Deep Learning Ensemble for Multiclass Recognition of Mature Leukocytes in Acute Myeloid Leukemia (AML)

Nicole E. M. Silvestre[1], Leonardo P. Sousa[1], Ana V. S. Coelho[1], Maíla L. Claro[2], André M. Santana[1], Rodrigo M. S. Veras[1]

[1]Universidade Federal do Piauí (UFPI) - Teresina, Brasil

[2]Instituto Federal do Piauí (IFPI) - Paulistana, Brasil

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

Abstract. *Manual classification of leukocytes in blood smear images is subjective, time-consuming, and prone to errors, especially in high-demand clinical contexts. In this context, this study proposes a Convolutional Neural Network (CNN)-based approach for multiclass classification of mature leukocytes to support the diagnosis of Acute Myeloid Leukemia (AML). Eight pre-trained CNNs were evaluated on a dataset of 30,929 images from six cellular subtypes. An ensemble model with majority voting (MobileNetV2, ResNet101, and MobileNet) achieved an accuracy of 93.18%. The results highlight the potential of CNNs and ensemble strategies for automated leukocyte identification in AML-related hematological examinations.*

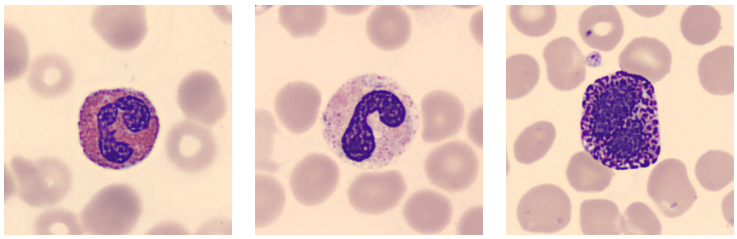
Resumo. *A classificação manual de leucócitos em esfregaços sanguíneos é subjetiva, demorada e propensa a erros, sobretudo em contextos clínicos de alta demanda. Diante disso, este estudo propõe uma abordagem com Redes Neurais Convolucionais (CNNs) para a classificação multiclasse de leucócitos maduros, visando apoiar o diagnóstico da Leucemia Mieloide Aguda (LMA). Oito CNNs pré-treinadas foram avaliadas num conjunto de 30.929 imagens de seis subtipos celulares. Um comitê com voto majoritário (MobileNetV2, ResNet101 e MobileNet) obteve acurácia de 93,18%. Os resultados destacam o potencial das CNNs e da estratégia de comitês na identificação automatizada de leucócitos em exames hematológicos voltados à LMA.*

1. Introdução

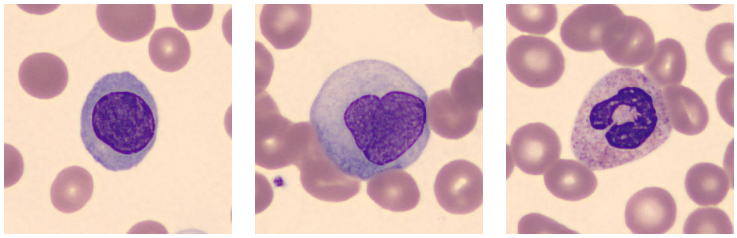
A leucemia é uma forma de câncer hegemônica no cenário hodierno que afeta diretamente a medula óssea e provoca o seu funcionamento anômalo [Travlos 2006]. A medula óssea é um tecido esponjoso situado no interior dos ossos, responsável pela geração de glóbulos vermelhos, plaquetas e glóbulos brancos, sendo estes últimos fundamentais para a defesa imunológica. Na leucemia, mutações genéticas causam a proliferação acelerada de células anormais, com ciclos de vida curtos que substituem as saudáveis. A doença pode se manifestar de forma aguda, com progressão rápida, ou crônica, com evolução mais lenta. É classificada em quatro subtipos principais: Leucemia Linfoide Aguda (LLA), Leucemia Linfoide Crônica (LLC), Leucemia Mieloide Crônica (LMC) e Leucemia Mieloide Aguda (LMA).

A LMA é considerada uma das expressões leucêmicas mais agressivas dentre as variantes. De acordo com um estudo realizado pelo *National Institute of Cancer* [American Cancer Society 2024], foram estimados 20.800 novos casos de leucemia mieloide aguda e 11.200 mortes causadas por essa doença nos Estados Unidos.

A identificação de leucócitos maduros auxilia no diagnóstico e no monitoramento da leucemia, uma vez que a doença compromete a maturação dos blastos, levando ao acúmulo de células imaturas [Döhner et al. 2010]. Assim, a análise da produção leucocitária é fundamental para detectar esse quadro. Nesse sentido, os tipos de leucócitos maduros incluem Eosinófilos, Neutrófilos de Banda, Basófilos, Linfócitos Típicos, Monócitos e Neutrófilos Segmentados, que estão ilustrados na Figura 1.



(a) Eosinófilo. (b) Neutrófilo em Banda. (c) Basófilo.



