

# Multi-Class Classification of Immature Leukocytes in AML Blood Smears Using an Ensemble of CNNs

Ana V. S. Coelho<sup>1</sup>, Leonardo P. Sousa<sup>1</sup>, Nicole E. M. Silvestre<sup>1</sup>,  
Maíla L. Claro<sup>2</sup>, André M. Santana<sup>1</sup>, Rodrigo M. S. Veras<sup>1</sup>

<sup>1</sup>Universidade Federal do Piauí (UFPI) - Teresina, Brasil

<sup>2</sup>Instituto Federal do Piauí (IFPI) - Paulistana, Brasil

{ana.coelho.ac, leonardosousa, nicole.silvestre,  
andremacedo, rveras}@ufpi.edu.br  
claromaila@gmail.com

**Abstract.** *Acute Myeloid Leukemia (AML) is a neoplasm characterized by the abnormal proliferation of immature leukocytes, whose identification is essential for diagnosis. This study evaluated the use of Convolutional Neural Networks (CNNs) for the automatic classification of five cell subtypes in blood smear images. Eight pre-trained architectures were tested using data augmentation, cross-validation, and transfer learning. Based on the top four networks, four majority voting committees were formed, with the best-performing one achieving an accuracy of 88.49% and outperforming individual models. The results indicate the potential of the approach to support AML diagnosis.*

**Resumo.** *A Leucemia Mieloide Aguda (LMA) é uma neoplasia caracterizada pela proliferação anormal de leucócitos imaturos, cuja identificação é essencial para o diagnóstico. Este estudo avaliou o uso de Redes Neurais Convolucionais (CNNs) na classificação automática de cinco subtipos celulares em esfregaços sanguíneos. Oito arquiteturas pré-treinadas foram testadas com aumento de dados, validação cruzada e aprendizado por transferência. A partir das quatro melhores redes foram formados quatro comitês por votação majoritária, sendo que o mais eficaz alcançou acurácia de 88,49% e superou os modelos individuais. Os resultados indicam o potencial da abordagem para auxiliar o diagnóstico da LMA.*

## 1. Introdução

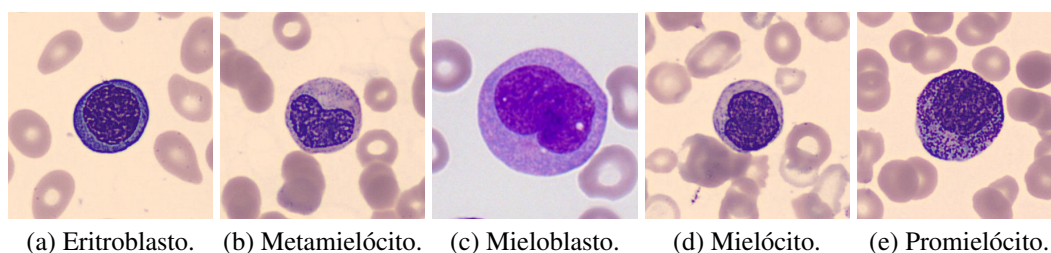
A leucemia compreende um grupo heterogêneo de neoplasias hematológicas caracterizadas pela proliferação descontrolada de leucócitos em diferentes estágios de maturação, afetando a medula óssea, o sangue periférico e os órgãos linfoides. A classificação clínica das leucemias baseia-se na velocidade de progressão, que pode ser aguda ou crônica, e no tipo celular predominante, sendo este mieloide ou linfóide. [Chennamadhavuni et al. 2023].

Entre os subtipos, a Leucemia Mieloide Aguda (LMA) é a forma mais prevalente em adultos, sendo responsável por aproximadamente 80% dos casos nessa população [Yamamoto and Goodman 2008]. Caracteriza-se pela produção excessiva de leucócitos imaturos, principalmente precursores mielóides, que se acumulam na medula óssea e no sangue periférico. Esse processo compromete a hematopoiese e resulta em

um quadro clínico agressivo [Saultz and Garzon 2016]. Segundo estimativas da American Cancer Society [American Cancer Society 2025], cerca de 22.010 novos casos de LMA são esperados nos Estados Unidos em 2025, com aproximadamente 11.090 óbitos atribuídos à doença.

O diagnóstico da LMA é tradicionalmente realizado por meio de exames laboratoriais, testes genéticos, exames de imagem e procedimentos invasivos, como a biópsia da medula óssea. Esses métodos, além de onerosos e demorados, podem causar desconforto físico e impacto emocional significativo. A presença de leucócitos imaturos em esfregaços sanguíneos é um dos principais marcadores morfológicos da doença, tornando sua identificação precisa uma alternativa promissora para triagens iniciais menos invasivas e mais acessíveis.

Dentre essas células, destacam-se cinco subtipos comumente associados à LMA: eritroblastos, metamielócitos, mieloblastos, mielócitos e promielócitos (Figura 1). A diferenciação morfológica entre esses subtipos é essencial para uma avaliação hematológica acurada, exigindo conhecimento especializado e tempo considerável por parte do patologista. Nesse cenário, ferramentas computacionais capazes de realizar essa tarefa de forma automática e precisa podem representar um avanço significativo na prática clínica.



