Image and Vision Computing 172 (2026) 106019

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

WBC-CLIP: A multimodal vision-language framework for morphology aware white blood cell analysis

Luca Zedda, Davide Antonio Mura *, Andrea Manzo, Cecilia Di Ruberto, Andrea Loddo

Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124, Cagliari, Italy

ARTICLE INFO

Keywords:
Deep learning
White blood cell analysis
Contrastive learning
LLM
Zero-shot classification
Image-text retrieval

ABSTRACT

Can the integration of vision and language representations advance artificial intelligence methods for automated white blood cell (WBC) analysis across heterogeneous clinical conditions? Motivated by this question, we present WBC-CLIP, a dual-encoder framework that enhances WBC classification and analysis by combining image data with rich textual descriptions derived from quantitative morphological features. Our method leverages multiple large language models to convert numerical and categorical cell attributes into diverse, semantically enriched textual descriptions. These captions are jointly embedded with their corresponding WBC images using a contrastive learning strategy inspired by the CLIP architecture, enabling the model to learn stable and meaningful cross-modal associations. We evaluate WBC-CLIP through zero-shot classification and image–text retrieval tasks across both in-distribution and out-of-distribution datasets. The framework advances automated WBC analysis while providing improved explainability by explicitly grounding visual representations in morphology-aware textual descriptors, addressing key challenges in computer-aided diagnostics.

1. Introduction

White blood cell analysis plays a crucial role in clinical diagnostics, contributing to the early detection and management of various hematological and immunological disorders [1,2].

The advent of computer-aided diagnostics has led to the development of automated frameworks designed to reduce human error and improve diagnostic efficiency [3–5]. However, despite these advancements, significant challenges persist due to variability in sample preparation, staining protocols, and the inherent morphological heterogeneity of WBCs [6–10]. A particularly persistent issue is the limited generalizability of existing models across datasets originating from different imaging sources [11,12]. These challenges underscore the need for advanced computer vision and deep learning strategies within modern artificial intelligence systems for hematology.

Recently, cross-modal learning approaches, particularly those leveraging Contrastive Language-Image Pre-training (CLIP) [13], have demonstrated remarkable success in aligning visual and textual representations, leading to improved model generalization and interpretability. Inspired by these advancements, we propose WBC-CLIP, a novel dual-encoder framework that integrates image data with semantically enriched textual descriptions derived from quantitative morphological features [14]. Our approach utilizes multiple large language models (LLMs) to transform numerical and categorical cell attributes into diverse textual captions. These textual descriptions are then embedded alongside their corresponding WBC images using a contrastive learning strategy, enabling the model to establish robust cross-modal associations and enhance interpretability. Our main contributions are as follows:

- We introduce a novel methodology for generating rich textual descriptions of WBC morphological features using multiple LLMs.
- We design a dual-encoder architecture that aligns visual and textual modalities through contrastive learning, improving model robustness and interpretability. Fig. 1 presents a compact representation of the proposed architecture.
- We demonstrate the effectiveness and generalizability of our approach through extensive evaluations of both in-distribution and out-of-distribution datasets.

Furthermore, our framework paves the way for future research in zero-shot learning applications for WBC analysis, facilitating adaptable and scalable AI-driven diagnostic tools.

The remainder of this paper is organized as follows. In Section 2, we review the most relevant literature on automated WBC analysis, contrastive and multimodal learning, and the use of large language models in biomedical imaging. In Section 3, we describe the datasets, the LLM-based textual generation process, and the evaluation metrics adopted for the downstream tasks. Next, Section 4 outlines the

* Corresponding author.
E-mail address: davideantonio.mura@unica.it (D.A. Mura).

https://doi.org/10.1016/j.imavis.2026.106019
Received 25 November 2025; Received in revised form 22 April 2026; Accepted 4 May 2026
Available online 5 May 2026

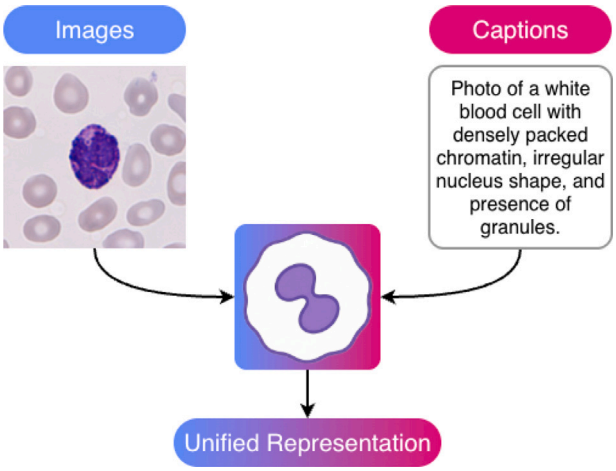

Fig. 1. WBC-CLIP provides a semantically rich unified representation of white blood cell visual and textual data. It consists of an image encoder shown in blue on the left and a text encoder shown in magenta on the right. The two encoders are trained with a contrastive objective to align their embedding spaces. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

proposed WBC-CLIP framework in detail, including the architectural design, contrastive learning formulation, and training pipeline. In Section 5, we present a comprehensive set of experiments assessing performance in zero-shot classification, retrieval tasks, and cross-dataset generalization. The results are analyzed and discussed in Section 6, highlighting both strengths and limitations of the proposed approach. Finally, Section 7 summarizes the main contributions of this work and outlines future research directions.

2. Related work

This section reviews prior research related to automated WBC analysis, cross-modal learning, and the integration of large language models in medical imaging. We categorize existing literature into four main areas: (i) traditional and deep learning-based WBC analysis, (ii) multimodal and contrastive learning frameworks, (iii) vision-language models in biomedical domains, and (iv) the use of LLMs for medical feature interpretation.

2.1. Traditional and deep learning-based WBC analysis

Early studies on WBC analysis predominantly relied on handcrafted features derived from cell morphology, texture, and color characteristics. Classical approaches employed segmentation followed by feature extraction pipelines for classification tasks [15]. While effective in controlled laboratory settings, these methods struggled with generalization due to variations in staining, illumination, and acquisition conditions [9].

The introduction of deep learning, particularly convolutional neural networks, marked a paradigm shift in automated hematology. Architectures such as ResNet, DenseNet, and EfficientNet have demonstrated superior performance in recognizing WBC subtypes [16] and detecting abnormalities [17]. However, despite improved accuracy, deep learning models remain susceptible to domain shifts across datasets, leading to a loss of robustness and interpretability [18].

2.2. Contrastive and cross-modal learning in biomedical imaging

Contrastive learning has recently emerged as a powerful approach for representation learning in both natural and biomedical images [19, 20]. By maximizing similarity between semantically aligned pairs and minimizing it for non-aligned pairs, contrastive methods learn robust latent spaces without extensive labeled data [21]. Several works have applied self-supervised contrastive paradigms, such as SimCLR and MoCo, to histopathology, radiology, and cytology [22–24]. More recent studies have extended these ideas to multimodal setups, integrating clinical metadata or textual annotations to enhance the generality of representation [25]. These approaches highlight the potential of cross-modal alignment to capture clinically meaningful relationships between modalities.

2.3. Vision-language models in the biomedical domain

Following the success of CLIP [13], vision-language models have been explored for medical imaging applications, including radiology report generation, image-text retrieval, and zero-shot disease classification [26]. Biomedical adaptations such as BioCLIP and MedCLIP [27, 28] have demonstrated that medical visual-textual alignment can improve both interpretability and performance under distributional shifts. However, most existing works rely on unstructured natural language reports rather than structured or quantitative morphological features, limiting their precision in capturing subtle cellular variations.

In the specific context of hematology and WBC analysis, the literature is still relatively limited. Recently, Dagnaw et al. similarly used CLIP as the base model for WBC classification by leveraging cell attributes [29]. They systematically created predefined text templates to represent both classes and cellular attributes, making their work, to the best of our knowledge, the closest published multimodal WBC study to our framework. While both studies adopt a CLIP-style dual-encoder formulation, the main difference lies in the textual supervision strategy. Dagnaw et al. rely on structured predefined prompts to encode class and attribute information, whereas our approach uses LLM-generated captions derived from the morphological attributes of each cell. In this way, all attribute values are integrated into a single holistic textual description, rather than being expressed through fixed template formulations.

More broadly, recent hematology-oriented multimodal works have started to explore related settings, including morphology-aware caption generation for leukemia cells, multimodal visual question answering on blood smear or peripheral blood cell images, and broader digital hematopathology reasoning frameworks [30–33]. Although these efforts are not always directly aligned with our experimental setting, they suggest an emerging interest in combining cell morphology with textual information for hematology analysis.

Building on this context, our work differs from prior hematology-oriented multimodal studies in two main respects. First, our approach employs several LLMs, each producing a distinct narrative of the associated image. Second, all attribute values of a given image are encapsulated into a single textual representation, generating a faithful description that is suitable for text embedding and capable of globally describing the image rather than focusing only on isolated attributes. We also evaluate our methodology in an information retrieval setting to more thoroughly assess the shared latent space created by the model during alignment.

