Soft Voting Vision Transformer Ensemble for Acute Myeloid Leukemia Cellular Classification

Marcelo H. de Almeida Lira, Elloá B. Guedes

[1]Grupo de Pesquisa em Sistemas Inteligentes
Universidade do Estado do Amazonas
Av. Darcy Vargas, 1200 – Manaus – Amazonas
{mhdal.snf23, ebgcosta}@uea.edu.br

Abstract. *This work proposes a comparative study among three Vision Transformer (ViT) models for blood cell classification aimed at the detection of Acute Myeloid Leukemia (AML). To this end, a literature benchmark consisting of single-cell images from blood smears was utilized. Data augmentation and ensemble learning techniques with soft voting were applied to enhance model robustness. The results demonstrate that the ensemble of the three ViTs with data augmentation exhibited the best performance across all evaluated metrics, achieving an average accuracy of* $96.95\,\%$ *and an average* F_1*-score of* $96.42\,\%$*. These results outperform the state-of-the-art in ViT-based architectures and deliver competitive performance against Convolutional Neural Network (CNN) models while drastically reducing the number of parameters used (from* 9 Bi *to* 113.8 Mi*).*

Resumo. *Este trabalho propõe um estudo comparativo entre três modelos de Vision Transformers (ViTs) para a classificação de células sanguíneas com vistas a detecção de Leucemia Mieloide Aguda (LMA). Para tanto, usou um* benchmark *da literatura composto por imagens de células isoladas em esfregaços de sangue. As técnicas de aumento de dados e* Ensemble Learning *com* soft voting *foram aplicadas para aumentar a robustez dos modelos. Os resultados demonstram que o comitê dos três ViTs com aumento de dados exibiu o melhor desempenho em todas as métricas calculadas, atingindo uma acurácia média de* $96{,}95\,\%$ *e* F_1-Score *médio de* $96{,}42\,\%$*. Estes resultados superam o estado da arte de arquiteturas baseadas em ViTs e apresentam um desempenho competitivo com modelos baseados em Redes Neurais Convolucionais (CNNs) enquanto reduzem drasticamente o número de parâmetros utilizados (de* 9 Bi *para* $113{,}8$ Mi*).*

1. Introdução

Em 2021, houve $460\,000$ novos casos de leucemia e $320\,000$ óbitos em decorrência dessa doença em âmbito global, figurando entre os cânceres de maior incidência e mortalidade no mundo [Hu et al. 2025, Du et al. 2022]. Tratando-se especificamente da Leucemia Mieloide Aguda (LMA), evidências demonstram um crescimento de $82\,\%$ nas ocorrências entre 1990 e 2021, afetando predominantemente populações idosas e apresentando taxas críticas de mortalidade infantil [Zhou et al. 2024]. O cenário brasileiro também reflete essa escalada alarmante, pois o Instituto Nacional de Câncer (INCA) estima $12\,220$ diagnósticos para cada ano do triênio 2026-2028 [INCA 2023], o que representa um aumento de $21\,\%$ em relação aos dados de 2016 reportados pela Abrale

[Associação Brasileira de Linfoma e Leucemia (ABRALE) 2026]. Mais grave ainda é o impacto da LMA, considerada o subtipo mais letal da doença. Conforme o Observatório de Oncologia, $36\,\%$ dos $63\,452$ óbitos por leucemia registrados no Brasil entre 2008 e 2017 foram causados pela LMA. Esses índices reforçam que a patologia constitui um grave problema de saúde pública, demandando o desenvolvimento de metodologias que otimizem a identificação de blastos leucêmicos.

Na LMA, mutações nas células-tronco mieloides interrompem o processo de diferenciação celular, resultando na formação de blastos imaturos. A proliferação descontrolada dessas células na medula óssea compromete a hematopoiese normal, tornando a análise microscópica de esfregaços de sangue uma etapa fundamental para a detecção da doença. Apesar de essencial, a classificação manual dessas células sanguíneas é um processo exaustivo e inerentemente subjetivo, com estudos indicando uma taxa de concordância de $76{,}6\,\%$ entre médicos especialistas [Bigorra et al. 2017]. Além da dependência de profissionais cuja formação exige longo tempo de capacitação, a lentidão do método tradicional pode atrasar o início do tratamento. Em um cenário em que a rapidez do diagnóstico é crucial para a eficácia do prognóstico, consolida-se a necessidade de adotar ferramentas automatizadas para otimizar essa análise clínica.

Frente a essas limitações, a Microscopia Inteligente tem sido explorada como alternativa tecnológica. A literatura propõe o uso de Redes Neurais Convolucionais (CNNs) na assistência ao diagnóstico médico [Sarvamangala and Kulkarni 2022], uma abordagem consolidada na análise de imagens histopatológicas e na classificação de células sanguíneas. A detecção da LMA por meio de leucócitos insere-se nessa tendência e, para este fim, destacam-se o emprego de CNNs pré-treinadas via *Transfer Learning* [Elhassan et al. 2024], o desenvolvimento de arquiteturas customizadas [Yadav et al. 2023] e a elaboração de comitês híbridos para otimizar a triagem celular [Moriconi et al. 2024].

