Proceedings of the 12th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS 2026)
August 16 – 18, 2026 – London, United Kingdom
Paper No. MVML 110
DOI: 10.11159/mvml26.110

Linear Probing Reveals Hidden Transferability: A Multi-Domain CLIP Encoder Benchmark for White Blood Cell Classification

Reyhan Hoşavcı[1], Sümeyye Zülal Dik[2], Zeliha Kaya Akçelik[2] , Zeki Kuş[3], Musa Aydin[4]
[1]Department of Biomedical Engineering/Fatih Sultan Mehmet Vakif University
Istanbul, Turkiye
rgurleyen@fsm.edu.tr
[2]Department of Computer Engineering/Fatih Sultan Mehmet Vakif University
Istanbul, Turkiye
szdik@fsm.edu.tr; zkaya@fsm.edu.tr
[3]Department of Artificial Intelligence and Data Engineering/Fatih Sultan Mehmet Vakif University
Istanbul, Turkiye
zkus@fsm.edu.tr
[4]Department of Artificial Intelligence and Data Engineering/Samsun University
Samsun, Turkiye
musa.aydin@samsun.edu.tr

Abstract **-** White blood cell (WBC) classification from peripheral blood smear images is essential for diagnosing infections, immune disorders, and hematological malignancies. Vision-language models based on Contrastive Language-Image Pretraining (CLIP) offer a promising approach by enabling inference with minimal or no labelled data, yet no prior study has jointly examined models across multiple pretraining domains, prompt strategies, and supervision levels for white blood cell classification. In this work, we present CLIPBlood, a comprehensive benchmark of six CLIP-based encoders for WBC classification on the WBCAtt dataset comprising 10,298 images across five leukocyte classes. The evaluated models span three pretraining domains: general (OpenCLIP ViT-B/32, ViT-L/14), biomedical (BiomedCLIP, PubMedCLIP), and pathology (PLIP, QuiltNet). We assess all models under two complementary settings: zero-shot classification with five prompt templates of increasing domain specificity, and few-shot linear probing with $k \in \{1, 5, 10, 25, 50, 100\}$ labelled samples per class. Our results show that zero-shot performance remains poor across all 30 model-prompt combinations, with no configuration exceeding 0.26 macro-F1, indicating that current CLIP encoders lack the text-image alignment necessary for fine-grained blood cell classification. In contrast, few-shot linear probing dramatically improves performance, confirming that frozen CLIP encoders capture discriminative visual structure that can be leveraged with minimal supervision. Among all models, BiomedCLIP consistently achieves the highest scores across all values of k, reaching 87.5% macro-F1 with only 100 samples per class, despite ranking among the lowest in zero-shot performance. This paradox highlights that visual representation quality, rather than text-image alignment, is the primary determinant of downstream performance. The t-SNE visualizations further confirm that BiomedCLIP produces the most separable embedding clusters. Overall, our findings demonstrate that pretraining domain alignment is a stronger predictor of performance than model scale, and that BiomedCLIP is the most suitable foundation encoder for WBC analysis under low-resource conditions.

Keywords**:** White Blood Cell Classification, Vision-Language Models, CLIP, Zero-Shot Learning, Few-Shot Linear Probing

1. Introduction

White blood cell (WBC) classification is central to the diagnosis of infections, immune system disorders, and hematological malignancies. Accurate identification of leukocyte subtypes from peripheral blood smear images is a fundamental component of clinical decision support systems. However, distinguishing between these cell types is inherently challenging, as it relies on subtle morphological differences such as nuclear structure, cytoplasmic granularity, and cell size. These fine-grained characteristics not only require expert-level interpretation but also introduce variability and potential diagnostic errors. Therefore, the development of automated and reliable WBC classification systems remains an important research direction.

While earlier approaches based on handcrafted features and convolutional neural networks have progressively improved WBC classification, these methods remain heavily dependent on large-scale annotated datasets, which are costly and time-

consuming to obtain in medical domains. More recently, multimodal approaches combining visual and textual modalities have been proposed to enhance both accuracy and interpretability [1], yet the systematic evaluation of such models in hematological analysis remains limited.

