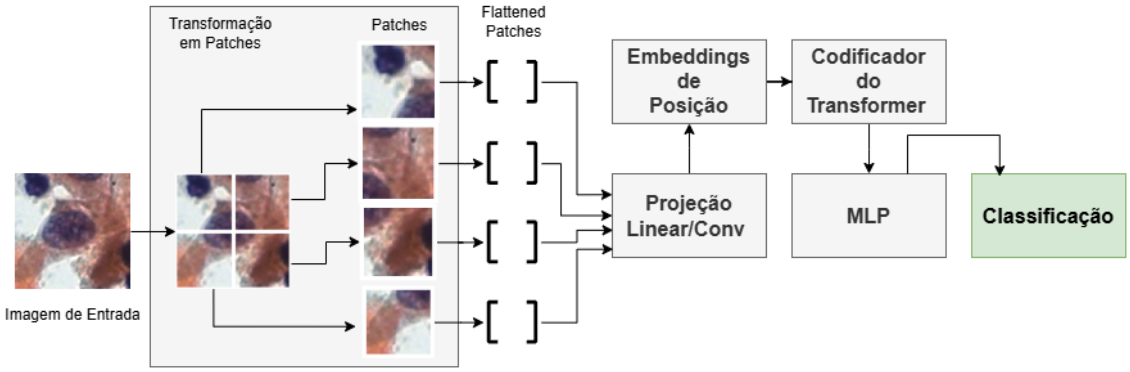


Figura 1. Fluxo de Entrada e Saída do Vison Transformer

A modificação da arquitetura do ViT [Dosovitskiy et al. 2021] é proposta neste trabalho por meio do ajuste fino (*fine-tuning*) de diferentes variações do modelo, a fim de comparar a acurácia de classificação entre células benignas e malignas. A alteração na estrutura do *Transformer* envolve ajustar dinamicamente o local de extração dos *patches* de entrada. Para isso, são apresentadas três abordagens desenvolvidas neste trabalho: extração de *patches* baseada em seleção randômica (SR), randômico aprimorado (RA) e seleção por segmentação (SS). O objetivo é otimizar a seleção dos *patches*, priorizando áreas que contenham mais objetos e menos fundo, com foco nas regiões de interesse da imagem. Dessa forma, é possível avaliar essa modificação em relação aos resultados obtidos a partir do *fine-tuning* do modelo sem alterações com o conjunto de dados de núcleos de células do Papanicolau.

2. Trabalhos Relacionados

Os avanços no uso de técnicas de aprendizado profundo para a automação da análise de imagens citológicas são exemplificados no projeto DeepCeLL [Fang et al. 2022]. Esse projeto tem como objetivo analisar recursos, com ênfase na avaliação de múltiplos *kernels*

de diferentes tamanhos. Para isso, propõe-se um novo modelo de CNN com três variantes para classificar imagens de citologia cervical. O treinamento do DeepCeLL foi realizado utilizando dados do Herlev Dataset, adquirido de um hospital universitário da Dinamarca, contendo 917 imagens individuais, e também utilizando o SIPaKMeD, que consiste em outro *dataset* com cerca de 4049 imagens capturadas de um microscópio óptico e uma câmera.

O projeto do CerviCell-Detector [Kalbhor et al. 2023] é um método automatizado de triagem do exame de Papanicolau que utiliza técnicas de *deep learning*. As redes que adotam a abordagem de detecção de objetos foram refinadas para realizar a classificação de imagens médicas. Esse método foi empregado para executar várias operações nos dados de imagem de entrada, como a aplicação de saturação e a adição de ruído de sal e pimenta. O estudo [Dodge and Karam 2016] descreveu o efeito adverso que imagens ruidosas têm na precisão da classificação, buscando fornecer insights sobre a robustez das RNPs (Redes Neurais Profundas) contra diferentes tipos de distorções visuais, realizando testes com diversas variações, como Blur, Noise, Contraste e Compressão, destacando a importância de tornar o modelo mais robusto e semelhante às amostras da realidade. No treinamento, foi utilizado o CRIC *dataset* [Rezende et al. 2021], e os resultados adquiridos utilizando a técnica de detecção de objeto mostram que pode ajudar na detecção e identificação de imagens médicas.

Um projeto que combina *Vision Transformers* (ViTs) e Redes Neurais Convolucionais (CNNs) para a identificação de tumores na glândula parótida é apresentado por [Dai et al. 2021]. Esse estudo destaca a integração de tecnologias de aprendizado profundo na medicina, aproveitando as capacidades das CNNs para extrair características de baixo nível e dos ViTs para capturar relações globais entre os dados. O objetivo é aprimorar a precisão na classificação de imagens médicas, superando os modelos baseados em CNNs de última geração, com ênfase na análise de imagens multimodais.

Os trabalhos de [Chen et al. 2022] e [Steiner et al. 2022] investigam o desempenho dos *Vision Transformers* (ViTs) em comparação com outras arquiteturas, como as ResNets, abordando o impacto de pré-processamento, aumento de dados e regularização no treinamento. O primeiro trabalho examina o desempenho dos ViTs sem pré-treinamento em larga escala ou uso extensivo de aumento de dados, destacando que, apesar dos problemas de otimização, causados por altas taxas de mínimos locais, os ViTs podem superar ResNets de tamanho de arquitetura semelhante quando treinados do zero ou com grande aumento de dados, especialmente com o uso do otimizador "*Sharpness-Aware Minimizer*"(SAM), que melhora a precisão. Já o segundo trabalho explora como técnicas de aumento e regularização de dados influenciam o desempenho dos ViTs, revelando que a combinação dessas abordagens pode compensar a necessidade de grandes conjuntos de dados. Para conjuntos menores, o estudo sugere que o ajuste fino de modelos pré-treinados em grandes *datasets*, como o *ImageNet-21k*, é mais eficiente do que o treinamento do zero. Ambos os estudos ressaltam a importância do ajuste adequado do treinamento e do uso de estratégias específicas para otimizar o desempenho dos ViTs.