Como alternativa às tradicionais CNNs, Dosovitskiy *et al.* [2021] propuseram o *Vision Transformer* (ViT), transpondo para a Visão Computacional a arquitetura baseada em atenção introduzida por Vaswani *et al.* [2017] no Processamento de Linguagem Natural (PLN). Atualmente, os ViTs apresentam desempenho competitivo ou superior frente às redes convolucionais em diversas tarefas. Conforme apontado por Maurício *et al.* [2023], esses modelos destacam-se pela elevada robustez e pelo potencial de maior eficiência computacional. Apesar dessas vantagens, o emprego de ViTs na Microscopia Inteligente voltada ao diagnóstico da LMA permanece limitado. Tal contexto configura uma lacuna na literatura, justificando a investigação dessa abordagem arquitetural para a mitigação do problema abordado.

Desta feita, o presente trabalho avalia a utilização de ViTs na tarefa de classificação de 15 tipos de células sanguíneas em um *benchmark* consolidado da literatura para LMA. Os resultados obtidos demonstraram a robustez das arquiteturas de ViTs escolhidas e a eficácia da combinação no uso das estratégias de aumento de dados e *Ensemble Learning* para a otimização das métricas de desempenho, em especial na estabilização do aprendizado e na redução de viéses individuais das redes.

Para apresentar o que se propõe, o artigo está organizado como segue: os trabalhos relacionados encontram-se dispostos na Seção 2; a visão geral da solução e a metodologia

para sua obtenção são descritos na Seção 3; os resultados experimentais são apresentados e discutidos na Seção 4; por fim, as conclusões e perspectivas futuras são dispostas na Seção 5.

2. Trabalhos Relacionados

Para compreender de forma abrangente as metodologias das soluções existentes, identificar tendências e lacunas, realizou-se uma análise dos resultados de buscas no Google Acadêmico entre 2023 e 2025. Os critérios de inclusão contemplaram trabalhos baseados em Inteligência Artificial para detecção de células de LMA utilizando a base de dados *AML Cytomorphology LMU* como fonte de experiência, por ser um *benchmark* robusto para a avaliação de soluções inteligentes em um cenário realístico da tarefa.

O trabalho de Yadav *et al.* [2023] propôs uma nova arquitetura de CNN para a classificação de células brancas, contemplando todas as classes inclusas no *dataset* original. No modelo proposto, uma CNN multi-entradas, considerou-se: a imagem original e dois descritores de imagens (Histograma de Gradiente e o Padrão Binário Local). Estas entradas são processadas paralelamente para novas etapas de extração automática de características pela CNN para que, ao final deste processo, seja realizada a fusão das características. Este resultado, então, é usado por uma rede neural *MultiLayer Perceptron* para a etapa de classificação. Esta arquitetura atingiu acurácia geral de $98{,}16\,\%$ ao custo de $9\,$Bi de parâmetros treináveis, além do custo computacional inerente à etapa de pré-processamento das entradas, o que representa um *tradeoff* sensível para aplicações de tempo real ou utilização em dispositivos com requisitos computacionais limitados.

Elhassan *et al.* [2024] avaliaram um *pipeline* diferente para abordar este problema: o *CAE-ResVGG FusionNet*. Este *framework* utiliza um *Autoencoder* Convolucional (CAE) e duas CNNs pré-treinadas, VGG19 e ResNet50, para fazer a classificação de quatro tipos de células brancas imaturas. A técnica de *data augmentation* é utilizada nas classes de blastos para fins de balanceamento. Primeiramente, as imagens passam pelo *Autoencoder* para a geração de representações internas de menor dimensionalidade que mantenham características relevantes. Estas representações são usadas como entrada dos modelos pré-treinados, que geram dois vetores de características para serem integrados em um único *set*. Então uma classificação binária é realizada para identificar células imaturas e uma classificação multiclasse determina o subtipo da célula. Para escolher a saída definitiva, uma votação ponderada é realizada nos resultados das duas CNNs. O trabalho alcançou uma acurácia geral de $99{,}99\,\%$, no entanto, assim como no trabalho anterior, este modelo faz uso de redes computacionalmente custosas, com a VGG19 que possui $140\,$Mi de parâmetros e a ResNet50, com $23\,$Mi de parâmetros. Além disso, a classificação das células brancas restringe-se a apenas quatro classes, limitando o uso deste *framework* em situações realísticas.

Aplicando outro tipo de classificador no problema, Lu *et al.* [2024] utilizaram um *Masked AutoEncoder* (MAE) juntamente a um ViT para a previsão das 15 classes. Além disso, para selecionar a amostra das imagens que seriam usadas como rótulos, a técnica de *Active Learning* foi aplicada, nomeando esta proposta de *MAE4AL*. O MAE é usado para aprender as características das imagens "mascarando-as" para então reconstruí-las, guardando-se a diferença entre a entrada e a saída por meio de uma função de perda. Então, o ViT do MAE é usado para treinar o ViT classificador, que determina as classes e

gera a entropia, incerteza da classe prevista, para ser armazenada. A perda do MAE e a entropia do ViT são usadas para definir as amostras mais valiosas para rótulo. Depois, a rede de classificação refaz suas predições com os novos dados rotulados. Estas etapas se repetem até que o número de imagens rotuladas atinja um determinado limiar. Usando apenas metade dos dados rotulados, *MAE4AL* conseguiu uma acurácia de $96{,}36\,\%$. Embora os autores não tenham especificado o tamanho final do modelo, sabe-se que ele é superior ao de um ViT base, o qual possui ao menos 86 Mi de parâmetros treináveis.

