Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 32, Num. 1 (2025) 128-135

RESEARCH ARTICLE

Classification of Cells Infected by Malaria with Vision Transformers Models

Classificação de Células Infectadas por Malária com Modelos de Vision Transformers

Eduardo David Campopiano[1], Iago Fonseca Marinho Pereira[1]*, Thomas William[1], Douglas Henrique Siqueira Abreu[2], Guilherme Ribeiro Sales[1], Valentino Corso[1], Cides S. Bezerra[1]

Abstract: This project develops a system for malaria classification using the Vision Transformer architecture on cellular tissue microscopy images. The implementation aims to assist in medical diagnosis by providing efficient and accurate information, with the potential to contribute to the control of endemic diseases. The system utilizes models from the vit-pytorch library, including the Vision Transformer, Patch Merger, and Vision Transformer for Small Datasets. The Vision Transformer for Small Datasets achieved the highest accuracy of 98.97% for an image size of [150,150].
Keywords: Vision Transformer — Malaria Classification — Cellular Tissue Microscopy — vit-pytorch

Resumo: Este projeto desenvolve um sistema para classificação de malária utilizando a arquitetura *Vision Transformer* em imagens de microscopia de tecido celular. A implementação visa auxiliar no diagnóstico médico ao fornecer informações eficientes e precisas, com potencial para contribuir no controle de doenças endêmicas. O sistema utiliza modelos da biblioteca vit-pytorch, incluindo o *Vision Transformer*, o *Patch Merger* e o *Vision Transformer for Small Datasets*. O modelo *Vision Transformer for Small Datasets* alcançou a maior precisão de 98,97% para uma dimensão de imagem de [150,150].
Palavras-Chave: *Vision Transformer* — Classificação de Malária — microscopia de Tecido Celular — vit-pytorch

[1] *CPQD - Research and Development Center, Brazil*
[2] *Pontifical Catholic University of Campinas (PUCC), Brazil*
***Corresponding author**: {davidcampopiano, iagofmpereira, thomaswpp02}@gmail.com, douglas.henrique@puc-campinas.edu.br, {guisales, valenti, cbezerra}@cpqd.com.br
DOI: http://dx.doi.org/10.22456/2175-2745.143548 • **Received:** 25/10/2024 • **Accepted:** 02/12/2024

1. Introdução

A malária é uma doença endêmica potencialmente fatal causada por parasitas do gênero *Plasmodium*, transmitidos aos humanos através da picada de mosquitos fêmeas infectadas do gênero Anopheles [1]. Segundo a Organização Mundial da Saúde (OMS), a malária continua sendo um dos maiores desafios de saúde pública global, com aproximadamente 241 milhões de casos e 627 mil mortes em 2020, principalmente na África Subsaariana [1].

A doença é causada principalmente por cinco espécies de *Plasmodium*: *Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium ovale*, *Plasmodium malariae* e *Plasmodium knowlesi* [2]. Entre elas, *P. falciparum* é responsável pela maioria dos casos graves e óbitos, especialmente em áreas endêmicas [2]. A identificação precoce e precisa dos parasitas é crucial para o tratamento eficaz e a redução da mortalidade associada à malária. No entanto, os métodos tradicionais de diagnóstico, como a microscopia convencional, são limitados por sua dependência da habilidade do operador e pela necessidade de equipamentos laboratoriais específicos, o que pode ser um desafio em regiões de recursos limitados [3].

Diversas aplicações de técnicas de segmentação de imagens e algoritmos classificadores baseados em inteligência artificial foram propostas para aprimorar o desempenho de redes neurais no diagnóstico de malária, abrangendo desde a classificação de parasitas até suas espécies e estágios de desenvolvimento. Dave [5] conduziu uma análise detalhada de passos de processamento de imagens para detecção de parasitas da malária a partir de imagens de esfregaço de sangue espesso. Ele utilizou técnicas como conversão para o espaço de cor matiz, saturação e valor (HSV) e limiarização adaptativa, resultando em uma boa distinção no número de parasitas, com valores de sensibilidade e especificidade de 86,34% e 96,60%, respectivamente.

Sifat e Islam [6] propuseram a identificação das espécies de parasitas da malária e dos estágios do ciclo de vida com

base em imagens de esfregaço de sangue fino. A metodologia proposta utilizou uma combinação de filtros de mediana e média geométrica para aprimorar a qualidade da imagem. Posteriormente, a imagem foi aprimorada aplicando uma técnica de contraste parcial antes de extrair os glóbulos vermelhos (RBC) das imagens de esfregaço de sangue utilizando a técnica de segmentação U-Net. Em seguida, uma rede neural convolucional (CNN) foi empregada para identificar os RBCs infectados. Por outro lado, a rede Visual Geometry Group 16 (VGG16) foi utilizada para distinguir entre as diferentes espécies e o ciclo de vida da malária. Este método proposto foi aplicado a 5512 imagens de malária. Utilizando o modelo CNN, a precisão e especificidade na detecção de RBCs infectados foram de 100% e 95%, respectivamente. Enquanto isso, uma precisão média de 95,55% e especificidade de 94,75% foram alcançadas para a identificação das espécies, e uma precisão geral de 96,25% e especificidade de 94,82% foram obtidas no estágio de vida em anel.