A técnica *Patch Sampling Schedule* (PSS) proposta por [McDanel and Ngoc Huynh 2023] foi desenvolvida para otimizar o tempo de treinamento e aumentar a eficiência dos *Vision Transformers* (ViTs). O objetivo é melhorar a precisão e a taxa de processamento ao ajustar dinamicamente a quantidade e o tamanho

dos *patches* utilizados durante o treinamento, diferenciando entre objeto e fundo das imagens. Para realizar essa distinção, são empregadas duas abordagens, uma baseada na magnitude dos pixels e outra em seleção aleatória. O método descarta *patches* que correspondem ao fundo e ajusta o tamanho dos *patches* de entrada, reduzindo a complexidade computacional sem comprometer o desempenho. Em paralelo, o método *Simple Dynamic Scanning Augmentation* [Kotyan and Vargas 2024] propõe o uso dinâmico de *patches* para aumentar a robustez dos *Vision Transformers* (ViTs), especialmente para resistir a ataques adversariais. A técnica utiliza a extração adaptativa de *patches* em diferentes regiões da imagem, propondo quatro algoritmos para identificar áreas de importância e distinguir entre fundo e objeto, priorizando a extração de *patches* em áreas de interesse. Duas abordagens adotam métodos aleatórios, *Random Patches* (RP) e *Random Tracing* (RT), que extraem *patches* de locais aleatórios da imagem sem buscar especificamente o objeto principal. As outras duas abordagens, não aleatórias, são baseadas em mapas de calor, *Salient Patches* (SP) e *Salient Tracing* (ST), que focam nas áreas mais relevantes da imagem. O estudo testa essas técnicas aplicando diversos ruídos na entrada do ViT para avaliar o impacto sobre a classificação final e a robustez do modelo. A análise revela que as abordagens aleatórias apresentam melhor desempenho em termos de acurácia e melhora principalmente a robustez em comparação com o método tradicional.

3. Vison Transformers (ViTs)

O *Vision Transformer* (ViT) representa uma abordagem de modelagem capaz de competir com as Redes Neurais Convolucionais (CNNs) em várias tarefas de visão computacional. Conforme introduzido por [Dosovitskiy et al. 2021], o ViT divide cada imagem em pequenos *patches*, que são posteriormente convertidos em um vetor de *embeddings*, de maneira análoga ao tratamento de palavras no NLP. Esses *embeddings* são complementados com informações posicionais, garantindo que a localização espacial de cada *patch* na imagem seja preservada. Em seguida, um token de classe é adicionado à sequência de *embeddings*, permitindo que o modelo aprenda uma representação global da imagem. Essa sequência de *embeddings* é processada por um codificador composto por múltiplos blocos, cada um integrando mecanismos de autoatenção e camadas MLP (Perceptron Multicamadas). Dependendo do tamanho do modelo, o codificador pode incluir um ou mais desses blocos, projetados para captar as relações entre os diferentes *patches* da imagem. Ao final, a arquitetura inclui um MLP de classificação, composto por uma ou mais camadas, onde a última camada é responsável por realizar a previsão da classe final.

Os *patches* são componentes que permitem ao modelo processar imagens de forma eficiente. Inicialmente, a imagem da entrada é dividida em *patches* não sobrepostos, de tamanho fixo. Por exemplo, uma imagem de 224x224 pode ser dividida em *patches* de 16x16 pixels, totalizando 196 *patches*, como ilustrado na figura 3b. Cada *patch* é, então, transformado em um vetor de *embedding*, que mapeia a matriz do *patch* para uma representação de menor dimensão, permitindo que os *patches* sejam tratados de maneira individual, semelhantes a tokens em tarefas de processamento de linguagem natural.

4. Materiais e Métodos

4.1. Base de Dados

O conjunto de dados do Centro de Reconhecimento e Inspeção de Células (CRIC) é uma base de dados disponibilizada pela Universidade Federal de Ouro Preto

[Rezende et al. 2021]. Composta por 400 imagens reais, essa base representa diversas lesões celulares analisadas por especialistas. Além das imagens, o CRIC inclui arquivos em formato CSV e JSON, que fornecem informações detalhadas sobre cada núcleo, incluindo a localização do pixel central e sua classificação conforme o sistema Bethesda. Introduzido em 1988, o sistema Bethesda [Rezende et al. 2021] é um método de classificação para exames citológicos cervicais, oferecendo uma terminologia padronizada que auxilia no diagnóstico, tratamento e acompanhamento de lesões pré-cancerosas e cancerosas. Este sistema abrange seis categorias de classificação: *Atypical squamous cells of undetermined signifcance* (ASC-US), *Atypical squamous cells cannot exclude a high-grade lesion* (ASC-H), *High-grade squamous intraepithelial lesion* (HSIL), *Low-grade squamous intraepithelial lesion* (LSIL), *Negative for intraepithelial lesion* (NFIL), e *Squamous cell carcinoma* (SCC).

O *dataset* do CRIC utilizado no estudo que implementou a CNN com *ensemble* por [N. Diniz et al. 2021] representa uma versão alternativa do conjunto de dados atual. A principal diferença entre a versão atual e a do *ensemble* está na quantidade de núcleos de células. O *dataset* utilizado na pesquisa do CRIC possui uma menor quantidade de núcleos, o que facilita o balanceamento das classes. Em contrapartida, o *dataset* atual contém um número significativamente maior de núcleos de células, o que resulta em um conjunto mais desbalanceado e, consequentemente, mais desafiador para as tarefas de classificação.