Moriconi *et al.* [2024] fizeram uso de oito redes pré-treinadas e de diversas técnicas, incluindo *data augmentation*, para fazer a classificação de leucócitos imaturos, que representam apenas 4 das 15 classes do problema. Além do *benchmark* de referência, foi utilizada experiência oriunda de outro *dataset* obtido junto ao Hospital Clínico de Barcelona, totalizando 5738 imagens, contabilizando apenas aquelas referentes às quatro classes de leucócitos imaturos. O método proposto consiste no treinamento e avaliação de desempenho individual de oito modelos de redes neurais selecionados pelos autores, das quais as três redes com maior acurácia irão compor um comitê classificador final. No *dataset AML Cytomorphology LMU*, os autores consideraram apenas 1443 imagens das quatro classes de células imaturas, em que a acurácia obtida foi de $82{,}05\,\%$. Não houve clareza quanto ao tamanho do comitê final, podendo estar limitado inferiormente a 44 Mi de parâmetros.

A Tabela 1 apresenta uma síntese comparativa dos trabalhos elencados anteriormente, ressaltando que o tamanho dos modelos em alguns casos é apenas estimado. A partir dessa revisão da literatura percebe-se que as soluções concentram-se principalmente no uso de CNNs e que modelos do estado da arte, como ViTs, ainda são pouco explorados na tarefa em questão. Além dos quesitos de reprodutibilidade e clareza no tamanho final dos modelos, aspectos críticos para implantação de soluções dessa natureza em ambientes realísticos, a lacuna mencionada motivou a realização deste trabalho.

Tabela 1: Comparação entre os trabalhos correlatos e o artigo proposto.

Trabalho	Modelo/Estratégia	Classes previstas	Imagens	Tamanho	Acurácia
[Yadav et al. 2023]	Deep CNN	15 classes	18 365	9 Bi	98,16 %
[Elhassan et al. 2024]	CAE + CNN hierárquica	4 classes	18 365	140 Mi	99,99 %
[Lu et al. 2024]	MAE + ViT + AL	15 classes	18 365	86 Mi	96,36 %
[Moriconi et al. 2024]	Comitê de CNNs	4 classes	1443	44 Mi	82,05 %

3. Material e Métodos

Visando apoiar no diagnóstico automatizado da LMA, o presente trabalho se propõe a classificar imagens microscópicas de esfregaços de sangue em quinze classes diferentes, das quais quatro contemplam células imaturas (ou blastos) cuja presença elevada pode ser um indicativo da doença. Deste modo, a identificação automática dessas células pode acelerar uma etapa importante do processo de diagnóstico.

A presente solução encontra-se no escopo da Microscopia Automatizada e Inteligente, cujo objetivo é colaborar diretamente no diagnóstico e prognóstico de patologias, mitigando a variabilidade intra e interobservador, além de aprimorar a objetividade e a reprodutibilidade de exames médicos. Para alcançar esse propósito, métodos automatizados baseados em *Deep Learning*, como é o caso dos ViTs a serem investigados no escopo

desse trabalho, são preferíveis em relação à avaliação manual, tanto pela velocidade de aquisição quanto pela redução no acúmulo de erros [Melanthota et al. 2022].

3.1. Dados Experimentais

Os dados utilizados como base para este estudo provêm do *dataset AML Cytomorphology LMU*, composto por 18 365 imagens de células individuais obtidas a partir de esfregaços de sangue periférico de 200 pacientes, dos quais 100 foram diagnosticados com LMA no Hospital Universitário de Munique entre 2014 e 2017. As imagens foram adquiridas com o microscópio M8 (Precipoint GmbH, Freising, Alemanha), com ampliação de 100x e imersão em óleo, e posteriormente rotuladas por especialistas. Segundo informações do provedor, a resolução dessas imagens é de 14,14 pixels por micrômetro. Cabe destacar que, nessa base de dados, as imagens não estão associadas a identificadores do paciente de origem, o que pode ensejar *data leakage*. Exemplos ilustrativos da base de dados podem ser encontrados na Figura 1, em que as classes MOB, MYO, PMB e PMO denotam blastos.

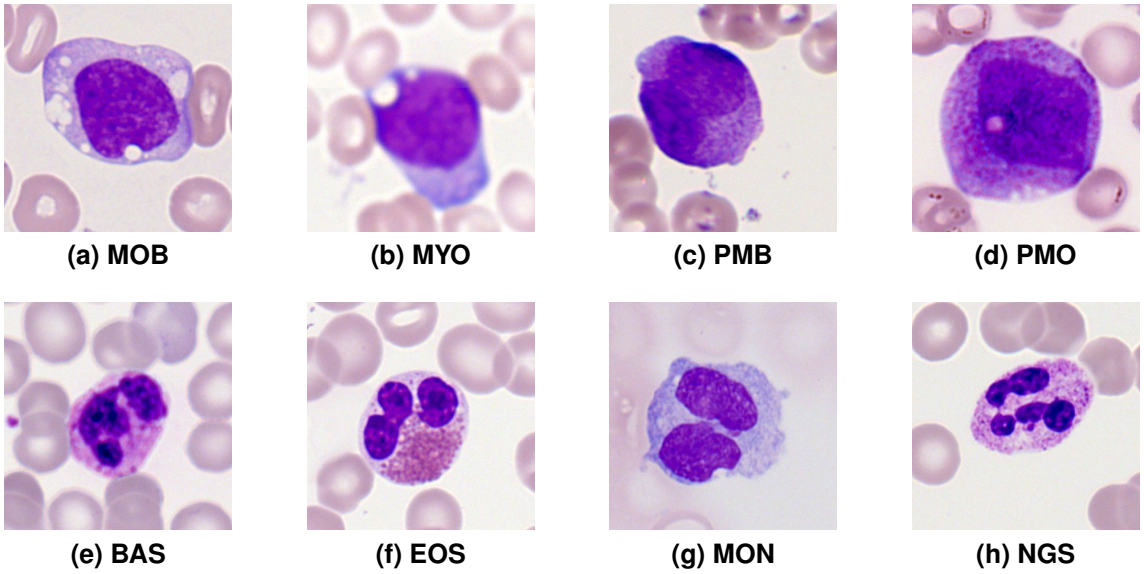


Figura 1: Exemplos de imagens do *dataset*.