To mitigate this limitation, foundation models, and in particular vision-language models (VLMs), have emerged as promising alternatives. These models are pretrained on large-scale image–text pairs, enabling them to learn transferable representations that can be applied to downstream tasks with minimal or no labelled data. A key milestone in this paradigm is Contrastive Language–Image Pretraining (CLIP), which aligns visual and textual representations in a shared embedding space, enabling zero-shot inference without task-specific training [2]. However, since CLIP models are predominantly trained on natural images, they often fail to capture the fine-grained morphological features required for hematological image analysis.

Recent studies have highlighted that model performance is strongly influenced not only by architectural capacity but also by the alignment between the pretraining data distribution and the target domain. Kukuljan et al. [3] demonstrated that domain-specific models outperform general-purpose vision-language models in cytomorphology tasks, though their evaluation was limited to a single model comparison without systematic prompt analysis. Similarly, Patel et al. [4] evaluated CLIP and its biomedical variants on hematology images but focused on a restricted set of models without exploring few-shot adaptation or representation quality. Neither study offers a systematic multi-model comparison spanning general, biomedical, and pathology pretraining domains, nor do they examine whether strong zero-shot performance is a reliable predictor of representational quality under few-shot adaptation.

Another key factor influencing the performance of vision-language models is the design of prompt templates. In zero-shot scenarios, textual prompts play a critical role in aligning visual and textual representations, thereby directly affecting classification performance and introducing significant variability across tasks. Recent studies have demonstrated that learnable prompt strategies can substantially improve model performance by adapting textual inputs to downstream tasks [5]. However, these approaches require task-specific optimization and have primarily been validated on natural image datasets. In contrast, our work investigates whether carefully designed hand-crafted prompt templates that require no additional training can provide meaningful performance gains in the hematological domain.

More specifically, medical imaging tasks, particularly hematological image analysis, require sensitivity to subtle, fine-grained morphological features that are rarely captured in generic text descriptions. Consequently, the role of prompt formulation in this domain remains a significantly under-researched area. In particular, the impact of domain-specific prompt templates on zero-shot performance in white blood cell classification has not been systematically investigated in the existing literature.

In this work, we present CLIPBlood, a comprehensive and systematic benchmark of vision-language encoders for white blood cell classification. We evaluate six CLIP-based models spanning three pretraining domains, namely general (OpenCLIP ViT-B/32, ViT-L/14), biomedical (BiomedCLIP, PubMedCLIP), and pathology (PLIP, QuiltNet), on the WBCAtt dataset under both zero-shot and few-shot linear probing settings. Furthermore, we analyse the effect of five domain-specific prompt templates on zero-shot performance and provide qualitative insights through t-SNE embedding visualizations. Our results reveal a striking contrast between zero-shot and few-shot performance, and demonstrate that pretraining domain alignment emerges as a stronger determinant of representation quality than model scale or architectural capacity.

2. Methods

We evaluate six CLIP-based models grouped by pretraining domain: general (OpenCLIP ViT-B/32, ViT-L/14 [6]), biomedical (BiomedCLIP [7], PubMedCLIP [8]), and pathology (PLIP [9], QuiltNet [10]). CLIP models learn a joint embedding space for images and text through contrastive pretraining on large-scale image–text pairs [2], enabling downstream inference without task-specific training. This domain grouping enables a systematic analysis of how pretraining domain proximity to the target task affects downstream performance. Each model is evaluated under two complementary strategies: zero-shot classification, which requires no labelled data, and few-shot linear probing, which trains a lightweight classifier on frozen embeddings.

2.1. Zero-Shot Classification

In the zero-shot setting, no labelled training data is used at any stage. At inference, each of the five WBCAtt class labels is inserted into a text template to produce class-level text embeddings; the predicted class is determined by the highest cosine similarity between the image embedding and these text embeddings [2]. This procedure is applied identically across all six models.

