Evaluating Vision-Language Models for hematology image Classification: Performance Analysis of CLIP and its Biomedical AI Variants

Tanviben Patel, Hoda El-Sayed, Md Kamruzzaman Sarker
Bowie State University, Bowie, MD, USA
patelt0902@student.bowiestate.edu, helsayed@bowiestate.edu, ksarker@bowiestate.edu

***Abstract*—Vision-language models (VLMs) have shown remarkable potential in various domains, particularly in zero-shot learning applications. This research focuses on evaluating the performance of notable VLMs—CLIP, PLIP, and BiomedCLIP—in the classification of blood cells, with a specific emphasis on distinguishing between normal and malignant (cancerous) cells datasets. While CLIP demonstrates robust zero-shot capabilities in general tasks, this study probes its biomedical adaptations, PLIP and BiomedCLIP, to assess their effectiveness in specialized medical tasks, such as hematological image classification. Additionally, we investigate the impact of prompt engineering on model performance, exploring how variations in prompt construction influence accuracy across these biomedical datasets. Extensive experiments were conducted on a variety of biomedical images, including microscopic blood cell images, brain MRIs, and chest X-rays, providing a comprehensive evaluation of the VLMs' applicability in medical imaging. Our findings reveal that while CLIP, trained on general datasets, performs well in broader contexts, PLIP and BiomedCLIP—optimized for medical imagery—demonstrate enhanced accuracy in medical settings, particularly in hematology. The results underscore the strengths and limitations of these models, offering valuable insights into their adaptability, precision, and potential for future applications in medical image classification.**

I. Introduction

The advent of vision-language models (VLMs) has revolutionized artificial intelligence by integrating visual perception with natural language processing, opening new possibilities across various domains. In healthcare, these multimodal models have shown tremendous promise, from generating diagnostic reports based on medical imaging to enhancing the functionality of medical chatbots. VLMs are particularly notable for their ability to perform zero-shot learning, where they successfully complete tasks without specific prior training on those tasks. This capability has the potential to significantly improve the accuracy and efficiency of medical image analysis, providing invaluable insights to healthcare professionals [1]–[5].

While VLMs have been applied successfully across multiple sectors, their effectiveness in medical applications often depends on the availability of large, annotated datasets. Such datasets, curated by human experts, are essential for training these models to ensure that they produce reliable and accurate results. Among the VLMs, CLIP (Contrastive Language–Image Pre-training) [6] stands out for its strong performance in general image classification using zero-shot learning. Building on this, specialized models like Pathology Language and Image Pre-Training (PLIP) [7] and BiomedCLIP [8] have been developed to better handle medical imaging tasks, particularly in fields such as pathology. These models are tailored to capture the complexities of medical images, potentially transforming diagnostic workflows. However, despite these advances, even medically pre-trained models like PLIP and BiomedCLIP may not encompass the full range of medical imaging domains, limiting their effectiveness, especially in the diagnosis of rare or complex conditions.

This study aims to evaluate the performance of leading VLMs—CLIP, PLIP, and BiomedCLIP—in the context of blood cell classification. Specifically, the research focuses on the models' ability to differentiate between normal and malignant (cancerous) blood cells across two distinct datasets. Although CLIP has established itself as a versatile tool for zero-shot learning, the study will investigate the effectiveness of its biomedical adaptations, PLIP and BiomedCLIP, in handling specialized medical tasks.

This research is guided by four main questions:

- How effective are vision-language models (VLMs) like CLIP, PLIP, and BiomedCLIP in classifying blood cells as normal or malignant?
- What are the strengths and limitations of general-purpose VLMs (e.g., CLIP) compared to biomedical-focused VLMs (e.g., PLIP, BiomedCLIP) for specialized medical image classification?
- How do variations in prompt construction affect the accuracy of VLMs in medical image classification, particularly in blood cell analysis?
- Can VLMs trained on general datasets, like CLIP, be effectively adapted for biomedical applications, particularly in hematology, as compared to models explicitly trained on medical data, such as PLIP and BiomedCLIP?

Through these questions, the study seeks to uncover insights into the adaptability and performance of VLMs in the biomedical domain, offering potential pathways for improving medical image classification workflows.

II. Related work

A. Dataset

In this study, we utilized the BloodMNIST dataset [9], which is also available via the GitHub

link https://github.com/MedMNIST/MedMNIST. The BloodMNIST dataset, sometimes referred to as "bloodmnist," consists of medical images depicting normal blood cells. These images were collected from individuals who were free from infections, hematologic disorders, oncological diseases, and had not undergone any pharmacological treatments at the time of blood collection. We selected 10,298 images, which are categorized into five distinct blood cell types: basophils, eosinophils, lymphocytes, monocytes, and neutrophils (as shown in Fig. 1).