(d) Linfócito Típico. (e) Monócito. (f) Neutrófilo Segmentado.

Figura 1. Exemplos de leucócitos maduros.

O diagnóstico da leucemia tradicionalmente é composto por análises de hemogramas, da coleta do material da medula óssea e do estudo de imagens da morfologia das células por um profissional da saúde, o que pode tornar esse processo demorado e dispendioso [Dwivedi 2018]. Diantes desses desafios, sistemas automatizados com aprendizado de máquina e visão computacional, como os de Diagnóstico Auxiliado por Computador (*Computer-Aided Diagnosis* – CAD), têm se destacado por oferecer uma segunda opinião médica, agilizando a análise de imagens e contribuindo para diagnósticos mais rápidos e precisos [Doi 2005].

A metodologia deste projeto visa utilizar técnicas de processamento de imagens e inteligência artificial para identificação e classificação de células leucêmicas maduras em imagens microscópicas de sangue. Serão exploradas e analisadas abordagens de aprendizagem profunda (*deep learning*), com ênfase em Redes Neurais Convolucionais (*Convolutional Neural Networks* - CNNs). Para isso, diversas CNNs serão avaliadas atuando sobre conjuntos de imagens para classificar leucócitos em seis classes: Basófilos, Eosinófilos, Linfócitos Típicos, Monócitos, Neutrófilos de Banda e Neutrófilos Segmentados. Com isso, pretende-se melhorar a eficiência dos diagnósticos de leucemia e viabilizar

o tratamento adequado aos pacientes.

O presente trabalho está organizado da seguinte forma: a Seção 2 apresenta a revisão da literatura relacionada ao tema abordado. A Seção 3 descreve a metodologia empregada no desenvolvimento do projeto, incluindo a caracterização da base de imagens médicas utilizadas e os métodos de avaliação adotados. Por fim, as Seções 4 e 5 expõem os resultados obtidos, bem como uma discussão acerca dos avanços do projeto e propostas para trabalhos futuros.

2. Trabalhos Relacionados

Diversas técnicas e classificadores foram desenvolvidos por pesquisadores profissionais com o intuito de facilitar o diagnóstico da leucemia em imagens de lâminas de sangue. Não obstante, a literatura a respeito da divisão de classes de leucócitos maduros ainda é limitada, com a maior parte dos trabalhos focada na identificação geral de leucócitos imaturos, e as pesquisas feitas não identificaram projetos que utilizem a mesma abordagem. Nesta seção, são abordadas algumas pesquisas relacionadas ao tema.

Thanh et al. [Thanh et al. 2018] desenvolveram um modelo de CNN com o objetivo de diferenciar células sanguíneas normais de anormais em imagens microscópicas. O treinamento da rede foi realizado utilizando a base de dados ALL-IDB1 [Labati et al. 2011], composta por 108 imagens, sendo 59 de indivíduos saudáveis e 49 de pacientes com leucemia. Diferentes técnicas de aumento de dados foram aplicadas, resultando num conjunto final de 1.188 imagens. O modelo proposto alcançou uma acurácia de 96,6%. No entanto, por se tratar de uma classificação binária, a abordagem não contemplou a distinção entre diferentes subtipos de leucócitos.

Rahman e Ahmad [Rahman and Ahmad 2023] propuseram uma abordagem para a classificação de leucócitos em células maduras e imaturas, utilizando arquiteturas de CNNs, tais como AlexNet, ResNet50, DenseNet161 e VGG-16. O conjunto de dados empregado foi extraído do *The Cancer Imaging Archive* [Clark et al. 2013] e compreendeu um total de 18.365 imagens de leucócitos. Entre os modelos avaliados, a melhor performance foi obtida por uma versão modificada da AlexNet, que alcançou uma precisão de 96,52% e um F1-*Score* de 97,00%. O experimento foi conduzido utilizando apenas um fold de validação, e técnicas de aumento de dados foram aplicadas para otimizar o treinamento. No entanto, de maneira semelhante ao estudo anterior, a classificação proposta não diferenciou os leucócitos em subtipos específicos.

Ding et al. [Ding et al. 2023] propuseram um sistema de classificação de leucócitos em 11 subtipos, utilizando uma base de dados com 11.102 imagens de esfregaços sanguíneos coletadas em hospitais locais. Além disso, foi utilizada uma abordagem híbrida de dois estágios, em que no primeiro os autores combinaram as redes ResNet34, ResNet50 e ResNet101 para classificação inicial baseada em características morfológicas, alcançando 97,03% de acurácia, e no segundo empregaram um algoritmo de aprendizado de máquina supervisionado para distinguir subtipos de linfócitos, obtendo 100% de precisão. Contudo, o trabalho utiliza imagens manualmente recortadas e não aborda a segmentação automática em lâminas completas, o que pode limitar sua aplicabilidade em cenários clínicos com grande volume de dados.