2.4. Large language models for morphological feature description

Large language models have recently shown impressive capabilities in understanding and generating medical text [34–36]. Their use in transforming structured diagnostic data into rich textual representations offers new avenues for interpretable learning [37]. Some studies have proposed leveraging LLMs to describe radiological findings or histopathological features [38], yet few have systematically explored their potential in hematology. In particular, the small number of recent hematology-specific multimodal studies tends to focus on related

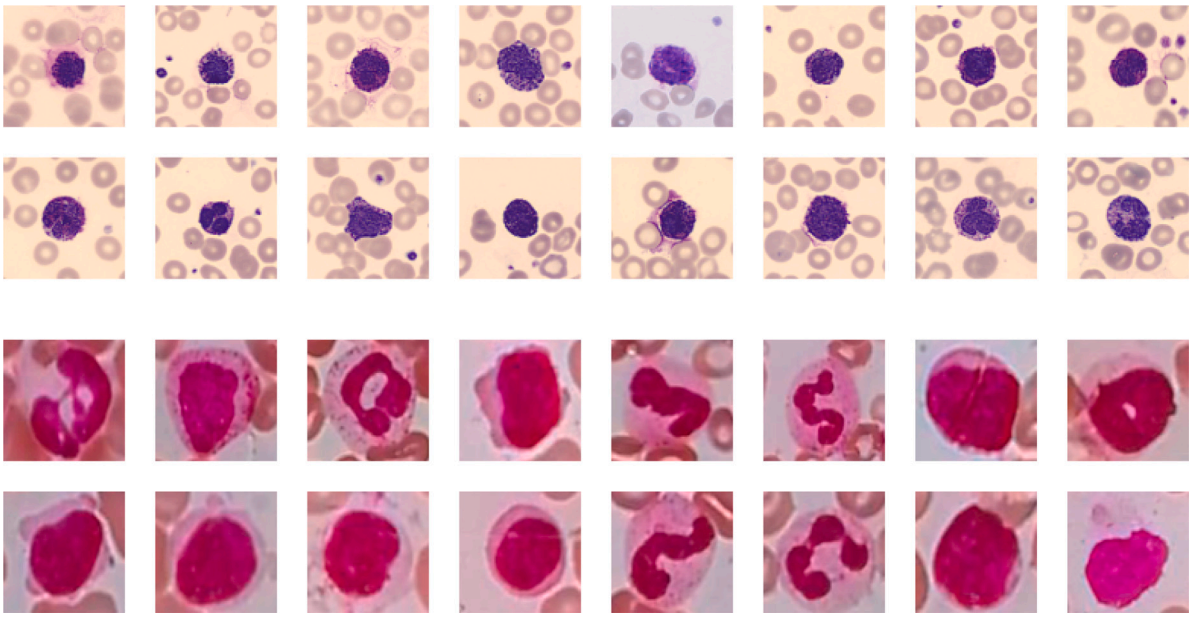

Fig. 2. Visual representation of the heterogeneity of the used datasets, showcasing clear differences in both color and size of the used samples. The first and second rows display samples from the WBCAtt dataset, whereas the third and fourth rows correspond to WBCs from the LeukemiaAttri dataset. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

objectives, such as morphology-aware caption generation, visual question answering, or broader multimodal reasoning, rather than direct contrastive alignment between single-cell images and morphology-aware textual descriptions [30–33]. In contrast, our work integrates LLM-generated morphological descriptions directly within a contrastive learning framework, enabling the creation of semantically aligned visual-textual embeddings for WBCs.

2.5. Research gap

Existing literature demonstrates strong progress in deep WBC analysis and cross-modal learning. At the same time, recent studies have begun to explore multimodal learning for WBC and hematology analysis [29–33], showing that visual and textual information can be combined in clinically meaningful ways. Nevertheless, this literature remains limited and heterogeneous. Many available approaches either rely solely on visual data, employ unstructured textual information, use predefined template-based prompts, or address related tasks such as caption generation or visual question answering rather than direct contrastive alignment between single-cell images and morphology-aware textual descriptions. Our proposed WBC-CLIP framework addresses this gap by employing multiple LLMs to generate diverse textual captions of WBC morphological features and aligning them with visual representations through contrastive pretraining. In this sense, our contribution is not to claim that multimodal hematology analysis is entirely absent from the literature, but rather to position WBC-CLIP within a still limited body of work that specifically targets morphology-aware image-text representation learning at the single-cell level.

3. Materials and methods

In this section, we describe the datasets and computational tools used in our experiments with WBC-CLIP. We first introduce the two white blood cell image collections employed, followed by an overview of the LLMs used to generate morphological descriptions. We then discuss the backbone architectures and the contrastive learning strategy at the core of our approach. Finally, we summarize the metrics used to evaluate classification performance.

3.1. Datasets

The datasets employed comprise two types of information: images representing white blood cells and corresponding tables detailing their morphological features.

The **WBCAtt** [39] dataset is an extensive repository of WBC images, each meticulously annotated with a broad array of morphological attributes. It comprises 10,298 high-resolution images accompanied by 113,278 attribute labels, spanning 11 critical morphological features, including cell size, cell shape, nucleus shape, nucleus-to-cytoplasm ratio, chromatin density, presence of cytoplasm vacuoles, cytoplasm texture, cytoplasm color, granule type, granule color, and granularity. In this study, the WBCAtt dataset serves as the primary training resource for the WBC-CLIP framework, facilitating the joint learning of robust visual and semantic representations from both the raw images and the enriched textual descriptions derived from these morphological features.

The **LeukemiaAttri** [40] dataset has been curated to reflect the inherent variability present in real-world WBC imaging conditions. This dataset contains 2400 images acquired using both high-cost and low-cost microscopes at multiple magnifications, and it features approximately 55,000 morphological labels annotated across 14 distinct WBC classes. The diverse imaging modalities and acquisition protocols inherent to LeukemiaAttri render it an ideal out-of-distribution evaluation set. In our experiments, we utilize only the images captured with high-cost, high-magnification equipment to rigorously assess the cross-dataset generalizability and robustness of the WBC-CLIP framework.

Fig. 2 presents sample images from the WBCAtt and LeukemiaAttri datasets, respectively. Clear differences can be observed between the two, including variations in color acquisition and white blood cell magnification.

3.2. LLM-based description

In our pipeline we used LLMs to generate description from morfological information. We used different type of model to provide different descriptions. We selected a set of predominantly instruction-tuned LLMs to balance three objectives: reliable prompt following,

diversity of linguistic realization, and manageable computational cost. Although several models fall within a comparable parameter range, the selection was not based on size alone. Rather, we intentionally included models with different pretraining backgrounds and adaptation styles, including general-purpose instruction-tuned LLMs and models with specialized refinements, in order to obtain multiple complementary textual descriptions of the same morphological input. Our objective was not to rank LLMs as standalone generators, but to use them as diverse semantic verbalizers that enrich the textual supervision available to the contrastive framework. Specifically we selected instruction-tuned models like Calme-3.1-Instruct,[1] Chocolate-3B-Instruct,[2] Granite-3-8B-Instruct,[3] and Qwen-2.5-Instruct,[4] Medit-mesh-3B-Instruct,[5] all designed to follow prompts effectively and generate contextually aligned responses. The GPT-3.5-turbo[6] model serves as a strong general-purpose baseline known for its balanced reasoning and generation capabilities. Models such as EPE-PRYMAL-YL-3B-SLERP-V3,[7] and Pancho-v1-qw25-3[8] represent specialized or hybrid architectures, integrating domain or fine-tuning adaptations. Finally, phi-3.5-mini[9] and phi-2[10] are lightweight transformer models optimized for efficiency and general text understanding. Before prompting the LLMs, the morphological annotations associated with each cell are converted into a standardized textual attribute list. In particular, categorical attributes are mapped to their human-readable labels, while binary attributes are represented through their explicit semantic state. The resulting attribute sequence is then inserted into a shared prompt template and provided to each LLM. In this way, all models receive the same normalized semantic input, while the linguistic realization of the description is allowed to vary across LLMs. This design ensures consistency at the attribute level while preserving textual diversity, which is beneficial for contrastive multimodal training. Although specialized biomedical foundation models are highly relevant to this domain, their role depends on the stage of the pipeline. In WBC-CLIP, the selected LLMs are used only to verbalize structured morphological annotations into short captions. For this reason, we prioritized instruction-tuned generators that reliably follow a fixed prompt, produce diverse linguistic realizations, and remain practical for large-scale caption generation. By contrast, models such as BioMedCLIP [41] are biomedical vision-language encoders rather than generative LLMs, and are therefore more appropriately treated as external multimodal baselines than as caption generators. To address this point, we additionally evaluated representative specialized biomedical foundation models, namely BioMedCLIP [41] and MedGemma [42], in the experimental section.

LLM prompt. Here are described the prompts used to instruct the various LLMs to generate cell descriptions based on the morphological features associated with each white blood cell. We employ two different prompts: a system prompt and a user prompt.

The system prompt defines the behavior of the LLM, in this case, guiding it to act as a hematology expert. The user prompt provides the specific input to which the LLM should respond.

The prompts are reported below.

