Vision Transformers com Patches Dinâmicos para a Análise de Lâminas Histológicas

Vinícius Henrique Giovanini [1], **Alexei Manso Correa Machado** [1,2]

Departamento de Ciência da Computação
Pontifícia Universidade Católica de Minas Gerais (PUC Minas), Belo Horizonte, Brasil
[2]Departamento de Anatomia e Imagem - Faculdade de Medicina
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brasil

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

Abstract. *This work presents an approach to improve Vision Transformers (ViT) by implementing dynamic patch input. Experiments were conducted with different types of models, performing fine-tuning and exploring multiple strategies to identify the most relevant issues related to patch extraction. The proposed approach was compared to the traditional ViT model in a study using 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 0.81 while the best dynamic approach obtained 0.78 of accuracy, due to the excessive overlapping of 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 0,81. Em contrapartida, a melhor abordagem dinâmica obteve 0,78, devido à excessiva sobreposição dos patches.*

1. Introdução

O uso de algoritmos de aprendizado de máquina para a análise de imagens de lâminas histológicas se intensificou na última década, com o desenvolvimento de hardware de alto desempenho e dos modelos de aprendizado profundo (*deep learning*) (DL). Um exemplo de aplicação relevante dessa tecnologia é no diagnóstico do câncer de colo de útero que 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 a doença mediante alterações nas células que possam indicar lesões pré-cancerígenas.

Visando aprimorar a precisão na identificação de células anormais, soluções baseadas em algoritmos de DL 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 serem realizados.

Este trabalho se baseia no modelo de *Vision Transformer* (ViT) [Dosovitskiy et al. 2021], uma técnica de DL recente 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.

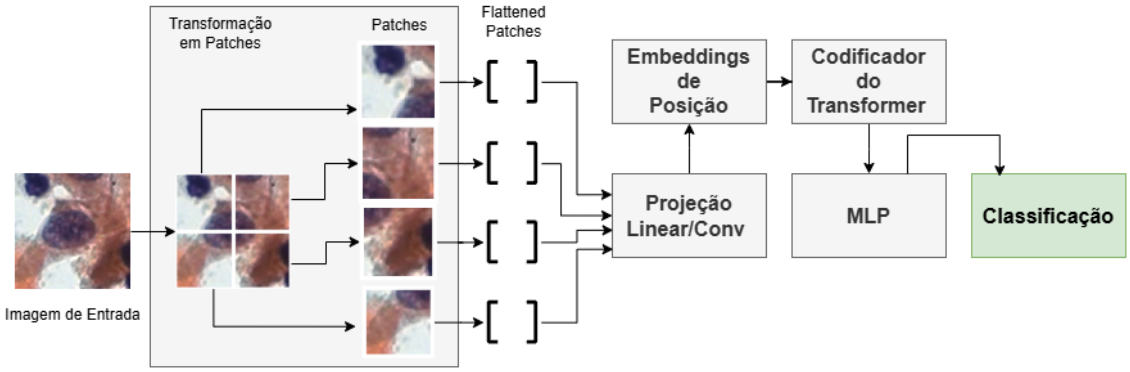


Figura 1. Fluxo de Entrada e Saída do Vison Transformer

A modificação da arquitetura do ViT [Dosovitskiy et al. 2021] proposta neste trabalho utiliza ajuste fino (*fine-tuning*) de diferentes variações do modelo, a fim de compreender seu impacto na 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 cinco abordagens: extração de *patches* baseada em seleção randômica (SR), randômico aprimorado (RA), seleção por segmentação (SS), seleção por zigue-zague (ZZ) e seleção por espiral (ES). O objetivo principal da seleção dos *patches* é priorizar áreas que contenham mais células e menos região de fundo. Dessa forma, é possível avaliar o impacto dessa estratégia em relação aos resultados obtidos a partir do *fine-tuning* do modelo original com *patches* fixos. As diversas abordagens são comparadas através de um estudo de caso contendo imagens de exames de Papanicolau.

2. Trabalhos Relacionados

[Dodge and Karam 2016] descrevem o efeito adverso de um conjunto de distorções espectrais na precisão de classificação de imagens histológicas, buscando fornecer *insights* sobre a robustez das Redes Neurais Profundas (RNP) nesses cenários. São analisados efeitos de suavização, adição de ruído, variações de contraste e compressão, destacando-se a importância de se tornar o modelo mais robusto a condições reais. No treinamento, foi utilizado o CRIC *dataset* [Rezende et al. 2021], e os resultados obtidos através da

detecção de objetos mostram que é possível auxiliar a identificação de células em imagens de exames histológicos.