**Figura 1. Exemplos dos cinco subtipos de leucócitos imaturos analisados.**

Sistemas de Diagnóstico Assistido por Computador (CAD), baseados em aprendizado de máquina, têm demonstrado resultados promissores na detecção automatizada de padrões morfológicos em exames laboratoriais [Agaian et al. 2018]. Dentre esses, destacam-se as Redes Neurais Convolucionais (CNNs), reconhecidas por seu alto desempenho na análise de imagens médicas, combinando elevada acurácia com baixa necessidade de intervenção manual [Vogado et al. 2021].

Diante desse panorama, este estudo propõe uma abordagem baseada em CNNs para a classificação automática de cinco subtipos de leucócitos imaturos em imagens de esfregaços sanguíneos de pacientes com LMA. A metodologia inclui o uso de modelos pré-treinados, técnicas de aumento de dados e *ensemble learning* (comitês de redes), com o objetivo de aprimorar a acurácia e a robustez da classificação, mesmo em cenários com dados limitados. Espera-se, assim, contribuir para o desenvolvimento de soluções computacionais voltadas à triagem automatizada da LMA, promovendo diagnósticos mais rápidos, acessíveis e precisos.

Este artigo está estruturado em quatro seções principais. A Seção 2 apresenta a revisão da literatura relacionada. A Seção 3 descreve a metodologia adotada, incluindo a base de dados e os critérios de avaliação. A Seção 4 apresenta os resultados obtidos e sua

análise. Por fim, a Seção 5 traz as conclusões, limitações do estudo e perspectivas para trabalhos futuros.

## 2. Trabalhos Relacionados

O diagnóstico automatizado da leucemia, por meio de imagens de esfregaços sanguíneos, tem sido amplamente explorado na literatura, com o uso de técnicas de processamento de imagens e algoritmos de aprendizado profundo. No entanto, a maioria das abordagens concentra-se em classificações binárias, como entre células normais e anormais ou entre leucócitos maduros e imaturos, deixando em aberto o desafio de diferenciar subtipos específicos de leucócitos imaturos, particularmente relevantes nos casos de LMA.

Thanh *et al.* [Thanh et al. 2018] propuseram uma abordagem baseada em CNNs para distinguir células normais de células leucêmicas, utilizando a base de dados ALL-IDB1. Os autores alcançaram uma acurácia de 96,6% com o uso de estratégias de aumento de dados, embora a proposta tenha se limitado à classificação binária, sem considerar diferentes tipos celulares.

Dasariraju *et al.* [Dasariraju et al. 2020] utilizaram o algoritmo *Random Forest* para classificar leucócitos imaturos em quatro categorias, a partir de um conjunto com 1.274 imagens. A acurácia global foi de 93,45%, porém o modelo apresentou alta variabilidade de desempenho entre as classes, evidenciando dificuldades na identificação de subtipos com características morfológicas semelhantes.

Santos *et al.* [Santos et al. 2022] aplicaram Support Vector Machines (SVM) para classificar células nucleadas do sangue em oito categorias, com base em 17.092 imagens. Embora tenham alcançado um F1-Score médio de 97%, os leucócitos imaturos foram tratados como uma única classe agregada, sem distinção entre eritroblastos, mieloblastos ou promielócitos, o que compromete a granularidade diagnóstica da abordagem.

No estudo de Rahman e Ahmad [Rahman and Ahmad 2023], foram testados modelos como AlexNet, ResNet50 e DenseNet161 para separar leucócitos maduros dos imaturos. A base de dados utilizada foi composta por 18.365 imagens provenientes do *The Cancer Imaging Archive*, resultando em um F1-Score de 97,0%. Contudo, não houve distinção entre os subtipos imaturos, o que limita sua aplicabilidade diagnóstica em cenários clínicos mais complexos.

Sousa et al. [Sousa et al. 2025] propuseram um sistema de comitês baseados em CNNs para a detecção da LMA, com foco na distinção entre leucócitos maduros e imaturos. Utilizando técnicas de *ensemble learning* como voto majoritário, voto ponderado e *bagging*, os autores alcançaram uma acurácia de 96,62% em um conjunto de dados composto por 48.100 imagens. Embora os resultados sejam promissores, a proposta se limita à classificação binária, sem contemplar a diferenciação entre os subtipos de leucócitos imaturos.

Diante dessas limitações, observa-se que, apesar do sucesso de métodos baseados em CNNs para tarefas gerais de classificação celular, a classificação multiclasse de leucócitos imaturos, com foco nos subtipos morfolologicamente relevantes para a LMA, ainda representa uma lacuna significativa na literatura.