Asghar et al.[Asghar et al. 2023] utilizaram CNNs para classificar células sanguíneas em 10 subtipos em duas etapas. Na primeira, testaram 8 arquiteturas pré-

treinadas com transferência de aprendizagem diante da base de dados *Peripheral Blood Cell* (PBC) [Acevedo et al. 2020], que contém 17.092 imagens de hemácias, plaquetas e leucócitos de pacientes saudáveis, obtendo acurácias entre 91,37% e 94,72%. Em seguida, propuseram uma nova CNN que alcançou 99,91% de acurácia na classificação das dez categorias presentes na base, incluindo tanto células maduras quanto imaturas. Contudo, o estudo não abordou em detalhe a distinção entre subtipos morfológicos de leucócitos maduros, aspecto importante na análise de LMA.

Sousa et al. [Sousa et al. 2025] propuseram uma abordagem com comitês de CNNs para classificar leucócitos em imagens de esfregaços sanguíneos, visando o diagnóstico de LMA. O estudo avaliou oito arquiteturas pré-treinadas, formando comitês por voto majoritário, votação ponderada e *bagging*. Utilizando 48.100 imagens de três bases públicas, a melhor performance foi obtida com o *bagging* da EfficientNetB3, que alcançou acurácia de 96,62% e índice Kappa de 92,27%. Apesar de abordar apenas a distinção entre leucócitos maduros e imaturos, os resultados demonstram o potencial da estratégia como ferramenta de apoio ao diagnóstico da leucemia.

Os estudos analisados evidenciam o uso bem-sucedido de CNNs na classificação de células sanguíneas, com ênfase em estratégias como aprendizado por transferência, aumento de dados e comitês de modelos. Contudo, a maioria das abordagens se restringe à distinção entre leucócitos maduros e imaturos ou à classificação binária, sem detalhar subtipos morfológicos de leucócitos maduros. Essa limitação representa uma lacuna relevante para o diagnóstico da LMA. Assim, o presente trabalho propõe uma abordagem inovadora baseada em CNNs para a classificação multiclasse desses subtipos, visando ampliar a aplicação de inteligência artificial em exames hematológicos.

3. Materiais e métodos

Esta seção tem como finalidade descrever, de forma detalhada, a metodologia adotada neste estudo, cujo objetivo central é a classificação de leucócitos maduros em imagens microscópicas sanguíneas, distribuídos em seis subtipos principais: Basófilos, Linfócitos Típicos, Monócitos, Eosinófilos, Neutrófilos Segmentados e Neutrófilos de Banda.

Para atingir tal objetivo, foram selecionados, ajustados e avaliados modelos de CNNs pré-treinadas. A implementação dessas redes foi feita mediante o uso da linguagem de programação Python, bem como das bibliotecas Keras e Tensorflow de aprendizagem profunda. Ademais, esta seção apresenta, de maneira estruturada, o método proposto, as características da base de imagens empregada, as CNNs selecionadas, a aplicação de um comitê de classificadores para a predição final, bem como as métricas adotadas para a avaliação de desempenho dos modelos.

3.1. Método Proposto

O método proposto, ilustrado na Figura 2, inicia com o pré-processamento das imagens de leucócitos maduros, incluindo redimensionamento e padronização. Em seguida, as imagens são classificadas por comitês de CNNs, formados por combinações de três redes entre as quatro com melhor desempenho individual. Foram avaliados quatro comitês distintos, todos utilizando voto majoritário para definir a classe final entre os seis subtipos de leucócitos. Por fim, o comitê com melhor desempenho foi selecionado para compor os resultados finais. Os modelos foram treinados por 65 épocas com *batch size* de 128, e avaliados pelas métricas: acurácia, precisão, *recall*, F1-*Score* e índice Kappa.

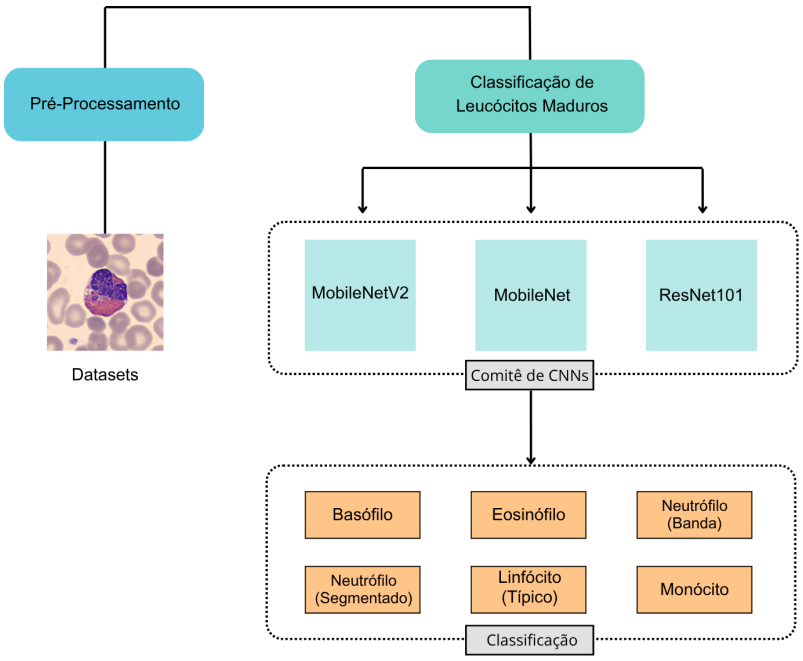


Figura 2. Fluxo do método proposto para a classificação automática de leucócitos maduros em imagens de esfregaços sanguíneos.