System Prompt

```
{"role": "system",
  "content": "You are a hematology expert specialized in white blood cell morphology.
Your task is to generate a short descriptive sentence describing how this cell might look under a microscope."}
```

User Prompt

```
{"role": "user",
    "content": "Below are the quantitative morphological features extracted from a microscopic image of a white blood cell.
Use this information to generate a short, coherent textual description of the cell's appearance including its size, nucleus shape, cytoplasm color/texture, and any distinctive morphological traits.
Avoid listing the features numerically; instead, integrate their meaning naturally into the description. Morphological features: f{list_of_morphological_features}"}
```

In our pipeline, LLMs are not asked to infer diagnoses or free-form findings directly from raw images. Instead, they verbalize pre-existing structured morphological attributes associated with each white blood cell. To assess how faithfully these attributes are preserved in the generated text, we compared the information expressed in the captions against the original attribute annotations across the LLMs used in this study. Exact agreement across all attributes was limited (full-match rate: 0.0875 ± 0.0873), but attribute-level agreement was substantially higher for several key descriptors, including cell shape (0.9052 ± 0.0228), nucleus shape (0.8113 ± 0.0582), nucleus-to-cytoplasm ratio (0.8948 ± 0.0404), cytoplasm texture (0.9579 ± 0.0211), and cytoplasm color (0.9392 ± 0.0450). At the same time, lower agreement for features such as cytoplasm vacuoles (0.5644 ± 0.0761) and granularity (0.2762 ± 0.1315) indicates that these descriptions should be regarded as auxiliary morphology-conditioned textual views rather than perfectly faithful clinical reports. This variability suggests that the generated descriptions introduce a mixture of faithful and imperfect textual supervision, which is useful for semantic enrichment but should also be considered a source of noise during multimodal training.

3.3. Backbone architectures

WBC-CLIP utilizes deep neural networks as feature extractors to encode white blood cell images into meaningful representations. Specifically, we experiment with both convolutional neural networks (ResNets) and transformer-based architectures (Vision Transformers, ViT).

ResNets rely on hierarchical convolutional layers with skip connections to capture local and global features efficiently, while ViT use self-attention mechanisms to model long-range dependencies and complex morphological patterns. By combining these backbones with contrastive learning, WBC-CLIP can generate robust embeddings that capture diverse structural variations of white blood cells.

3.4. Contrastive learning

WBC-CLIP adopts a contrastive image-text learning strategy inspired by CLIP, with the goal of aligning white blood cell images and morphology-aware textual descriptions within a shared embedding

[1] https://huggingface.co/MaziyarPanahi/calme-3.1-instruct-3b.
[2] https://huggingface.co/jpacifico/Chocolatine-3B-Instruct-DPO-Revised.
[3] https://huggingface.co/ibm-granite/granite-3.0-8b-instruct.
[4] https://huggingface.co/Qwen/Qwen2.5-3B-Instruct.
[5] https://huggingface.co/meditsolutions/MedIT-Mesh-3B-Instruct.
[6] https://huggingface.co/Xenova/gpt-3.5-turbo.
[7] https://huggingface.co/LilRg/10PRYMMAL-3B-slerp.
[8] https://huggingface.co/fblgit/pancho-v1-qw25-3B-UNAMGS.
[9] https://huggingface.co/microsoft/Phi-3.5-mini-instruct.
[10] https://huggingface.co/microsoft/phi-2.

space. Unlike standard self-supervised contrastive learning based only on image augmentations, our framework operates on paired multimodal samples, where each image is associated with a textual description generated from structured morphological attributes.

Given a mini-batch of paired samples, the image encoder extracts a visual representation from each white blood cell image, while the text encoder processes the corresponding tokenized description. These modality-specific features are then mapped into a common latent space through dedicated projection heads, producing image and text embeddings with the same dimensionality. During batch construction, samples are selected so as to avoid repeated image identities within the same mini-batch. This prevents trivial duplication effects and allows the multiple captions associated with a given image, encountered across training, to act as alternative textual views, i.e., textual augmentations, of the same morphology-conditioned sample.

Let $\mathbf{z}_i^{(I)}$ and $\mathbf{z}_i^{(T)}$ denote the projected image and text embeddings of the ith sample in a batch of size B. A cross-modal similarity matrix is computed through the dot product between all text and image embeddings:

$$\mathbf{L}_{ij} = \frac{\mathbf{z}_i^{(T)} \cdot \mathbf{z}_j^{(I)}}{\tau}, \tag{1}$$

where τ is a temperature parameter controlling the concentration of the similarity distribution. This matrix defines the image-text matching logits used during training.

In standard CLIP, supervision is typically defined through hard one-to-one correspondences, where only the matched image-text pair is treated as positive and all remaining pairs in the batch are treated as equally negative. However, this assumption is overly restrictive for our scenario. First, different white blood cells may share highly similar morphology, so semantically related samples can appear within the same mini-batch. Second, the textual modality is generated by multiple LLMs from the same structured attributes, meaning that captions behave as alternative semantic views and may differ in phrasing, granularity, or faithfulness. As a result, non-diagonal pairs are not always true negatives, and enforcing a purely hard target may over-penalize morphologically related image-text pairs.

For this reason, WBC-CLIP computes intra-modal similarity matrices for images and texts separately and uses them to construct soft targets that reflect the relational structure of the batch. Formally, let $\mathbf{S}^{(I)}$ and $\mathbf{S}^{(T)}$ denote the image-image and text-text similarity matrices, respectively. The target distribution is then defined as:

$$\mathbf{Y} = \mathrm{softmax}\left(\frac{\mathbf{S}^{(I)} + \mathbf{S}^{(T)}}{2} \cdot \tau\right). \tag{2}$$

In all experiments, the temperature was set to $\tau = 1$. We adopted this neutral setting because our training relies on soft targets derived from intra-modal similarities and on LLM-generated captions, whose partial variability and noise make overly sharp contrastive distributions less desirable. Accordingly, $\tau = 1$ provides a stable compromise between alignment strength and robustness to imperfect textual supervision. The training objective is computed symmetrically in both directions, from text to image and from image to text. More precisely, the model minimizes the average of the cross-entropy loss between the image-text logits and the soft targets, and the corresponding transposed formulation:

$$\mathcal{L}_{\text{text}} = \mathrm{CE}(\mathbf{L}, \mathbf{Y}), \tag{3}$$

$$\mathcal{L}_{\text{image}} = \mathrm{CE}(\mathbf{L}^\top, \mathbf{Y}^\top), \tag{4}$$

$$\mathcal{L} = \frac{\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{image}}}{2}. \tag{5}$$

This formulation is particularly suitable for WBC-CLIP because it preserves the main alignment signal of matched image-text pairs while softening the treatment of semantically neighboring samples. In practice, it encourages the model to align each cell image with its corresponding morphology-aware description, but without forcing all other cells in the batch to be equally dissimilar. This is beneficial in our setting, where morphology-aware captions are informative yet imperfect, and where cells with overlapping attributes should remain relatively close in the shared embedding space.

Consequently, the learned contrastive space captures not only pairwise correspondence, but also broader neighborhood relationships across both modalities. This improves robustness to linguistic variability in LLM-generated descriptions, reduces the impact of partial caption noise, and yields embeddings that are better suited to downstream zero-shot classification and retrieval. By aligning images with semantically enriched textual descriptions derived from morphological annotations, WBC-CLIP learns representations that are more robust to visual variability and more appropriate for cross-dataset generalization.

3.4.1. Classification measures

The classification performance is evaluated using several measures, including *accuracy*, *recall*, *precision*, and the *F1-score*.

In the following subsections, we provide straightforward definitions of these metrics as they pertain to binary classification problems, followed by their generalizations for multiclass scenarios.

Standard definitions for binary classification problems. An example, denoted as e, is characterized by a pair $\langle i, t\rangle$, where i represents a list of feature values and t denotes the assigned category (i.e., the target category). A dataset D is defined as a collection of such examples. When the dataset D contains two target categories, it constitutes a binary classification problem. In this context, the categories are referred to as *negative* and *positive*.

To assess the performance of a binary classifier on the dataset D, each instance is labeled as either *negative* or *positive* based on the classifier's output. Depending on the classification outcome and the actual target value, an instance will contribute to one of the following counts:

- *True Negatives (TN)*: The number of instances belonging to the *negative* class that have been accurately predicted;
- *False Positive (FP)*: The number of instances belonging to the *negative* class that have been incorrectly predicted as positive;
- *False Negative (FN)*: The number of instances belonging to the *positive* class that have been incorrectly predicted as negative;
- *True Positive (TP)*: The number of instances belonging to the *positive* class that have been accurately predicted.

Based on these quantities, the measures to evaluate the classification performance can be defined as follows:

- *Accuracy* (ACC) The ratio of correctly classified instances to the total number of instances:
$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$
- *Precision (PRE)* The fraction of positive instances that are correctly classified among all instances classified as positive:
$$PRE = \frac{TP}{TP + FP} \tag{7}$$
- *Recall* (REC) It measures the classifier's ability to correctly identify the positive class, calculated against FN:
$$REC = \frac{TP}{TP + FN} \tag{8}$$
- *F1-score* (F1) Defined as the harmonic mean of *precision* and *recall*:
$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{9}$$

Standard definitions for multiclass classification problems. As previously mentioned, the measures outlined can also be generalized for multiclass classification scenarios. A straightforward approach to achieve this is to calculate the metrics for each category using a One-vs-Rest (OvR) strategy. Following this process, the average value of each binary measure is computed, yielding an informative metric for the multiclass model.

Three distinct averaging methods can be employed: micro, macro, and weighted. In this study, macro averaging has been adopted.

In summary, for a classification problem involving K classes, the metrics with macro averaging are calculated as follows:

- *Macro Average Precision* (where P_k denotes the precision for class k):

$$MacroPrecision = \frac{\sum_{k=1}^{K} P_k}{K} \tag{10}$$

- *Macro Average Recall* (where SEN_k denotes the sensitivity for class k):

$$MacroRecall = \frac{\sum_{k=1}^{K} SEN_k}{K} \tag{11}$$

- *Macro Average F1-score* (where P and R denote the macro average precision and recall, respectively):

$$MacroF1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{12}$$

3.4.2. Retrieval measures

The retrieval performance is evaluated by analyzing the relevance of the cases returned by the system for each query. In medical image retrieval, the practical value of a system depends on its ability to provide at least one clinically relevant case within the first retrieved samples and on how early such a case appears in the ranked list. Since only the top–k retrieved items are considered, we adopt a set of retrieval measures that operate on truncated rankings.