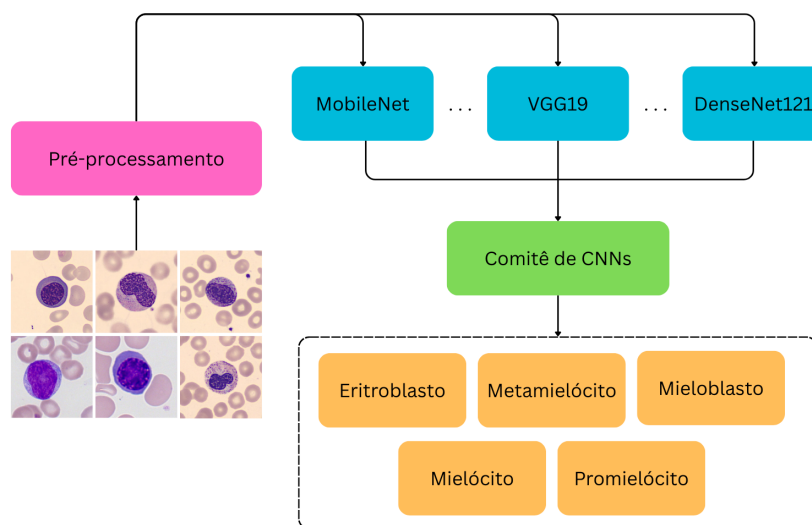
### 3. Metodologia

Nesta seção são apresentados os recursos e o método utilizados no desenvolvimento deste trabalho, cujo objetivo é classificar imagens de células leucêmicas imaturas em cinco subtipos principais encontrados nas bases de dados selecionadas: eritroblasto, metamielócito, mieloblasto, mielócito e promielócito. Para essa tarefa, foram avaliadas oito arquiteturas de CNNs pré-treinadas: MobileNet, ResNet50, DenseNet121, MobileNetV2, ResNet101, VGG19, DenseNet201 e EfficientNetB3, implementadas em Python com o uso das bibliotecas Keras e TensorFlow.

Nos tópicos a seguir, são detalhados o conjunto de dados utilizado e suas características, as técnicas de aumento de dados aplicadas, as CNNs selecionadas, a estratégia de comitê adotada para a classificação final, bem como as métricas utilizadas para avaliação de desempenho.

#### 3.1. Estrutura metodológica

A metodologia adotada está organizada em etapas sequenciais, conforme ilustrado no fluxograma da Figura 2. Inicialmente, as imagens foram coletadas a partir de três bases de dados médicas e organizadas por subtipo celular. Em seguida, foi realizado o pré-processamento das imagens, incluindo redimensionamento, normalização e filtragem de ruído.



**Figura 2. Fluxograma do método proposto para classificação de subtipos celulares em imagens de esfregaços sanguíneos.**

Para mitigar a limitação de dados e melhorar a generalização dos modelos, aplicaram-se técnicas de aumento de dados (*data augmentation*) de forma dinâmica durante o treinamento, utilizando transformações geométricas controladas, como rotações, translações, zoom e espelhamento horizontal. O conjunto ampliado foi utilizado no treinamento de oito arquiteturas de CNNs pré-treinadas.

O treinamento foi conduzido por meio de validação cruzada estratificada com cinco partições (*k-fold*), garantindo o balanceamento entre as classes em cada divisão. Após a avaliação dos desempenhos individuais, as quatro arquiteturas com melhor desempenho foram combinadas três a três, resultando em quatro comitês distintos. Cada

comitê foi avaliado por meio de votação majoritária, com o objetivo de aumentar a confiabilidade e a estabilidade das predições. Todos os treinamentos foram realizados por 65 épocas, com *batch size* de 128, parâmetros definidos empiricamente para assegurar a convergência dos modelos sem comprometer a estabilidade numérica ou o uso eficiente de recursos computacionais.

### 3.2. Base de imagens

Para o desenvolvimento deste projeto, foi utilizado um conjunto de dados composto por 9.229 imagens de lâminas de sangue, distribuídas em cinco classes principais de leucócitos imaturos: Eritroblasto, Metamielócito, Mieloblasto, Mielócito e Promielócito.

As imagens foram organizadas em um único conjunto, formado a partir de três bases de dados distintas. A primeira base de dados é proveniente do Hospital Universitário de Munique [Matek et al. 2019] e contém 1.443 imagens de esfregaços sanguíneos com leucócitos imaturos de 100 pacientes diagnosticados com LMA. As imagens foram coletadas entre 2014 e 2017 e abrangem quatro classes de leucócitos imaturos: Promielócitos, Mielócitos, Metamielócitos e Eritroblastos.

O segundo conjunto de dados foi obtido no Hospital Clínico de Barcelona [Boldú et al. 2021], na Espanha, e contém 4.295 imagens de leucócitos imaturos, seguindo a mesma subdivisão adotada na base de Munique.

Por fim, a terceira base de dados foi coletada no Hospital Johns Hopkins [Sidhom et al. 2021] e inclui imagens de esfregaços sanguíneos contendo leucócitos individuais. Esse conjunto abrange informações de 106 pacientes diagnosticados com LMA e Leucemia Linfoblástica Aguda (LLA), considerando diferentes faixas etárias e gêneros. Originalmente, os dados estavam organizados por paciente, mas foram reorganizados por tipo de leucócito, uma vez que a identificação celular já estava presente nas pastas individuais. As imagens da classe Mieloblasto foram extraídas dessa base. A Tabela 1 ilustra a combinação das imagens utilizadas.