A partir das informações presentes no arquivo CSV, é possível realizar extrações das células com base nas coordenadas dos núcleos. Utilizando essas coordenadas como referência central, cada célula foi recortada em uma região de 90x90 pixels, resultando em um total de 11.534 núcleos, como ilustrado na figura 2.

Outras bases de dados para o problema de análise de imagens de exames de Papanicolau também são amplamente utilizadas, como a Herlev e a SIPaKMeD. No entanto, todos esses conjuntos de dados apresentam inconsistências na nomenclatura das classes, o que impossibilita a sua combinação com o CRIC [Kalbhor et al. 2023].

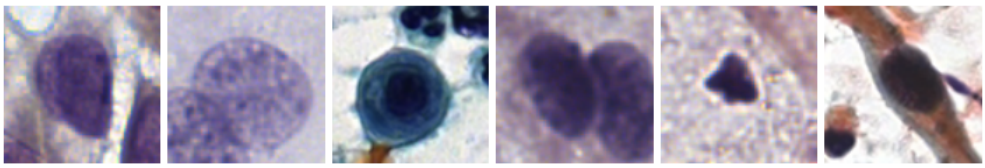


Figura 2. Recorte das células 90x90

4.2. Balanceamento de dados

O conjunto de dados utilizado neste trabalho contém exatamente 11.534 núcleos de células, apresentando um desbalanceamento significativo entre as classes: ASC-H possui 925 amostras, ASC-US 606, HSIL 1.703, LSIL 1.360, NFIL 6.779 e SCC apenas 161 imagens. Conforme um estudo conduzido pelo Centro de Reconhecimento e Inspeção de Células [N. Diniz et al. 2021], o balanceamento do conjunto de dados é essencial devido ao significativo desequilíbrio entre as classes, especialmente a classe SCC. Portanto, será adotada uma abordagem combinada de subamostragem e sobreamostragem, uma vez que

apenas aumentar os dados não é viável para uma classe com apenas 161 exemplos. O balanceamento será obtido por meio da redução de algumas classes e do aumento de outras, visando um conjunto de dados mais equilibrado e representativo.

O balanceamento de dados foi realizado usando a biblioteca *Albumentations* [Buslaev et al. 2020], que oferece uma variedade de métodos de ampliação de dados. O aumento foi aplicado especificamente às classes SCC e ASC-US, abrangendo técnicas como corte aleatório, inversão horizontal e rotação. Quando a rotação não correspondia a um ângulo que produziria um complemento apropriado, o algoritmo preenchia a área restante com base no contexto da imagem. Ambas as classes foram aumentadas para 1.000 amostras. Em contraste, as demais classes passaram por uma redução no número de amostras, realizada por meio da seleção e remoção aleatória de um subconjunto dos dados. A distribuição dos dados antes das transformações pode ser vista na tabela 1.

Para atingir o balanceamento de dados ideal, vários testes foram conduzidos, reduzindo-se todas as classes para 400 amostras, equilibrando-as para 1.500, 2.000 e até mesmo aumentando para o número máximo de imagens na classe predominante, que é 6.779. Portanto, o balanceamento final foi realizado dividindo os dados em 80% para treinamento e 20% para validação, dos 80% alocados para treinamento, 20% foram alocados para teste, permitindo que a precisão final fosse avaliada usando dados que nunca tinham sido vistos durante o processo de treinamento, realizando assim um nivelamento dos dados em 1.000 amostras para todas as 6 classes. A divisão final é apresentada na tabela 1.

Tabela 1. Quantidade de Imagens antes(A) e depois(D) do balanceamento

Divisão	ASC-H	ASC-US	HSIL	LSIL	NFIL	SCC
Treino (A/D)	592/1000	388/1000	1.090/1000	871/1000	4.339/1000	103/1000
Validação	185	122	341	272	1.356	33
Teste	148	96	272	217	1.084	25

4.3. Metodologia

O conceito de *patches* dinâmicos envolve explorar a sobreposição de *patches* [Kotyan and Vargas 2024], extraindo-os de áreas de interesse, evitando assim a extração de regiões de fundo da imagem. Este trabalho propõe diversos métodos de extração, incluindo seleção randômica (SR), randômico aprimorada (RA) e seleção por segmentação (SS). A técnica SR envolve selecionar *patches* randomicamente da imagem, já a abordagem RA seleciona *patches* de maneira randômica, porém com uma proposta de diminuir sobreposições, enquanto a SS foca na extração de regiões de interesse com base em máscaras de segmentação.

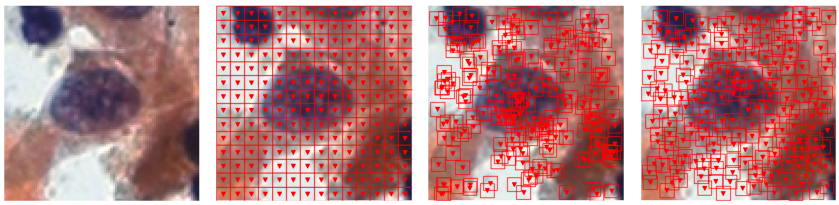
Inicialmente, este estudo foca na análise e seleção do melhor modelo sem modificações de ViT, visando otimizar a acurácia mediante *fine-tuning* utilizando o conjunto de dados do CRIC. Para isso, são explorados modelos pré-treinados, nas variantes *Tiny*, *Small* e *Base*, que operam com *patches* 16x16, e foram pré-treinados na base de dados ImageNet-21k. Essas arquiteturas são utilizadas como ponto de partida para um modelo derivado que modifica a camada de extração de *patches*. Em [Dosovitskiy et al. 2021], é proposto que a entrada para o ViT seja a própria imagem,