The choice of text template is known to substantially affect zero-shot performance [2]. To quantify this sensitivity in the blood cell domain, we design five prompt templates spanning a specificity gradient from generic visual description to domain-specific clinical language: (1) *"a photo of a {label}"*, (2) *"a microscopy image of a {label}"*, (3) *"a peripheral blood smear image showing a {label}*", (4) *"a hematology laboratory image of a {label} white blood cell"*, and (5) *"a Romanowsky-stained blood smear image of a {label}"*. Templates (1) and (2) serve as domain-agnostic baselines, while templates (3)–(5) progressively incorporate hematology-specific context. This design is further motivated by domain-specific prompting practices in medical CLIP models [7, 9]. All 30 model–prompt combinations are evaluated on the WBCAtt test set.

2.2. Few-Shot Linear Probing

Few-shot linear probing evaluates the quality of the visual representations learned by each CLIP encoder when a small amount of labelled data is available. In contrast to full fine-tuning, encoder weights are kept entirely frozen throughout; only a lightweight linear classifier trained on top of the frozen embeddings is optimized. This design isolates performance differences to the representations acquired during pretraining, rather than to each model's capacity to adapt to new data.

We follow a three-stage evaluation protocol. First, each frozen encoder extracts L2-normalized embeddings for the entire WBCAtt train, validation, and test sets once. Second, a multinomial logistic regression classifier (C=1) is trained on $k \in \{1, 5, 10, 25, 50, 100\}$ examples per class sampled from the training set. Third, the best k is selected based on validation set macro-F1; the test set is evaluated exactly once using this configuration and is never consulted during any model selection decision, strictly preventing data leakage.

To account for sampling variability at small k, each configuration is repeated with three random seeds (0,1,2), and test macro-F1 is reported as mean ±standard deviation. Due to class imbalance in WBCAtt, macro-F1 is adopted as the primary metric throughout, assigning equal weight to each class regardless of frequency.

3. Experimental Study

3.1. Dataset

We conduct all experiments on the WBCAtt dataset [11], a peripheral blood smear benchmark comprising 10,298 microscopic cell images across five white blood cell types: basophil, eosinophil, lymphocyte, monocyte, and neutrophil. The images originate from the PBC collection [12], acquired using a CellaVision DM96 analyser with May Grünwald- Giemsa staining. While WBCAtt additionally provides expert-verified morphological attribute annotations covering nuclear structure, cytoplasmic properties, and granule characteristics, our study uses only the cell-type labels, as the benchmark targets encoder-level discriminability rather than attribute prediction. We adopt the official three-way split of 6,169 training, 1,030 validation, and 3,099 test images [11]. Training samples serve exclusively for few-shot sampling, the validation set guides model selection, and the test set is evaluated only once per final configuration to prevent data leakage. The class distribution is imbalanced—neutrophils comprise approximately 33% of samples while lymphocytes and basophils each represent roughly 12%—which motivates the use of macro-F1 as the primary metric, assigning equal weight to each class regardless of frequency. The standardized acquisition protocol and clinically grounded five-class taxonomy make WBCAtt a suitable testbed for evaluating vision-language encoders on fine-grained hematological distinctions, particularly between morphologically similar types such as monocytes and lymphocytes.

3.2. Implementation Details

All experiments were conducted on an NVIDIA RTX A5000 (24 GB) GPU (CUDA 11.8) using the PyTorch framework. Model loading and inference were handled through two different APIs: models based on the OpenCLIP library (OpenCLIP

ViT- B/32, ViT-L/14, BiomedCLIP, QuiltNet) were run with a batch size of 64, while models using the HuggingFace Transformers API (PLIP, PubMedCLIP) were run with a batch size of 32. During data loading, all images were converted to RGB format; resizing and normalization were delegated entirely to each model's own preprocessing pipeline. This ensures that no externally imposed preprocessing is applied, preserving the input conditions consistent with each model's original training distribution. All source code and experimental configurations are publicly available at GitHub to ensure full reproducibility.