Let the retrieval output for a query be represented by a relevance vector $r = \langle r_1, r_2, \ldots, r_k \rangle$, where $r_i = 1$ if the ith retrieved case is clinically relevant and 0 otherwise. The following measures are used in this study:

Hit@k. Indicates whether at least one relevant case is present among the top–k retrieved samples and is defined as:

$$Hit@k = \begin{cases} 1 & \text{if } \exists\, i \le k \text{ such that } r_i = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

Precision@k. This measure quantifies the proportion of relevant cases among the first k retrieved items:

$$Precision@k = \frac{1}{k} \sum_{i=1}^{k} r_i. \tag{14}$$

Mean Reciprocal Rank (MRR@3). The Mean Reciprocal Rank reflects the rank of the first relevant case in the retrieved list. When limited to the top three retrieved samples, it is computed as:

$$MRR@3 = \begin{cases} \frac{1}{i} & \text{if } r_i = 1 \text{ for the smallest } i \le 3, \\ 0 & \text{if no relevant item is found within the top 3.} \end{cases} \tag{15}$$

4. Proposed approach

Our proposed methodology is founded on a dual-encoder architecture that draws inspiration from the CLIP framework. CLIP uses two different types of encoders, one designed to process images and the other to process text. It generates embeddings for both modalities and, through a contrastive learning approach, successfully creates a shared embedding space that represents both image and text. This shared space enables improved performance on downstream tasks such as classification. Based on this, we propose an innovative approach that integrates both visual and morphological information, processed as text, to perform a comprehensive analysis of WBCs.

Each WBC image is systematically accompanied by a set of morphological features such as cytoplasm ratio, nuclear size, granularity, etc... Since these features are typically stored in a tabular format, it is necessary to convert this information into a form suitable for the text encoder. Therefore, this tabular data is used as foundation for generating detailed textual descriptions that encapsulate the subtle and complex characteristics of each cell. To convert these raw morphological measurements into expressive descriptive text, we employ multiple large language models operating in parallel. For a fixed set of morphological features, all LLMs process them in parallel, each producing its own textual description of the WBC image. This process yields a collection of diverse captions, that reflect the unique characteristics of each LLM, each producing its own distinctive description. The ensemble of these descriptions enriches the semantic embedding space used during training.

For the experiments, we selected two different types of image encoder architectures: one based on convolutional neural networks and the other on vision transformers, which is built upon the transformer architecture. For the text encoder, we used only models based on the transformer architecture. Both encoders are trained jointly via a contrastive loss function, which aligns image and text embeddings within a shared latent space, and allows to maximize the similarity between correctly paired image-text embeddings and minimize the similarity with non-matching pairs, thereby fostering robust cross-modal associations and enhancing the model's discriminative capability.

In Fig. 3 we present an overview of the training pipeline. The first part (above) illustrates the process used to generate descriptions of the morphological features. These features are concatenated with a textual prompt and fed into "n" numbers of LLMs, each producing its own textual description. The resulting texts are then passed to the text encoder, which generates corresponding text embeddings. Simultaneously, the image associated with its morphological features is processed by the image encoder, producing an image embedding. This embedding is paired with each description produced by the LLMs. Finally, a contrastive loss is applied to both embeddings in order to maximize the similarity between matching image–text pairs and minimize the dissimilarity between non-matching pairs.

The proposed training pipeline yields a pretrained model capable of processing multimodal data for downstream tasks. The model was evaluated on two selected downstream tasks: the first assessed its ability to generalize to unseen samples, while the second examined the fidelity of the learned embeddings.

An overview of these evaluation tasks is presented in Figs. 4 and 5. On the left, we illustrate the image–text retrieval task, which consists of retrieving the top-k images that best match a given text prompt. The prompt may describe one or more characteristics or classes of WBCs. Conversely, the same process can be applied in the opposite direction, given an image, the model retrieves the top-k textual descriptions that correspond to its characteristics. On the right, we present the zero-shot classification task, which involves classifying WBC images solely based on text prompts, without any fine-tuning. In this task, both the image and the classes prompts are processed by their respective encoders, and classification is performed using cosine similarity between the resulting embeddings.

5. Experiments and results

5.1. WBC-CLIP development

Setup. The experimental evaluation has been carried out on a high-performance workstation equipped with two NVIDIA RTX 4090 GPUs, ensuring efficient training and evaluation of our proposed framework. Additionally, we set the learning rate to $5e{-}4$ and configured the

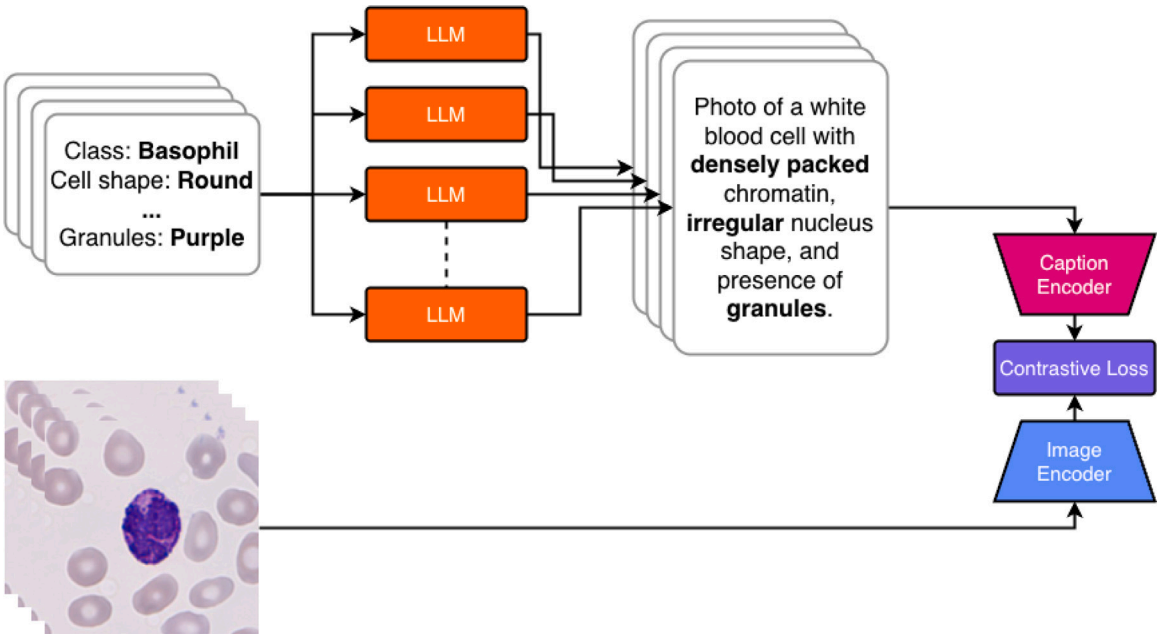

Fig. 3. Overview of the training pipeline. Morphological features are parsed and converted into multiple textual descriptions using different LLMs, which are then jointly embedded with the corresponding WBC images via dual encoders trained with contrastive learning.

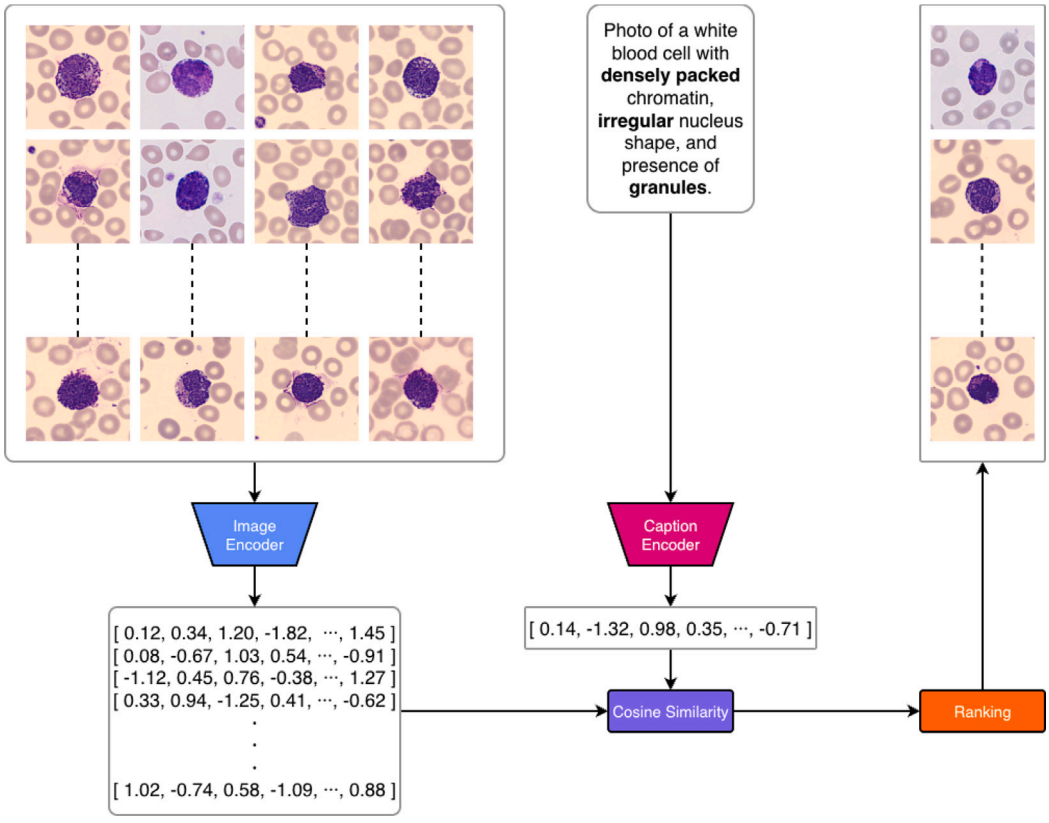

Fig. 4. Retrieval approach using WBC-CLIP. Images are embedded using the image encoder, and the retrieval prompt is embedded by the text encoder. Then, cosine similarity is applied between each possible pair of image-text, and finally, a value-based ranking is applied.