que é dividida em *patches* diretamente, através da projeção linear. Porém, no modelo pré-treinado disponibilizado pelo *Hugging Face*, inspirado no método proposto por [Wu et al. 2020], a extração dos *patches* é realizada por meio de uma camada convolucional (Conv2D), um método denominado projeção convolucional, isso significa que, ao invés de utilizar a imagem diretamente, o modelo trabalha com um mapa de características. Foi implementado o *fine-tuning* desse modelo pré-treinado com a projeção convolucional sem modificações, e os resultados foram comparados com as três abordagens dinâmicas. Esses métodos têm como objetivo explorar diferentes formas de entrada, sendo a própria imagem utilizando projeção linear e também com a extração de *patches* via projeção convolucional.

A implementação do modelo de extração dinâmica de *patches* exige a criação de uma classe de *embeddings* personalizada. A ideia central é que, ao importar o modelo ViT pré-treinado, a classe de *embeddings* original contém a chamada para a classe *patch-embeddings*. Nesse contexto, preserva-se a classe original e apresenta-se uma nova classe para *patch-embeddings*. Assim, o token de classificação e a criação dos *embeddings* de posição são mantidos como na classe padrão. No entanto, a extração dos *patches* e sua transformação em vetores são modificadas para ocorrer dentro da classe personalizada.

Para o desenvolvimento do método de extração de *patches* por aleatoriedade (SR), é proposto o conceito de centros, em que a extração de *patches* é baseada em uma lista de tuplas que contêm as coordenadas x e y centrais de cada *patch*. Essa lista é retornada pelo método de extração para a classe personalizada de *patch-embeddings*. Considerando uma imagem de 224x224 pixels, com *patches* de 16x16 pixels, tamanho usado consistentemente em todos os testes, o método resulta em um total de 196 *patches*. O objetivo da extração dinâmica randômica é selecionar aleatoriamente 196 posições centrais para os *patches*, sem qualquer restrição de localização. Esse processo está ilustrado na figura 3c.

A implementação do método denominado randômico aprimorado (RA) baseia-se na seleção aleatória de pontos centrais, mas com uma técnica para evitar sobreposições. Quando um centro é escolhido, ele é verificado para se garantir que não esteja dentro dos pixels adjacentes de outro centro já existente. Assim, a posição central selecionada não poderá estar dentro de outro *patch* quando adicionada à lista final de centros. Dessa forma, cada novo ponto é colocado em uma região não extraída da imagem, como ilustrado na figura 3d.

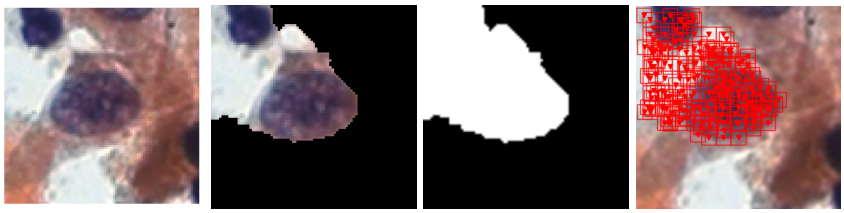


(a) Original **(b)** Grid **(c)** Randômico **(d)** R. Aprimorado

Figura 3. Métodos de Extração de Patches

O método de extração por segmentação (SS) identifica as áreas de interesse da imagem, adotando o GrabCut para gerar uma máscara da imagem. Essa máscara revela as regiões de maior relevância, onde são extraídos seguindo o stride definido pelo tamanho do *patch*, assim os centros são obtidos somente em pixels brancos da máscara,

garantindo a captura máxima de informações e detalhes. Para os *patches* remanescentes, a seleção ocorre de forma aleatória dentro da zona de interesse demarcada pela máscara de segmentação, considerando que esses *patches* serão completamente sobrepostos, como ilustrado na figura 4. A qualidade da segmentação desempenha um papel crucial nesse método. O GrabCut foi utilizado com seus parâmetros padrões, o que pode resultar ocasionalmente em falhas de segmentação, gerando máscaras totalmente pretas. Para esses casos, uma exceção foi implementada, e se a segmentação cobrir menos de 10% da imagem (isto é, se a máscara for composta de 90% ou mais de pixels pretos), a extração de *patches* será feita de forma convencional na imagem original.



(a) Original **(b)** Recorte **(c)** Máscara **(d)** S. Segmentação

Figura 4. Sequência usada para a segmentação e extração de patches

Durante a extração dinâmica de *patches*, quando realizada ao decorrer do treinamento, o método randômico ocorre como esperado, mantendo o tempo de treinamento dentro dos parâmetros normais. No entanto, observa-se que ao aplicar os métodos randômicos aprimorados e o de segmentação, o tempo por época apresenta um aumento significativo. Isso se deve, principalmente, ao processo de conversão de dados (cast) necessário para realizar a segmentação, um processo que aumenta consideravelmente o tempo de execução. Mesmo com o método RA, que verifica os pixels adjacentes, o processo ainda demora muito tempo. Para solucionar esse problema, foi implementada uma etapa de pré-processamento. Utiliza-se o *dataset* balanceado para gerar uma lista de centros para todas as imagens de cada abordagem específica. Essa lista foi armazenada em um dicionário, onde a chave era o nome da imagem e o valor corresponde à lista de centros. O dicionário é salvo em um arquivo e, durante o treinamento, esse arquivo é carregado em uma variável. Dessa forma, a geração dos centros passa a ser realizada por meio de uma simples busca no dicionário, o que resultou em uma significativa melhoria no tempo total de treinamento.