Maqsood et al. [7] analisaram a eficiência de diversos modelos de *deep learning* atuais na detecção de malária usando imagens de esfregaço sanguíneo. Além disso, os autores também propuseram um método eficiente de *deep learning* para classificar células de malária infectadas e não infectadas utilizando a CNN sem o uso de características feitas manualmente. Este estudo empregou técnicas de aumento de imagem e filtragem bilateral para destacar as características dos glóbulos vermelhos (RBCs) antes do treinamento dos conjuntos de dados. Aqui, o modelo CNN modificado é generalizado e o sobreajuste pode ser evitado devido às abordagens de aumento de imagem. A análise foi conduzida em um total de 27.588 conjuntos de dados de malária obtidos do site do Instituto Nacional de Saúde (NIH). Os resultados indicam que o algoritmo proposto pode realizar a detecção de malária com uma precisão de 96,82%.

Gummadi, Ghoosh e Vootla [8] empregaram uma arquitetura de CNN baseada em transfer learning para diferenciar entre imagens de malária não infectadas e infectadas usando agrupamento médio global (GAP - global average pooling) e heat map. Neste estudo, um total de 27.588 imagens de esfregaço sanguíneo fino foram utilizadas em três arquiteturas de redes neurais baseadas em transfer learning, conhecidas como VGG16, VGG19 e InceptionResNetV2. Esses três classificadores foram treinados e comparados. Com base nos resultados do estudo, a arquitetura InceptionResNetV2 obteve uma precisão máxima de 96,88%, com sensibilidade, especificidade, pontuação F1 e precisão de 97,35%, 96,41%, 96,89% e 96,44%, respectivamente.

Razin et al. [9] desenvolveram um modelo para detecção de parasitas da malária implementando o modelo CNN e YOLOv5 para detectar os parasitas da malária e classificar as imagens de malária não infectadas e infectadas. Esta pesquisa foi aplicada a um total de 27.588 imagens de esfregaço sanguíneo fino obtidas do site do NIH, um conjunto de dados de acesso público. De acordo com os resultados experimentais, o modelo YOLOv5 obteve uma precisão de 95,37% para amostras infectadas e 97,05% para amostras não infectadas. Além disso, o modelo CNN pode obter a maior precisão com um valor de 96,21% para amostras de sangue infectadas e não infectadas. Além disso, este modelo CNN também alcançou altas taxas de precisão e recall de 95,42% e 97,05%, respectivamente.

Diante desse cenário, o objetivo deste projeto é desenvolver um sistema de identificação de malária em células utilizando a arquitetura *Vision Transformer* (ViT) [4]. Essa abordagem visa otimizar a detecção da malária, tornando o processo mais eficiente e acessível, especialmente em áreas remotas, onde o acesso a especialistas é escasso. Além disso, os desafios existentes incluem a falta de conjuntos de dados rotulados em quantidade suficiente e a complexidade na identificação de padrões sutis nas imagens das células infectadas. O uso do *Vision Transformer* é especialmente relevante devido à sua capacidade de lidar eficientemente com grandes conjuntos de dados de imagens, capturando relações complexas entre pixels [4]. O projeto também envolve a exploração de técnicas de pré-processamento de imagens, treinamento de modelos de *Machine Learning* e avaliação de desempenho para garantir resultados confiáveis e precisos.

1.1 Estudos Iniciais

Inicialmente, o estudo empregou um conjunto de dados (dataset) do National Institutes of Health (NIH) contendo 27.558 imagens de lâminas de malária préprocessadas e anotadas como ”não infectadas” e ”infectadas” [10]. Experimentos preliminares com Redes Neurais Convolucionais (CNNs) e *Vision Transformers* foram conduzidos para estabelecer um entendimento fundamental da classificação de imagens de parasitas, permitindo uma análise inicial das características distintivas entre espécies e estágios de desenvolvimento do parasita.

Diante dos resultados obtidos nas etapas iniciais, decidiu-se expandir o escopo do projeto para um cenário mais próximo da prática clínica. Assim, um novo conjunto de dados do NIH foi utilizado, composto por imagens de microscopia não segmentadas, ou seja, com células saudáveis e parasitárias coexistindo na mesma imagem. Essa abordagem, embora mais desafiadora, possibilitou o desenvolvimento de um algoritmo capaz de realizar tanto a segmentação de células quanto a classificação de parasitas, simulando o fluxo de trabalho de um laboratório de diagnóstico. A implementação bem-sucedida dessa etapa representa um avanço significativo, pois permite a aplicação do modelo em novos conjuntos de dados, tornando-o uma ferramenta potencial para auxiliar no diagnóstico de malária em diferentes contextos.

