Implementation of Transformer-Based Model for Acute Lymphoblastic Leukemia Segmentation

Phumiphat Charoentananuwat
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
phumiphatcn@gmail.com

Suree Pumrin
Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University
Bangkok, Thailand
suree.p@chula.ac.th

***Abstract*—The examination of peripheral blood smear images for acute lymphoblastic cells represents a diagnostic approach for leukemia. The utilization of semantic segmentation of acute lymphoblastic cells can be employed in the development of a computer-aided analysis system. In the realm of peripheral blood smear analysis, deep learning methods, particularly convolutional neural networks, are commonly utilized. Currently, transformer-based models have emerged as the state-of-the-art approach for semantic segmentation tasks. In this study, SegFormer, a transformer-based model for semantic segmentation, was utilized to segment and classify acute lymphoblastic cells using four distinct training strategies. The optimal outcome was achieved with a mean intersection-over-union (IoU) of 0.821 and a mean accuracy of 0.917.**

***Index Terms*—Acute Lymphoblastic Leukemia, Semantic Segmentation, Transformer.**

I. Introduction

It is necessary to diagnose the symptoms to medicate the illness. In this process, physicians must use several information to accurately identify such as blood smear analysis. Blood smear is composed of red blood cells (RBCs), white blood cells (WBCs), and platelets. Acute Lymphoblastic Leukemia (ALL) is a cancer of the blood and bone marrow that affects mainly children and young adults. Approximately 6000 cases of ALL, half of the cases occur in children and teenagers [1]. The manifestation of cancer can result in various symptoms, including but not limited to fatigue, weakness, fever, bone pain, and anemia, which may have an impact on the physical health and development of a child. The accurate and efficient classification of ALL is crucial for proper treatment and prognosis of the disease.

The utilization of computer-assisted classification systems has the potential to enhance the efficiency of the peripheral blood sample analysis procedure, especially in terms of speed. Image analysis is a promising approach for the automatic classification of ALL, but the segmentation of lymphoblasts, the abnormal blood cells observed in ALL patient blood sample, remains a challenging task.

To develop an algorithm for semantic segmentation, various types of algorithms, including conventional image processing, machine learning, and deep learning, have been utilized. In deep learning methods, algorithms typically make use of Convolutional Neural Network (CNN) architecture.

In this article, the proposed method is deep learning semantic segmentation based on transformer architecture, which has recently achieved great success in various computer vision tasks. The outcome of this article is a model for semantic segmentation of lymphoblasts in blood smear image. The model was trained and validated utilizing the ALL-IDB database [2].

The subsequent sections of the paper, namely Section II, III, IV, V, and VI, respectively provide a literature review, related theory and principle, methodology, results, and conclusion. The literature review section provides an overview of how previous articles have approached the task. Related theory and principle elucidate attention mechanism, vision transformer, and evaluation metrics. The methodology section provides an overview of the ALL-IDB dataset, as well as the data augmentation and preprocessing techniques employed. Additionally, the proposed model is described in detail. The findings of the experiment and subsequent analysis are presented in the Results section. The final section of the paper presents the conclusion and potential avenues for future research.

II. Literature Review

Within the realm of literature, various kinds of segmentation techniques have been employed for the purpose of segmenting lymphoblasts within blood smear images. Several techniques utilize image thresholding for the purpose of segmentation. The authors of [3] employ a multilevel thresholding technique utilizing the Otsu and Kapur method for image segmentation after preprocessing. Furthermore, they enhance the algorithm's performance through the utilization of learning enthusiasm-based teaching-learning-based optimization (LebTLBO). The authors in [4] utilize a single-threshold approach on a preprocessed image, a contrast-stretched gray image, and the hue component extracted from an HSV image. The implementation of the golden section search method was utilized to determine the optimal threshold, utilizing the Dice Similarity Coefficient as a cost function.

Another kind of algorithm is deep learning. The authors of [5] have utilized deep semantic segmentation, SegNet, to perform segmentation of red blood cells (RBCs) and white blood cells (WBCs) in images of blood smears. The SegNet architecture is a type of convolutional encoder-decoder model. Its encoder component is responsible for extracting feature maps, while the decoder component performs upsampling of the feature maps to generate a predicted segmentation mask. The network utilized VGG-16 as its backbone. For [6], the authors applied CNN to extract hierarchical features from

979-8-3503-0219-6/23/$31.00 ©2023 IEEE

each layer of the proposed model. The feature was used to generate object masks, and contour masks. The masks were combined to a create segmentation mask as an output of this article.