Para viabilizar a busca pela lista de centros, foi necessário ajustar o método de carregamento de dados da classe *ImageFolder*, do PyTorch, responsável por importar imagens e seus respectivos rótulos a partir de um diretório. Criou-se uma classe customizada que, além de carregar a imagem e seu rótulo, também carrega o nome da imagem. O dicionário, então, é armazenado em uma variável global, acessada no método de extração personalizada. Dessa forma, durante a iteração do *batch*, é possível utilizar os nomes das imagens para consultar o dicionário e obter a lista de centros associada a cada imagem, otimizando significativamente o processo de treinamento.

Na abordagem empregada com os modelos customizados de extração de *patches* via projeção linear, a imagem original foi utilizada como entrada e ajustada para um formato adequado. Nesse processo, os *patches* são extraídos e os *embeddings* são gerados diretamente a partir da imagem, utilizando uma projeção linear.

Para utilizar a técnica de extração mediante convolução, um ajuste é realizado aplicando a convolução diretamente aos *patches* extraídos. Diferentemente do método padrão que utiliza uma camada convolucional para gerar automaticamente um mapa de características, o método customizado que combina a convolução com a extração dinâmica de *patches* inclui uma etapa adicional, a extração dos *patches* seguida pela aplicação da convolução, o que aumenta consideravelmente o tempo de treinamento do modelo. Por conta disso, os métodos dinâmicos com convolução podem ser implementados apenas no ViT-Small, cuja arquitetura é mais simples em comparação com o modelo ViT-Base.

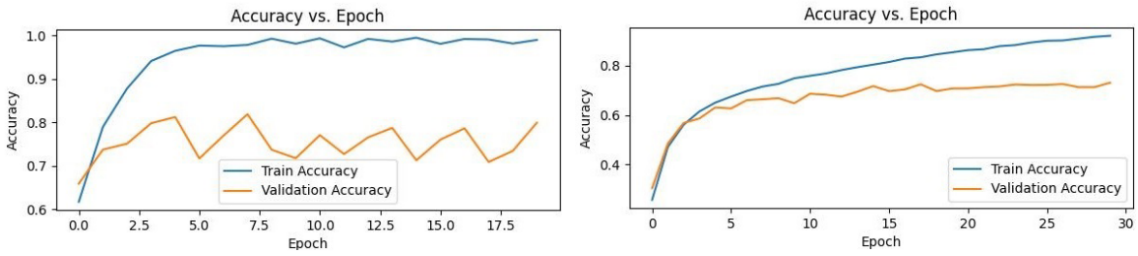
5. Experimentos e Resultados

Para realizar o fine-tuning dos modelos, o treinamento foi conduzido em um computador local com uma GPU RTX 2060 (6 GB), processador Ryzen 5 3600x e 32 GB de RAM. Essa configuração foi suficiente para processar modelos de menor complexidade, como o *Tiny* e o *Small*. No entanto, devido à maior demanda computacional do modelo *Base*, não foi possível processá-lo localmente com as camadas descongeladas do MLP do codificador. Para contornar essa limitação, foi utilizada uma assinatura Colab PRO, que oferece acesso a recursos de computação em nuvem. No Colab PRO, o treinamento foi realizado em uma GPU T4 com 15 GB e 13 GB de RAM.

Para determinar o modelo de *baseline*, foram realizados diversos testes com diferentes arquiteturas e hiperparâmetros. Um dos testes, ilustrado na figura 5, envolveu o treinamento dos MLPs presentes nos últimos dois blocos do codificador, especificamente os blocos 10 e 11. Além disso, foram adicionadas três camadas lineares ao MLP do classificador, enquanto o restante da arquitetura foi congelada. Utilizando um conjunto de dados balanceado com 1.000 amostras por classe e uma taxa de aprendizado de 1e-4, e um tamanho de *batch* de 16, observou-se que o modelo *Small* se adaptou bem aos dados de treinamento. No entanto, ele apresentou dificuldades para generalizar nos dados de validação. Notou-se que, com essa taxa de aprendizado, o modelo exibiu variações significativas na validação, com oscilações expressivas tanto em termos de piora quanto de melhora, como pode ser analisado na figura 5a.

Para estabilizar o conjunto de validação do modelo *Small*, diversas taxas de aprendizado foram experimentadas, sendo incorporada uma taxa de 1e-5. Outra alteração em relação à estrutura do modelo *Small* anterior foi a remoção das três camadas lineares no MLP do classificador, substituindo-as por uma única camada. Como resultado, a validação apresentada na figura 5b mostrou uma leve melhoria nas oscilações. No entanto, o problema de *overfitting* persistiu em ambos os testes. Experimentos também foram realizados com o modelo *Tiny*, no entanto, devido à sua arquitetura mais simples, a extração adequada das características das células mostrou-se mais desafiadora, resultando em dificuldades de convergência.