A partir de uma análise exploratória, foi constatado o alto desbalanceamento do *dataset*, conforme histograma ilustrado na Figura 2. O baixo número de amostras é particularmente desfavorável aos blastos MOB e PMB.

3.2. Modelos e Parametrização

Considerando a proposta de expandir o campo de pesquisa para avaliar o uso de ViTs no problema, foram testados os modelos ViT-Base, MobileViT e ViT Small, os quais são descritos a seguir. Todos eles baseiam-se, de forma direta ou híbrida, na arquitetura de *Vision Transformers* (ViT), uma mudança de paradigma em *Deep Learning* na modelagem de dados visuais através do mecanismo de auto-atenção. [Dosovitskiy et al. 2021].

O ViT-Base, proposto por Dosovitskiy *et al.* [2021], contém 86 Mi de parâmetros e foi pré-treinado no *ImageNet-21k*, *dataset* com 21 mil classes e 14 Mi imagens. O ViT-Base tem a sua arquitetura baseada no codificador do *Transformer*, no qual cada *token*

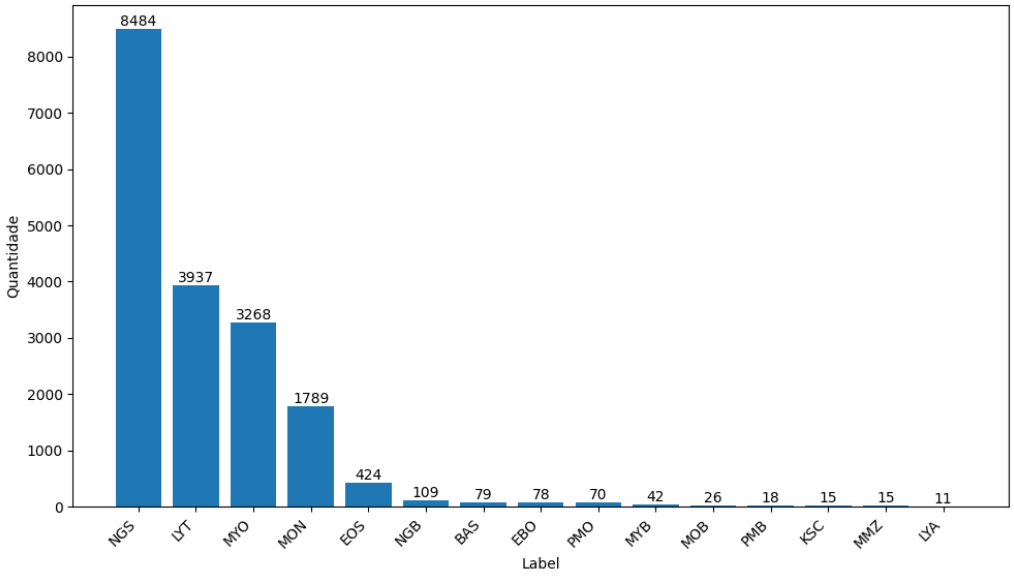


Figura 2: Distribuição das classes do *dataset*.

de entrada é convertido em um vetor que, por sua vez, é submetido ao mecanismo de autoatenção de múltiplas cabeças. Este mecanismo faz com que o vetor agregue informações do contexto global da entrada a si mesmo, enriquecendo o entendimento de sua função em relação ao resto dos *tokens*. No ViT-Base, o modelo de *Transformer* original proposto para tarefas de NLP foi adaptado para o uso de imagens, e assim, divide a entrada em *patches*, que são posteriormente achatados em vetores unidimensionais usados na entrada. O ViT-Base possui 12 camadas, 12 cabeças de atenção e vetores de entrada com dimensão 768.

O modelo proposto por Mehta e Rastegari [2022] consiste em uma combinação entre CNN e ViT, tendo em sua arquitetura um bloco do tipo MobileViT em meio às convoluções tradicionais de CNNs com o objetivo de inserir o entendimento do contexto global da imagem durante este processo. O diferencial deste modelo está presente no bloco MobileViT, que recebe como entrada um mapa de características da imagem que já passou por convoluções 3×3 e MV2, que referencia o bloco *MobileNetv2*. Dentro do bloco MobileViT, a imagem passa por uma convolução $n \times n$ para agregar informações locais aos vetores e por uma convolução 1×1 para aumentar a sua dimensionalidade. Terminados esses passos, a imagem é dividida em *patches* e é processada pelo *Transformer*. A diferença entre esta arquitetura, denominada MobileViT, e a ViT-Base é que enquanto esta última calcula as relações entre cada *patch*, o MobileViT calcula as relações entre as sequências de pixels da mesma posição em cada *patch*. Desta forma, o modelo consegue capturar as informações globais por meio do mecanismo de atenção sem se preocupar com informações locais, pois estas foram capturadas pelas convoluções feitas anteriormente. Então, o modelo intercala entre este bloco e outros dois blocos MV2 e segue com a classificação da imagem.