Vision Transformer (ViT) [8] is a deep learning architecture designed for image recognition tasks. Unlike traditional Convolutional Neural Networks (CNNs), ViT employs a transformer encoder for feature extraction. The utilization of patch embedding, self-attention, and skip connection in ViT enables it to effectively preserve spatial information and provide enhanced representation of global features compared to CNN [9]. The ability of ViT to excel in segmentation tasks can be attributed to understanding the relationship of global context. This is particularly important in establishing contextual dependencies between various sections of an image.

III. Methodology

In this study, the implementation procedure has been divided into four steps: data acquisition, data augmentation, and model development. Overall processes are visualized as Fig. 1.

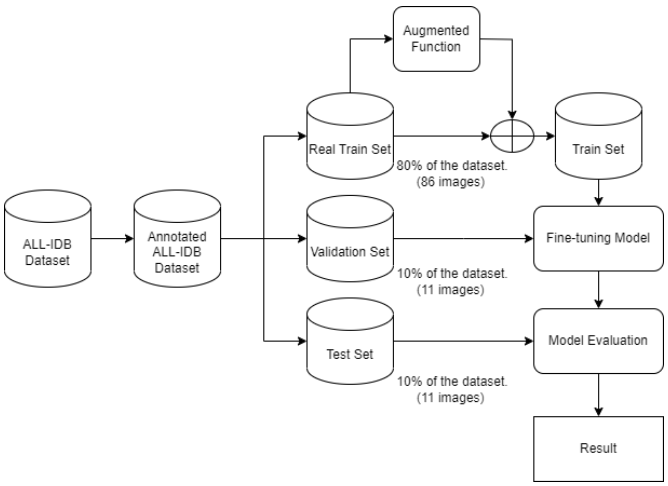


Fig. 1: The overall process of our proposed method.

A. Data Acquisition

For the development of a model, we utilize the public dataset ALL-IDB. The dataset contains 108 blood smear images obtained from healthy individuals and ALL patients, with a resolution of 2592 x 1944 pixels. Each lymphoblast was annotated with the coordinates of its center. To prepare the dataset for semantic segmentation, the images were annotated with CVAT and exported in COCO format [10].

Using original images and a COCO format file, a ground truth mask was created. The dataset was divided into a training set, a validation set, and a test set using a ratio of 80:10:10 percents, respectively.

The sample images and masks from healthy person and ALL patient are shown in Fig. 2.

B. Data Augmentation

80 percents of ALL-IDB consists of only 86 images. To enhance performance and generalize the model, it must learn additional data patterns. The purpose of data augmentation is to enhance the quantity and diversity of training data. In our

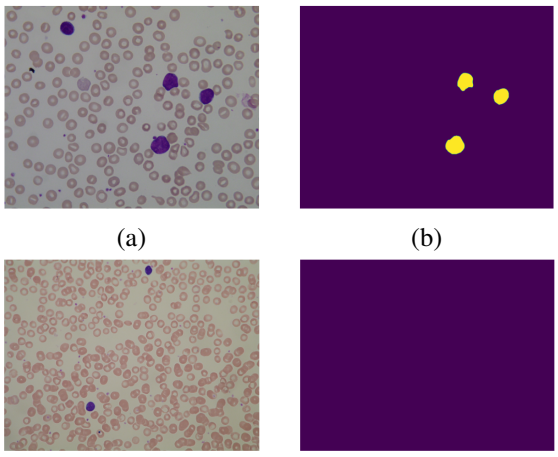


Fig. 2: (a) The original image, and (b) segmentation mask from ALL patients, (c) the original image, and (d) segmentation mask from healthy person.

implementation, the augmented function consists of a vertical flip, a random rotate, one of elastic transform, grid distortion, and optical distortion, contrast Limited adaptive histogram equalization (CLAHE), a random brightness contrast, and a random gamma. The details of augmented strategies are shown in Table. I and the sample result is shown in Fig. 3.