3.2. Base de imagens

Para o desenvolvimento do seguinte estudo, utilizou-se uma base de dados com 30.929 imagens de lâminas de sangue redimensionadas para 224×224 pixels, formato compatível com as CNNs pré-treinadas utilizadas. Essa base de dados originou-se da combinação de três conjuntos de imagens. O primeiro foi o conjunto de dados do Hospital Johns Hopkins [Sidhom et al. 2021], o qual contém imagens de esfregaços sanguíneos (extensão .jpg, 360 × 360 pixels) com leucócitos individuais de 106 pacientes diagnosticados com LMA Leucemia Linfoblástica Aguda (LLA), que foram organizadas por tipo de leucócito, totalizando 14.822 imagens.

O segundo foi o conjunto de imagens do Hospital Universitário de Munique [Matek et al. 2019], o qual é composto por imagens de esfregaços sanguíneos (extensão .tiff, 400 × 400 pixels) contendo leucócitos individuais de 100 pacientes diagnosticados com LMA. Os dados foram coletados entre 2014 e 2017, sendo considerados para este projeto as 6 principais classes de leucócitos maduros: Basófilos, Neutrófilos Segmentados, Neutrófilos de Banda, Eosinófilos, Monócitos e Linfócitos Típicos, resultando em 7.073 imagens. O terceiro foi o conjunto de dados Laboratório Central do Hospital Clínico de Barcelona [Boldú et al. 2021], que possui 9.034 imagens (extensão .jpg, 360 × 363 pixels) referentes a 5 categorias celulares anotadas por patologistas especializados: neutrófilos, eosinófilos, basófilos, linfócitos e monócitos.

Ao restringir a análise a seis classes, priorizou-se os tipos celulares mais relevantes para o diagnóstico da LMA, especialmente os mieloides maduros. Essa seleção também favorece uma base mais diversa e representativa, aumentando a confiabilidade da classificação. A Tabela 1 apresenta a organização da base empregada.

Tabela 1. Resumo da combinação das bases de imagens utilizadas.

Leucócitos Maduros	Johns Hopkins	Munique	Barcelona	Total por Classes
Basófilo	79	51	1218	1348
Eosinófilo	424	107	3117	3648
Linfócito (Típico)	3937	3412	0	7349
Monócito	1789	1311	1420	4520
Neutrófilo (Banda)	109	170	1633	1912
Neutrófilo (Segmentado)	8484	2022	1646	12152
Total	**14822**	**7073**	**9034**	**30929**

3.3. Aumento de dados

O aumento de dados consiste em técnicas que ampliam a diversidade e a qualidade do conjunto de dados, melhorando o desempenho dos modelos de aprendizado de máquina ao gerar novas amostras a partir das existentes. Segundo Mumuni e Mumuni [Mumuni and Mumuni 2022], conjuntos de dados maiores, mais diversificados e representativos melhoram a eficácia dos modelos de aprendizado profundo em dados não observados. Para evitar problemas como *overfitting*, durante o treinamento, transformações dinâmicas foram aplicadas em cada época, incluindo rotação de 20 graus, deslocamento horizontal e vertical de até 20% da largura e altura da imagem, variação de zoom em 20% e inversão horizontal para espelhamento. Essas técnicas foram aplicadas na base de dados completa, o que contribui para uma maior robustez do modelo.

3.4. Redes Neurais Convolucionais

As Redes Neurais Convolucionais (CNNs) constituem uma abordagem amplamente empregada no campo do aprendizado profundo, notadamente eficaz no tratamento de dados com estrutura espacial, como imagens. Essas redes são compostas por camadas convolucionais que operam aplicando filtros sobre os dados de entrada, possibilitando a extração automática de atributos relevantes por meio da identificação de padrões visuais.

No contexto de classificação de imagens, as CNNs demonstram elevado desempenho ao distinguir elementos com base em suas características morfológicas [Tajbakhsh et al. 2016]. No presente estudo, foram analisadas diferentes arquiteturas de CNNs previamente treinadas sobre o banco de dados ImageNet, que reúne milhões de imagens categorizadas. As arquiteturas exploradas incluíram MobileNet, ResNet50, DenseNet121, MobileNetV2, ResNet101, VGG19, DenseNet201 e EfficientNetB3, selecionadas por sua expressiva acurácia nos cenários de validação da base ImageNet.