For zero-shot inference, both image and text embeddings were L2-normalized to unit length prior to cosine similarity computation, ensuring that similarity scores reflect angular distance rather than magnitude. Class assignment was performed by selecting the text embedding with the highest cosine similarity to the image embedding. All inference operations were executed without gradient computation, minimizing memory overhead and processing time.

For few-shot experiments, embeddings were extracted once per model and cached for reuse during the linear probing stage, avoiding redundant forward passes across different k configurations. The logistic regression classifier was optimized using the L-BFGS method with a maximum of 1000 iterations. To ensure reproducibility, all stochastic processes, including training set sampling and classifier initialization, were initialized with fixed random seeds prior to each experiment.

4. Results and Discussion

4.1. Zero-Shot Classification

Zero-shot performance across all six models and five prompt templates is summarized in Fig. 1. All models perform poorly in this setting, with macro-F1 scores remaining below 0.26 across all 30 model–prompt combinations. This result is not entirely surprising. CLIP models are trained to align natural language descriptions with visual content at a coarse semantic level. White blood cell classification requires sensitivity to fine-grained morphological differences, such as nuclear shape, cytoplasmic granularity, and cell size. These features are rarely described in the image-text pairs used during general or biomedical pretraining. As a result, the learned text-image alignment is insufficient to resolve fine-grained intra-class variation in peripheral blood smear images.

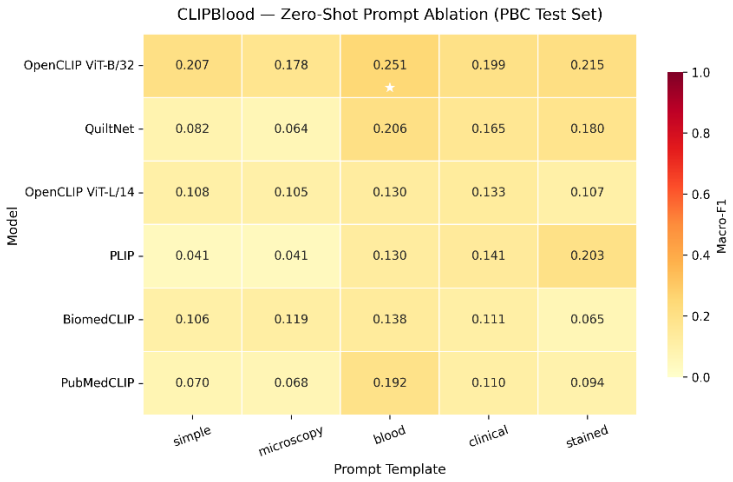


Fig. 1: Zero-shot prompt ablation heatmap showing macro-F1 scores for each model–prompt combination on the WBCAtt test set (3,099 images, 5 classes). The best overall result is marked with ⋆.

Among all 30 combinations, OpenCLIP ViT-B/32 paired with the blood prompt achieves the highest zero-shot macro-F1 of 0.251. This is somewhat counterintuitive, as one might expect biomedical or pathology-domain models to benefit more from domain-specific prompts. A reasonable explanation is that general-domain models are trained on broader and more diverse text corpora, which may include a wider range of descriptive language for biological entities. Biomedical models

such as BiomedCLIP and PubMedCLIP, although trained on medical image-text pairs, may have been primarily exposed to radiological or histological descriptions. This limits their zero-shot sensitivity to blood cell morphology.

Domain-specific prompts (blood, clinical, stained) generally outperform generic alternatives (simple, microscopy), though the optimal template varies across models. This sensitivity to prompt formulation highlights a fundamental limitation of zero-shot CLIP inference: performance depends heavily on alignment between the prompt vocabulary and the model's pretraining distribution. The lack of a universally optimal prompt across models suggests that prompt engineering alone is insufficient to bridge the gap between general visual-language pretraining and the specialized demands of hematological image classification.