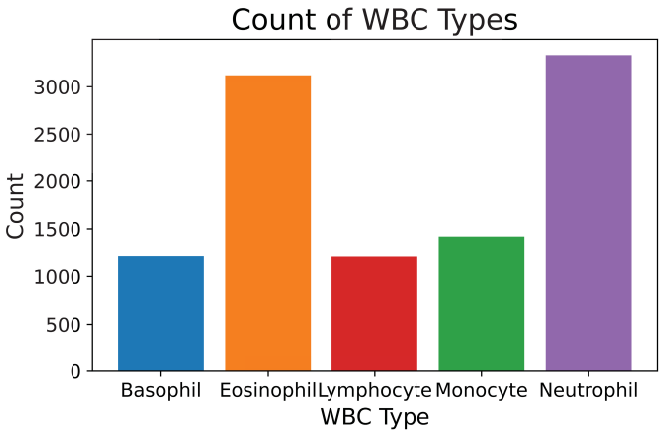

Fig. 1. White Blood Cell (WBC) dataset

The second dataset used in this study is the Blood Cancer dataset [10], which was prepared at the bone marrow laboratory of Taleqani Hospital in Tehran, Iran. This dataset comprises 3,256 peripheral blood smear (PBS) images from 89 patients suspected of having acute lymphoblastic leukemia (ALL). The blood samples were prepared and stained by experienced laboratory staff. The dataset is divided into two main categories: benign and malignant. The benign class consists of hematogones, while the malignant class includes the ALL group, further subdivided into three types of malignant lymphoblasts: Early Pre-B, Pre-B, and Pro-B ALL. All images were captured using a Zeiss microscope with 100x magnification and saved as JPG files. For this study, we focused on malignant cells, specifically early, pre, and pro cancer stages (as shown in Fig. 2).

B. Background

In a recent study [11], the authors evaluated CLIP in both open- and closed-world settings, where the closed-world scenario provides a predefined set of labels. CLIP processes the image and the prompt, selecting the class with the highest cosine similarity to the image feature. In the closed-world setting, CLIP outperformed various vision-language models (VLMs) across multiple classification tasks, including fine-grained datasets like Flowers102 and StanfordCars. However, in the open-world setting, CLIP's performance was significantly lower, achieving only 32% accuracy. The study also

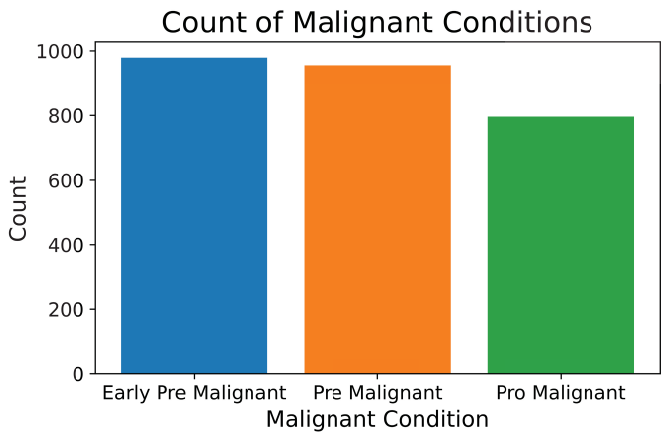

Fig. 2. Acute Lymphoblastic Leukemia (ALL) image dataset

emphasized the importance of prompt engineering as a major limitation for VLMs [12].

Vision-Language Models (VLMs) have emerged as powerful tools for bridging the gap between visual and textual information. CLIP [6], a pioneering VLM developed by OpenAI, has demonstrated impressive performance in various tasks, including zero-shot learning. However, its application in the medical domain, particularly for medical image classification, remains relatively unexplored. To address the challenges of applying general-purpose VLMs to medical imaging, domain-specific adaptations like PLIP and BiomedCLIP have been developed. These models incorporate medical-specific datasets to improve performance in healthcare tasks, such as disease classification from imaging data [3] [13] [14]..

Previous studies on blood cell classification have largely relied on convolutional neural networks (CNNs) with extensive labeled datasets [15]–[18]. The introduction of VLMs in medical imaging offers a new approach, enabling models to generalize from broader knowledge to specialized tasks without extensive retraining [19]. Recent advancements in prompt engineering have further underscored its impact on model performance, particularly in specialized fields like medicine.

In this study, we evaluate the performance of CLIP, PLIP, and BiomedCLIP on blood cell datasets, assessing their ability to distinguish between normal and cancerous cells. We also explore the influence of prompt variation on VLM accuracy in biomedical tasks.

A recent paper by Kakkar et al. [20] addresses the gap in automated image description generation for whole-body multimodal clinical scans, specifically MR and CT radiological images. While previous research focused on generating clinical descriptions for specific body regions or modalities, this study presents a method for generating standardized descriptions of body stations and organs across the entire body using the CLIP model. With refinements such as fine-tuning the model, augmenting the prompt structure, and leveraging domain-specific data, the approach achieved a 47.6% performance improvement over the baseline PubMedCLIP.