**Tabela 1. Distribuição das imagens de leucócitos imaturos por base de dados.**

Leucócitos Imaturos	Johns Hopkins	Munique	Barcelona	Total por classe
Eritroblasto	78	520	1551	2149
Metamielócito	15	75	1015	1105
Mieloblasto	3268	0	0	3268
Mielócito	42	181	1137	1360
Promielócito	88	667	592	1347
<b>Total</b>	<b>3491</b>	<b>1443</b>	<b>4295</b>	<b>9229</b>

### 3.3. Aumento de dados

O aumento de dados consiste em estratégias computacionais que ampliam e enriquecem conjuntos de dados originais, gerando novas variações a partir das amostras disponíveis para otimizar o desempenho de modelos de aprendizado de máquina. Como destacam Mumuni e Mumuni [Mumuni and Mumuni 2022], bases de dados ampliadas e diversificadas potencializam a capacidade de generalização dos algoritmos em cenários reais. Durante o treinamento do modelo, implementou-se aumento dinâmico de dados (*online*)

a cada época para prevenir *overfitting*, aplicando transformações geométricas controladas: rotações de 20 graus, translações horizontais e verticais de até 20% das dimensões da imagem, variações de zoom na mesma proporção e espelhamento horizontal.

### 3.4. Redes Neurais Convolucionais

As CNNs constituem uma classe de algoritmos de *deep learning* originalmente concebida para o processamento de dados estruturados em grade, como imagens bidimensionais [LeCun et al. 2015]. Sua arquitetura é particularmente adequada para esse tipo de dado, pois emprega filtros convolucionais locais capazes de preservar a coerência espacial entre os pixels. Esses filtros permitem a extração progressiva de características hierárquicas relevantes, como contornos, texturas e formas, ao longo das diferentes camadas da rede, mantendo a topologia espacial da imagem. Devido a essa estrutura, as CNNs têm se mostrado altamente eficazes em tarefas de classificação visual, apresentando capacidade de aprendizado automático de padrões complexos e discriminativos [Tajbakhsh et al. 2016].

No contexto médico, as CNNs têm se mostrado particularmente eficazes na análise de imagens diagnósticas, incluindo a detecção automatizada de condições hematológicas como a leucemia. Nesta pesquisa, investigamos oito arquiteturas de CNN pré-treinadas no banco de dados ImageNet, amplamente utilizado em tarefas de visão computacional por conter milhões de imagens rotuladas em milhares de categorias. As redes avaliadas — MobileNet, ResNet50, DenseNet121, MobileNetV2, ResNet101, VGG19, DenseNet201 e EfficientNetB3 — foram escolhidas com base em seu desempenho consistentemente alto em benchmarks de validação, bem como por suas diferentes profundidades, complexidades e capacidades de generalização.

### 3.5. Transferência de Aprendizado

A transferência de aprendizado consiste em adaptar um modelo já treinado em um conjunto extenso de dados (como o ImageNet) para executar uma nova tarefa relacionada, aproveitando o conhecimento previamente adquirido. Neste trabalho, adotamos a abordagem de *Shallow Fine-Tuning* (SFT), na qual apenas as camadas finais da rede são reajustadas, mantendo-se as camadas convolucionais iniciais fixas. Conforme demonstrado por Tajbakhsh et al. [Tajbakhsh et al. 2016], essa estratégia reduz significativamente a demanda por dados e recursos computacionais, além de acelerar o processo de treinamento, permitindo que o modelo alcance um desempenho satisfatório de forma mais eficiente. No experimento realizado, todas as camadas da arquitetura base foram mantidas congeladas, sendo treinadas apenas as três camadas adicionadas ao topo da rede: uma de *pooling*, uma camada densa com 256 unidades e a camada de saída.

### 3.6. Validação Cruzada

A validação cruzada é uma das técnicas mais utilizadas para avaliar modelos preditivos e mitigar o risco de sobreajuste, permitindo estimar sua capacidade de generalização [Berrar 2019]. Neste estudo, adotou-se a validação cruzada estratificada com 5 *folds* ( $k=5$ ), na qual o conjunto de dados foi particionado em cinco subconjuntos com distribuição proporcional das classes. Em cada ciclo, um *fold* foi usado como validação (20%), enquanto os demais foram destinados ao treinamento (80%). Esse processo permitiu uma avaliação mais consistente do desempenho dos modelos e maior controle sobre possíveis vieses.