O DeiT (*Data-efficient image Transformer*) de Touvron *et al.* [2021] teve como objetivo aplicar a arquitetura ViTs de forma mais eficiente e em *datasets* menores. Para tal, é adicionado um *token* de destilação ao início da sequência de *patches* do modelo, junto ao *token* de classificação. Ambos os *tokens* interagem com os outros *embeddings*

por meio do mecanismo de auto-atenção. No entanto, enquanto o vetor de classificação tem o objetivo de aprender contexto geral da imagem, o vetor de destilação tem como objetivo prever o rótulo escolhido por um modelo professor, normalmente uma CNN. O trabalho de Steiner *et al.* [2022] dá continuidade à busca de eficiência do modelo, testando diversos parâmetros e contextos como aumento de dados, diversas técnicas de regularização, pré-treinamento e *fine-tuning* em diferentes *datasets*. O modelo escolhido para a presente pesquisa é o ViT Small (derivado do DeiT Small), que foi pré-treinado no *ImageNet-21k* com aumento de dados e regularização e passou pelo *fine-tuning* com o *ImageNet-1k*. O modelo possui 12 camadas, vetores de entrada com dimensão de 384 e 6 cabeças de atenção, contendo um total de 22 Mi de parâmetros.

Todos os modelos citados foram treinados com 200 épocas e *batches* de tamanho 16. O otimizador utilizado foi o *Stochastic Gradient Descent* (SGD) com *momentum* de $0{,}7$. A taxa de aprendizagem utilizada foi de $0{,}001$. Para evitar *overfitting*, foi implementado um *weight decay* de $0{,}001$, além de *early stopping* com paciência de 20 épocas. Também com este fim e devido à característica *data-hungry* dos ViTs, a técnica de aumento de dados online foi aplicada ao *dataset* de treino. Transformações das imagens incluíram a inversão horizontal aleatória, o rotacionamento aleatório e o recorte aleatório.

3.3. Análise de Desempenho

As métricas de desempenho selecionadas para avaliar os modelos foram a acurácia, a precisão, a revocação e o F_1-*Score* conforme validação cruzada *holdout* $80:20$ de forma estratificada, em que 80% das imagens foram usadas no treino e 20% no teste, mantendo-se as proporções das classes nos dois conjuntos. Três particionamentos dos dados foram considerados utilizando as sementes 42, 43 e 44. O resultado final de cada modelo foi calculado pela média e desvio padrão de suas métricas em cada. Separando as previsões de um modelo de Aprendizado de Máquina em quatro tipos, a citar Verdadeiro Positivo (VP), Verdadeiro Negativo (VN), Falso Positivo (FP) e Falso Negativo (FN), as métricas de desempenho foram calculadas conforme Eqs. (1)-(4).

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}}, \tag{1}$$

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}, \tag{2}$$

$$\text{Revocação} = \frac{\text{VP}}{\text{VP} + \text{FN}}, \tag{3}$$

$$F_1\text{-Score} = 2 \cdot \left(\frac{\text{Precisão} \cdot \text{Revocação}}{\text{Precisão} + \text{Revocação}} \right). \tag{4}$$

No entanto, uma vez que o problema em questão é multiclasse, foi necessário usar a versão ponderada das métricas descritas anteriormente, excetuando-se a acurácia, conforme generalização proposta na Eq. (5)

$$M_{weighted} = \frac{\sum_{c=1}^{C} M_c \cdot I_c}{N_t}, \tag{5}$$

em que $M_{weighted}$ é a versão ponderada de uma das três métricas citadas, $C = 15$ é o número de classes do problema, M_c é a métrica calculada para a classe correspondente, I_c é o número de imagens de cada classe e N_t é o número total de imagens a serem testadas. Nesta versão das fórmulas considera-se a classe atual na qual se calcula a métrica como positivo e todas as demais classes são consideradas negativo.

Considerando que a LMA é discriminada pela superprodução de células imaturas das classes MOB, MYO, PMB e PMO, elencou-se uma métrica específica para a capacidade dos modelos em classificar corretamente tais células a despeito das demais, definida como:

$$\text{Acurácia}_{\text{LMA}} = \frac{\text{VP}_{\text{MOB}} + \text{VP}_{\text{MYO}} + \text{VP}_{\text{PMO}} + \text{VP}_{\text{PMB}}}{N_{\text{LMA}}}, \tag{6}$$

em que N_{LMA} é o total de amostras de teste cuja classe verdadeira está nessas quatro categorias { MOB, MYO, PMB, PMO}, formando um subproblema fechado de 4 classes.

4. Resultados e Discussão

Os experimentos foram conduzidos em um *workstation* com processador NVIDIA GB10 Grace Blackwell com 20 núcleos ARM, *clock* de até $2{,}52\,\text{GHz}$, memória principal unificada de $128\,\text{GB}$, $4\,\text{TB}$ de memória secundária e placa gráfica integrada NVIDIA Blackwell GB20B, para promover a aceleração em hardware do treino dos modelos. Os resultados obtidos encontram-se dispostos na Tabela 2.

Tabela 2: Resultados dos experimentos.