In this study, we also compare CLIP, PLIP, and BiomedCLIP in the context of blood cell classification, investigating how these models handle complex medical imagery and the extent to which prompt engineering can enhance their performance in this domain.

C. CLIP

CLIP, developed by OpenAI, is a vision-language model designed to understand both images and text by jointly training on image-text pairs. It works by learning to associate images with their corresponding textual descriptions through a contrastive learning approach. CLIP excels in zero-shot learning, meaning it can classify images without being explicitly trained on specific tasks, leveraging its broad knowledge across domains. This allows it to generalize well across different datasets and perform a variety of visual tasks without task-specific retraining. CLIP has shown impressive performance in object detection, image classification, and other general-purpose vision-language tasks [21]–[23]. It is often used in contexts requiring flexibility, such as in creative applications, automated content moderation, and more [24].

D. PLIP

PLIP is a biomedical adaptation of CLIP designed to address the specific needs of medical image analysis. While CLIP is trained on general datasets, PLIP uses medical image-text pairs to focus on pathology and related tasks. PLIP is tailored for medical imagery, offering improved accuracy and performance in tasks like medical image classification, segmentation, and diagnosis [7], [25]. By incorporating domain-specific data, PLIP bridges the gap between general-purpose models and specialized medical applications [26]. PLIP is primarily used in pathology and related fields, aiding in tasks like disease detection, diagnosis support, and interpretation of histopathology images [27]. Its ability to understand medical terminology and complex visual features makes it highly useful in clinical settings.

E. BiomedCLIP

BiomedCLIP is another specialized adaptation of CLIP, but it is even more focused on a broader range of biomedical tasks beyond pathology. It is trained on biomedical text and images to improve performance in clinical and diagnostic tasks across various medical domains. By leveraging biomedical corpora and extensive image-text pairs from medical contexts, BiomedCLIP enhances the model's ability to interpret and classify complex medical images, such as MRIs, CT scans, and X-rays, alongside their associated reports [28]–[30]. BiomedCLIP is used in various medical imaging tasks, including disease classification, image annotation, and clinical report generation [31]. Its domain-specific training allows it to excel in healthcare applications where general models like CLIP may lack the necessary depth of understanding.

III. Methodology

We utilized three different vision-language models (VLMs)—CLIP, PLIP, and BiomedCLIP—to evaluate their performance in classifying both malignant blood cells and normal white blood cells (WBCs) using three distinct text prompts. The dataset, which included images in formats such as .png, .jpg, .jpeg, .bmp, and .tiff, was loaded from a directory. Each image was preprocessed by resizing it to 224x224 pixels to match the input requirements of the models. The images were processed in batches of 32, and once a batch reached this size, it was fed into the model for inference.

For each batch, the model computed image-text similarity scores (logits) between the images and the given prompts. These logits were then converted into probabilities using the softmax function. The predicted classes and their corresponding probabilities were recorded for each image. This procedure was repeated for all three models—CLIP, PLIP, and BiomedCLIP—using the same dataset and prompts. The resulting predictions were analyzed to compare each model's performance in classifying different stages of malignant blood cells.

Each model was loaded, and the PLIP model was initialized using the `vinid/plip` pre-trained model from HuggingFace's model hub.

The figure illustrates (Fig. 3) the process of zero-shot image classification using a Vision-Language Model (VLM). On the left, a set of textual prompts is provided, describing different scenarios for white blood cell (WBC) images. These prompts include:

- "A photo of c"
- "A microscopic hematology image of c"
- "A microscopic image of a white blood cell, surrounded by red blood cells, classified as a c."

Here, c represents the possible classes of white blood cells, such as neutrophils, eosinophils, lymphocytes, etc. The input images (I1, I2, ..., IN) correspond to different WBC images.

On the right side of the figure, the VLM performs zero-shot predictions by combining each image (I1, I2, ..., IN) with the textual prompts (T1, T2, ..., TN). Each combination is assessed based on the model's ability to match the visual content of the image with the most relevant class described in the text. The VLM generates predictions by selecting the image-text pair with the highest similarity, leading to the final zero-shot classification for each image.

This method allows VLMs to classify images without requiring explicit training on the specific task, showcasing their generalization capabilities across various medical image classification challenges, such as identifying different white blood cell types.