4.2. Few-Shot Linear Probing

Few-shot linear probing results are presented in Fig. 2 and Table 1. Even a small number of labelled samples significantly enhances performance across all models. This indicates that the frozen CLIP encoders effectively capture meaningful visual structure that can be leveraged with minimal supervision. This transition from near-random zero-shot performance to competitive few-shot accuracy underscores the importance of representation quality over zero-shot text alignment for this task.

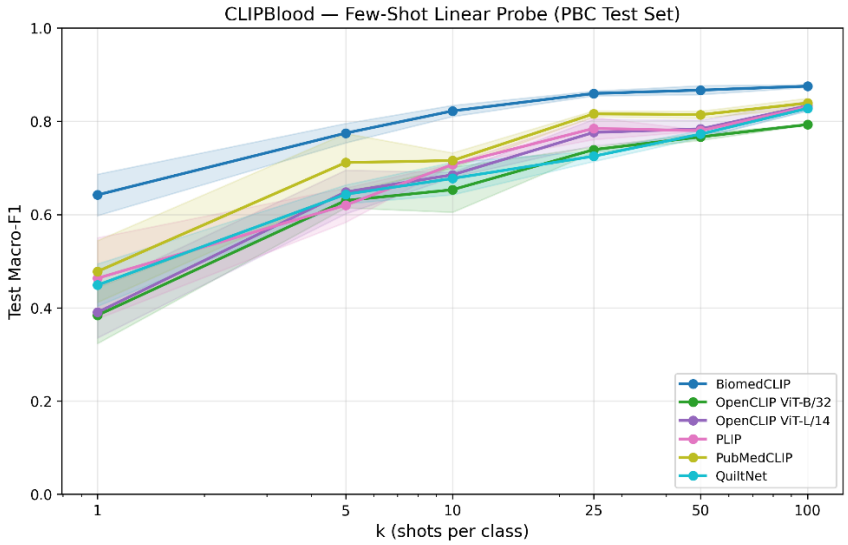


Fig. 2: Few-shot linear probe results (test macro-F1) on the WBCAtt dataset across varying numbers of labelled samples per class ($k \in \{1, 5, 10, 25, 50, 100\}$). Shaded regions indicate ±1 standard deviation over 3 random seeds. BiomedCLIP consistently outperforms all other encoders, achieving 0.64 macro-F1 with only one sample per class and reaching 0.875 at *k=100*.

BiomedCLIP consistently outperforms all other encoders across all values of k by a substantial margin. With only one labelled sample per class, BiomedCLIP achieves a macro-F1 of 64.2 ± 4.4, compared to 47.8 ± 6.6 for PubMedCLIP and 46.3 ± 8.7 for PLIP. At k=100, BiomedCLIP reaches a macro-F1 of 87.5 ± 0.3, demonstrating that its representations are highly discriminative even in the low-data regime. For context, a fully supervised ResNet50 trained on the complete training set (6,169 samples) achieves 99.54% macro-F1 [11], indicating that the remaining gap is attributable to the substantially smaller labelled budget rather than encoder quality. The primary reason for BiomedCLIP's superiority likely lies in the nature of its pretraining data. BiomedCLIP is trained on PMC-15M, a large-scale dataset of biomedical figure-caption pairs sourced from PubMed Central, which encompasses a broad spectrum of medical imaging modalities, including microscopy, histology, and hematology. This diversity of biomedical visual content encourages the encoder to learn representations sensitive to fine-grained morphological differences, precisely the type of discriminability required for white blood cell classification.

Yet this representational richness does not translate to zero-shot performance. BiomedCLIP exhibits the most striking contrast between zero-shot and few-shot settings. Despite ranking among the lowest in zero-shot performance (best macro-F1 of 0.138), it achieves the highest few-shot scores across all values of k. This paradox suggests that BiomedCLIP's encoder captures highly discriminative visual features for blood cell morphology, but its text encoder lacks alignment with the specific vocabulary used in our prompt templates. This decoupling between visual representation quality and text-image alignment further supports our conclusion that representation quality, rather than zero-shot text alignment, is the primary determinant of downstream performance.