2. Métodos

A originalidade deste trabalho reside na aplicação de técnicas avançadas de visão computacional para a análise de imagens de malária. Para garantir a robustez e a acurácia dos modelos propostos, utilizamos um conjunto de dados de alta qualidade proveniente do National Institutes of Health (NIIH) [10]. As imagens neste dataset foram anotadas por especialistas, assegurando a precisão das informações sobre a presença e o tipo

de parasita em cada célula. Essa base de dados cuidadosamente rotulada serve como referência para o treinamento de nossos modelos de *deep learning*, permitindo que eles aprendam a extrair características visuais discriminativas e realizar diagnósticos com alta confiabilidade.

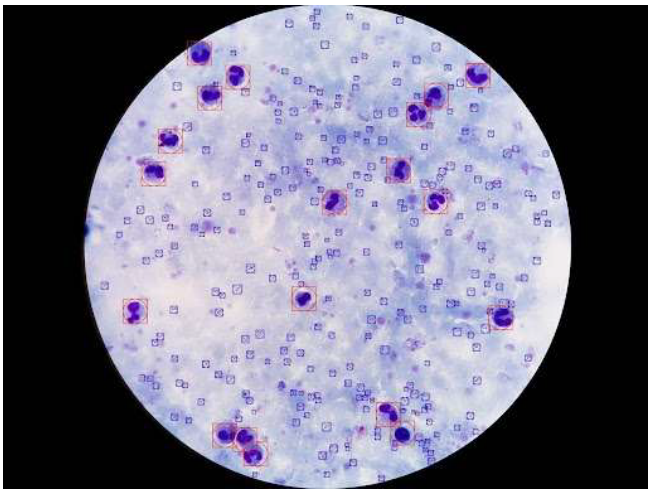

Figure 1. Esfregaço Sanguinio Grosso, com marcação de bounding box

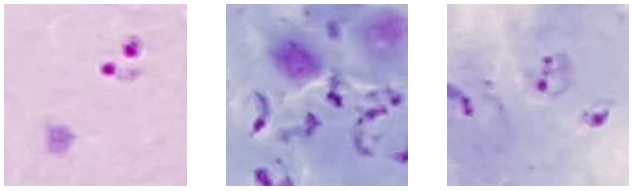

Figure 2. Três imagens representando exemplos da espécie P. falciparum : (a) Primeira imagem, (b) Segunda imagem, (c) Terceira imagem.

Figure 3. Três imagens representando exemplos da espécie *P. vivax* : (a) Primeira imagem, (b) Segunda imagem, (c) Terceira imagem.

Figure 4. Três imagens representando exemplos de células brancas : (a) Primeira imagem, (b) Segunda imagem, (c) Terceira imagem.

O treinamento dos modelos de visão computacional requer a apresentação de um grande número de imagens rotuladas, nas quais as regiões de interesse são delimitadas por bounding boxes, como ilustrado na (Fig.1). Essas anotações fornecem uma supervisão robusta, direcionando o aprendizado dos modelos para as características mais discriminativas. A partir do conjunto de dados, foram segmentadas 10 mil imagens de células para cada gênero: células infectadas pelo *Plasmodium falciparum*, células infectadas pelo *Plasmodium vivax* e células brancas, totalizando 30 mil imagens. Para este projeto, dividimos os dados em 80% para treinamento, 10% para validação e 10% para teste. Essas anotações fornecem uma supervisão robusta, direcionando o aprendizado dos modelos para as características mais discriminativas. Com base em anotações precisas, como as mostradas nas Fig 2(a)-(c), Fig 3(a)-(c), Fig 4(a)-(c), garantimos que os modelos aprendam a identificar os parasitas com alta acurácia. Isso contribui para a obtenção de diagnósticos mais precisos.

A presente pesquisa contou com a utilização de um conjunto diversificado de bibliotecas de programação, cada qual com um papel específico no desenvolvimento e implementação dos modelos de aprendizado profundo. O ecossistema PyTorch, composto pelas bibliotecas Torch, torchvision e vit-pytorch, proporcionou uma plataforma flexível e poderosa para a implementação de arquiteturas de redes neurais complexas, como os *Vision Transformers*. Por fim, a biblioteca OpenCV foi empregada para o pré-processamento das imagens, oferecendo um conjunto abrangente de ferramentas para manipulação e análise de imagens digitais. Para o treinamento e avaliação dos modelos de Visão Computacional aplicados à classificação de imagens de malária, foram utilizadas várias técnicas e hiperparâmetros para otimizar o desempenho e garantir a robustez dos modelos. Nesta seção, descrevemos em detalhes os métodos e hiperparâmetros utilizados, com foco especial nos mecanismos de ajuste de taxa de aprendizado (*ReduceLROnPlateau*), parada antecipada (*Early Stopping*) e salvamento de pontos de verificação do modelo (*Model Checkpoint*).

A taxa de aprendizado é um dos hiperparâmetros mais críticos em algoritmos de otimização. No contexto de redes neurais, uma taxa de aprendizado muito alta pode causar um comportamento instável e impedir a convergência do modelo, enquanto uma taxa muito baixa pode tornar o treinamento muito lento e potencialmente levar a uma convergência para um mínimo local não ótimo. Para mitigar esses problemas, uti-

lizamos a técnica de ajuste de taxa de aprendizado conhecida como *ReduceLROnPlateau* [11].