Figure 4(a) showcases examples of the same microscopic blood cell images, where prediction probabilities fluctuate based on different prompt formulations. The labels and probabilities for "Early Pre-B malignant" cells vary significantly with each prompt, highlighting the model's sensitivity to

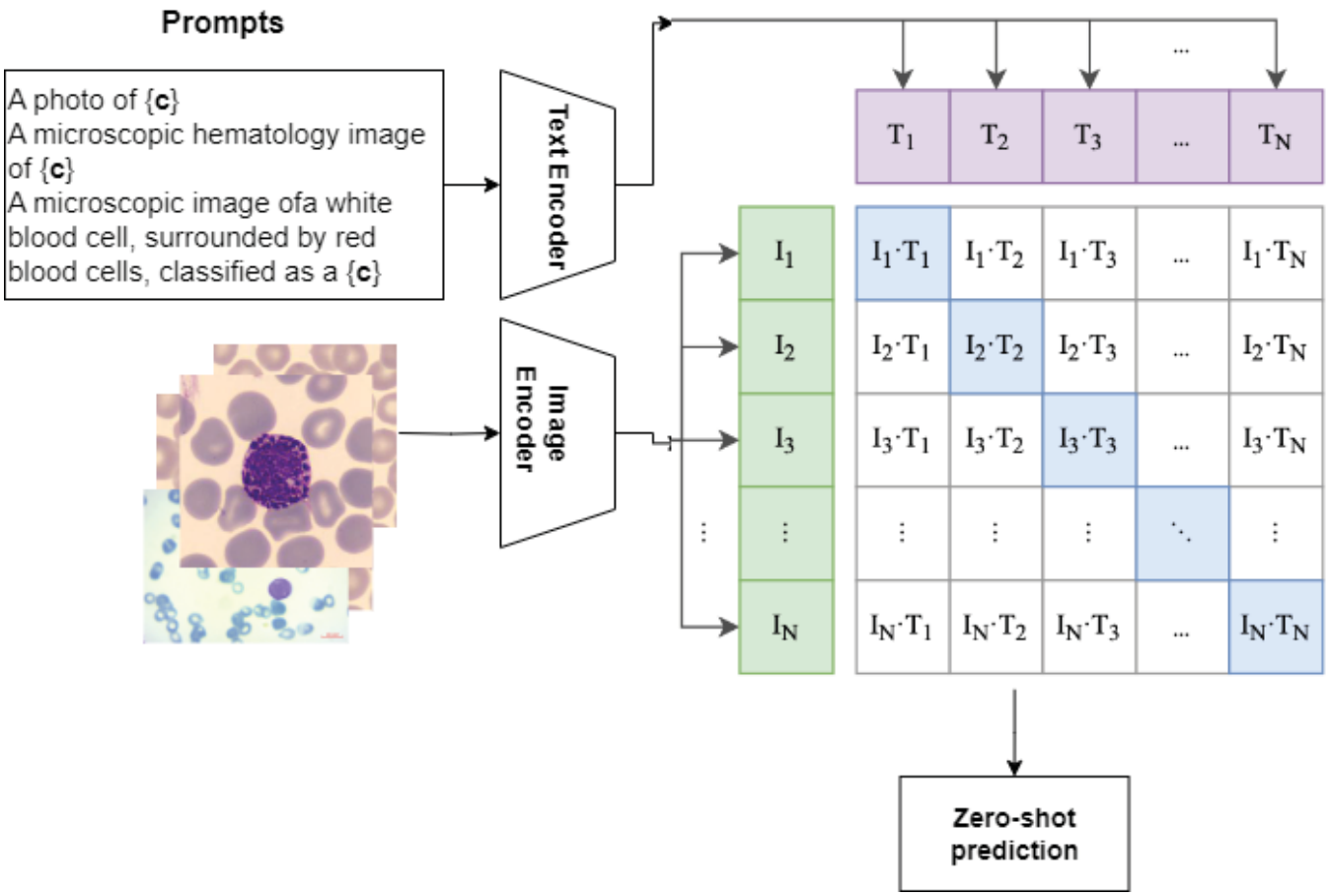

Fig. 3. Zero-shot prediction using Vision-Language Models (VLMs) on Blood Cell images

changes in textual descriptions. The green-highlighted sections represent the correct class and probability predictions for each prompt with that class.

Figure 4(b) displays three microscopic blood cell images, each classified using distinct textual prompts. The predictions, color-coded for clarity, indicate the likelihood of three malignancy stages: early pre-malignant, pre-malignant, and pro-malignant. The left image, using the prompt "A photo of class", showed the highest probability for the pro-malignant stage. The middle image, using the prompt "A microscopic hematology image of class", was most likely classified as early pre-malignant. The right image, prompted with "A microscopic image of a white blood cell, surrounded by red blood cells, classified as class", was predicted to be pre-malignant. These distinct prompts were designed to evaluate the effect of descriptive language on the model's performance in classifying blood cell malignancy stages accurately.

IV. Results

A. WBC dataset probability without fine-tuning

In this section, we present the results of the three models (PLIP, BiomedCLIP, and CLIP) across different prompts, analyzing their class probabilities for basophils, eosinophils, lymphocytes, monocytes, and neutrophils.