Para a definição do *baseline* de comparação, os modelos *Base* e *Small* foram identificados como os melhores candidatos. As taxas de aprendizado de 1e-4 e 1e-5 demonstraram ser as mais adequadas, sendo escolhida para os testes 1e-5, utilizando um tamanho de *batch* de 32, para todos os testes, enquanto a regularização com *weight decay* não apresentou impacto significativo e, por isso, foi descartada. No *fine-tuning* do modelo ViT-Small, optou-se por descongelar mais blocos do codificador, que possui um total de 12. Foram liberados os dois primeiros blocos e os três últimos, resultando em uma



(a) Taxa de Aprendizado 1e-4 **(b) Taxa de Aprendizado 1e-5**

Figura 5. Acurácia e Loss do modelo Small Alterando Taxa de Aprendizado

acurácia de 79% no conjunto de validação e 81% no conjunto de teste, como mostrado na figura 6. Ao realizar o *fine-tuning* com o modelo ViT-Base, não foi possível replicar a mesma configuração de descongelamento do modelo *Small* devido à alta complexidade do modelo. Dessa forma, para o *fine-tuning* do modelo ViT-Base, apenas o último MLP do último bloco de codificador foi descongelado. Como ilustrado na figura 7, essa abordagem resultou em uma acurácia de 76% nos dados de validação e 75% nos dados de teste.

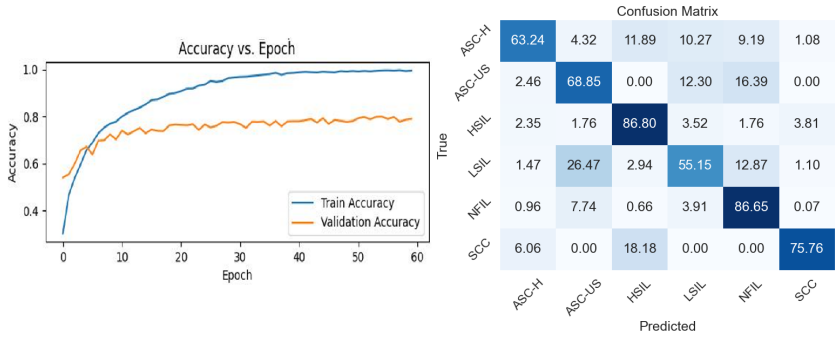


Figura 6. Gráfico de Acurácia e Matriz de Confusão - Modelo Small

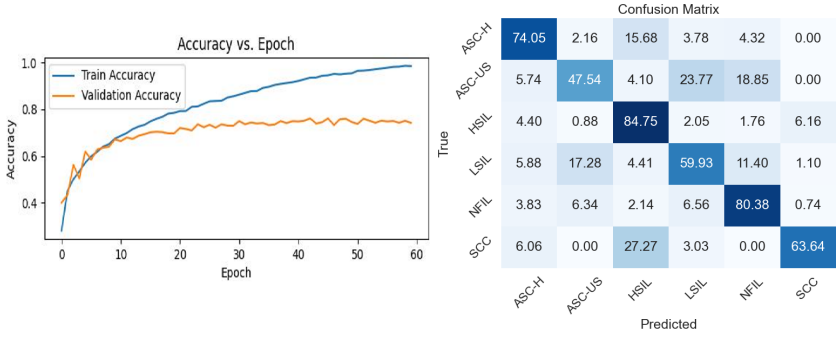


Figura 7. Gráfico de Acurácia e Matriz de Confusão - Modelo Base

Para o treinamento dos modelos com extração dinâmica, foram utilizadas as mesmas arquiteturas dos modelos *Small* e *Base*, ambos configurados com tamanho de *batch* de 32 e taxa de aprendizado de 1e-5. Para os testes com o modelo *Small*, foram aplicadas tanto a projeção linear quanto a convolucional em todas as três abordagens dinâmicas, totalizando 6 experimentos, cujos resultados de acurácia final são apresentados na tabela 2. No caso do modelo *Base*, devido à sua maior complexidade, os testes foram limitados

às abordagens dinâmicas com projeção linear, totalizando 3 experimentos, ilustrados na figura 8.

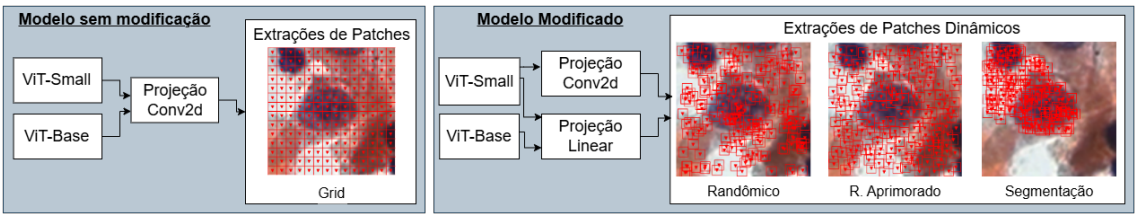


Figura 8. Fluxo de Testes: Modelos Sem Modificações vs. Modificados

Analisando-se os resultados obtidos a partir do fine-tuning dos modelos *Small* e *Base* do ViT, sem modificações, em comparação com os resultados da abordagem de ensemble de Redes Neurais Convolucionais (CNNs) proposta por [N. Diniz et al. 2021], observa-se que os modelos ViT não atingiram a mesma acurácia obtida pela CNN com ensemble. Enquanto a CNN alcançou uma acurácia de 95%, o modelo *ViT Small* obteve apenas 81%. Esse desempenho pode ser atribuído à arquitetura do ViT, que tende a exigir uma quantidade significativa de dados para atingir seu potencial máximo. Com uma arquitetura robusta, mas sem o volume ideal de dados, os modelos mais simples, como o *Small*, demonstraram desempenho superior em comparação com modelos mais complexos, como o *Base*.