3.5. Transferência de Aprendizagem e Validação Cruzada

A técnica de transferência de aprendizagem consiste na reutilização de modelos previamente treinados em grandes bases de dados, como a ImageNet, para aplicação em novas tarefas, aproveitando o conhecimento previamente adquirido durante o treinamento original. No presente estudo, empregou-se a abordagem conhecida como *Shallow Fine-Tuning* (SFT), onde apenas as camadas finais da rede são ajustadas para a nova tarefa, enquanto as camadas convolucionais iniciais permanecem congeladas [Tajbakhsh et al. 2016]. Essa estratégia se mostra eficiente do ponto de vista computacional e é particularmente útil quando o volume de dados disponíveis é limitado, permitindo que o modelo alcance desempenho satisfatório com menor tempo de treinamento e menor demanda de recursos.

Além disso, o projeto empregou a técnica de validação cruzada k-*fold* para a avaliação da performance dos modelos, mediante a divisão do conjunto de dados em 5 partes (*folds*). Em cada iteração, um *fold* foi utilizado como conjunto de teste, enquanto os demais foram destinados ao treinamento (80%) e validação (20%). O processo foi repetido cinco vezes, assegurando que cada fold fosse utilizado uma vez como teste. Essa abordagem preserva a proporção das classes em cada subdivisão, o que é essencial em conjuntos de dados desbalanceados, minimizando vieses e evitando *overfitting*.

3.6. Comitês

A abordagem de comitês combina as forças de diferentes classificadores, aumentando a precisão e a robustez das previsões, e melhorando a capacidade de generalização [Dietterich 2000]. Este estudo utilizou as quatro CNNs com melhor desempenho individual para gerar todas as combinações possíveis de comitês de três redes, totalizando quatro grupos distintos. Ademais, para cada comitê foi aplicada a técnica de voto majoritário, na qual a decisão final é tomada baseando-se na maioria das previsões dos modelos. No comitê de voto majoritário, cada CNN realiza sua previsão e a decisão final é tomada baseada na maioria dos votos desses modelos, o que proporciona uma classificação mais fidedigna, uma vez que combina diversas perspectivas.

3.7. Métricas de Avaliação

Para avaliar o desempenho do modelo, foram utilizadas as métricas de acurácia, precisão, *recall*, F1-*Score* e o índice Kappa [Powers 2020]. A acurácia representa a proporção de previsões corretas em relação ao total de previsões realizadas, proporcionando uma visão ampla da eficácia do modelo. A precisão, por sua vez, expressa a taxa de verdadeiros positivos entre todas as previsões positivas, indicando o nível de exatidão do modelo ao classificar uma instância como positiva. O *recall*, também chamado de sensibilidade, corresponde à fração de verdadeiros positivos em relação ao total de casos que realmente pertencem à classe positiva, demonstrando a capacidade do modelo de identificar corretamente essas ocorrências.

Já o F1-*Score*, calculado como a média harmônica entre precisão e *recall*, busca equilibrar essas duas métricas, sendo especialmente relevante quando há um desbalanceamento entre as classes. Além disso, o índice Kappa, conforme definido por [Rosenfield and Fitzpatrick-Lins 1986] é um coeficiente que mede a concordância entre as previsões do modelo e os valores reais, considerando a influência do acaso e da discordância esperada. O cálculo dessa métrica é apresentado na Fórmula 1, na qual o valor observado representa a acurácia real do modelo, enquanto o valor esperado corresponde à acurácia que seria obtida ao acaso. Valores de Kappa acima de 80% são interpretados como indicando forte concordância, enquanto valores entre 60% e 80% representam concordância substancial, e valores abaixo de 40% indicam concordância fraca.

$$\text{Kappa} = \frac{observado - esperado}{1 - esperado} \times 100 \tag{1}$$

4. Resultados e Discussão

Os experimentos foram conduzidos com as técnicas previamente descritas, como transferência de aprendizagem utilizando redes pré-treinadas e o *Shallow Fine Tuning* (SFT),

técnicas de validação cruzada estratificada como o k-*fold Cross Validation* com k = 5 *folds* e o uso de oito CNNs para treinamento, além da seleção das redes de melhores métricas para composição e avaliação de comitês. Os experimentos detalhados nesta Seção foram alcançados utilizando um computador com processador AMD Ryzen 5 5600 (3.50 GHz, 6 núcleos), 16 GB de RAM e GPU RTX 3060 Ti (8 GB, 4864 núcleos CUDA).

Os modelos foram treinados individualmente e os resultados das métricas estão expressos na Tabela 2. Nesse sentido, nota-se que a MobileNetV2 obteve a melhor acurácia dentre as redes individuais (91,50% ± 0,44) e um coeficiente Kappa de 89,00% ± 0,57, destacando-se pela alta precisão e estabilidade na classificação. O destaque dessa performance pode ser atribuído à sua arquitetura com blocos residuais invertidos, que favorecem a extração eficiente de padrões visuais relevantes. Além disso, por possuir menos parâmetros, apresenta menor risco de *overfitting* e maior eficiência computacional, características vantajosas tanto no treinamento quanto na aplicação clínica.

Tabela 2. Comparação dos resultados na identificação de leucócitos maduros, com desvios padrões.