The **PLIP model** exhibited considerable variability across the prompts, with significant shifts in class probabilities. In **Prompt-1**, the model assigned similar probabilities to basophils (25.8%) and neutrophils (25.7%), while eosinophils, lymphocytes, and monocytes had more evenly distributed probabilities ranging from 13% to 18%. This distribution suggests that the model lacked strong confidence in distinguishing between specific classes, particularly showing ambiguity between basophils and neutrophils. In **Prompt-2**, neutrophil confidence increased to 29%, with lymphocytes following closely at 23%. There was a marked decrease in basophil and eosinophil probabilities, indicating that the model had shifted its focus towards neutrophils and lymphocytes. In **Prompt-3**, the model displayed a dramatic shift, with basophil probability increasing significantly to 60.4%. This sharp rise suggests that the model became highly confident in identifying basophils. However, probabilities for other classes decreased, with monocytes maintaining a notable presence, while eosinophils, neutrophils, and lymphocytes were deprioritized.

The **BiomedCLIP** model demonstrated more stability but showed a heavy bias towards neutrophils and monocytes across all prompts. In **Prompt-1**, neutrophils (33.7%) and monocytes (31.1%) dominated the predictions, while lower probabilities were assigned to basophils, eosinophils, and lymphocytes. This

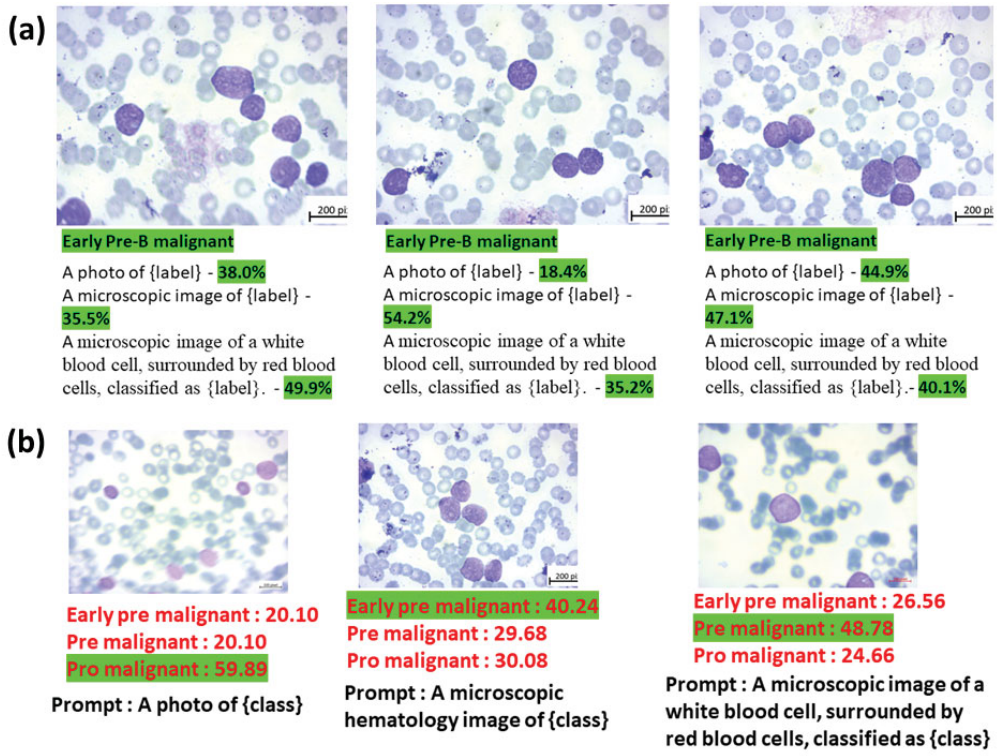

Fig. 4. Example classification of malignant blood cells using three different prompts with corresponding prediction probabilities for early pre-malignant, pre-malignant, and pro-malignant conditions using BiomedCLIP pre-trained model. (a) illustrates examples of the same microscopic blood cell images where prediction probabilities vary with changes in prompt formulations. (b) displays the impact of using different prompts on the classification accuracy of the malignant cells.

suggests that the model was more confident in distinguishing neutrophils and monocytes. In **Prompt-2**, neutrophil confidence soared to 72%, overwhelming the predictions for other classes. This strong bias towards neutrophils indicates the model's preference for identifying neutrophils in this prompt. In **Prompt-3**, neutrophil confidence decreased to 44.4%, but the model still favored neutrophils and monocytes. Basophil probability increased slightly to 9.6%, although neutrophils and monocytes remained the dominant classes.