### 3.7. Comitês

A estratégia de *ensemble learning* integra as capacidades preditivas de múltiplos classificadores, potencializando a precisão, adaptabilidade e generalização do sistema [Dietterich 2000]. Neste trabalho, foram formados quatro comitês distintos, combinando três a três as quatro arquiteturas de melhor desempenho individual, com decisão baseada em votação majoritária. Essa abordagem coletiva favorece decisões mais equilibradas e confiáveis, ao explorar diferentes perspectivas de análise e reduzir a influência de vieses individuais. Além disso, a comparação entre diferentes formações de comitês permitiu uma avaliação empírica da combinação mais eficaz, fortalecendo a escolha da configuração final adotada.

### 3.8. Métricas de Avaliação

O desempenho dos modelos foi avaliado por meio de cinco métricas fundamentais [Powers 2020]: acurácia, precisão, *recall*, *F1-Score* e índice Kappa. A acurácia representa a taxa global de acertos do modelo, calculada pela razão entre previsões corretas e o total de predições realizadas. A precisão quantifica a confiabilidade das classificações positivas, demonstrando quanto dessas previsões eram realmente verdadeiras.

O *recall* (ou sensibilidade) avalia a capacidade do modelo de detectar casos positivos, medindo a proporção de verdadeiros positivos identificados em relação ao total existente. O *F1-Score* harmoniza precisão e *recall* em um único valor, sendo particularmente relevante em conjuntos de dados desbalanceados. Complementarmente, o índice Kappa avalia a concordância entre as previsões e os valores reais, considerando e corrigindo os acertos que poderiam ocorrer aleatoriamente, proporcionando assim uma medida mais rigorosa de desempenho [Rosenfield and Fitzpatrick-Lins 1986].

## 4. Resultados e Discussão

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da avaliação das arquiteturas de CNNs utilizadas para a classificação dos subtipos de leucócitos imaturos, considerando também a contribuição de comitês formados por combinações de redes.

A Tabela 2 apresenta os resultados médios obtidos pelas oito arquiteturas avaliadas. Observa-se que a EfficientNetB3 se destacou entre os modelos testados, alcançando acurácia de 86,89% e índice Kappa de 82,80%. Esse desempenho expressivo motivou sua inclusão em todas as combinações de comitês avaliadas, com o objetivo de investigar os ganhos decorrentes da integração dessa arquitetura a diferentes composições de redes.

**Tabela 2. Comparativo de desempenho entre arquiteturas CNN avaliadas.**

Arquitetura	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)	Kappa (%)
ResNet50	85,06 ± 0,55	85,25 ± 0,46	85,06 ± 0,55	84,94 ± 0,48	80,41 ± 0,71
ResNet101	85,30 ± 0,33	85,26 ± 0,22	85,30 ± 0,33	85,23 ± 0,26	80,71 ± 0,41
VGG19	83,13 ± 0,97	83,46 ± 0,62	83,13 ± 0,97	83,12 ± 0,92	77,93 ± 1,26
DenseNet121	84,90 ± 1,05	84,93 ± 1,09	84,90 ± 1,05	84,81 ± 1,08	80,20 ± 1,38
DenseNet201	85,77 ± 1,34	86,01 ± 1,15	85,77 ± 1,34	85,81 ± 1,31	81,37 ± 1,76
EfficientNetB3	<b>86,89 ± 0,70</b>	<b>87,02 ± 0,67</b>	<b>86,89 ± 0,70</b>	<b>86,90 ± 0,68</b>	<b>82,80 ± 0,93</b>
MobileNet	86,12 ± 0,48	86,36 ± 0,43	86,12 ± 0,48	86,14 ± 0,45	81,83 ± 0,61
MobileNetV2	86,63 ± 0,58	86,83 ± 0,60	86,63 ± 0,58	86,69 ± 0,59	82,49 ± 0,76

Com base nos resultados obtidos pelas arquiteturas individuais, foram selecionadas as quatro com melhor desempenho para compor comitês de decisão por votação majoritária, combinando-as três a três. As combinações geradas envolveram as redes EfficientNetB3, MobileNetV2, MobileNet e DenseNet201. A Tabela 3 apresenta a comparação entre os desempenhos obtidos por cada comitê formado.