O *ReduceLROnPlateau* é uma técnica de agendamento da taxa de aprendizado que reduz automaticamente a taxa de aprendizado quando uma métrica específica para de melhorar. Em nosso caso, a métrica monitorada foi a perda de validação. A lógica por trás dessa abordagem é simples: se o modelo não está melhorando na métrica de validação, isso pode ser um sinal de que a taxa de aprendizado atual é muito alta e está impedindo a convergência. Reduzindo a taxa de aprendizado, o modelo tem uma chance melhor de encontrar o mínimo global da função de perda. A técnica de *Early Stopping* [12] é utilizada para evitar o overfitting durante o treinamento do modelo. Overfitting ocorre quando o modelo aprende demasiado bem os detalhes e o ruído do conjunto de treinamento, a ponto de seu desempenho em dados não vistos (como o conjunto de validação ou teste) piorar. A parada antecipada interrompe o treinamento quando a métrica monitorada para de melhorar por um número especificado de épocas, conhecido como paciência. No nosso experimento, a métrica de validação foi monitorada e, se não houvesse melhora após um certo número de épocas (definido pela paciência), o treinamento era interrompido. Este número foi definido como 10 épocas. A *Early Stopping* não apenas ajuda a evitar o overfitting, mas também economiza tempo de computação, já que o treinamento é interrompido assim que o modelo para de melhorar.

A biblioteca ViT-PyTorch oferece uma variedade de arquiteturas de *Vision Transformers* (ViT) projetadas para lidar com diferentes escalas de dados e complexidades de tarefas. Neste trabalho, exploramos as arquiteturas ViT padrão [4], *Patch Merger* [13], e *ViT for Small Datasets* [14], selecionadas por sua eficácia em tarefas de classificação de imagens. Essas arquiteturas são notáveis por sua capacidade de capturar longas dependências espaciais e por sua adaptabilidade a diferentes tamanhos de imagem, tornando-as particularmente adequadas para a análise de imagens médicas.

3. Resultados e Discussão

Após os resultados iniciais, conseguimos aperfeiçoar o modelo, aprimorando hiperparâmetros e implementando callbacks com eficiência. Realizamos comparação de desempenho com diferentes dimensões de imagens e avaliamos os modelos por meio da acurácia e perda dos conjuntos de treino, validação e teste. Conforme os gráficos das fig 2, 3, 4, 5 e a tabela da figura 6.

O gráfico da Fig: 5 (esquerda) mostra a perda de treinamento e validação ao longo das épocas com a dimensão de imagens [110,110]. Observa-se uma redução consistente da perda de treinamento, começando em aproximadamente 0,40 na época 0 e diminuindo para cerca de 0,05 na última época. A perda de validação, embora inicialize em um valor próximo a 0,20, exibe uma flutuação mais pronunciada ao longo das épocas, refletindo uma variabilidade maior em comparação com a perda de treinamento.

O gráfico da Fig: 5 (direita) ilustra a acurácia de treinamento e validação ao longo das épocas. Nota-se uma melhoria constante na acurácia de treinamento, iniciando em cerca de 84% na primeira época e alcançando aproximadamente 98% na última época. A acurácia de validação também mostra uma tendência de aumento, embora com flutuações, atingindo valores próximos aos 97% nas épocas finais.

Conforme ilustrado no gráfico da Fig: 6 (esquerda), a perda de treinamento diminui consistentemente de aproximadamente 0,35 na época 0 para cerca de 0,05 na última época. A perda de validação, que começa em cerca de 0,20, exibe flutuações mais notáveis ao longo das épocas, similarmente aos outros modelos.

O gráfico da Fig: 6 (direita) retrata a acurácia de treinamento e validação ao longo do treinamento. A acurácia de treinamento cresce de cerca de 84% na primeira época para aproximadamente 98% na última época. A acurácia de validação também melhora ao longo do tempo, apesar das oscilações, atingindo valores próximos a 97-98% nas épocas finais.

No gráfico da Fig: 7 (esquerda), a perda de treinamento inicia em aproximadamente 0,40 e reduz consistentemente ao longo das épocas, chegando a cerca de 0,05. A perda de validação, similarmente ao modelo ViT, começa em torno de 0,20 e apresenta flutuações durante o treinamento, mas demonstra uma tendência de redução global.

O gráfico da Fig: 7 (direita) mostra a acurácia de treinamento e validação ao longo do tempo. A acurácia de treinamento aumenta de aproximadamente 84% na primeira época para cerca de 98% na última. A acurácia de validação também segue uma tendência de melhoria, com oscilações, alcançando valores próximos a 96-97% nas épocas finais.