The **CLIP model** exhibited a more balanced response between eosinophils and basophils, although it demonstrated moderate variability across prompts. In **Prompt-1**, eosinophil probability was the highest (41.7%), followed by basophils at 20.2%, while monocytes, lymphocytes, and neutrophils had lower probabilities. This indicates that the model was more confident in identifying eosinophils in this prompt. In **Prompt-2**, eosinophil confidence increased further to 45.6%, with basophil probability rising to 31.7%. Probabilities for other classes decreased, reflecting a stronger focus on eosinophils and basophils. In **Prompt-3**, a more balanced distribution was observed, with basophils (33.6%) and eosinophils (21.8%) sharing prominence. Monocyte and lymphocyte probabilities also increased slightly, while eosinophil confidence decreased compared to the earlier prompts.

TABLE I. CLASS PROBABILITIES ACROSS DIFFERENT PROMPTS FOR PLIP, BIOMEDCLIP, AND CLIP MODELS FROM WBC DATASET.

Class	Prompt-1	Prompt-2	Prompt-3
PLIP			
Basophil Probability	25.81%	19.45%	60.46%
Eosinophil Probability	13.55%	10.69%	29.20%
Lymphocyte Probability	18.76%	23.45%	95.00%
Monocyte Probability	16.13%	17.33%	13.41%
Neutrophil Probability	25.75%	29.07%	13.72%
BiomedCLIP			
Basophil Probability	45.50%	0.00%	95.90%
Eosinophil Probability	22.30%	4.00%	54.70%
Monocyte Probability	31.14%	23.55%	31.71%
Lymphocyte Probability	83.50%	40.10%	88.10%
Neutrophil Probability	33.67%	72.03%	44.41%
CLIP			
Basophil Probability	20.18%	31.74%	33.64%
Eosinophil Probability	41.67%	45.64%	21.81%
Monocyte Probability	88.30%	73.70%	12.26%
Lymphocyte Probability	11.56%	33.80%	13.20%
Neutrophil Probability	17.76%	11.87%	19.09%

B. Results for Malignant Blood Cells without fine-tuning

The table compares the predictions of the PLIP, BiomedCLIP, and CLIP models for malignant blood cell classification across three prompts, focusing on three categories: Early, Pre, and Pro-malignant stages.

For **PLIP**, in **Prompt-1**, the model assigns the highest probability to the ”Pre” stage (52.3%), with moderate confidence in the ”Early” stage (23.98%) and lower confidence for ”Pro” (14.78%). In **Prompt-2**, there is a noticeable shift, with the ”Pro” stage now having the highest probability (41.3%), followed by ”Pre” (29.8%), while the confidence in ”Early” remains lower (21.0%). By **Prompt-3**, the model shows an increase in the probability for the ”Early” stage (33.3%) compared to previous prompts, while ”Pre” remains relatively high (34.6%), and ”Pro” declines to 23.8%.

For **BiomedCLIP**, in **Prompt-1**, the model demonstrates strong confidence in identifying the ”Pro” stage (64.4%), with significantly lower probabilities for ”Early” (17.6%) and ”Pre” (17.5%). In **Prompt-2**, there is a shift toward higher confidence in the ”Early” stage (42.5%), while ”Pro” confidence decreases (32.9%) and ”Pre” remains moderate (29.6%). In **Prompt-3**, the ”Pre” stage receives the highest probability (40.3%), followed by ”Early” (37.6%) and ”Pro” (27.6%), suggesting a more balanced distribution of confidence across the malignant stages.

For CLIP, in **Prompt-1**, the model exhibits very high confidence in the ”Pre” stage (86.7%), while the ”Early” and ”Pro” stages are assigned much lower probabilities (12.1% and 11.8%, respectively). In **Prompt-2**, confidence in the ”Pre” stage decreases to 34.5%, while confidence in the ”Early” stage rises significantly to 36.5%, and confidence in the ”Pro” stage also increases to 26.5%. By **Prompt-3**, the ”Early” stage takes the highest probability (50.6%), followed by ”Pre” (40.6%), while the confidence in the ”Pro” stage drops significantly to 6.96%.

TABLE II. CLASS PROBABILITIES ACROSS DIFFERENT PROMPTS FOR PLIP, BIOMEDCLIP, AND CLIP MODELS FOR BLOOD CELL MALIGNANT DATASET.

Class	Prompt-1	Prompt-2	Prompt-3
PLIP			
Early	23.98%	21.05%	33.28%
Pre	52.34%	29.78%	34.58%
Pro	14.78%	41.27%	23.81%
BiomedCLIP			
Early	17.62%	42.54%	37.63%
Pre	17.54%	29.60%	40.27%
Pro	64.44%	32.95%	27.55%
CLIP			
Early	12.07%	36.46%	50.61%
Pre	86.74%	34.53%	40.56%
Pro	11.80%	26.50%	69.60%

V. DISCUSSION

This analysis highlights the varying performance and behavior of the PLIP, BiomedCLIP, and CLIP models across different malignant blood cell stages and white blood cell (WBC) classification tasks. Across both datasets, these models demonstrated different strengths and biases depending on the prompts and the specific classes involved.