**Tabela 3. Resultados consolidados dos comitês de classificação.**

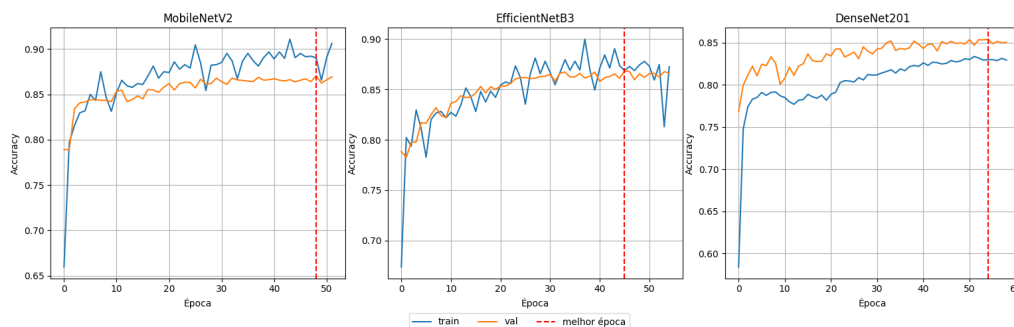
Comitê	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)	Kappa (%)
EfficientNetB3, MobileNet, DenseNet201	88,07 $\pm$ 0,73	88,04 $\pm$ 0,61	88,07 $\pm$ 0,73	88,00 $\pm$ 0,71	84,35 $\pm$ 0,96
EfficientNetB3, MobileNetV2, DenseNet201	<b>88,49 <math>\pm</math> 0,73</b>	<b>88,48 <math>\pm</math> 0,58</b>	<b>88,49 <math>\pm</math> 0,73</b>	<b>88,45 <math>\pm</math> 0,70</b>	<b>84,90 <math>\pm</math> 0,96</b>
EfficientNetB3, MobileNetV2, MobileNet	88,48 $\pm$ 0,39	88,46 $\pm$ 0,44	88,48 $\pm$ 0,39	88,43 $\pm$ 0,40	84,89 $\pm$ 0,51
MobileNetV2, MobileNet, DenseNet201	88,02 $\pm$ 0,64	88,11 $\pm$ 0,58	88,02 $\pm$ 0,64	87,99 $\pm$ 0,64	84,29 $\pm$ 0,84

O comitê com melhor desempenho foi composto pela EfficientNetB3, MobileNetV2 e DenseNet201, superando todos os modelos individuais em todas as métricas avaliadas. Ele alcançou uma acurácia de 88,49% e um índice Kappa de 84,90%, indicando uma melhora consistente na qualidade da classificação. A Tabela 4 apresenta as métricas desse comitê em comparação com aquelas obtidas por cada uma de suas arquiteturas componentes.

**Tabela 4. Desempenho do comitê em relação às arquiteturas CNN avaliadas individualmente.**

Arquitetura	Acurácia (%)	Precisão (%)	Recall (%)	F1-Score (%)	Kappa (%)
EfficientNetB3	86,89 $\pm$ 0,70	87,02 $\pm$ 0,67	86,89 $\pm$ 0,70	86,90 $\pm$ 0,68	82,80 $\pm$ 0,93
MobileNetV2	86,63 $\pm$ 0,58	86,83 $\pm$ 0,60	86,63 $\pm$ 0,58	86,69 $\pm$ 0,59	82,49 $\pm$ 0,76
DenseNet201	85,77 $\pm$ 1,34	86,01 $\pm$ 1,15	85,77 $\pm$ 1,34	85,81 $\pm$ 1,31	81,37 $\pm$ 1,76
<b>Comitê</b>	<b>88,49 <math>\pm</math> 0,73</b>	<b>88,48 <math>\pm</math> 0,58</b>	<b>88,49 <math>\pm</math> 0,73</b>	<b>88,45 <math>\pm</math> 0,70</b>	<b>84,90 <math>\pm</math> 0,96</b>

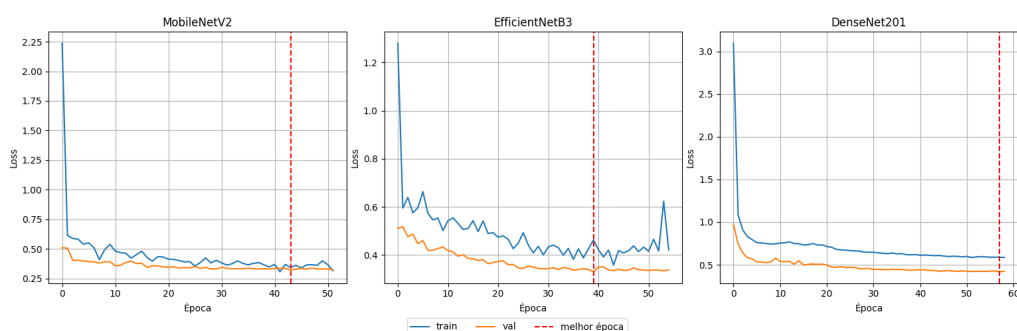
Para avaliar a estabilidade dos modelos durante o treinamento e investigar a ocorrência de *overfitting*, foram geradas curvas de acurácia e função de perda (*loss*) ao longo das épocas, considerando a média dos cinco folds da validação cruzada. As Figuras 3 e 4 ilustram o comportamento dessas métricas para as arquiteturas MobileNetV2, EfficientNetB3 e DenseNet201.



**Figura 3. Curvas da função de acurácia para os modelos que compõem o comitê.**

Entre os modelos avaliados, a DenseNet201 apresentou o comportamento mais estável, com curvas de acurácia e perda suaves e consistentes ao longo do treinamento, sem oscilações abruptas ou divergência entre treino e validação. A EfficientNetB3, embora tenha exibido flutuações mais acentuadas, especialmente na perda de treino, obteve





**Figura 4. Curvas da função de perda para os modelos que compõem o comitê.**

o melhor desempenho geral em termos de acurácia e índice Kappa, indicando forte capacidade de generalização mesmo diante da variabilidade. Já a MobileNetV2 também apresentou bom desempenho, com acurácia de validação alta e trajetória estável, embora com leve variação nas últimas épocas. Esses resultados confirmam que os três modelos foram treinados de forma eficaz, sem indícios significativos de sobreajuste, e com desempenho notáveis nos dados de validação.