PubMedCLIP, despite also being a biomedical model, performs considerably worse than BiomedCLIP at low k, achieving only 47.8 ± 6.6 at k=1. This gap is likely attributable to differences in the scale and diversity of the pretraining data. PubMedCLIP is trained on a smaller subset of medical image-text pairs, which may not provide sufficient coverage of hematological imaging to produce representations as discriminative as those of BiomedCLIP. Nevertheless, PubMedCLIP recovers substantially at higher values of k, reaching 83.9 ± 1.0 at k=100. This suggests that its encoder captures relevant visual features, but it requires more labelled examples to reveal them through linear probing.

Among pathology-domain models, PLIP and QuiltNet perform comparably across most values of k, with PLIP achieving slightly higher scores at larger k (83.1 ± 0.4 vs. 82.8 ± 0.5 at k=100). Both models are pretrained on histopathology image-text pairs, which share some visual characteristics with blood smear microscopy—such as staining artifacts and cellular texture—but differ substantially in tissue structure and cell morphology. This partial domain overlap may explain why pathology models outperform general-domain models at low k.

Table 1: Few-shot linear probe macro-F1 scores (mean ± std over 3 random seeds) on the WBCAtt dataset (5 classes, 3,099 test images). Models are grouped by pretraining domain. For each domain group, the best result per column is shown in **bold**. The overall best per column is underlined.

Pretraining Domain	**Model**	***k*=1**	***k*=5**	***k*=10**	***k*=25**	***k*=50**	***k*=100**
General	ViT-B/32	38.4±6.0	63.0±1.6	65.3±4.8	73.9±0.6	76.7±0.7	79.3±0.1
	ViT-L/14	**39.0±5.5**	**64.8±4.7**	**68.5±0.5**	**77.6±2.8**	**78.4±1.9**	**83.3±1.0**
Biomedical	BiomedCLIP	**64.2±4.4**	**77.5±2.1**	**82.2±1.2**	**86.0±0.5**	**86.7±1.0**	**87.5±0.3**
	PubMedCLIP	47.8±6.6	71.1±6.3	71.6±1.7	81.6±0.6	81.4±0.8	83.9±1.0
Pathology	PLIP	**46.3±8.7**	62.0±3.7	**70.7±0.2**	**78.5±2.4**	**78.0±0.1**	**83.1±0.4**
	QuiltNet	44.9±4.5	**64.3±2.0**	67.7±3.5	72.5±1.1	77.3±0.8	82.8±0.5

Within the general-domain group, ViT-L/14 outperforms ViT-B/32 at most values of k, with the gap widening at higher sample counts (83.3 ± 1.0 vs. 79.3 ± 0.1 at k=100). This is consistent with the well-established scaling behaviour of vision transformers, where larger models tend to learn richer and more transferable representations. However, the performance gap between ViT-L/14 and the biomedical models remains large, indicating that model scale alone cannot compensate for the absence of domain-relevant pretraining data.

4.3. Representation Quality

The t-SNE visualizations in Fig. 3 support the quantitative results and reveal structural properties of each encoder's embeddings. BiomedCLIP produces the most clearly separated clusters across the five WBCAtt classes, with minimal inter-class overlap. This well-structured embedding geometry directly explains its strong few-shot performance: a linear classifier can easily find decision boundaries in a space where classes are already well-separated. PubMedCLIP and PLIP exhibit partial separation, with notable confusion between monocytes and lymphocytes. This overlap aligns with the morphological

similarity between these two cell types, both of which are mononuclear and similar in size. It also suggests that these encoders have not fully captured the fine-grained differences in cytoplasmic and nuclear features that distinguish them.