training parameters as follows: 20 epochs and batch sizes per GPU of 128 for ResNet50, 512 for ViT tiny, 256 for ViT small, 128 for ViT base, and 64 for ViT large. Initially, we trained the architecture on the WBCAtt dataset, utilizing both the original high-resolution images and the accompanying morphological attributes parsed by multiple LLMs, as detailed in Section 4. This integrated approach enabled the model to learn robust visual and semantic representations from complementary sources of information.

Experiment #1: text encoder investigation. Following training, we performed a first evaluation of the model's class and attribute predic-

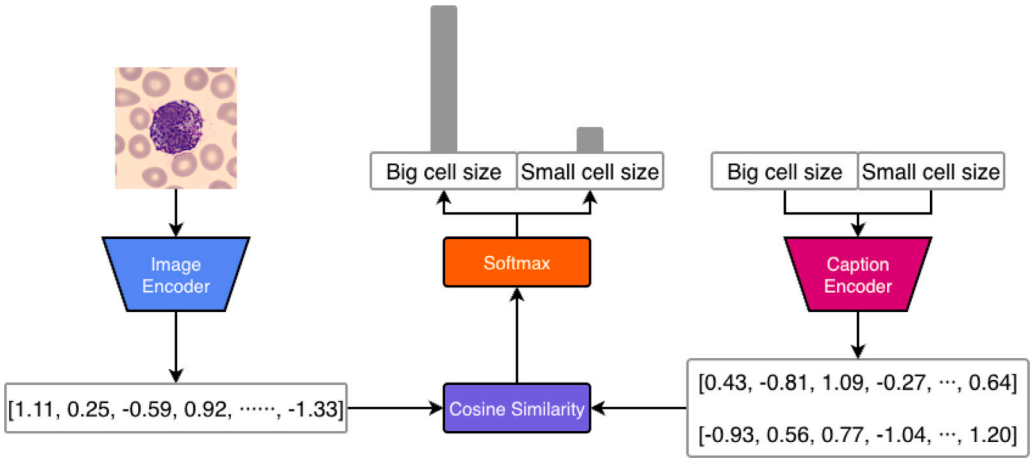

Fig. 5. Zero-shot approach using WBC-CLIP. Images are embedded using the image encoder, and the class-related descriptions are embedded via the text encoder. After computing the cosine similarity between pairs, softmax is then used to calculate class assignment probabilities.

Table 1
Bio Reddit BERT emerges as the best text encoder by a large margin for the WBC-CLIP text encoder selection. The models share the same training configuration and number of epochs.

Text encoder	Image encoder	F1
Bio DistilBERT	ViT Small	24.30
Bio DistilBERT	ViT Base	26.89
ALBERT Base v2	ViT Small	30.16
ALBERT Base v2	ViT Base	21.71
Bio Reddit BERT	ViT Small	60.34
Bio Reddit BERT	ViT Base	**65.99**

tion capabilities in a zero-shot setting, as illustrated in Fig. 5. For this purpose, a subset of the WBCAtt dataset was employed. This evaluation not only assessed the model's predictive performance but also served as a comparative investigation to determine the most effective text encoder. For the purpose of this investigation, we used ViT architecture as image encoder, exploring both the small and base configurations with a patch size of 16, thereby providing insights into potential scaling laws. Table 1 summarizes the performance outcomes of the text encoder comparative investigation, revealing that the Bio Reddit BERT text encoder outperforms alternative text encoders by over 30% in terms of F1-score (F1).

Table 1 summarizes the performance outcomes of this internal text encoder investigation, revealing that Bio Reddit BERT outperforms the alternative text encoders considered within the WBC-CLIP architecture by over 30% in terms of F1-score. We therefore adopt it in the subsequent experiments. Comparisons against external specialized biomedical foundation models are reported separately, as they do not represent interchangeable text-encoder components of the proposed framework.

Experiment #2: image encoder investigation. Afterwards, we set Bio Reddit BERT as optimal text encoder and systematically vary the image encoders to establish a more diverse model family. This variation is motivated by two primary objectives. First, it enables a comprehensive analysis of the relationship between image encoder model size and performance in real-world evaluation scenarios. Second, it ensures greater flexibility for future applications, accommodating institutions with different computational resources. Given that not all hospitals and clinics have access to high-end hardware, our approach provides an equitable and adaptable solution by offering both lightweight and computationally intensive models.

A summary of the results is presented in Table 2. The findings reveal a pronounced disparity in specialized performance across different image encoder architectures. Specifically, ResNet50 demonstrates superior proficiency in recognizing cytoplasm-related attributes, whereas ViT-based encoders excel at capturing granularity-related properties of WBCs. This performance divergence underscores the complementary strengths of convolutional and transformer-based approaches in WBC morphology analysis.

The comparison on WBCAtt shows that specialized biomedical foundation models alone do not match the proposed morphology aware multimodal training strategy. In particular, BioMedCLIP and MedGemma achieve average macro F1 scores of 24.11% and 28.92%, respectively, remaining far below all WBC-CLIP variants. Although these baselines benefit from biomedical pretraining, the results suggest that explicitly aligning cell images with morphology-conditioned textual descriptions provides a substantially stronger supervisory signal for fine-grained WBC attribute prediction.

Experiment #3: out-of-distribution evaluation. This WBC-CLIP evaluation focuses on demonstrating its generalizability across different sources. To this end, we assess various model sizes on the LeukemiaAttri dataset, which represents a truly out-of-distribution scenario. We evaluate the predictions of our WBC-CLIP models on a subset of attributes that align with those used during training, as reported in Table 3. The comparison also includes specialized biomedical foundation models, namely BioMedCLIP and MedGemma, in order to verify whether generic biomedical pretraining alone is sufficient in this cross-dataset setting. Our analysis reveals that WBC-CLIP remains clearly superior, with the best average macro F1 achieved by ViT tiny (52.72%), followed closely by ViT small (51.94%), whereas BioMedCLIP and MedGemma obtain substantially lower averages of 24.95% and 8.56%, respectively. Among the WBC-CLIP variants, ViT-based encoders maintain superior performance in the cross-dataset scenario, whereas the ResNet-based model, despite achieving the highest performance on the WBCAtt test set, lacks generalizability. Moreover, smaller vision encoders, such as ViT tiny and ViT small, outperform their larger counterparts by a margin exceeding 10%. Notably, the attributes most frequently mispredicted are those related to cytoplasmic properties. This outcome is anticipated, as the considerable variations in acquisition methods and staining techniques tend to compromise resolution-dependent features, including the detection of vacuoles and the assessment of cytoplasmic basophilia.

5.2. *Use case: Image retrieval quality assessment*

Our WBC-CLIP models are primarily designed to assess morphological attributes in entirely unseen datasets, thereby supporting robust

Table 2
Comparison on the WBCAtt test set. WBC-CLIP consistently outperforms specialized biomedical foundation models, while ResNet50 remains the best-performing configuration overall in terms of average macro F1.

Attribute	Choices	BioMedCLIP (%)	MedGemma (%)	WBC-CLIP ViT tiny (%)	WBC-CLIP ViT small (%)	WBC-CLIP ViT base (%)	WBC-CLIP ViT large (%)	WBC-CLIP ResNet50 (%)
Class	5	10.66	4.16	97.16	98.46	97.71	**99.38**	87.12
Cell size	2	31.11	42.45	65.96	64.57	64.49	54.59	**69.09**
Cell shape	2	18.43	40.72	78.80	64.59	76.73	**81.24**	70.33
Nucleus shape	6	2.63	12.88	40.17	48.36	50.58	51.66	**52.34**
NC ratio	2	11.63	28.35	30.47	40.53	47.49	37.83	**65.05**
Chromatin density	2	47.37	47.57	49.12	23.00	57.86	52.03	**76.93**
Cytoplasm texture	2	44.24	61.23	82.33	76.37	57.07	78.57	**84.61**
Cytoplasm color	3	6.59	8.31	56.56	36.71	40.16	57.64	**63.61**
Cytoplasm vacuole	2	7.33	9.04	47.83	34.31	29.90	34.96	**49.07**
Granularity	2	42.49	45.18	74.50	64.43	**98.31**	57.83	65.99
Granule color	4	31.88	9.60	98.12	98.27	**98.22**	97.79	96.34
Granule type	4	35.00	37.52	61.26	**74.54**	73.45	61.82	58.60
Average	–	24.11	28.92	65.19	60.34	65.99	63.77	**69.92**

Table 3
Comparison on the out-of-distribution LeukemiaAttri dataset. Only the five attributes shared with the manuscript evaluation are reported. WBC-CLIP remains clearly stronger than BioMedCLIP and MedGemma in this cross-dataset scenario.