O gráfico comparativo apresentado na Figura 5 reforça visualmente o desempenho superior do comitê, evidenciando com clareza sua consistência em todas as métricas avaliadas. Destaca-se, em especial, o ganho expressivo no índice Kappa, métrica que considera os acertos ao acaso e possui relevância particular em cenários com desbalanceamento entre as classes.

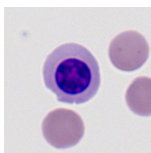
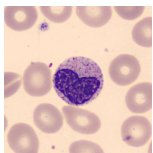
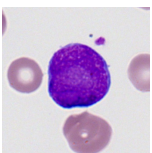
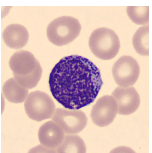
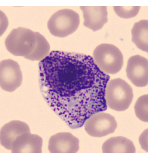
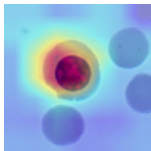
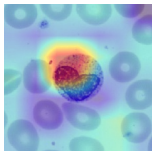
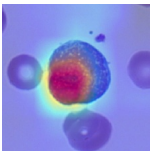
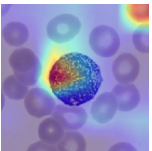
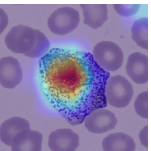
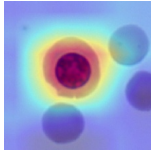
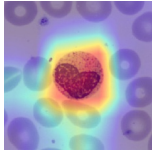
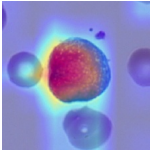
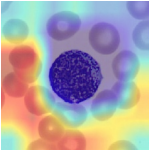
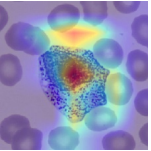
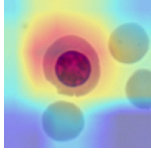
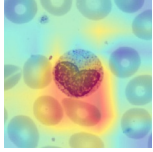
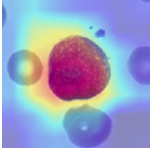
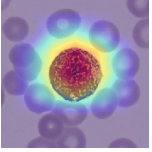
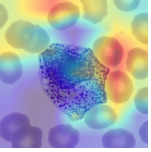


**Figura 5. Comparação do desempenho entre o comitê selecionado e suas arquiteturas componentes.**

A Tabela 5 apresenta os mapas de calor gerados pelos três modelos do comitê para cada subtipo de leucócito imaturo, com base em exemplos representativos. Observa-se que todas as arquiteturas do comitê tendem a concentrar as ativações nas regiões nucleares das células, o que está alinhado aos critérios morfológicos utilizados por especialistas.

Entretanto, subtipos como mielócito e promielócito apresentam variações entre os modelos, com algumas redes destacando também regiões periféricas. Essa diversidade nos focos de atenção sugere que cada arquitetura extrai diferentes aspectos visuais das imagens, o que contribui para decisões mais completas e fundamentadas no comitê.

**Tabela 5. Mapas de calor por subtipo celular gerados pelos modelos do comitê.**

	Eritroblasto	Metamielócito	Mieloblasto	Mielócito	Promielócito
Original					
EfficientNetB3					
DenseNet201					
MobileNetV2					

Essa variação nos mapas de calor sugere que cada arquitetura adota estratégias distintas de extração de atributos morfológicos. A EfficientNetB3 e a DenseNet201 exibiram ativações mais concentradas no núcleo celular, indicando foco nos detalhes nucleares como principal critério de classificação. Em contraste, a MobileNetV2 apresentou regiões de ativação mais amplas, abrangendo também o citoplasma, especialmente no subtipo promielócito. Essa diversidade de atenção reforça a complementaridade entre os modelos, aspecto que justifica o desempenho superior obtido pela combinação em comitê, ao integrar diferentes perspectivas na análise morfológica dos leucócitos.

Por fim, vale ressaltar que o ganho obtido pelo comitê foi alcançado sem aumento significativo da complexidade computacional, uma vez que as inferências são realizadas a partir da média das decisões dos modelos participantes, sem necessidade de retreinamento. Esse aspecto torna a abordagem não apenas eficaz em termos de desempenho, mas também viável para aplicações práticas em contextos com restrições de recursos, reforçando seu potencial de utilização em ambientes clínicos ou laboratoriais com infraestrutura limitada.

## 5. Conclusão

Este trabalho apresentou uma abordagem baseada em CNNs para a classificação automática de cinco subtipos de leucócitos imaturos em imagens de esfregaços sanguíneos de pacientes com diagnóstico de LMA. Foram avaliadas oito arquiteturas pré-treinadas,

integrando técnicas de aumento de dados, validação cruzada estratificada e transferência de aprendizagem. Como principal resultado, um comitê formado pelas redes EfficientNetB3, MobileNetV2 e DenseNet201 alcançou acurácia de 88,49%, superando os modelos individuais e evidenciando os ganhos obtidos com a estratégia de combinação.