Os gráficos da Fig: 8 apresentados ilustram a evolução da perda de treinamento (*Train Loss*) e da acurácia de treinamento (*Train Accuracy*) ao longo das épocas para três modelos diferentes: *Vision Transformer* (ViT), *Vision Transformer for Small Datasets* e *Patch Merger*, utilizando imagens de dimensão [150, 150]. No Gráfico da Fig: 8 (superior), que representa a perda de treinamento ao longo das épocas, observa-se uma redução contínua da perda para todos os modelos, com o *Vision Transformer for Small Datasets* demonstrando uma superioridade clara, apresentando as menores perdas na maioria das épocas, seguido pelo ViT e, por último, o *Patch Merger*. This tendência suggests that the *Vision Transformer for Small Datasets* is more efficient in minimizing the perda funcionario during the trinamento.

No Gráfico da Fig: 8 (inferior), que ilustra a acurácia de treinamento ao longo das épocas, todos os modelos apresentam um aumento constante da acurácia, convergindo para valores elevados à medida que as épocas avançam. O *Vision Transformer for Small Datasets* novamente lidera em desempenho, alcançando uma acurácia ligeiramente superior à dos outros modelos, seguido de perto pelo ViT, enquanto o *Patch Merger* apresenta um desempenho um pouco inferior.

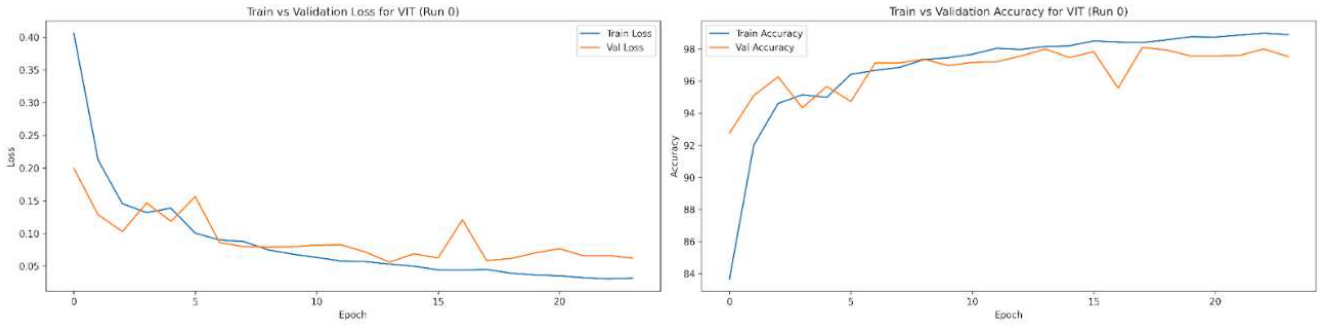

Figure 5. Resultados de avaliação(perda na esquerda e acurácia na direita) da arquitetura ViT [110,110]. Dados de treino em cor azul e dados de validação em cor laranja.

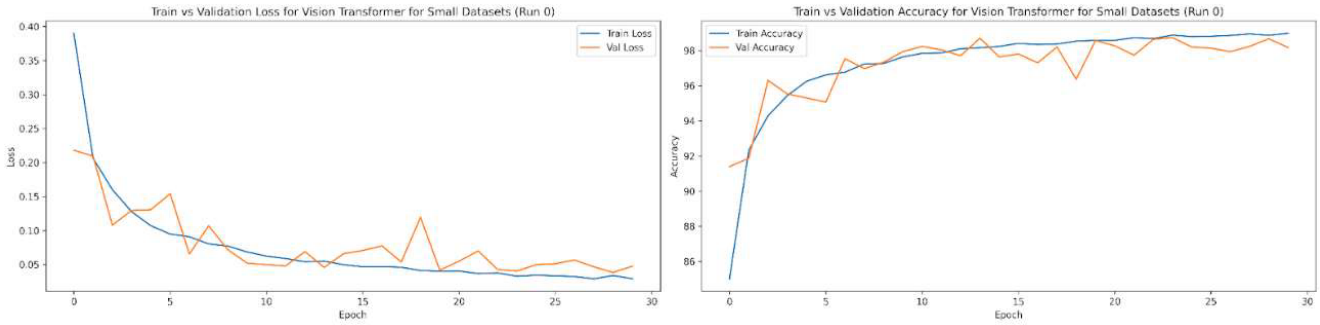

Figure 6. Resultados de avaliação(perda na esquerda e acurácia na direita) da arquitetura *Vision Transformers for Smalls Datasets* [110,110]. Dados de treino em cor azul e dados de validação em cor laranja.