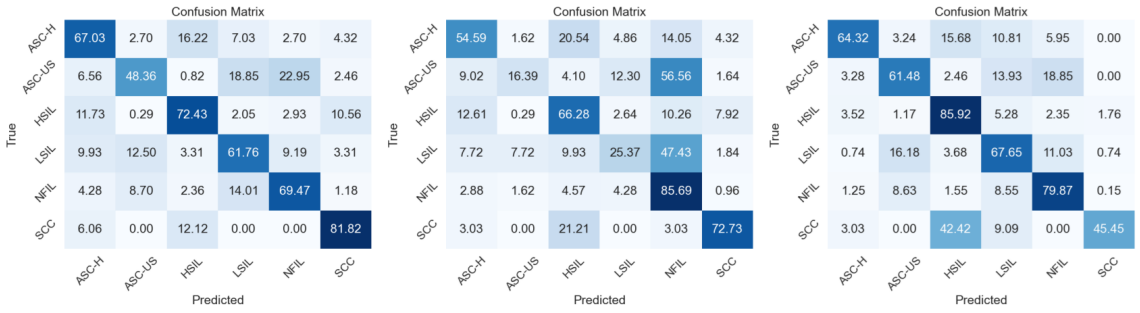
O *fine-tuning* comparando os modelos *Small* e *Base* sem modificação com os de extração dinâmica utilizando projeção linear, conforme ilustrado na figura 8, mostraram que a aplicação da extração de *patches* impactou a acurácia e a perda no modelo *Base*, como pode ser analisado na tabela 2. No entanto, esses impactos foram menores no modelo *Small*, ainda que nenhum dos métodos com *patches* dinâmicos tenha superado o modelo sem modificação na qual utiliza de técnica convolucional.

Os testes de extração dinâmica com operação convolucional foram realizados somente no modelo *Small*, como pode ser analisado no fluxo de teste na figura 8. Comparando com a extração dinâmica baseada em projeção linear, a abordagem convolucional demonstrou uma melhora na acurácia no conjunto de testes para as três técnicas dinâmicas. Esse ganho pode ser atribuído à própria convolução, que realiza uma extração inicial de características e preserva as relações espaciais, aumentando o poder de convergência do modelo. Entretanto, ao comparar a abordagem dinâmica convolucional com a extração por *Grid*, que também utiliza a convolução, observou-se que nenhuma das três técnicas dinâmicas superou o desempenho da extração de *patchs* por *Grid*, como demonstrado na figura 3b. Isso pode ser explicado pelas características do conjunto de dados, onde as áreas de interesse são relativamente pequenas e não favorecem a extração dinâmica.

As matrizes de confusão ilustradas na figura 9 correspondem aos melhores métodos de cada modelo e projeção. No caso do modelo *Base* com projeção linear e o método SR, as classes mais desafiadoras são a ASC-US e LSIL, que apresentam dificuldades de classificação, com muitas amostras dessas classes sendo erroneamente classificadas como Negativa (NFIL), uma classe com maior variabilidade de dados. Esse comportamento pode ser explicado pelo fato de o modelo *Base* possuir um número maior de parâmetros, o que o torna mais robusto, mas também exige uma quantidade maior de

dados para convergência. A classe ASC-US, que passou por aumento de dados, apresenta menos variabilidade, o que contribui para uma maior dificuldade na classificação correta dessas amostras. Para o modelo *Small* com projeção linear, as abordagens de extração randômica aprimorada (RA) e segmentação (SS) apresentam resultados praticamente idênticos, com uma leve melhoria no método SS. No entanto, para o modelo *Small* com projeção convolucional, a abordagem RA obteve o melhor desempenho, com uma acurácia de 77%, como pode ser analisado na tabela 2. Esse resultado se deve, em parte, ao uso da projeção convolucional, que mostrou mais eficaz para este conjunto de dados, em combinação com a exploração da extração dinâmica, limitando a sobreposição de *patches*.

Pode-se observar que a sobreposição de *patches* é um fator importante para o desempenho do modelo com esse conjunto de dados. Entre os métodos dinâmicos, a abordagem randômica aprimorada (RA) com projeção convolucional apresentou a acurácia mais próxima da extração tradicional, sendo o melhor de todos os testes dinâmicos. Para este conjunto de dados, percebe-se que métodos baseados em áreas de interesse sofrem com sobreposição excessiva de *patches*, o que resulta na perda de características importantes da imagem, especialmente em regiões menores. Observou-se que, quando a segmentação contorna perfeitamente o núcleo da célula, a sobreposição de *patches* aumenta significativamente. Em células menores, como as células negativas, essa sobreposição torna-se especialmente alta, resultando em *patches* contendo informações redundantes.



(a) SR Base Linear **(b) SS Small Linear** **(c) RA Small Conv**

Figura 9. Matriz de Confusão - Melhores Resultados Dinâmicos

Tabela 2. Resultados Finais Extração de Patches Dinâmicos

ViT	Projeção	Seleção Randômica	Randômica Aprimorada	Seleção por Segmentação
Base	Linear	69.54	64.92	59.50
Small	Linear	63.73	70.14	70.73
Small	Conv2d	75.02	77.30	75.78

6. Conclusão

Com base nos resultados obtidos, pode-se inferir que o fine-tuning do modelo sem modificação de *Vision Transformers*, que incorpora a extração de *patches* via convolução (Conv2d), demonstrou-se otimizado e eficaz na captura de características iniciais da imagem. Nenhuma abordagem com projeção linear com extração dinâmica superou o desempenho da extração baseada em convolução. Nos testes com extração dinâmica aplicando convolução, observou-se uma melhora em comparação ao método dinâmico com

projeção linear, entretanto, a sobreposição excessiva de *patches* impediu que as abordagens dinâmicas superassem o modelo sem alterações.