Attribute	Choices	BioMedCLIP (%)	MedGemma (%)	WBC-CLIP ViT tiny (%)	WBC-CLIP ViT small (%)	WBC-CLIP ViT base (%)	WBC-CLIP ViT large (%)	WBC-CLIP ResNet50 (%)
Cell size	2	19.09	4.29	59.24	48.63	43.69	**62.03**	48.73
Nucleus shape	2	17.83	2.82	46.27	**48.18**	26.27	18.17	14.09
Chromatin density	2	28.35	8.93	**73.61**	65.91	30.35	51.91	38.21
Cytoplasm vacuoles	2	29.80	16.71	44.16	**50.77**	49.29	45.99	47.14
Cytoplasmic basophilia	2	29.71	10.05	40.36	**46.21**	41.31	**46.21**	27.26
Average	–	24.95	8.56	**52.72**	51.94	38.18	44.86	35.08

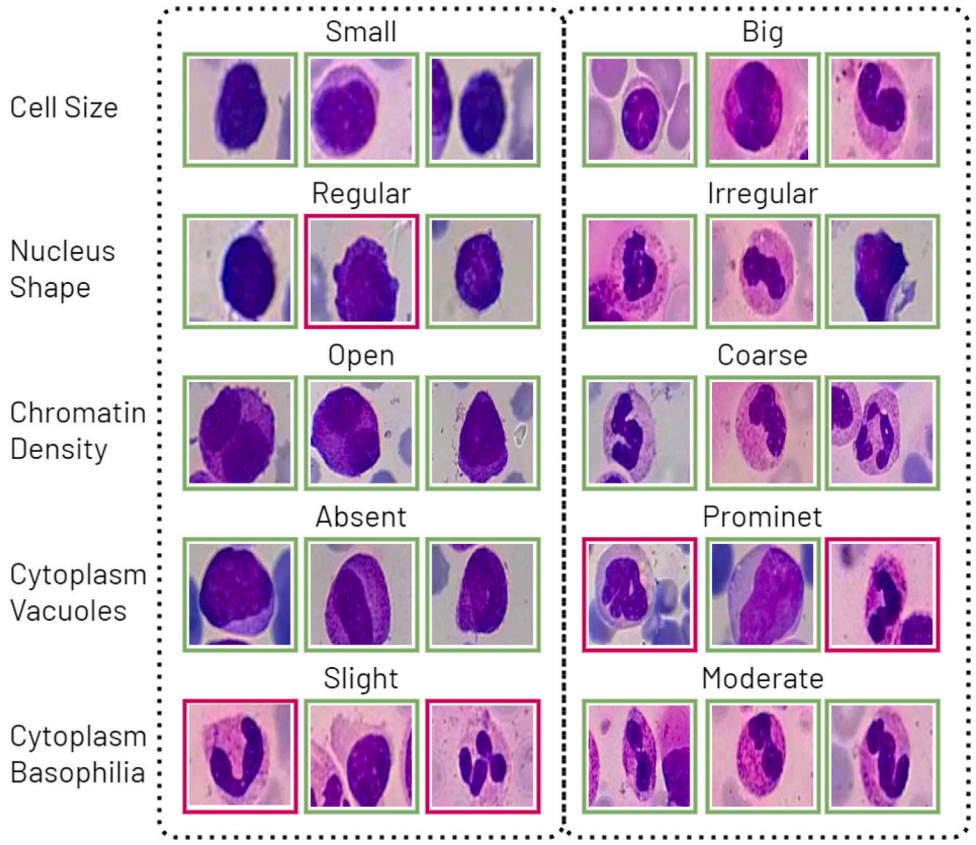

Fig. 6. Visualization of the top-3 retrieval hits obtained on the LeukemiaAttri dataset. The correct predictions are highlighted with green boxes while mispredictions with red ones. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Retrieval measures computed on the LeukemiaAttri dataset. Results are the average over the five attributes. All metrics are expressed as percentages.

Hit@1	Hit@3	Precision@1	Precision@3	MRR@3
80.00	100.00	80.00	79.99	90.00

generalization across diverse real-world conditions. In practical applications, we envision a two-stage pipeline: first, a detection algorithm identifies all WBCs within a blood smear; then, our model operates in a retrieval-based manner to isolate and rank cells exhibiting relevant pathological morphologies. This pipeline may support rapid, training-free screening within diagnostic workflows.

From a clinical perspective, the retrieval setting should be interpreted as a decision-support mechanism rather than a direct diagnostic endpoint. In practice, retrieving morphologically similar cells can assist rapid smear screening, support the review of suspicious or rare patterns, and facilitate comparison against representative examples during expert inspection. This is particularly relevant in hematology, where morphological interpretation is often comparative and context-dependent.

Moreover, the heterogeneity between WBCAtt and LeukemiaAttri makes this evaluation especially informative. Since the external dataset differs in staining, acquisition protocol, and visual appearance, strong retrieval performance suggests that the learned shared embedding space captures morphology-aware relationships that extend beyond dataset-specific style. For this reason, we regard retrieval on an out-of-distribution dataset as a useful proxy for the practical value of the learned multimodal representation. Nevertheless, these retrieved matches are intended to support, not replace, expert clinical assessment.

Furthermore, the text prompting mechanism enhances retrieval capabilities by allowing the incorporation of custom information into queries. This flexibility facilitates the identification of specific attribute combinations and specialized morphological traits that would otherwise be difficult to capture using predefined classification labels.

In high-throughput smear analysis, such a retrieval-based tool could help prioritize candidate abnormal cells before full expert examination, thereby reducing review burden while preserving the central role of the hematologist.

Fig. 6 illustrates the retrieval process, displaying the top three ranked WBCs according to the selected prompts. In the visualization, correct predictions are marked in green, while mispredictions are highlighted in red.

For this evaluation, we selected the ViT base model as the image encoder. Although it was not the best-performing ViT variant in the considered setting, it was chosen to show that retrieval remains feasible even when using a less favorable encoder configuration. Given the large number of WBCs typically present in blood smears, this result further supports the practical applicability of the proposed framework.

In our experimental setting, retrieval performance was evaluated using $k = 1$ and $k = 3$ in order to provide a deliberately strict and informative assessment. This choice reflects the importance of surfacing clinically relevant matches within the first retrieved samples, especially when rare pathological WBCs are involved. In addition, this evaluation setting is consistent with the zero-shot scenario considered for LeukemiaAttri as an unseen dataset. The corresponding results are reported in Table 4.

5.3. Explainability through cross-modal alignment

The WBC-CLIP framework provides an interpretable representation of white blood cell morphology by aligning visual features with semantically rich textual descriptions, enabling a transparent link between pixel-level patterns and clinically meaningful attributes. Instead of relying on post-hoc explanations, the model offers intrinsic interpretability: image–text contrastive learning organizes cells in a shared space shaped by morphological concepts such as nuclear structure, chromatin appearance, granularity, and cytoplasmic traits. This alignment also supports a natural form of self-explanation through bidirectional retrieval, where images retrieve the most relevant descriptions and textual queries surface representative cells, revealing how the model internally conceptualizes specific characteristics. Embedding similarities further quantify the strength of each attribute in a given sample, providing a direct and intuitive indicator of morphological reasoning. Although inherently interpretable, the architecture remains compatible with transformer-based attention inspection or Grad-CAM-style visualizations for additional scrutiny when required. Overall, this cross-modal grounding improves transparency, facilitates clinical validation, and supports the integration of multimodal models into diagnostic workflows where interpretability is essential.

6. Discussion and limitations

Despite the strong performance and inherent interpretability of the WBC-CLIP framework, several limitations warrant consideration. The approach remains sensitive to the quality, completeness, and consistency of the morphological attributes used during training; noisy annotations or stylistic differences in staining and acquisition protocols may bias the learned representations, particularly for cytoplasmic features that are more susceptible to color and illumination variability. Although our evaluation on out-of-distribution data demonstrates encouraging robustness, achieving consistent generalization across highly heterogeneous clinical settings remains challenging. The model also relies on a predefined set of morphological attributes, which, while comprehensive, does not fully capture the breadth of subtle or rare variations encountered in hematology practice. Additionally, transformer-based encoders introduce non-trivial computational demands, potentially limiting deployment in low-resource laboratories without further optimization.

An additional finding of this study is that specialized biomedical foundation models, although relevant and competitive baselines, were not sufficient to match the performance of WBC-CLIP in either the in-distribution or out-of-distribution setting. This suggests that the main advantage of the proposed framework stems not only from biomedical prior knowledge, but from the explicit alignment between cell images and morphology-aware textual supervision.

Additional limitations arise from the use of LLM-generated descriptions, which may introduce stylistic noise, omissions, paraphrastic distortions, or partial mismatches with the original attributes. In our setting, this risk is mitigated to some extent by constraining generation to structured morphological annotations rather than allowing open-ended medical interpretation, and by quantitatively measuring attribute consistency between the generated captions and the original labels. However, this mitigation is only partial. Our consistency analysis indicates that, although several key attributes are preserved reliably, others are less consistently reflected in the generated text. As a result, the textual modality may contain imperfect or partially inconsistent supervision signals.