Embora promissora, a abordagem apresenta limitações técnicas que indicam caminhos relevantes para aprimoramento. O desbalanceamento entre as classes pode comprometer a sensibilidade para subtipos menos representados, enquanto a elevada variabilidade morfológica intra-classe dificulta a definição de padrões discriminativos robustos. Além disso, a votação majoritária utilizada no comitê, apesar de eficaz, não explora plenamente o potencial de combinações mais flexíveis entre os modelos. Como direções futuras, propõe-se a adoção de técnicas de aumento de dados *offline*, com geração controlada de variações morfológicas mais diversificadas, e a aplicação de *Deep Fine Tuning*, permitindo o ajuste integral das redes ao domínio-alvo. Essas estratégias têm potencial para elevar a precisão, a generalização e a aplicabilidade clínica da solução desenvolvida.

## Referências

- Agaian, S., Madhukar, M., and Chronopoulos, A. T. (2018). A new acute leukaemia automated classification system. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):303–314.
- American Cancer Society (2025). Key statistics for acute myeloid leukemia (aml). <https://www.cancer.org/cancer/types/acute-myeloid-leukemia/about/key-statistics.html>. Accessed on: 29 Mar. 2025.
- Berrar, D. (2019). Cross-validation. In Ranganathan, S., Gribskov, M., Nakai, K., and Schönbach, C., editors, *Encyclopedia of Bioinformatics and Computational Biology*, pages 542–545. Academic Press, Oxford.
- Boldú, L., Merino, A., Acevedo, A., Molina, A., and Rodellar, J. (2021). A deep learning model (alnet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images. *Computer Methods and Programs in Biomedicine*, 202:105999.
- Chennamadhavuni, A., Iyengar, V., Mukkamala, S. K. R., and Shimanovsky, A. (2023). Leukemia. *StatPearls [Internet]*. Available at: <https://pubmed.ncbi.nlm.nih.gov/32809325/>.
- Dasariraju, S., Huo, M., and McCalla, S. (2020). Detection and classification of immature leukocytes for diagnosis of acute myeloid leukemia using random forest algorithm. *Bioengineering*.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Matek, C., Schwarz, S., Spiekermann, K., and Marr, C. (2019). Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. *Nature Machine Intelligence*, 1(11):538–544.

- Mumuni, A.-R. and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.
- Powers, D. M. W. (2020). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.
- Rahman, J. and Ahmad, M. (2023). Detection of acute myeloid leukemia from peripheral blood smear images using transfer learning in modified cnn architectures. pages 447–459.
- Rosenfield, G. H. and Fitzpatrick-Lins, K. (1986). A coefficient of agreement as a measure of thematic classification accuracy. *Photogrammetric Engineering and Remote Sensing*.
- Santos, M., Bertemes, W., de Souza, , Andrades, M., Barros, D., and Patto, V. (2022). Algoritmos de aprendizado de máquina para classificação de células nucleadas do sangue periférico - uma experiência do projeto hemovision. In *Anais da X Escola Regional de Informática de Goiás*, pages 130–140. SBC.
- Saultz, J. N. and Garzon, R. (2016). Acute myeloid leukemia: a concise review. *Journal of Clinical Medicine*, 5(3):33.
- Sidhom, J.-W., Siddharthan, I. J., Lai, B. S., Luo, A., Hambley, B. C., Bynum, J., Duffield, A. S., Streiff, M. B., Moliterno, A. R., Imus, P., et al. (2021). Deep learning for diagnosis of acute promyelocytic leukemia via recognition of genomically imprinted morphologic features. *NPJ Precision Oncology*, 5(1):38.
- Sousa, L. P., Silva, R. R. V., Claro, M. L., Araújo, F. H. D., Borges, R. N., Machado, V. P., and Veras, R. M. S. (2025). Ensemble of cnns for enhanced leukocyte classification in acute myeloid leukemia diagnosis. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 399–413, Cham. Springer Nature Switzerland.
- Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., and Liang, J. (2016). Convolutional neural networks for medical image analysis: Full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312.
- Thanh, T. T. P., Vununu, C., Atoev, S., Lee, S.-W., and Kwon, K.-R. (2018). Leukemia blood cell image classification using convolutional neural network. *International Journal of Computer Theory and Engineering*, 10:54–58.
- Vogado, L. H. S., Veras, R. M. S., and Aires, K. R. T. (2021). Leuknet: Um modelo de rede neural convolucional para o diagnóstico de leucemia. In *Anais do Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*, pages 85–90. Sociedade Brasileira de Computação.
- Yamamoto, J. F. and Goodman, M. T. (2008). Patterns of leukemia incidence in the united states by subtype and demographic characteristics, 1997–2002. *Cancer Causes & Control*, 19(4):379–390.