Arquitetura	Acurácia(%)	Precisão(%)	*Recall*(%)	F1-*Score*(%)	Kappa(%)
MobileNetV2	**91,50 ± 0,44**	**92,18 ± 0,38**	**91,50 ± 0,44**	**91,66 ± 0,43**	**89,00 ± 0,57**
ResNet50	91,32 ± 0,65	91,87 ± 0,54	91,32 ± 0,65	91,42 ± 0,63	88,72 ± 0,84
MobileNet	91,19 ± 0,71	91,91 ± 0,46	91,19 ± 0,71	91,37 ± 0,65	88,60 ± 0,90
ResNet101	90,99 ± 0,40	91,86 ± 0,25	90,99 ± 0,40	91,19 ± 0,37	88,33 ± 0,50
DenseNet201	89,92 ± 0,79	91,13 ± 0,45	89,92 ± 0,79	90,16 ± 0,72	86,94 ± 1,01
EfficientNetB3	89,75 ± 0,61	91,05 ± 0,52	89,75 ± 0,61	90,05 ± 0,60	86,79 ± 0,78
DenseNet121	87,40 ± 0,81	88,48 ± 0,71	87,40 ± 0,81	87,51 ± 0,82	83,69 ± 1,03
VGG19	84,64 ± 0,52	85,75 ± 0,29	84,64 ± 0,52	84,74 ± 0,43	80,14 ± 0,61

Em seguida, visando investigar se a combinação dos modelos poderia trazer ganho de performance, foram formados quatro comitês compostos por combinações de três redes entre as quatro CNNs de melhor desempenho individual, sendo elas a MobileNetV2, ResNet50, MobileNet e ResNet101. A composição de cada comitê foi a seguinte: c1 (MobileNetV2, ResNet101 e MobileNet), c2 (MobileNetV2, ResNet50 e MobileNet), c3 (MobileNetV2, ResNet50 e ResNet101) e c4 (ResNet50, ResNet101 e MobileNet).

Com base na análise da Tabela 3, nota-se que a combinação do processamento dos modelos trouxe aumentos em todas as métricas quando comparado aos modelos individuais. Dá-se destaque ao comitê c1, o qual obteve o melhor desempenho com valores de acurácia de 93,18% ± 0,42 e índice Kappa de 91,16% ± 0,53, indicando maior precisão e estabilidade. Esses resultados demonstram que a abordagem de comitês melhora a robustez e a acurácia do sistema de classificação. A complementaridade entre as arquiteturas combinadas permitiu atenuar limitações individuais, o que contribuiu para decisões mais confiáveis e consistentes, um aspecto essencial em contextos clínicos.

Tabela 3. Desempenho dos comitês formados pelas combinações entre as quatro CNNs com melhor desempenho individual, com desvios padrões.

Comitê	Acurácia(%)	Precisão(%)	*Recall*(%)	F1-*Score*(%)	Kappa(%)
c1	**93,18 ± 0,42**	**93,62 ± 0,31**	**93,18 ± 0,42**	**93,28 ± 0,40**	**91,16 ± 0,53**
c2	93,15 ± 0,41	93,51 ± 0,33	93,15 ± 0,41	93,23 ± 0,39	91,11 ± 0,52
c3	92,92 ± 0,24	93,35 ± 0,21	92,92 ± 0,24	93,01 ± 0,23	90,81 ± 0,31
c4	92,79 ± 0,30	93,22 ± 0,22	92,79 ± 0,30	92,88 ± 0,28	90,64 ± 0,38

Devido ao desempenho expressivo do comitê c1, também foi investigada a possibilidade de ocorrência de vestígios de *overfitting* durante o treinamento dos modelos que o compõem. As Figuras 3 e 4 apresentam os comportamentos das curvas de acurácia e *loss* das redes ResNet101, MobileNet e MobileNetV2.

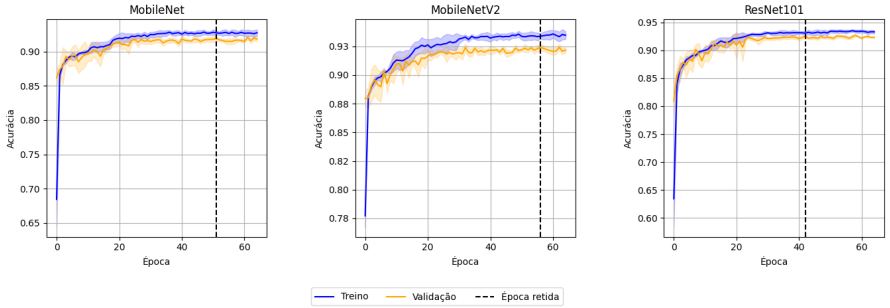


Figura 3. Evolução da acurácia por época no treinamento do comitê c1, formado por ResNet101, MobileNet e MobileNetV2.