From a training perspective, such noise may weaken image-text alignment for some samples and can introduce additional variability into the optimization process. At the same time, exposure to multiple non-identical textual realizations of the same morphology-conditioned information may provide a degree of robustness to superficial linguistic variation, encouraging the model to align with shared semantic content rather than a single phrasing style. Nevertheless, this should not be interpreted as a guarantee of robustness to erroneous descriptions. Automatic consistency analysis is not equivalent to clinical validation, and we did not perform a formal hematology-expert review protocol for all generated captions in the present study. Accordingly, these captions should be regarded as auxiliary morphology-conditioned training signals for multimodal alignment, not as clinically validated reports.

Future research should therefore focus on expanding the diversity and granularity of morphological features, refining the quality and consistency of textual descriptions, and incorporating more heterogeneous datasets to better reflect real-world laboratory variability. Methods such as domain adaptation, continual learning, or attribute-aware alignment could enhance stability across different imaging conditions. Additionally, introducing lightweight or distilled variants of WBC-CLIP may facilitate adoption in resource-constrained environments. Finally, systematic assessment of bias, transparency, and auditability as well as the development of human-in-the-loop interfaces will be essential for supporting safe, equitable, and clinically responsible deployment of multimodal AI systems in hematology.

6.1. Clinical impact

The WBC-CLIP framework has the potential to support several aspects of hematological diagnostics by providing interpretable, morphology-aware representations that complement existing laboratory workflows. Its ability to align images with descriptive attributes enables more transparent triage of peripheral blood smears, facilitating rapid screening for morphological abnormalities, early detection of atypical or leukemic patterns, and more efficient prioritization of cases requiring expert review. The bidirectional retrieval capabilities can assist clinicians in comparing ambiguous cells against prototypical examples or in surfacing consistent textual descriptions that clarify subtle traits, thereby reducing inter-observer variability. Moreover, the zero-shot and attribute-centric learning paradigm allows the system to adapt to new or rare morphological categories without retraining, which is particularly valuable in scenarios where annotated data are scarce. By integrating the model within digital hematology platforms, laboratories could benefit from enhanced quality control, more consistent morphological reporting, and decision-support tools that strengthen diagnostic confidence while preserving the central role of expert interpretation. Ultimately, WBC-CLIP offers a scalable foundation for multimodal clinical support systems that augment, rather than replace, professional expertise in routine and specialized hematological assessment.

7. Conclusion

In this work, we proposed WBC-CLIP, a dual-encoder framework that effectively integrates visual and textual data to enhance the analysis of WBCs. By leveraging contrastive learning in conjunction with multiple large language models, our approach generates detailed textual descriptions that, when paired with image embeddings, provide reliable and interpretable diagnostic insights, as suggested by the image retrieval quality assessment use case.

Our extensive evaluations on both in-distribution and out-of-distribution datasets demonstrate that WBC-CLIP not only advances the state of automated WBC analysis but also facilitates rapid, training-free screening in clinical environments. The ability to retrieve and rank cells based on morphological attributes further underscores its potential for assisting pathologists in identifying diagnostically relevant patterns with greater efficiency.

Despite its promising performance, our approach still faces challenges, particularly concerning data quality, potential annotation biases, and computational demands. However, WBC-CLIP represents a significant step toward more transparent, interpretable, and adaptable AI-driven diagnostic tools. Future research should focus on mitigating these limitations by improving dataset diversity, refining retrieval mechanisms, and optimizing computational efficiency to ensure broader accessibility in real-world clinical applications.

Acronyms

See Table 5.

Table 5
List of acronyms used in this manuscript.

Acronym	Meaning
WBC	White Blood Cell
CLIP	Contrastive Language–Image Pretraining
LLM	Large Language Model
NLP	Natural Language Processing
CNN	Convolutional Neural Network
ViT	Vision Transformer
AUC	Area Under the Curve
TP	True Positives
FP	False Positives
FN	False Negatives
TN	True Negatives
ACC	Accuracy
PRE	Precision
REC	Recall
F1	F1-score
OOD	Out-of-Distribution

CRediT authorship contribution statement

Luca Zedda: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Davide Antonio Mura:** Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Andrea Manzo:** Validation, Software, Methodology, Data curation, Conceptualization. **Cecilia Di Ruberto:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Formal analysis, Conceptualization. **Andrea Loddo:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Formal analysis, Conceptualization.

Code availability

The code for this study is available at the following GitHub repository: https://github.com/unica-visual-intelligence-lab/WBC-CLIP.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No. 3277 published on December 30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR), and by the GNCS Project 2025 “Metodi di approssimazione globale per operatori integrali e applicazioni alle equazioni funzionali” (CUP E53C24001950001).

Data availability

Data will be made available on request.

References

[1] N.M. Deshpande, S. Gite, R. Aluvalu, A review of microscopic analysis of blood cells for disease detection with AI perspective, PeerJ Comput. Sci. 7 (2021) e460.

[2] S. Doulatov, F. Notta, E. Laurenti, J.E. Dick, Hematopoiesis: A human perspective, Cell Stem Cell 10 (2) (2012) 120–136, http://dx.doi.org/10.1016/j.stem.2012.01.006, URL https://www.sciencedirect.com/science/article/pii/S1934590912000082.

[3] E. Cuevas, M. Díaz, M. Manzanares, D. Zaldivar, M. Perez-Cisneros, An improved computer vision method for white blood cells detection, Comput. Math. Methods Med. 2013 (1) (2013) 137392, http://dx.doi.org/10.1155/2013/137392, URL https://onlinelibrary.wiley.com/doi/abs/10.1155/2013/137392. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1155/2013/137392.

[4] A. Sadafi, M. Bordukova, A. Makhro, N. Navab, A. Bogdanova, C. Marr, RedTell: an AI tool for interpretable analysis of red blood cell morphology, Front. Physiol. 14 (2023) 1058720.

[5] C. Stringer, M. Pachitariu, Cellpose3: one-click image restoration for improved cellular segmentation, Nature Methods (2025) 1–8.

[6] M. Chossegros, F. Delhommeau, D. Stockholm, X. Tannier, Improving the generalizability of white blood cell classification with few-shot domain adaptation, J. Pathol. Inform. 15 (2024) 100405, http://dx.doi.org/10.1016/j.jpi.2024.100405, URL https://www.sciencedirect.com/science/article/pii/S2153353924000440.

[7] L. Fu, J. Chen, Y. Zhang, X. Huang, L. Sun, CNN and transformer-based deep learning models for automated white blood cell detection, Image Vis. Comput. 161 (2025) 105631, http://dx.doi.org/10.1016/j.imavis.2025.105631, URL https://www.sciencedirect.com/science/article/pii/S0262885625002197.

[8] J. Ferdousi, S.I. Lincoln, M.K. Alom, M. Foysal, A deep learning approach for white blood cells image generation and classification using SRGAN and VGG19, Telemat. Inform. Rep. 16 (2024) 100163, http://dx.doi.org/10.1016/j.teler.2024.100163, URL https://www.sciencedirect.com/science/article/pii/S2772503024000495.

[9] O. Saidani, M. Umer, N. Alturki, A. Alshardan, M. Kiran, S. Alsubai, T.-H. Kim, I. Ashraf, White blood cells classification using multi-fold pre-processing and optimized CNN model, Sci. Rep. 14 (1) (2024) 3570, http://dx.doi.org/10.1038/s41598-024-52880-0, URL https://www.nature.com/articles/s41598-024-52880-0. Publisher: Nature Publishing Group.

[10] L. Putzu, S. Porcu, A. Loddo, Distributed collaborative machine learning in real-world application scenario: A white blood cell subtypes classification case study, Image Vis. Comput. 162 (2025) 105673, http://dx.doi.org/10.1016/j.imavis.2025.105673, URL https://www.sciencedirect.com/science/article/pii/S0262885625002616.

[11] C. Jung, M. Abuhamad, D. Mohaisen, K. Han, D. Nyang, WBC image classification and generative models based on convolutional neural network, BMC Med. Imaging 22 (1) (2022) 94, http://dx.doi.org/10.1186/s12880-022-00818-1, URL https://doi.org/10.1186/s12880-022-00818-1.

[12] A. Panthakkan, S.M. Anzar, W. Mansoor, H.A. Ahmad, A new frontier in hematology: Robust deep learning ensembles for white blood cell classification, Biomed. Signal Process. Control. 100 (2025) 106995, http://dx.doi.org/10.1016/j.bspc.2024.106995, URL https://www.sciencedirect.com/science/article/pii/S174680942401053X.

[13] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 8748–8763, URL https://proceedings.mlr.press/v139/radford21a.html. (ISSN: 2640-3498).

[14] S. Tsutsui, W. Pang, B. Wen, Wbcatt: A white blood cell dataset annotated with detailed morphological attributes, Adv. Neural Inf. Process. Syst. 36 (2023) 50796–50824.

[15] S. Khan, M. Sajjad, T. Hussain, A. Ullah, A.S. Imran, A review on traditional machine learning and deep learning models for WBCs classification in blood smear images, IEEE Access 9 (2021) 10657–10673, http://dx.doi.org/10.1109/ACCESS.2020.3048172, URL https://ieeexplore.ieee.org/document/9311202/.

[16] M. Toptaş, B. Toptaş, D. Hanbay, Classifying white blood cells using combining different convolutional neural networks, Multimedia Tools Appl. 84 (35) (2025) 44089–44112, http://dx.doi.org/10.1007/s11042-025-20879-y, URL https://link.springer.com/article/10.1007/s11042-025-20879-y. Publisher: Springer.