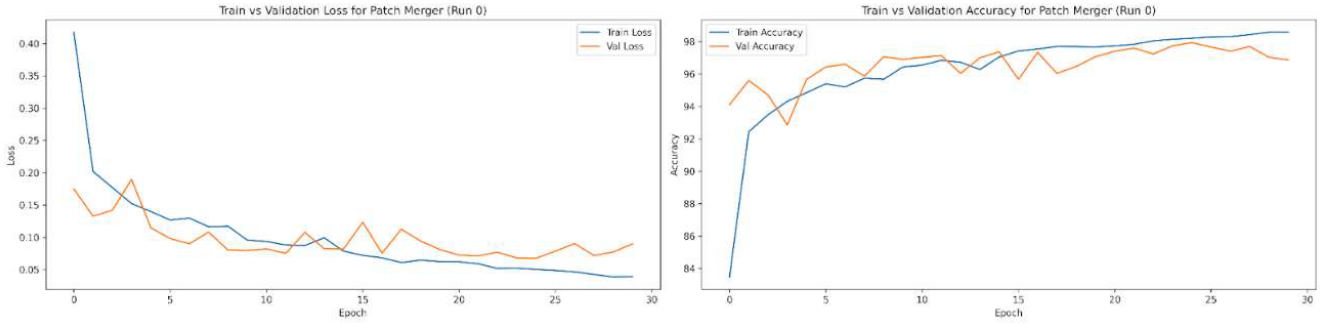

Figure 7. Resultados de avaliação(perda na esquerda e acurácia na direita) da arquitetura *Patch Merger* [110,110]. Dados de treino em cor azul e dados de validação em cor laranja.

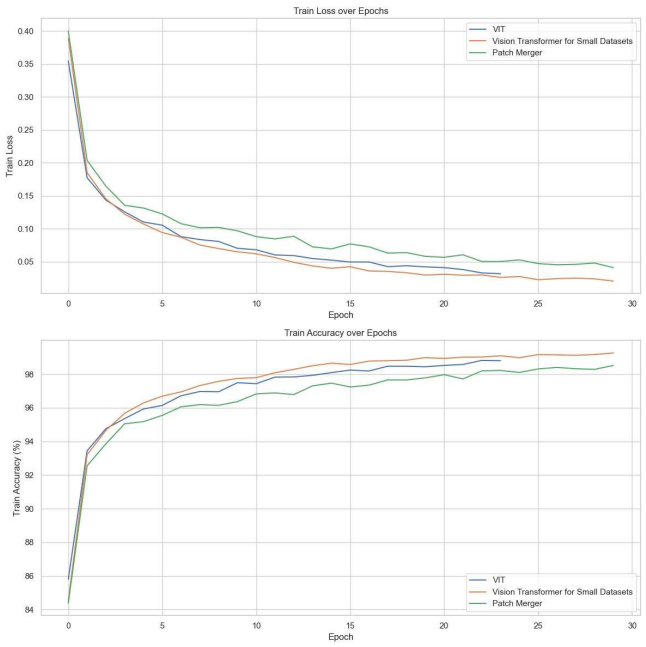

Figure 8. Resultados de avaliação por arquitetura(perda superior e acurácia inferior) para a dimensão de imagem [150,150].

Esses resultados indicam que o *Vision Transformer for Small Datasets* não apenas minimiza a perda de maneira mais eficaz, mas também atinge uma acurácia mais alta em comparação com os outros modelos, confirmando sua adequação para a tarefa de classificação de células infectadas por malária com imagens de dimensão [150, 150].

Modelo	Dimensão da Imagem	Tempo de Treinamento (min)	Épocas de Treino	Acurácia de Teste (%)	Loss de Teste (%)
Vision Transformer (ViT)	[50, 50]	13.2	30	98.73	0.0303
Vision Transformer (ViT	[110, 110]	41.18	30	97.57	0.0717
Vision Transformer (ViT)	[150, 150]	79.27	24	97.57	0.0881
Vision Transformer for Small Datasets	[50, 50]	13.84	30	98.77	0.0410
Vision Transformer for Small Datasets	[110, 110]	46.36	30	98.43	0.0496
Vision Transformer for Small Datasets	[150, 150]	122.36	30	98.97	0.0246
Patch Merger	[50, 50]	13.29	30	98.00	0.0553
Patch Merger	[110, 110]	31.28	24	94.07	0.1767
Patch Merger	[150, 150]	99.67	30	97.70	0.0775

Table 1. Resultados de avaliação dos dados de Teste

A Tabela 1 apresenta uma comparação entre diferentes configurações do modelo *Vision Transformer* (ViT) e suas variantes, incluindo o *Vision Transformer for Small Datasets* e o *Patch Merger*, aplicados na tarefa de classificação de malária. As variáveis analisadas incluem a dimensão da imagem de entrada, o tempo de treinamento, o número de épocas de treino, a acurácia e o valor da função de perda (loss). Os resultados indicam que o modelo *Vision Transformer for Small Datasets* com imagens de dimensão [150, 150] apresentou a melhor acurácia, alcançando 98.97% com uma perda de 0.0246. O tempo de treinamento para essa configuração foi o mais alto entre todos os modelos testados, totalizando 122.36 minutos. O *Vision Transformer* (ViT) tradicional também obteve bons resultados, especialmente para imagens de dimensão [50, 50], com uma acurácia de 98.73e a menor perda de 0.0303, necessitando de apenas 13.2 minutos para o treinamento.

O modelo *Patch Merger* mostrou uma variação maior na acurácia e perda dependendo da dimensão da imagem. Com imagens de [110, 110], a acurácia foi a menor entre todas as configurações testadas, atingindo apenas 94.07% com uma perda de 0.1767. Esses resultados sugerem que, embora o aumento da dimensão da imagem possa melhorar a acurácia em alguns modelos, ele também pode aumentar significativamente o tempo de treinamento e nem sempre resulta em melhorias consistentes na performance do modelo. O *Vision Transformer for Small Datasets* demonstrou ser uma opção robusta para classificação de malária, especialmente com imagens de maiores dimensões.