Os avanços no uso de técnicas de aprendizado profundo na Citologia são exemplificados no projeto DeepCeLL [Fang et al. 2022]. Esse projeto tem como objetivo analisar recursos computacionais, 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, adquiridos de um hospital universitário da Dinamarca, contendo 917 imagens individuais, além do SIPaKMeD, um *dataset* com cerca de 4.049 imagens capturadas de microscópio óptico. Um projeto semelhante, o CerviCell-Detector [Kalbhor et al. 2023] propõe um método automatizado de triagem do exame de Papanicolau que utiliza redes profundas de detecção, robustas a mudanças de saturação e adição de ruídos.

Uma combinação de ViTs e CNNs para a identificação de tumores na glândula parótida é apresentada 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 não-locais 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 ViTs, comparados a outras arquiteturas, como as *ResNets*, com ênfase no impacto do pré-processamento, aumento de dados e regularização. O primeiro trabalho examina o desempenho dos ViTs sem pré-treinamento ou uso extensivo de aumento de dados, destacando que esses modelos podem superar *ResNets* de tamanhos semelhantes sem transferência de aprendizado ou com grande número de dados aumentados, especialmente com o uso do otimizador *Sharpness-Aware Minimizer* (SAM). Já o segundo trabalho analisa 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 *datasets*. 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 a partir de pesos aleatórios. Ambos os estudos ressaltam a importância do ajuste adequado do treinamento e do uso de estratégias específicas para otimizar o desempenho dos ViTs.

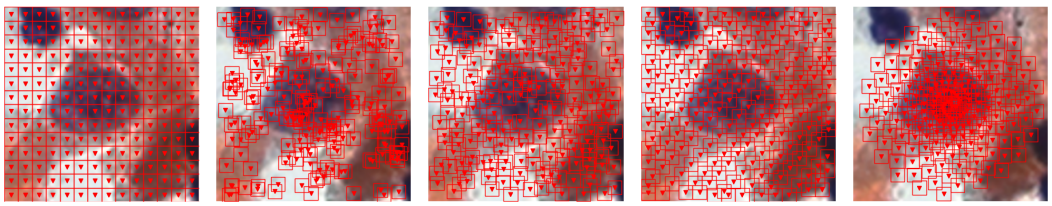
A técnica *Patch Sampling Schedule* (PSS) proposta por [McDanel and Ngoc 2023] foi desenvolvida para otimizar o tempo de treinamento e aumentar a eficiência dos ViTs. O objetivo é melhorar a precisão e a taxa de processamento ao ajustar dinamicamente a quantidade e o tamanho dos *patches* utilizados durante o treinamento. O método seleciona e ajusta o tamanho dos *patches* baseado nos valores dos *pixels* ou de forma aleatória, mas descartando aqueles que correspondem ao fundo, reduzindo a complexidade computacional sem comprometer o desempenho. De modo similar, o método *Simple Dynamic Scanning Augmentation* [Kotyan and Vargas 2024] propõe o uso dinâmico de *patches* para aumentar a robustez dos 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

objetos. Duas abordagens adotam métodos aleatórios: *Random Patches* (RP) e *Random Tracing* (RT), enquanto as demais são baseadas em mapas de calor: *Salient Patches* (SP) e *Salient Tracing* (ST), que buscam regiões de interesse na imagem. Ruídos são adicionados à entrada do ViT para analisar seu impacto sobre a classificação final e a robustez do modelo. A análise revela que as abordagens aleatórias apresentam maior acurácia e robustez quando comparadas com o método original.

3. Vison Transformers

O *Vision Transformer* representa uma abordagem de modelagem capaz de competir com as CNNs em várias tarefas de visão computacional. Conforme proposto 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 *Perceptrons* Multicamadas (MLP). 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. No modelo original, a imagem da entrada é dividida em *patches* não sobrepostos, de tamanho fixo. Por exemplo, uma imagem de 224×224 pode ser dividida em *patches* de 16×16 *pixels*, totalizando 196 *patches*, como ilustrado na Figura 2a. 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.



(a) Grid **(b)** Randômico **(c)** R. Aprimorado **(d)** Zigue-Zague **(e)** Espiral

Figura 2. Métodos de Extração de Patches

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), como ilustrados na Figura 3.

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.