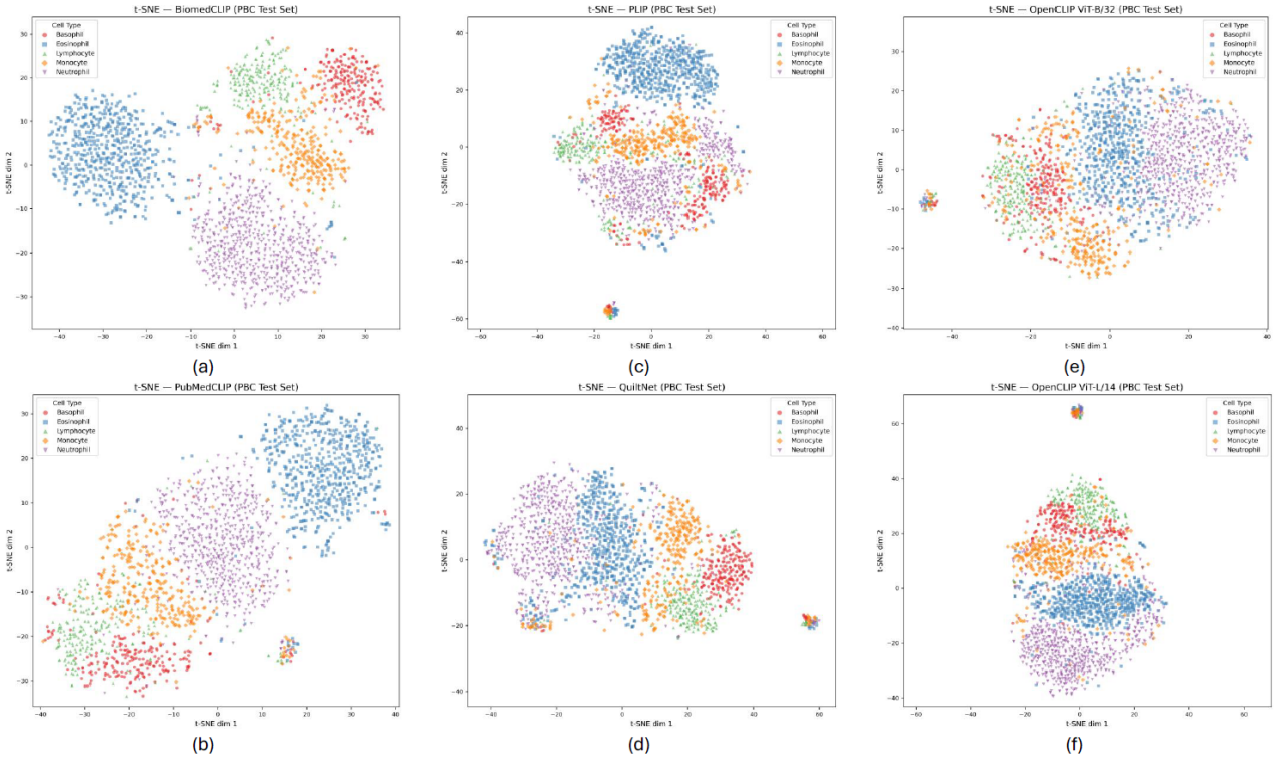


Fig. 3: t-SNE visualizations of frozen image embeddings from six CLIP encoders on the WBCAtt test set (5 classes). BiomedCLIP (a) produces the most separable clusters. PubMedCLIP (b) and PLIP (c) show partial separation with monocyte–lymphocyte overlap. QuiltNet (d) and OpenCLIP ViT-B/32 (e) exhibit considerable class mixing. OpenCLIP ViT-L/14 (f) displays a collapsed cluster structure suggesting poor discriminability.

QuiltNet and OpenCLIP ViT-B/32 show substantial class mixing, suggesting limited discriminability of their representations for this task. OpenCLIP ViT-L/14 displays a particularly striking pattern: a collapsed cluster structure in which most samples are concentrated in a single dense region. This collapse indicates that the encoder has learned representations that are mostly invariant to the morphological features important for white blood cell classification. This is likely because such features are not well represented in its natural image pretraining distribution.