Modelo	Acurácia	Precisão	Revocação	F_1-*Score*	$\text{Acurácia}_{\text{LMA}}$
ViT-Base	$95{,}60\,\% \pm 0{,}16$	$94{,}50\,\% \pm 0{,}29$	$95{,}60\,\% \pm 0{,}16$	$95{,}00\,\% \pm 0{,}22$	$94{,}13\,\% \pm 0{,}60$
ViT-Base-aug	$96{,}02\,\% \pm 0{,}05$	$94{,}85\,\% \pm 0{,}12$	$96{,}02\,\% \pm 0{,}05$	$95{,}42\,\% \pm 0{,}07$	$94{,}18\,\% \pm 0{,}43$
ViT Small	$96{,}07\,\% \pm 0{,}89$	$95{,}05\,\% \pm 1{,}25$	$96{,}07\,\% \pm 0{,}89$	$95{,}43\,\% \pm 1{,}02$	$93{,}64\,\% \pm 1{,}26$
ViT Small-aug	$96{,}14\,\% \pm 0{,}15$	$95{,}56\,\% \pm 0{,}16$	$96{,}14\,\% \pm 0{,}15$	$95{,}72\,\% \pm 0{,}17$	$94{,}53\,\% \pm 1{,}21$
MobileViT	$95{,}15\,\% \pm 0{,}43$	$93{,}73\,\% \pm 0{,}63$	$95{,}15\,\% \pm 0{,}43$	$94{,}31\,\% \pm 0{,}62$	$94{,}97\,\% \pm 0{,}82$
MobileViT-aug	$94{,}75\,\% \pm 0{,}46$	$93{,}11\,\% \pm 0{,}65$	$94{,}75\,\% \pm 0{,}46$	$93{,}85\,\% \pm 0{,}62$	$92{,}36\,\% \pm 1{,}93$

Ao analisar as versões originais dos modelos, observa-se um desempenho robusto perante a tarefa, acima dos $94\,\%$ em termos de F_1-*Score* e acurácia. Dentre eles, o ViT Small apresentou o melhor desempenho, atingindo a acurácia de $96{,}07\,\%$, porém com maior desvio padrão, o que é sugestivo de instabilidade nos resultados entre as execuções.

Os modelos ViT Small-aug e ViT-Base-aug, além de apresentarem ganhos nas métricas de classificação, exibiram redução no desvio padrão, o que corrobora a efetividade da técnica de regularização adotada. O modelo MobileViT-aug, por outro lado, sofreu degradação de desempenho ao receber os dados aumentados, provavelmente em decorrência de *underfitting* dada a sua natureza arquitetural híbrida, a qual já embute vieses indutivos fortes, tornando a distorção adicional dos dados uma regularização excessiva para a sua capacidade paramétrica [Mehta and Rastegari 2022]. Ao analisar a $\text{Acurácia}_{\text{LMA}}$, os resultados evidenciam que os ViTs mostraram uma eficácia alta na classificação dos blastos LMA, ressaltando a robustez do modelo, uma vez que o aprendizado das características

Tabela 3: Análise comparativa de eficiência.

Modelo	Parâmetros	Tempo de Treino (h)	Latência (ms)
ViT-Base	86 Mi	1,49	5,57
ViT-Base-aug	86 Mi	4,42	7,74
ViT Small	22,2 Mi	0,29	3,09
ViT Small-aug	22,2 Mi	0,73	5,24
MobileViT	5,6 Mi	2,24	5,29
MobileViT-aug	5,6 Mi	3,45	5,93

discriminativas desses blastos ocorreu simultaneamente ao das demais células, sob um cenário de desbalanceamento de dados desfavorável.

Ao analisar a Tabela 3, é possível comparar os modelos utilizados no que tange ao tamanho e velocidade de processamento. O ViT-Base, sendo o maior ViT utilizado (86 Mi de parâmetros), possui o maior tempo de latência, porém o tempo de treino é inferior ao MobileViT, embora este possua uma quantidade significativamente menor de parâmetros (5,6 Mi). Isso se dá por causa da ativação do *Early Stopping*, que costuma ser mais atrasada no MobileViT, mostrando que este necessitou de mais tempo para aprender a classificar as imagens. O ViT Small apresentou o menor tempo de treino e a menor latência dentre os modelos usados. Ao aplicar o aumento de dados, todos os ViTs apresentaram uma elevação substancial no tempo de treino e na latência, mas o ViT Small continuou sendo o mais rápido.