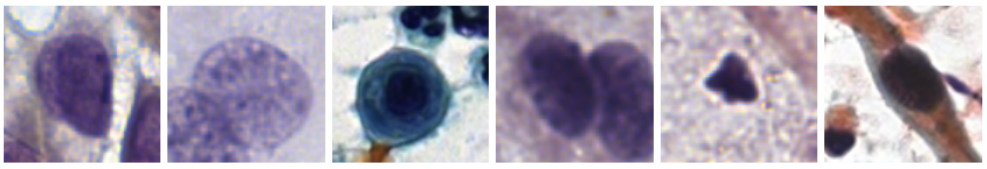


Figura 3. Exemplos de recorte de células com 90×90 pixels

4.2. Balanceamento de dados

O conjunto de dados utilizado neste trabalho contém 11.534 núcleos de células, apresentando um desbalanceamento significativo entre as classes: ASC-H com 925 amostras, ASC-US com 606, HSIL com 1.703, LSIL com 1.360, NFIL com 6.779 e SCC com 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, foi adotada uma abordagem combinada de subamostragem e sobreamostragem, uma vez que apenas aumentar os dados não era viável para uma classe com apenas 161 exemplos. O balanceamento foi obtido por meio da redução de algumas classes e do aumento de outras, visando um conjunto de dados de treino mais equilibrado e representativo.

O balanceamento do conjunto de treino 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.

O conjunto de dados foi dividido em 64% para treinamento, 20% para validação e 16% para teste. A separação foi feita com base em pacientes, de modo que cada paciente contribuiu com imagens para apenas um dos conjuntos, evitando contaminação de dados entre treino, validação e teste. 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 controlada de *patches* [Kotyan and Vargas 2024], focando na extração de áreas de interesse e evitando regiões de fundo. Este trabalho propõe cinco métodos de extração: seleção randômica (SR), randômica aprimorada (RA), segmentação (SS), zigue-zague (ZZ) e espiral (ES). A SR seleciona *patches* aleatoriamente, enquanto a RA reduz sobreposições. A SS utiliza máscaras de segmentação, a ZZ segue um percurso alternado, e a ES organiza os *patches* em espiral.

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* 16×16, 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. No presente estudo, foi implementado o *fine-tuning* desse modelo pré-treinado com a projeção convolucional sem modificações e os resultados foram comparados com as cinco 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.

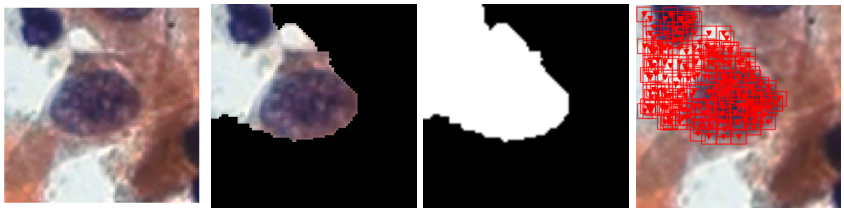
Para o desenvolvimento do método de extração de *patches* por aleatoriedade (SR), é proposto o conceito de centros, em que a extração de *patches* é baseada em uma lista de tuplas que contêm as coordenadas x e y centrais de cada *patch*. Essa lista é retornada pelo método de extração para a classe personalizada de *patch-embeddings*. Considerando

uma imagem de 224×224 *pixels*, com *patches* de 16×16 *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 2b.

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 2c.

O desenvolvimento dos métodos Zigue-Zague (ZZ) e Espiral (ES) foi motivado pela ideia de se extraírem sequências de *patches* adjacentes, evitando saltos como nos métodos de grid (entre uma linha e a próxima) e randômico (saltos entre cada *patch* e o seguinte). O método ZZ, como ilustrado na Figura 2d, segue um traçado em zigue-zague, explorando a sobreposição de *patches* de maneira controlada. Por outro lado, o método ES organiza os *patches* em um padrão espiralado, iniciando a partir do centro da imagem, região onde, por convenção, encontra-se o núcleo da célula e expandindo-se progressivamente para a periferia. Dessa forma, o método apresenta alta sobreposição de *patches* na região central, que vai diminuindo à medida que a espiral se expande, como demonstrado na Figura 2e.

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 os *patches* são extraídos seguindo o *stride* definido pelo seu próprio tamanho. 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 *default*, o que pode resultar ocasionalmente em falhas de segmentação, gerando máscaras totalmente vazias. Para esses casos, uma exceção foi implementada, e se a segmentação cobrir menos de 10% da imagem, a extração de *patches* será feita de forma linear convencional.