Este estudo representa um avanço significativo na aplicação das arquiteturas *Vision Transformers* (ViTs) para a classificação de malária em tecidos celulares. Os resultados se demonstram promissores e sinalizam o potencial dessas arquiteturas em tarefas desafiadoras de visão computacional, principalmente na detecção de padrões complexos em imagens de microscopia.

Os resultados obtidos neste estudo destacam a eficácia dos modelos *Vision Transformer* (ViT) e suas variantes na classificação de células infectadas por malária. A análise comparativa mostrou que o modelo *Vision Transformer for Small Datasets*, especialmente com imagens de dimensão [150, 150], alcançou a melhor acurácia, atingindo 98.97% com uma perda de 0.0246. Este desempenho robusto, demonstra que os *Vision Transformers* são altamente competentes para tarefas de classificação de imagens médicas, mesmo em cenários onde a quantidade de dados pode ser limitada. A análise dos gráficos de evolução da perda e acurácia ao longo das épocas confirma ainda mais essa competência, evidenciando que o *Vision Transformer for Small Datasets* não apenas minimiza a função de perda de maneira mais eficaz, mas também atinge uma acurácia superior em comparação com os outros modelos testados.

No âmbito da saúde pública, a implementação de modelos *Vision Transformers* pode revolucionar a triagem e diagnóstico de malária em áreas endêmicas. Hospitais e centros de saúde podem adotar essas soluções para agilizar o diagnóstico, auxiliando os profissionais de saúde e melhorando a precisão dos diagnósticos. Além disso, programas de saúde pública podem utilizar essa tecnologia para monitorar e controlar surtos de

malária de forma mais eficaz, permitindo intervenções mais rápidas e direcionadas. A capacidade de realizar diagnósticos precisos e rápidos em campo, utilizando dispositivos portáteis equipados com modelos *Vision Transformer*, pode ser particularmente valiosa em regiões remotas com acesso limitado a laboratórios sofisticados.

Empresas de tecnologia médica e startups podem explorar essas soluções para desenvolver dispositivos portáteis de diagnóstico que utilizam *Vision Transformers* para detectar malária de maneira rápida e precisa. Laboratórios de análises clínicas também podem integrar esses modelos em suas operações para aumentar a eficiência e reduzir o tempo de espera para resultados. Além disso, essas tecnologias podem ser incorporadas em plataformas de telemedicina, permitindo diagnósticos remotos e acessíveis em regiões carentes. A precisão e velocidade desses modelos podem também abrir oportunidades para serviços de diagnóstico como serviço, onde imagens médicas são processadas na nuvem para fornecer resultados rápidos e precisos a clientes em todo o mundo.

Este estudo abre várias direções para futuras pesquisas. A aplicação dos *Vision Transformers* pode ser expandida para a classificação de outras doenças infecciosas, como tuberculose e pneumonia, utilizando imagens médicas. A adaptação e aprimoramento dos modelos para lidar com conjuntos de dados ainda menores ou mais desequilibrados também são áreas promissoras. Por exemplo, a integração de técnicas de data augmentation mais avançadas e métodos de transferência de aprendizado pode melhorar ainda mais a performance dos modelos em situações com dados limitados.

Considerar dados multimodais pode enriquecer ainda mais as capacidades dos modelos de classificação. Além das imagens de microscopia, incorporar informações adicionais, como dados clínicos, históricos de pacientes ou dados genéticos, pode proporcionar um contexto mais abrangente e melhorar a precisão dos diagnósticos. A fusão de múltiplas fontes de dados pode resultar em modelos mais robustos e completos, que capturam uma gama mais ampla de sinais relevantes para o diagnóstico da malária. Além disso, a integração de *Vision Transformers* com outras técnicas de aprendizado profundo, como redes neurais convolucionais (CNNs), pode potencialmente combinar o melhor de ambos os mundos, aumentando ainda mais a precisão e a robustez dos modelos. Outra área de interesse é a exploração de *Vision Transformers* em outras áreas da saúde, como a análise de imagens de ressonância magnética (MRI) e tomografia computadorizada (CT) para diagnósticos mais complexos. A criação de modelos híbridos que utilizam tanto Transformers quanto CNNs pode oferecer novas soluções para problemas de classificação em domínios com grandes volumes de dados visuais complexos. Tais perspectivas poderão direcionar novos estudos e projetos na área da biologia e saúde, considerando a visão computacional uma ferramenta importante para evoluir diagnósticos com precisão e eficiência.

A arquitetura de *Vision Transformers* se demonstrou bastante promissora e poderá proporcionar soluções reais, conciliando o desenvolvimento tecnológico e aplicações que proporcionam bem estar e preservação da saúde pública, evitando inclusive surtos endêmicos de doenças como a malária. Em resumo, este estudo demonstra que os *Vision Transformers* são uma ferramenta poderosa para a classificação de imagens médicas e possuem um potencial significativo para transformar várias indústrias. A continuação da pesquisa nesta área promete não apenas melhorar os sistemas de diagnóstico existentes, mas também abrir novas oportunidades para a inovação tecnológica em diversos setores. O avanço na aplicação de *Vision Transformers* pode levar a uma nova era de diagnósticos médicos mais precisos e acessíveis, bem como melhorias significativas em outras áreas que dependem da análise precisa de imagens.