Em relação à arquitetura dos *Vision Transformers*, observou-se que, para esse conjunto de dados, quanto mais camadas são treinadas, maiores são as chances de alcançar uma acurácia elevada e uma perda (loss) menor. Esse resultado evidencia o forte impacto do conjunto de dados na acurácia, e fica evidente que o desbalanceamento dos dados afeta significativamente a qualidade do treinamento. O ViT, mesmo com *fine-tuning*, é um modelo que demanda uma grande quantidade de dados para alcançar seu potencial. Quando o balanceamento do conjunto de dados foi ajustado para 1.000 amostras por classe, a classe SCC, que teve um aumento expressivo de 850 imagens, acabou prejudicada, pois essa expansão reduziu a diversidade de características representadas, limitando a eficácia do treinamento.

Ao se analisar a implementação dos *patches* dinâmicos, observa-se que, para este conjunto de dados em que o objeto de interesse ocupa uma área pequena da imagem, essa abordagem não supera a eficácia do modelo sem modificações, em grande parte devido à sobreposição excessiva de *patches*. No entanto, é possível que um modelo mais robusto que o *Small*, aliado a uma extração dinâmica com sobreposições controladas e utilizando convolução, possa superar o método convencional. Essa abordagem exigiria um maior poder computacional para realizar o *fine-tuning* de um modelo com mais parâmetros.

Apesar das contribuições apresentadas, diversos aspectos do problema ainda merecem investigação, o que abre caminho para futuros trabalhos. A melhoria no balanceamento dos dados se mostra uma estratégia eficaz para aprimorar os testes realizados, incluindo abordagens com problemas binários e com três classes, como demonstrado no estudo [N. Diniz et al. 2021]. Além disso, o uso de redes generativas para o aumento de dados da classe SCC, que possui uma quantidade baixa expressivamente de imagens, pode ser uma alternativa promissora para melhorar o desempenho do modelo.

Este trabalho abre possibilidades para otimizar e melhorar a aplicação de *patches* dinâmicos com projeção utilizando convolução, explorando alternativas que substituam o método atual de extração seguido pela aplicação do *kernel*. Além disso, propõe-se a aplicação de outros métodos para a detecção da zona de interesse, como o uso de modelos pré-treinados de redes neurais convolucionais. Esses modelos poderiam fornecer mapas de características que, por meio de técnicas como o *Grad-CAM*, permitiriam identificar as zonas de interesse com maior precisão.

Outro ponto a ser explorado seria testar o treinamento com mais camadas de codificadores descongeladas no modelo *Base*. Nos dois hardwares utilizados neste projeto, o computador local e o do Google Colab, houve limitações quanto ao treinamento mais extenso do modelo *Base*. Assim, ao realizar o fine-tuning do modelo com um conjunto maior de parâmetros treináveis, é possível obter um desempenho superior.

Referências

Barcelos, M. R. B., Lima, R. d. C. D., Tomasi, E., Nunes, B. P., Duro, S. M. S., and Facchini, L. A. (2017). Quality of cervical cancer screening in brazil: external assessment of the pmaq. *REV SAUDE PUBL*, 51:67.

Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). Albumentations: Fast and flexible image augmentations. *Information*, 11(2).

Chen, X., Hsieh, C.-J., and Gong, B. (2022). When vision transformers outperform resnets without pretraining or strong data augmentations. *ArXiv*, abs/2106.01548.

Dai, Y., Gao, Y., and Liu, F. (2021). Transmed: Transformers advance multi-modal medical image classification. *Diagnostics*, 11(8).

Dodge, S. and Karam, L. (2016). Understanding how image quality affects deep neural networks. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Fang, M., Lei, X., Liao, B., and Wu, F.-X. (2022). A deep neural network for cervical cell classification based on cytology images. *IEEE Access*, 10:130968–130980.

Kalbhor, M., Shinde, S., Wajire, P., and Jude, H. (2023). Cervicell-detector: An object detection approach for identifying the cancerous cells in pap smear images of cervical cancer. *Heliyon*, 9(11):e22324.

Kotyan, S. and Vargas, D. V. (2024). Improving robustness for vision transformer with a simple dynamic scanning augmentation. *Neurocomputing*, 565:127000.

McDanel, B. and Ngoc Huynh, C. P. (2023). Dynamic patch sampling for efficient training and dynamic inference in vision transformers. In *International Conference on Machine Learning and Applications*, pages 83–9.

Michalas, S. (2000). The pap test: George n. papanicolaou (1883–1962): a screening test for the prevention of cancer of uterine cervix. *EUR J OBSTET GYN R B*, 90(2):135–138.

N. Diniz, D., T. Rezende, M., G. C. Bianchi, A., M. Carneiro, C., J. S. Luz, E., J. P. Moreira, G., M. Ushizima, D., N. S. de Medeiros, F., and J. F. Souza, M. (2021). A deep learning ensemble method to assist cytopathologists in pap test image classification. *J IMAGING SCI*, 7(7).

Rezende, M. T., Silva, R., Bernardo, F. d. O., Tobias, A. H. G., Oliveira, P. H. C., Machado, T. M., Costa, C. S., Medeiros, F. N. S., Ushizima, D. M., Carneiro, C. M., and Bianchi, A. G. C. (2021). Cric searchable image database as a public platform for conventional pap smear cytology data. *Scientific Data*, 8(1):151.

Steiner, A. P., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., and Beyer, L. (2022). How to train your vit? data, augmentation, and regularization in vision transformers. *Trans. Mach. Learn. Res.*, 2022.

Wu, B., Xu, C., Dai, X., Wan, A., Zhang, P., Tomizuka, M., Keutzer, K., and Vajda, P. (2020). Visual transformers: Token-based image representation and processing for computer vision. *ArXiv*, abs/2006.03677.