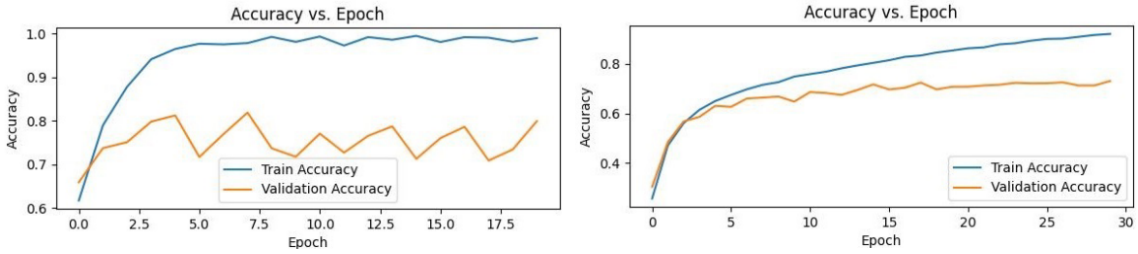
(a) Original **(b)** Recorte **(c)** Máscara **(d)** Patches

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

5. Resultados Experimentais

O *fine-tuning* dos modelos 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.

Ao se definir o modelo de *baseline*, foram testadas diversas arquiteturas e hiperparâmetros. Um dos experimentos treinou os MLPs dos blocos 10 e 11 do codificador, adicionando três camadas lineares ao MLP do classificador e congelando o restante da arquitetura. Com um conjunto balanceado de 1.000 amostras por classe, taxa de aprendizado de 1e-4 e *batch* de 16, o modelo *Small* apresentou boa adaptação ao treino, mas dificuldades na generalização, com oscilações significativas na validação (Figura 5a). Para mitigar esse problema, reduziu-se a taxa para 1e-5 e substituíram-se as três camadas lineares do classificador por uma única, resultando em leve estabilidade na validação (Figura 5b), mas com *overfitting* ainda presente. Já o modelo *Tiny*, devido à sua arquitetura simplificada, teve dificuldades na extração de características das células, prejudicando sua convergência.



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

Figura 5. Acurácia do modelo Small para diferentes taxa de aprendizado

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 escolhido 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 acurácia de 0,81 no conjunto de validação e 0,79 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 0,76 nos dados de validação e 0,75 nos dados de teste.

No treinamento dos modelos com extração dinâmica, foram utilizadas as mesmas arquiteturas dos modelos *Small* e *Base*, ambos configurados com tamanho de *batch* de

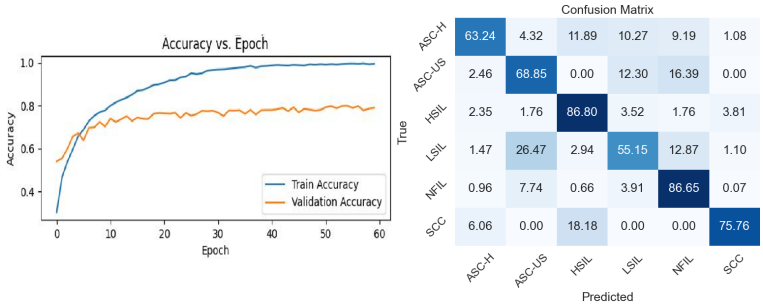


Figura 6. Gráfico de Acurácia e Matriz de Confusão - Modelo Small

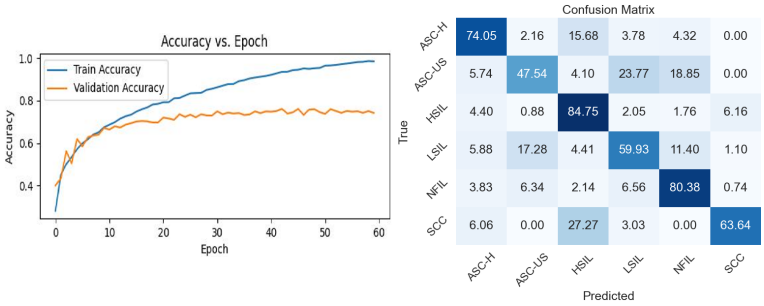


Figura 7. Gráfico de Acurácia e Matriz de Confusão - Modelo Base

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 cinco abordagens dinâmicas, totalizando 10 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 5 experimentos, ilustrados na Figura 8.