[17] A. Girdhar, H. Kapur, V. Kumar, Classification of white blood cell using convolution neural network, Biomed. Signal Process. Control. 71 (2022) 103156, http://dx.doi.org/10.1016/j.bspc.2021.103156, URL https://www.sciencedirect.com/science/article/pii/S1746809421007539.

[18] V. Koch, S.J. Wagner, S. Kazeminia, E. Sancar, M. Hehr, J.A. Schnabel, T. Peng, C. Marr, DinoBloom: A foundation model for generalizable cell embeddings in hematology, in: Medical Image Computing and Computer Assisted Intervention – MICCAI, 2024, pp. 520–530, http://dx.doi.org/10.1007/978-3-031-72390-2_49, URL https://link.springer.com/chapter/10.1007/978-3-031-72390-2_49.

[19] Y. Wu, D. Zeng, Z. Wang, Y. Shi, J. Hu, Distributed contrastive learning for medical image segmentation, Med. Image Anal. 81 (2022) 102564, http://dx.doi.org/10.1016/j.media.2022.102564, URL https://www.sciencedirect.com/science/article/pii/S1361841522002079.

[20] Y. Wang, T. Wang, X. Shu, Y. Zheng, J. Ding, X. Fu, Z. Zheng, Structure-aware contrastive learning for glomerulus segmentation in renal pathology, Image Vis. Comput. 162 (2025) 105698, http://dx.doi.org/10.1016/j.imavis.2025.105698, URL https://www.sciencedirect.com/science/article/pii/S0262885625002860.

[21] H. Hu, X. Wang, Y. Zhang, Q. Chen, Q. Guan, A comprehensive survey on contrastive learning, Neurocomput. 610 (C) (2024) http://dx.doi.org/10.1016/j.neucom.2024.128645.

[22] H. Yang, W. Zeng, K. Chen, Z. Hua, Y. Zhuang, L. Han, G. Liao, Y. Zhang, H. Li, Z. Li, J. Lin, SPM-CyViT: A self-supervised pre-trained cycle-consistent vision transformer with multi-branch for contrast-enhanced CT synthesis, Image Vis. Comput. 164 (2025) 105802, http://dx.doi.org/10.1016/j.imavis.2025.105802, URL https://www.sciencedirect.com/science/article/pii/S0262885625003907.

[23] L. Zedda, A. Loddo, C. Di Ruberto, C. Marr, RedDino: A foundation model for red blood cell analysis, in: Medical Image Computing and Computer Assisted Intervention – MICCAI, 2025, pp. 445–455, http://dx.doi.org/10.1007/978-3-032-04965-0_42, URL https://link.springer.com/chapter/10.1007/978-3-032-04965-0_42.

[24] L. My, B. Chen, D. Williamson, R.J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. Le, G. Gerber, A. Parwani, A. Zhang, F. Mahmood, A visual-language foundation model for computational pathology, Nature Med. (2024) URL https://pubmed.ncbi.nlm.nih.gov/38504017/.

[25] T. Ding, S.J. Wagner, A.H. Song, R.J. Chen, M.Y. Lu, A. Zhang, A.J. Vaidya, G. Jaume, M. Shaban, A. Kim, D.F.K. Williamson, H. Robertson, B. Chen, C. Almagro-Pérez, P. Doucet, S. Sahai, C. Chen, C.S. Chen, D. Komura, A. Kawabe, M. Ochi, S. Sato, T. Yokose, Y. Miyagi, S. Ishikawa, G. Gerber, T. Peng, L.P. Le, F. Mahmood, A multimodal whole-slide foundation model for pathology, Nature Med. (2025) 1–13, http://dx.doi.org/10.1038/s41591-025-03982-3, URL https://www.nature.com/articles/s41591-025-03982-3. Publisher: Nature Publishing Group.

[26] D. Mahapatra, B. Bozorgtabar, Z. Ge, Medical image classification using generalized zero shot learning, in: 2021 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW, 2021, pp. 3337–3346, http://dx.doi.org/10.1109/ICCVW54120.2021.00373, URL https://ieeexplore.ieee.org/document/9607746. (ISSN: 2473-9944).

[27] S. Stevens, J. Wu, M.J. Thompson, E.G. Campolongo, C.H. Song, D.E. Carlyn, L. Dong, W.M. Dahdul, C. Stewart, T. Berger-Wolf, W.-L. Chao, Y. Su, BioCLIP: A vision foundation model for the tree of life, 2024, pp. 19412–19424, CoRR.

[28] Z. Wang, Z. Wu, D. Agarwal, J. Sun, MedCLIP: Contrastive learning from unpaired medical images and text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, Vol. 2022, 2022, pp. 3876–3887, http://dx.doi.org/10.18653/v1/2022.emnlp-main.256, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11323634/.

[29] G.H. Dagnaw, Y. Zhu, M.H. Maqsood, X. Yin, A.W. Liew, Explainable multimodal hematology analysis for white blood cell classification and attribute prediction, Comput. Biol. Med. 196 (2025) 110734, http://dx.doi.org/10.1016/J.COMPBIOMED.2025.110734.

[30] J.v. Logtestijn, P. Manescu, HemBLIP: A vision-language model for interpretable leukemia cell morphology analysis, 2026, http://dx.doi.org/10.48550/arXiv.2601.03915, URL http://arxiv.org/abs/2601.03915. arXiv:2601.03915 [cs].

[31] A. Lubna, S. Kalady, A. Lijiya, Visual question answering on blood smear images using convolutional block attention module powered object detection, Vis. Comput. 41 (1) (2024) 739–757, http://dx.doi.org/10.1007/s00371-024-03359-6, URL https://link.springer.com/article/10.1007/s00371-024-03359-6.

[32] F. Shehzad, C. Mennella, M. Esposito, A. Minutolo, Efficient multimodal learning using BERT and vision transformers for visual question answering on peripheral blood cells, Discov. Artif. Intell. 6 (1) (2026) 265, http://dx.doi.org/10.1007/s44163-026-01011-x, URL https://link.springer.com/article/10.1007/s44163-026-01011-x.

[33] A. Rehman, I. Rasool, A. Imran, M. Ali, W. Sultani, Uni-Hema: Unified model for digital hematopathology, 2025, http://dx.doi.org/10.48550/ARXIV.2511.13889, URL https://arxiv.org/abs/2511.13889. Version Number: 2.

[34] Y. Guo, A. Ovadje, M.A. Al-Garadi, A. Sarker, Evaluating large language models for health-related text classification tasks with public social media data, J. Am. Med. Inform. Assoc. : JAMIA 31 (10) (2024) 2181–2189, http://dx.doi.org/10.1093/jamia/ocae210, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11413434/.

[35] S.R. Pulari, M. Umadevi, S.K. Vasudevan, Optimizing multimodal personalized disease prediction accuracy using generated prompts and large language models, Image Vis. Comput. 161 (2025) 105649, http://dx.doi.org/10.1016/j.imavis.2025.105649, URL https://www.sciencedirect.com/science/article/pii/S0262885625002379.

[36] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth, X. Zhou, L. Qian, H. He, D. Shung, L. Ohno-Machado, Y. Wu, H. Xu, J. Bian, Medical foundation large language models for comprehensive text analysis and beyond, Npj Digit. Med. 8 (1) (2025) 141, http://dx.doi.org/10.1038/s41746-025-01533-1, URL https://www.nature.com/articles/s41746-025-01533-1. Publisher: Nature Publishing Group.

[37] M. Spotnitz, B. Idnay, E.R. Gordon, R. Shyu, G. Zhang, C. Liu, J.J. Cimino, C. Weng, A survey of clinicians' views of the utility of large language models, Appl. Clin. Inform. 15 (2) (2024) 306–312, http://dx.doi.org/10.1055/a-2281-7092, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11023712/.

[38] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C.S. Schmidt, S. Koitka, O. Pelka, A.B. Abacha, A.G.S.d. Herrera, H. Müller, P.A. Horn, F. Nensa, C.M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Sci. Data 11 (1) (2024) 688, http://dx.doi.org/10.1038/s41597-024-03496-6, URL http://arxiv.org/abs/2405.10004. arXiv:2405.10004 [eess].

[39] S. Tsutsui, W. Pang, B. Wen, WBCAtt: A white blood cell dataset annotated with detailed morphological attributes, in: A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, S. Levine (Eds.), Advances in Neural Information Processing Systems, Vol. 36, 2023, pp. 17637–17654, URL https://nips.cc.

[40] A. Rehman, T. Meraj, A.M. Minhas, A. Imran, M. Ali, W. Sultani, A large-scale multi domain leukemia dataset for the white blood cells detection with morphological attributes for explainability, in: M.G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, J.A. Schnabel (Eds.), Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part III, in: Lecture Notes in Computer Science, vol. 15003, Springer, 2024, pp. 553–563, http://dx.doi.org/10.1007/978-3-031-72384-1_52.

[41] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, M.P. Lungren, T. Naumann, H. Poon, Large-scale domain-specific pretraining for biomedical vision-language processing, 2023, http://dx.doi.org/10.48550/ARXIV.2303.00915, CoRR abs/2303.00915. arXiv:2303.00915.

[42] G. Research, G. DeepMind, Medgemma technical report, 2025, http://dx.doi.org/10.48550/ARXIV.2507.05201, CoRR abs/2507.05201. arXiv:2507.05201.