PLIP showed significant variability across prompts in both datasets, indicating that the model is highly sensitive to the input prompts. In the malignant blood cell dataset, PLIP shifted its focus between the ”Pre” and ”Pro” stages, with notable fluctuations in confidence for the ”Early” stage. In the WBC dataset, the model was initially uncertain but gradually developed a strong focus on basophils in later prompts. This variability suggests that PLIP can adapt to the prompts but lacks stability in its class predictions across tasks.

BiomedCLIP exhibited more stability across prompts but demonstrated a strong bias toward certain classes, particularly neutrophils and monocytes in the WBC dataset and the ”Pro” stage in the malignant dataset. While this consistency can be beneficial in scenarios where these classes are dominant, it limits the model’s flexibility and may lead to overconfidence in certain predictions. This model performed well when focusing on specific classes but struggled to balance predictions across all categories.

CLIP demonstrated more balanced responses, particularly in the malignant dataset, where it shifted from high confidence in the ”Pre” stage in earlier prompts to a more even distribution between ”Early” and ”Pre” stages in later prompts. In the WBC dataset, CLIP showed strong confidence in eosinophil classification across prompts. CLIP’s relatively balanced performance makes it more adaptable to different classes, but its variability across prompts may lead to reduced confidence in distinguishing between similar classes. Recent study [14] also support Inconsistency between pre-training and application, biasing in prediction and still a challenges like domain-specific CLIP models are tailored specifically for Chest X-rays within medical imaging, leaving other prevalent image types like mammography, knee MRI, and histology without adequate research. This limitation is primarily attributed to the scarcity of publicly available medical datasets.

VI. CONCLUSION AND LIMITATION

This study highlights the strengths of vision-language models in zero-shot learning, particularly within the realm of medical image classification. Models like PLIP and BiomedCLIP demonstrated remarkable potential in distinguishing between normal and cancerous blood cells using textual prompts, all without requiring additional training on specific domain-related data. Their ability to perform accurately in such specialized tasks emphasizes the utility of vision-language models in the medical field, showing that with the right design, these models can effectively handle complex classifications.

However, there are some limitations to this approach. While CLIP has proven to be powerful in general contexts, its performance lagged behind PLIP and BiomedCLIP, models that have been specifically trained for medical purposes. This indicates that vision-language models may require specialized training data to achieve optimal performance in domain-specific tasks like hematology. Additionally, the accuracy of the classification results was found to be highly dependent on the design of the textual prompts. This suggests that further refinement in prompt engineering is needed to enhance the models’ performance in accurately classifying medical images.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my professors for their unwavering support and guidance throughout this study. I am also thankful to the Bowie State University Graduate Resource Center, particularly Mr. Christopher Beck (Mr. B), for his valuable assistance in refining the manuscript.

REFERENCES

[1] I. Hartsock and G. Rasool, "Vision-language models for medical report generation and visual question answering: A review," *arXiv preprint arXiv:2403.02469*, 2024.

[2] A. K. Tanwani, J. Barral, and D. Freedman, "Repsnet: Combining vision with language for automated medical reports," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 714–724.

[3] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, and Y. Zhang, "Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge," *Cureus*, vol. 15, no. 6, 2023.

[4] Z. Yang, D. Wang, F. Zhou, D. Song, Y. Zhang, J. Jiang, K. Kong, X. Liu, Y. Qiao, R. T. Chang *et al.*, "Understanding natural language: Potential application of large language models to ophthalmology," *Asia-Pacific Journal of Ophthalmology*, p. 100085, 2024.

[5] X. Meng, X. Yan, K. Zhang, D. Liu, X. Cui, Y. Yang, M. Zhang, C. Cao, J. Wang, X. Wang *et al.*, "The application of large language models in medicine: A scoping review," *Iscience*, vol. 27, no. 5, 2024.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[7] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[8] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri *et al.*, "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," *arXiv preprint arXiv:2303.00915*, 2023.

[9] 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 in brief*, vol. 30, 2020.

[10] M. Ghaderzadeh, M. Aria, A. Hosseini, F. Asadi, D. Bashash, and H. Abolghasemi, "A fast and efficient cnn model for b-all diagnosis and its subtypes classification using peripheral blood smear images," *International Journal of Intelligent Systems*, vol. 37, no. 8, pp. 5113–5133, 2022.

[11] Y. Zhang, A. Unell, X. Wang, D. Ghosh, Y. Su, L. Schmidt, and S. Yeung-Levy, "Why are visually-grounded language models bad at image classification?" *arXiv preprint arXiv:2405.18415*, 2024.