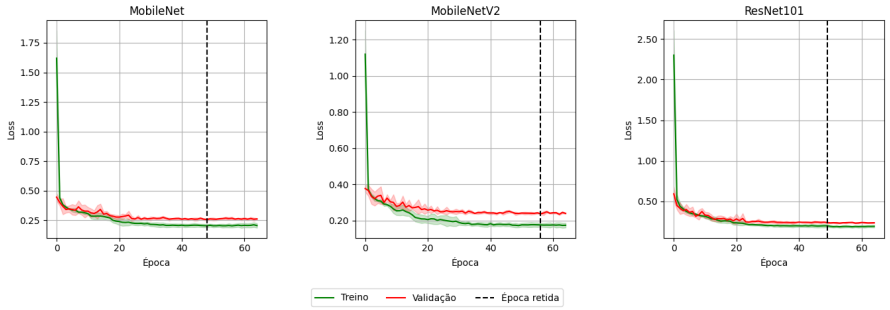


Figura 4. Evolução da *loss* por época no treinamento do comitê c1, formado por ResNet101, MobileNet e MobileNetV2.

A MobileNet apresentou comportamento estável ao longo do treinamento, com curvas de acurácia e *loss* próximas entre os conjuntos de treino e validação, indicando baixo risco de *overfitting*. Já a MobileNetV2 obteve a maior acurácia entre as redes e manteve os valores de *loss* baixos, reforçando sua efetividade na tarefa de classificação. Por fim, a ResNet101 demonstrou aprendizado consistente, com curvas suavemente crescentes e pouca diferença entre treino e validação, o que evidencia um treinamento equilibrado. Tais análises indicam que o treinamento foi conduzido de modo adequado e estratégico, ocasionando em alto potencial de generalização pela complementaridade dos modelos.

Além disso, foram gerados mapas de calor Grad-CAM para a análise dos padrões de atenção dos modelos do comitê c1, sendo que a seleção das classes foi baseada nas características visuais das imagens. A Figura 5 mostra que a ResNet101 e a MobileNetV2 concentram suas atenções nas regiões nucleares das células, especialmente em eosinófilos e neutrófilos segmentados, onde os gradientes coincidem com áreas morfologicamente relevantes. Já a MobileNet exibiu ativações mais difusas em alguns casos, o que pode explicar seu desempenho inferior. A diversidade nos padrões de ativação entre os modelos, evidenciada pelos mapas Grad-CAM, sugere que cada rede extrai informações distintas das imagens. Esse direcionamento complementar da atenção fortalece a decisão final do comitê, pois combina múltiplas perspectivas sobre regiões morfologicamente relevantes.

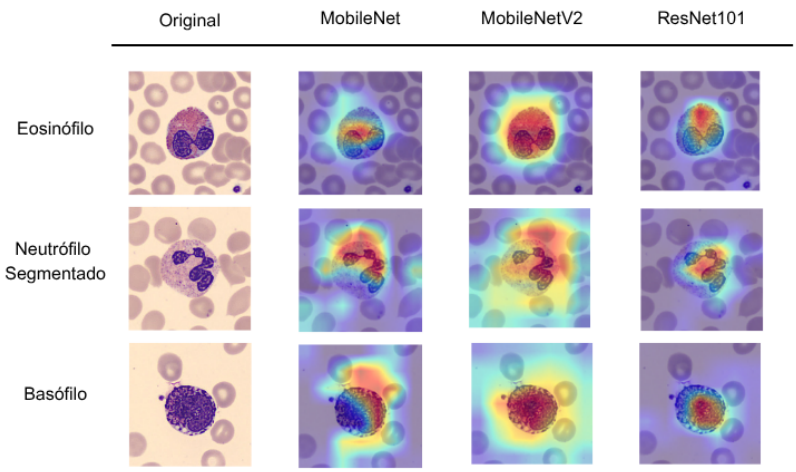


Figura 5. Visualizações Grad-CAM para três subtipos de leucócitos geradas pelos modelos do comitê c1: MobileNet, MobileNetV2 e ResNet101.

A Figura 6 apresenta a comparação do desempenho do comitê em relação aos seus integrantes por meio de um gráfico radar. Percebe-se que o comitê c1 supera os demais em todos os indicadores avaliados. Cabe ressaltar que o aumento expressivo no valor Kappa destaca uma melhoria significativa na concordância entre as previsões do sistema e as classificações de referência, reforçando a confiabilidade do modelo. Esses resultados evidenciam que o uso de comitês não apenas eleva o desempenho geral, mas também reduz a variabilidade e promove maior robustez e capacidade de generalização.