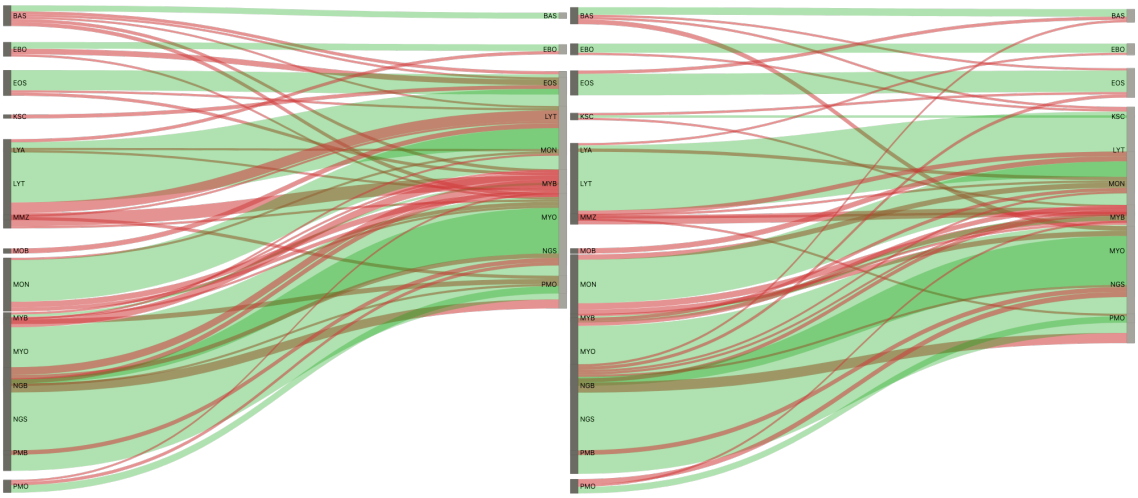
Utilizando as configurações dos ViTs obtidas na etapa de treinamento, explorou-se a técnica de *Ensemble Learning*. Conforme Polikar [2006], a garantia de diversidade em um comitê atua na mitigação de viéses individuais, favorecendo a capacidade de generalização. Nesse contexto, foram elaborados dois comitês baseados em *soft voting*, em que os vetores de distribuição de probabilidades perante as classes do problema produzidos pelos ViTs a partir de uma mesma entrada são somados e a classificação resultante é atribuída à classe correspondente ao argumento máximo (*argmax*) do vetor acumulado, refletindo a confiança agregada dos modelos. Os resultados dos comitês de modelos encontram-se dispostos na Tabela 4.

Tabela 4: Resultados dos comitês de modelos.

Modelo	Acurácia	Precisão	Revocação	F_1-*Score*	$\text{Acurácia}_{\text{LMA}}$
Comitê	$95{,}59\,\% \pm 0{,}28$	$94{,}83\,\% \pm 0{,}21$	$95{,}59\,\% \pm 0{,}28$	$95{,}04\,\% \pm 0{,}30$	$96{,}63\,\% \pm 0{,}23$
Comitê-aug	$96{,}95\,\% \pm 0{,}21$	$95{,}96\,\% \pm 0{,}23$	$96{,}95\,\% \pm 0{,}21$	$96{,}42\,\% \pm 0{,}22$	$96{,}82\,\% \pm 0{,}60$

O Comitê-aug alcançou uma acurácia de $96{,}95\,\%$, superando o Comitê padrão em $1{,}36\,\text{p.p.}$, bem como o desempenho de todas as arquiteturas ViT individuais, independentemente da aplicação prévia de aumento de dados. A métrica de F_1-*Score* exibiu um incremento análogo, evidenciando a maior robustez do modelo, sobretudo na classificação das instâncias minoritárias. Em suma, os resultados experimentais atestam a eficácia da integração entre estratégias de *Ensemble Learning* e aumento de dados para a maximização do desempenho preditivo. Em comparação com a literatura, os resultados obtidos superam o trabalho de Lu *et al.* [2024] usando apenas ViTs. Embora considerem

o mesmo número de classes que o trabalho de Yadav *et al.* [2023], o número de parâmetros do Comitê-aug é de $113{,}8$ Mi frente a 9 Bi da contrapartida da literatura, o que pode ensejar mais facilidade de adoção em contextos práticos.



(a) Comitê **(b) Comitê-aug**

Figura 3: Diagramas de Sankey dos Comitês. A largura do fluxo usa a raiz quadrada da escala original.

Ao contrastar os diagramas de Sankey dos comitês presentes na Figura 3, percebe-se que a aplicação do aumento de dados é refletida na melhor classificação de certas células, como NGS, MYO, MYB, LYT, EOS, EBO e BAS, e em um recuo em PMO e, em menor grau, em MON. Destaca-se que o Comitê-aug obtém o melhor índice de classificação da célula crítica MYO entre todos os modelos avaliados ($99{,}08\,\%$ de acerto), o que é particularmente relevante do ponto de vista clínico: MYO corresponde aos mieloblastos, cuja contagem é o critério morfológico central para o diagnóstico de LMA, de modo que mesmo um ganho percentualmente modesto nessa classe tem impacto direto na utilidade diagnóstica do modelo. Ressalta-se, ainda, que o Comitê-aug possui boa taxa de acerto para a célula MYB, que dificilmente foi classificada corretamente pelos ViTs, obtendo $88\,\%$.

5. Considerações Finais

Este trabalho demonstrou resultados experimentais acerca da eficácia e eficiência do uso de ViTs no contexto da Microscopia Inteligente aplicada ao rastreio da LMA. A avaliação empírica permitiu concluir que a integração de técnicas de aumento de dados conjuntamente com *Ensemble Learning* mediante *soft voting* resultou em um comitê de classificadores com acurácia igual a $96{,}95\,\%$, mesmo em face do forte desbalanceamento presente na base de dados do problema. Embora os resultados obtidos não superem totalmente as métricas de referência do estado da arte, há um menor número de parâmetros, o que enseja maior proximidade com cenários de utilização prática.

Com o intuito de favorecer a reprodutibilidade dos resultados aqui reportados, os artefatos de código utilizados neste trabalho, incluindo os *scripts* de pré-processamento,

treinamento dos modelos e geração das análises apresentadas, estão disponíveis publicamente no GitHub, no seguinte repositório: https://github.com/elloa/eniac2026-vitmla.

Em trabalhos futuros almeja-se realizar a validação externa do comitê com outras bases de dados para LMA, avaliar estratégias de *fine tuning* para o treinamento de ViTs e explorar outras arquiteturas de ViTs perante o problema.