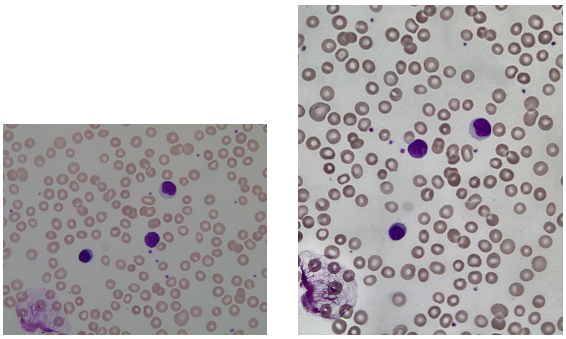
TABLE I: A list of transform functions utilized by this augmentation strategy, where p represents the probability of applying each transform function to an image.

Strategy	Implementation detail
Vertical flip (p=0.5)	Flip the input vertically around the x-axis
Random rotate 90 (p=0.5)	Randomly rotate the input by 90 degrees zero or more times.
Elastic transform (p=0.5)	Randomly apply elastic distortions to input images
Grid distortion (p=0.5)	Randomly apply grid distortion to input images
Optical distortion (p=0.5)	Randomly apply optical distortion to input images
CLAHE (p=0.8)	Enhance the contrast of input images with noise over-amplified prevention
Random brightness contrast (p=0.8)	Randomly change brightness and contrast of the input images
Random gamma (p=0.8)	Adjust the brightness of an image by modifying the pixel's intensity values.

C. Model development

The process of semantic segmentation entails assigning a categorical tag to every single pixel present in an image. For several years, semantic segmentation has been widely employed in various applications, including augmented reality, autonomous driving, and medical image analysis. The present study employs SegFormer [7] to perform segmentation of acute lymphoblastic cells in an image of a peripheral blood smear. The architectural design comprises of an encoder and decoder, as depicted in Fig. 4.

The encoder architecture of SegFormer is stacked of a series of transformer blocks and pooling layers, which are



(a) (b)

Fig. 3: (a) The sample original image, and (b) the result from proposed augmentation strategy.

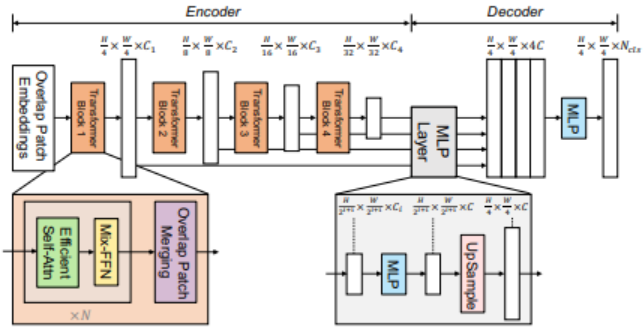


Fig. 4: SegFormer architecture [7].

arranged in four different scales ($\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, $\frac{1}{32}$ relative to the input size) to represent hierarchical feature maps. The transformer blocks comprise an Efficient Self-Attention layer, a Mix-FFN layer, and an Overlapped Patch Merging layer. The utilization of Efficient Self-Attention has been proposed as a solution to address the computational bottleneck issue that arises from the complexity of regular multi-head attention. This approach involves the projection of K and V tensors to reduce their dimensionality with a reduction ratio. The Overlapped Patch Merging technique is intended to extract the contiguous local feature from adjacent patches through the utilization of a stride value, S, which is smaller than the patch size, K, yet larger than the padding size, P. In this implementation, we use K=4, S=3, P=2. Mix-FFN has been employed as a positional encoding layer in the ViT [8]. Its implementation enhances the capacity to extract positional information without the need for interpolation, even when the input resolution differs from the training resolution. The issue at hand is addressed by MixFFN through the incorporation of a 3x3 convolution layer and multilayer perceptron (MLP) into each feed-forward neural network (FFN), as demonstrated in equation (1).

$$x_{out} = MLP(GELU(Conv_{3x3}(MLP(X_{in})))) + x_{in} \quad (1)$$

SegFormer employs a MLP exclusively for the decoder component, with the aim of mitigating computational complexity associated with the production of segmentation maps. The decoder of SegFormer comprises four sequential stages. The initial step involves standardizing the channel dimension of multi-level features using MLP. Subsequently, performing upsampling to achieve identical resolution. Subsequently, the procedure involves merging all feature tensors and feeding them into a MLP to obtain the segmentation map.