Altogether, these findings support a clear conclusion. Pretraining domain proximity to the target task is a stronger determinant of representation quality than model scale or architectural capacity. BiomedCLIP's broad biomedical pretraining enables it to produce embeddings that are inherently well-suited to morphological discrimination, while general and pathology-domain models require more labelled data to achieve comparable performance. These findings have practical implications for the deployment of vision-language models in clinical hematology: when labelled data is scarce, the choice of a pretrained encoder matters far more than that of a classifier or prompt strategy.

5. Conclusion

This paper presented a systematic benchmark of CLIP-based vision-language encoders for white blood cell classification on the WBCAtt dataset. We evaluated six models spanning three pretraining domains under both zero-shot and few-shot linear probing settings, and analysed the effect of prompt template design on zero-shot performance. The results lead to several clear conclusions. First, zero-shot CLIP inference is insufficient for reliable white blood cell classification. All 30 model-prompt combinations fail to exceed a macro-F1 of 0.26, regardless of pretraining domain or prompt specificity. This demonstrates that current vision-language models do not acquire the fine-grained morphological sensitivity required for

hematological image analysis through pretraining alone. Second, few-shot linear probing dramatically recovers performance, confirming that frozen CLIP encoders do encode discriminative visual structure that can be exploited with minimal labelled supervision. Third, and most importantly, pretraining domain proximity to the target task is the dominant factor determining representation quality. BiomedCLIP, pretrained on the large-scale and diverse PMC-15M biomedical corpus, consistently outperforms all other models across all sample sizes, achieving a macro-F1 of 87.5% with only 100 labelled samples per class. General and pathology-domain models, despite their architectural capacity, require substantially more labelled data to reach comparable performance. These findings have practical implications for deploying vision-language models in clinical hematology. When labelled data is scarce, selecting a pretrained encoder is the most consequential design decision, outweighing the choice of classifier architecture or prompt strategy. BiomedCLIP emerges as the most suitable foundation model for white blood cell analysis under low-resource conditions. Future work should investigate full fine-tuning and domain-adaptive pretraining strategies to further close the gap between vision-language representations and the demands of clinical hematological image analysis.

References

[1] G. H. Dagnaw, Y. Zhu, M. H. Maqsood, X. Yin, and A. W.-C. Liew, "Explainable multimodal hematology analysis for white blood cell classification and attribute prediction," *Comput. Biol. Med.*, vol. 196, p. 110734, 2025.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, Virtual, 2021, vol. 139, pp. 8748-8763.

[3] I. Kukuljan, M. F. Dasdelen, J. Schäfer, M. Buck, K. S. Götze, and C. Marr, "Domain specific models outperform large vision language models on cytomorphology tasks," *medRxiv*, 2025.

[4] T. Patel, H. El-Sayed, and M. K. Sarker, "Evaluating vision-language models for hematology image classification: Performance analysis of CLIP and its biomedical AI variants," in *Proceedings of the 36th Conference of Open Innovations Association (FRUCT)*, Lappeenranta, Finland, 2024, pp. 578-584.

[5] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *Int. J. Comput. Vis.*, vol. 130, no. 9, pp. 2337-2348, 2022.

[6] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023, pp. 2818-2829.

[7] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, M. P. Lungren, T. Naumann, S. Wang, and H. Poon, "BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," arXiv preprint arXiv:2303.00915, 2023.

[8] S. Eslami, C. Meinel, and G. de Melo, "PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain?" in *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia, 2023, pp. 1181-1193.

[9] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nat. Med.*, vol. 29, pp. 2307-2316, 2023.

[10] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, and L. Shapiro, "Quilt-1m: One million image-text pairs for histopathology," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS)*, New Orleans, LA, 2023, vol. 36, pp. 37995-38017.

[11] S. Tsutsui, W. Pang, and B. Wen, "WBCAtt: A white blood cell dataset annotated with detailed morphological attributes," in *Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), Track on Datasets and Benchmarks*, New Orleans, LA, 2023.

[12] A. Acevedo, A. Merino, S. Alférez, Á. Molina, L. Boldú, and J. Rodellar, "A dataset of microscopic peripheral blood cell images for development of automatic recognition systems," *Data Brief*, vol. 30, p. 105474, 2020.