[12] L. Li, H. Guan, J. Qiu, and M. Spratling, "One prompt word is enough to boost adversarial robustness for pre-trained vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 408–24 419.

[13] J. Liu, Y. Zhang, J.-N. Chen, J. Xiao, Y. Lu, B. A Landman, Y. Yuan, A. Yuille, Y. Tang, and Z. Zhou, "Clip-driven universal model for organ segmentation and tumor detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 152–21 164.

[14] Z. Zhao, Y. Liu, H. Wu, Y. Li, S. Wang, L. Teng, D. Liu, X. Li, Z. Cui, Q. Wang *et al.*, "Clip in medical imaging: A comprehensive survey," *arXiv preprint arXiv:2312.07353*, 2023.

[15] M. A. Parab and N. D. Mehendale, "Red blood cell classification using image processing and cnn," *SN Computer Science*, vol. 2, no. 2, p. 70, 2021.

[16] M. Xu, D. P. Papageorgiou, S. Z. Abidi, M. Dao, H. Zhao, and G. E. Karniadakis, "A deep convolutional neural network for classification of red blood cells in sickle cell anemia," *PLoS computational biology*, vol. 13, no. 10, p. e1005746, 2017.

[17] T. Patel, H. El-Sayed, and M. K. Sarker, "Efficientswin: A hybrid model for blood cell classification with saliency maps visualization," in *2024 35th Conference of Open Innovations Association (FRUCT)*. IEEE, 2024, pp. 544–551.

[18] M. Sharma, A. Bhave, and R. R. Janghel, "White blood cell classification using convolutional neural network," in *Soft Computing and Signal Processing: Proceedings of ICSCSP 2018, Volume 1*. Springer, 2019, pp. 135–143.

[19] N. Hussein, F. Shamshad, M. Naseer, and K. Nandakumar, "Promptsmooth: Certifying robustness of medical vision-language models via prompt learning," *arXiv preprint arXiv:2408.16769*, 2024.

[20] M. Kakkar, D. Shanbhag, C. Aladahalli *et al.*, "Language augmentation in clip for improved anatomy detection on multi-modal medical images," *arXiv preprint arXiv:2405.20735*, 2024.

[21] M. Aono, H. Shinoda, T. Asakawa, K. Shimizu, T. Togawa, and T. Komoda, "Multi-stage medical image captioning using classification and clip." in *CLEF (Working Notes)*, 2023, pp. 1387–1395.

[22] S. Baliah, F. A. Maani, S. Sanjeev, and M. H. Khan, "Exploring the transfer learning capabilities of clip in domain generalization for diabetic retinopathy," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2023, pp. 444–453.

[23] Y. Lei, Z. Li, Y. Shen, J. Zhang, and H. Shan, "Clip-lung: Textual knowledge-guided lung nodule malignancy prediction," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 403–412.

[24] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie, "Pmc-clip: Contrastive language-image pre-training using biomedical documents," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 525–536.

[25] L. Zhang, B. Yun, X. Xie, Q. Li, X. Li, and Y. Wang, "Prompting whole slide image based genetic biomarker prediction," *arXiv preprint arXiv:2407.09540*, 2024.

[26] S. Zheng, X. Cui, Y. Sun, J. Li, H. Li, Y. Zhang, P. Chen, X. Jing, Z. Ye, and L. Yang, "Benchmarking pathclip for pathology image analysis," *Journal of Imaging Informatics in Medicine*, pp. 1–17, 2024.

[27] T. Gonçalves, D. Pulido-Arias, J. Willett, K. V. Hoebel, M. Cleveland, S. R. Ahmed, E. Gerstner, J. Kalpathy-Cramer, J. S. Cardoso, C. P. Bridge *et al.*, "Deep learning-based prediction of breast cancer tumor and immune phenotypes from histopathology," *arXiv preprint arXiv:2404.16397*, 2024.

[28] S. Denner, M. Bujotzek, D. Bounias, D. Zimmerer, R. Stock, P. F. Jäger, and K. Maier-Hein, "Visual prompt engineering for medical vision language models in radiology," *arXiv preprint arXiv:2408.15802*, 2024.

[29] H. Wei, B. Liu, M. Zhang, P. Shi, and W. Yuan, "Visionclip: An medaigc based ethical language-image foundation model for generalizable retina image analysis," *arXiv preprint arXiv:2403.10823*, 2024.

[30] S. Mohammed, J. Fiaidhi, and A. S. Martinez, "Using meta-transformers for multimodal clinical decision support and evidence-based medicine," *medRxiv*, pp. 2024–08, 2024.

[31] B. Yang, Y. Yu, Y. Zou, and T. Zhang, "Pclmed: Champion solution for imageclefmedical 2024 caption prediction challenge via medical vision-language foundation models," in *CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS. org, Grenoble, France*, 2024.