Agradecimentos

Este projeto foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei no 8.248, de 23 de outubro de 1991, no âmbito do PPI-SOFTEX, coordenado pela Softex e publicado PDI 03, DOU 01245.023862/2022-14.

Contribuições dos autores

Os autores Eduardo David Campopiano, Iago Fonseca Marinho Pereira e Thomas William, desempenharam papéis fundamentais na implementação do sistema de classificação de malária, utilizando a arquitetura Vision Transformer. Eles foram responsáveis pela aquisição do conjunto de dados e pré-processamento das imagens de microscopia celular, análise de dados, otimização e avaliação do modelo, além de documentar o progresso do projeto e redigir as seções do manuscrito. Tal colaboração foi essencial para a validação do desempenho do sistema, garantindo resultados precisos e confiáveis.

Em complemento, os autores Douglas Henrique Siqueira Abreu, Guilherme Ribeiro Sales, Valentino Corso e Cides S. Bezerra ofereceram orientação estratégica e suporte técnico ao longo do desenvolvimento do projeto. Eles contribuíram com *expertise* na definição dos objetivos da pesquisa, revisão das metodologias e análise dos dados, além de proporcionar feedback contínuo. Está presença foi crucial para o sucesso do trabalho, permitindo que os autores executores tomassem decisões informadas e garantissem a qualidade do manuscrito final.

References

[1] World Health Organization, "World Malaria Report 2021," Geneva, Switzerland: WHO, 2021.

[2] N. J. White and J. G. Breman, "Malaria Parasites," in *Encyclopedia of Malaria*, J. G. Breman, M. S. Alilio, and N. J. White, Eds. Springer, 2013.

[3] M. Amexo, R. Tolhurst, G. Barnish, and I. Bates, "Malaria misdiagnosis: effects on the poor and vulnerable," *The Lancet*, vol. 364, no. 9448, pp. 1896-1898, 2004.

[4] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., ”An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *arXiv preprint arXiv:2010.11929*, 2020.

[5] I. R. Dave, ”Image analysis for malaria parasite detection from microscopic images of thick blood smear,” in *Proceedings of the 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, Mar. 22–24, 2017.

[6] M. M. H. Sifat and M. M. Islam, ”A Fully Automated System to Detect Malaria Parasites and their Stages from the Blood Smear,” in *Proceedings of the 2020 IEEE Region 10 Symposium (TENSYMP)*, Dhaka, Bangladesh, Jun. 5–7, 2020.

[7] A. Maqsood, M. S. Farid, M. H. Khan, and M. Grzegorzek, ”Deep malaria parasite detection in thin blood smear microscopic images,” *Applied Sciences*, vol. 11, no. 5, p. 2284, 2021. [Online]. Available: ⟨https://www.mdpi.com/2076-3417/11/5/2284⟩

[8] S. D. Gummadi, A. Ghosh, and Y. Vootla, ”Transfer Learning based Classification of *Plasmodium Falciparum* Parasitic Blood Smear Images,” in *Proceedings of the 2022 10th International Symposium on Digital Forensics and Security (ISDFS)*, Istanbul, Turkey, Jun. 6–7, 2022.

[9] W. R. W. M. Razin, T. S. Gunawan, M. Kartiwi, and N. M. Yusoff, ”Malaria Parasite Detection and Classification using CNN and YOLOv5 Architectures,” in *Proceedings of the 2022 IEEE 8th International Conference on Smart Instrumentation, Measurement and Applications (ICSIMA)*, Melaka, Malaysia, Sep. 26–28, 2022.

[10] Centro Nacional de Informação sobre Biotecnologia (s.d.). Dataset de imagens de malária. Recuperado de ⟨https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/image-processing/malaria-datasheet.html⟩.

[11] PYTORCH. `ReduceLROnPlateau`. Disponível em: ⟨https://pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.ReduceLROnPlateau.html⟩. Acesso em: 23 jul. 2024.

[12] PYTORCH. EarlyStopping. Disponível em: ⟨https://pytorch.org/docs/stable/generated/torch.nn.EarlyStopping.html⟩. Acesso em: 23 jul. 2024.

[13] J. Wang, W. Wang, X. Yang, and W. Zhang, ”Attention-based Temporal Context Aggregation for Video Action Recognition,” arXiv preprint arXiv:2202.12015, 2022. [Online]. Available: ⟨https://arxiv.org/abs/2202.12015⟩.

[14] X. Zhang, Y. Zhang, Z. Zhang, and Y. Liu, ”Dual Encoder Network for Zero-Shot Video Retrieval,” arXiv preprint arXiv:2112.13492, 2021. [Online]. Available: ⟨https://arxiv.org/abs/2112.13492⟩.