Declaração de Ética no Uso de IA Generativa

A revisão gramatical e ortográfica foi realizada com auxílio dos modelos GPT-5.5 Thinking e Gemini 3.5 Flash. Além disso, todas as revisões automáticas foram complementadas por revisão humana, a fim de garantir a precisão e a qualidade do texto.

Agradecimentos

Os autores agradecem o apoio material recebido do Laboratório de Sistemas Inteligentes (LSI) da Universidade do Estado do Amazonas (UEA). EBG agradece o apoio financeiro da CAPES por meio do Projeto COFECUB-PJ3126170P.

Referências

Associação Brasileira de Linfoma e Leucemia (ABRALE) (2026). Casos diagnosticados de leucemia no Brasil disparam em 10 anos. Disponível em https://abrale.org.br/noticias/casos-diagnosticados-de-leucemia-no-brasil-disparam-em-10-anos/. Acesso: 12 de junho de 2026.

Bigorra, L., Merino, A., Alférez, S., and Rodellar, J. (2017). Feature Analysis and Automatic Identification of Leukemic Lineage Blast Cells and Reactive Lymphoid Cells from Peripheral Blood Cell Images. *J. Clin. Lab. Anal.*, 31(2):e22024.

Correa, N. (2020). Tendências da Mortalidade por Leucemia no Brasil. Observatório de Oncologia. Disponível em https://observatoriodeoncologia.com.br/estudos/cancer-de-sangue/leucemias/2020/mortalidade_leucemias/. Acesso: 3 de março de 2026.

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. In *International Conference on Learning Representations*.

Du, M., Chen, W., Liu, K., Wang, L., Hu, Y., Mao, Y., Sun, X., Luo, Y., Shi, J., Shao, K., Huang, H., and Ye, D. (2022). The Global Burden of Leukemia and its Attributable Factors in 204 Countries and Territories: Findings from the Global Burden of Disease 2019 Study and Projections to 2030. *J. Oncol.*, 2022:1612702.

Elhassan, T., Osman, A. H., Mohd Rahim, M. S., Mohd Hashim, S. Z., Ali, A., Elhassan, E., Elkamali, Y., and Aljurf, M. (2024). CAE-ResVGG FusionNet: A Feature Extraction Framework integrating Convolutional Autoencoders and Transfer Learning for immature White Blood Cells in Acute Myeloid Leukemia. *Heliyon*, 10(19):e37745.

Hu, C., Chen, W., Zhang, P., Shen, T., and Xu, M. (2025). Global, Regional, and National Burden of Leukemia: Epidemiological Trends Analysis from 1990 to 2021. *PLoS One*, 20(6):e0325937.

INCA (2023). Síntese de Resultados e Comentários — Estimativa. Disponível em `https://www.gov.br/inca/pt-br/assuntos/cancer/numeros/estimativa/sintese-de-resultados-e-comentarios`. Acesso: 4 de março de 2026.

Lu, Q., Wang, B., He, Q., Zhang, Q., Guo, L., Li, J., Li, J., and Ma, Q. (2024). A Blood Cell Classification Method based on MAE and Active Learning. *Biomed. Signal Process. Control*, 90(105813):105813.

Maurício, J., Domingues, I., and Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Appl. Sci. (Basel)*, 13(9):5521.

Mehta, S. and Rastegari, M. (2022). MobileViT: Light-weight, general-purpose, and mobile-friendly Vision Transformer. In *International Conference on Learning Representations*, Virtual. Disponível em `https://openreview.net/forum?id=vh-0sUt8HlG`. Acesso: 12 de junho de 2026.

Melanthota, S. K., Gopal, D., Chakrabarti, S., Kashyap, A. A., Radhakrishnan, R., and Mazumder, N. (2022). Deep Learning-Based Image Processing in Optical Microscopy. *Biophysical Reviews*, 14(2):463–481.

Moriconi, É. F., Sousa, L. P., Claro, M., Rocha, A. A. F., and Veras, R. M. S. (2024). Refinamento e Combinação de CNNs para Melhoria na Classificação de Leucócitos Imaturos no Diagnóstico de Leucemia Mieloide Aguda. In *Anais da XII Escola Regional de Computação do Ceará, Maranhão e Piauí (ERCEMAPI 2024)*, pages 129–138. Sociedade Brasileira de Computação - SBC.

Sarvamangala, D. R. and Kulkarni, R. V. (2022). Convolutional Neural Networks in Medical Image Understanding: a Survey. *Evol. Intell.*, 15(1):1–22.

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. *Transactions on Machine Learning Research*. Disponível em `https://arxiv.org/abs/2106.10270`. Acesso: 15 de julho de 2026.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jegou, H. (2021). Training Data-Efficient Image Transformers & Distillation through Attention. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.

Yadav, D. P., Kumar, D., Jalal, A. S., Kumar, A., Singh, K. U., and Shah, M. A. (2023). Morphological Diagnosis of Hematologic Malignancy using Feature Fusion-based Deep Convolutional Neural Network. *Sci. Rep.*, 13(1):16988.

Zhou, Y., Huang, G., Cai, X., Liu, Y., Qian, B., and Li, D. (2024). Global, Regional, and National Burden of Acute Myelyoid Leukemia, 1990-2021: a Systematic Analysis for the Global Burden of Disease Study 2021. *Biomark. Res.*, 12(1):101.