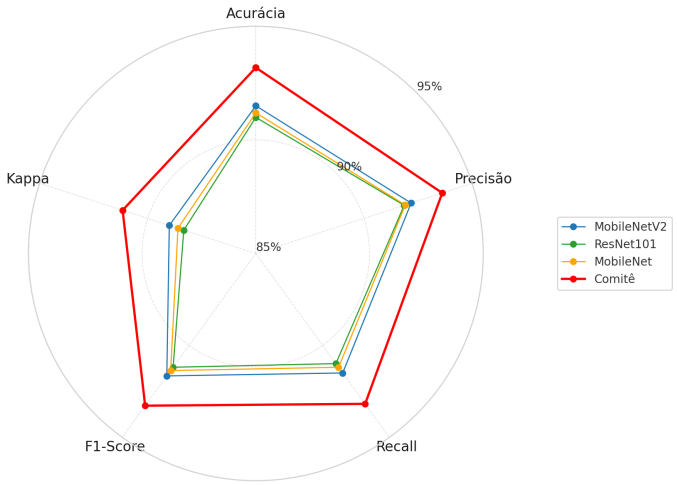


Figura 6. Desempenho do comitê c1, composto por ResNet101, MobileNet e MobileNetV2, em comparação com cada rede isoladamente.

Embora o uso de comitês represente um custo computacional mais elevado em comparação aos modelos individuais, a inclusão de arquiteturas leves, como a MobileNet e a MobileNetV2, contribuiu para reduzir esse impacto, que permitiu uma execução eficiente, garantindo um equilíbrio adequado entre desempenho e economia de recursos. Tais características tornam a abordagem relevante para aplicações clínicas, em que a agilidade na resposta e a viabilidade computacional são fatores determinantes para a adoção de sistemas automatizados de apoio ao diagnóstico.

5. Conclusão

O presente estudo apresentou uma abordagem baseada em CNNs para a classificação multiclasse de leucócitos maduros em imagens de esfregaços sanguíneos, com o objetivo de apoiar o diagnóstico da Leucemia Mieloide Aguda (LMA). Foram avaliadas oito arquiteturas de CNNs pré-treinadas, além da formação de quatro comitês com voto majoritário, compostos por combinações de três redes entre os quatro modelos de melhor desempenho individual. O comitê formado pelas redes MobileNetV2, ResNet101 e MobileNet obteve os melhores resultados, alcançando acurácia de 93,18% e superando o desempenho isolado dos modelos. Esses achados reforçam o potencial das técnicas de aprendizado profundo como ferramentas eficazes na análise morfológica automatizada em exames hematológicos.

Apesar dos avanços, o estudo evidenciou desafios importantes. Entre eles, destaca-se o desbalanceamento entre as classes do conjunto de dados, fator que pode prejudicar o desempenho em categorias com menor representatividade. Além disso, embora a base utilizada seja numerosa, a limitação na diversidade de fontes pode restringir a generalização dos modelos em diferentes contextos clínicos.

Como direções futuras, recomenda-se a adoção de estratégias para mitigar o desbalanceamento, como técnicas de *oversampling*, aumento de dados direcionado e realizado de forma *offline*, bem como o uso de funções de perda ponderadas, como a *focal loss*. Ademais, a aplicação do *Deep Fine-Tuning* (DFT), com o ajuste completo de todas as camadas das redes convolucionais, surge como uma abordagem promissora para ampliar a capacidade de adaptação dos modelos à tarefa específica de classificação de leucócitos.

Referências

Acevedo, A., Merino, A., Alférez, S., Molina, , Boldú, L., and Rodellar, J. (2020). A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. *Data in Brief*, 30:105474.

American Cancer Society (2024). Key statistics for acute myeloid leukemia (aml). `https://www.cancer.org/cancer/types/acute-myeloid-leukemia/about/key-statistics.html`. Accessed: 2025-03-16.

Asghar, R., Kumar, S., and Mahfooz, A. (2023). Classification of blood cells using deep learning models. arXiv preprint arXiv:2308.06300. Accessed: 2025-03-16.

Boldú, L., Merino, A., Acevedo, A., Molina, , 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.

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., and et al. (2013). The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of Digital Imaging*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pages 1–15. Springer.

Ding, Y., Tang, X., Zhuang, Y., Mu, J., Chen, S., Liu, S., Feng, S., and Chen, H. (2023). Leukocyte subtype classification with multi-model fusion. *Medical & Biological Engineering & Computing*, 61:2305–2316.

Doi, K. (2005). Current status and future potential of computer-aided diagnosis in medical imaging. *The British Journal of Radiology*, 78:S3–S19.

Dwivedi, A. K. (2018). Artificial neural network model for effective cancer classification using microarray gene expression data. *Neural Computing and Applications*, 29(12):1545–1554.

Döhner, H., Estey, E. H., Amadori, S., Appelbaum, F. R., Buchner, T., Burnett, A. K., and et al. (2010). Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the european leukemianet. *Blood*, 115:453–474.

Labati, R. D., Piuri, V., and Scotti, F. (2011). All-idb: the acute lymphoblastic leukemia image database for image processing. In *IEEE International Conference on Image Processing (ICIP)*.

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. 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*.

Sidhom, J.-W., Siddarthan, I. J., Lai, B.-S., Luo, A., Hambley, B. C., Bynum, J., and 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.-H., and Kwon, K.-R. (2018). Leukemia blood cell image classification using convolutional neural network. *International Journal of Computer Theory and Engineering*, 10:54–58.

Travlos, G. S. (2006). Normal structure, function, and histology of the bone marrow. *Toxicologic Pathology*, 34(5):548–565.