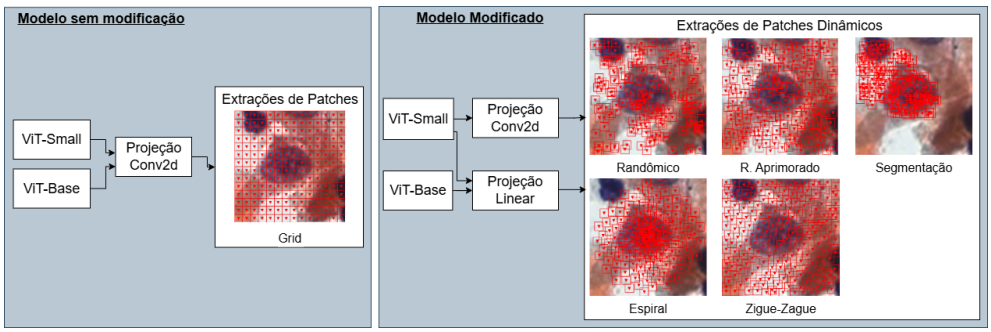


Figura 8. Fluxo de Testes: Modelos Sem Modificações vs. Modificados

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 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 0,95, o modelo *ViT Small* obteve apenas 0,81. 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* realizado para se compararem os modelos *Small* e *Base* sem modificação com os de extração dinâmica utilizando projeção linear, conforme ilustrado na Figura 8, mostrou 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 no qual utiliza-se a 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. Comparado com a extração dinâmica baseada em projeção linear, a abordagem convolucional demonstrou uma melhora na acurácia no conjunto de testes para as cinco 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 se comparar a abordagem dinâmica convolucional com a extração por *Grid*, que também utiliza a convolução, observou-se que nenhuma das cinco técnicas dinâmicas superou o desempenho da extração de *patchs* por *Grid*, como demonstrado na Figura 2a. Isso pode ser explicado pelas características do conjunto de dados, onde as áreas de interesse são relativamente pequenas e não favorecem a extração dinâmica.

No modelo *Small* com projeção linear, as abordagens RA e SS tiveram resultados quase idênticos, com leve vantagem para SS. Já com projeção convolucional, a abordagem ES obteve a melhor acurácia (0,78), como mostrado na Tabela 2. Esse desempenho se deve à combinação da projeção convolucional com a extração em espiral. O modelo *Base* teve desempenho inferior ao *Small* em todas as metodologias dinâmicas, devido à menor robustez do modelo.

A sobreposição de *patches* desempenhou um papel fundamental no desempenho do modelo para o conjunto de dados usado neste estudo. Entre os métodos dinâmicos, a abordagem de Seleção por Espiral (ES) com projeção convolucional apresentou a acurácia mais próxima da extração por *Grid*, sendo a mais eficaz entre os testes propostos. Isso ocorre porque, neste conjunto de dados, onde a área de interesse é pequena e a quantidade de *patches* gerados para cada imagem é grande, a sobreposição é bastante elevada. Nos métodos como o zigue-zague, essa sobreposição acontece principalmente na geração de *patches* ao longo das diagonais. Já no método de espiral, a extração começa pelo centro da imagem, onde está o núcleo da célula, uma região de alta sobreposição, e se expande progressivamente para áreas menos relevantes, capturando melhor as características essenciais do objeto de interesse.

Tabela 2. Resultados finais de acurácia para a extração de patches dinâmicos

ViT	Projeção	Seleção Randômica	Randômica Aprimorada	Seleção por Segmentação	Zigue Zague	Espiral
Base	Linear	0,70	0,65	0,60	0,67	0,67
Small	Linear	0,64	0,70	0,71	0,68	0,69
Small	Conv2d	0,75	0,77	0,76	0,76	0,78

6. Discussão

Com base nos resultados obtidos, pode-se inferir que o *fine-tuning* do modelo de *Vision Transformer* sem modificação, 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 este conjunto de dados, quanto mais camadas são treinadas, maiores são as chances de se alcançar uma acurácia elevada e uma perda (*loss*) menor. Esse resultado evidencia o forte impacto do conjunto de dados na acurácia, e 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.

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 de amostras, pode ser uma alternativa promissora para melhorar o desempenho do modelo.

Outro ponto a ser explorado é o treinamento com mais camadas de codificadores descongeladas no modelo *Base*. Nos dois hardwares utilizados neste projeto, 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.

Agradecimentos — AMC Machado agradece o auxílio financeiro do Fundo de Incentivo à Pesquisa FIP-PUCMinas 2025/32467 e da FAPEMIG através dos projetos APQ-02753-24 e APQ-06556-24.

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, 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.

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.