IV. Results

In this experiment, we trained 4 models with different augmentation strategy and size from 86 images and 86 pixel labeled images which were splited from ALL-IDB dataset. The first used image size of 128x128 without augmentation strategy. The second used image size of 128x128 with augmentation strategy. The third used image size of 256x256 without augmentation strategy. And the last one used image size of 256x256 with augmentation strategy. After passing the training image through augmentation strategy, number of trainings images is 172 images and 172 pixel labeled images. Then, the 11 images and 11-pixel labels images were used to evaluate the models with mean intersection-over-union (IoU) and mean accuracy. The experiment results are shown in Table II.

TABLE II: A comparative analysis of the evaluation outcomes derived from four distinct models, with a focus on the mean IoU and mean accuracy metrics.

	Mean IoU	Mean accuracy
Size 128x128 without augmentation	0.809	0.897
Size 128x128 with augmentation	0.821	0.917
Size 256x256 without augmentation	0.721	0.809
Size 256x256 with augmentation	0.740	0.832

According to the findings presented in Table II, the model trained using images of size 128x128 and an augmentation strategy exhibited the most superior performance in both mean IoU and mean accuracy. The segmentation map obtained from the optimal model is presented in Fig. 6, as depicted in the sample segmentation map outcome of Fig. 5.

From the Fig. 5, the white blood cells located in the left corner exhibit similarities to acute lymphoblastic leukemia cells. Consequently, the model generated an inaccurate prediction. However, the model is capable of distinguishing platelets that share similar color characteristics with white blood cells from acute lymphoblastic leukemia cells.

V. Conclusion

In this paper, we propose transformer-based semantic segmentation model to segment and classify acute lymphoblastic cells from peripheral blood smear images. Our approach get high accuracy when classify the region of acute lymphoblastic cells. However, there are instances of incorrect predictions with certain white blood cells which have a resemblance to our intended target cells. This issue might be attained through the incorporation of certain image processing methodologies in conjunction with the utilization of this deep learning model. In forthcoming research, our intention is to enhance the model's capacity to segment and classify RBCs, subtype of WBCs, and platelets. This will enable the implementation of the model for diagnosing a wider range of diseases.

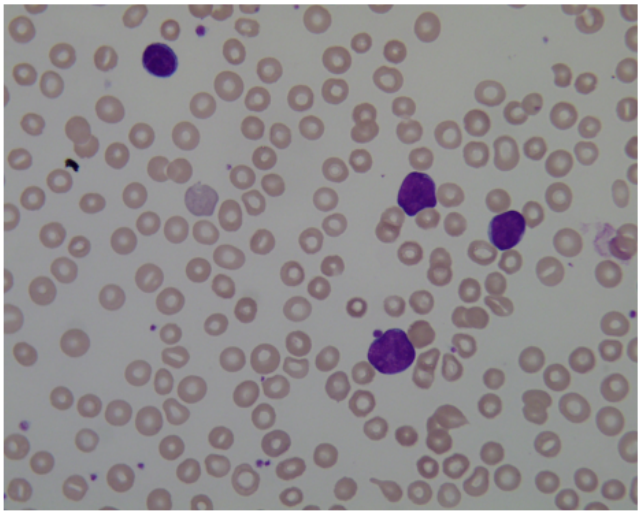


Fig. 5: Sample evaluation image.

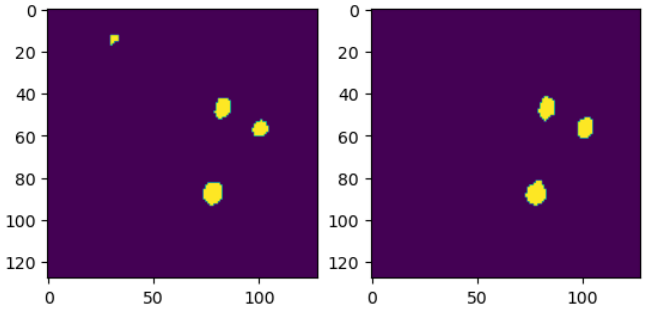


Fig. 6: A comparison of predicted mask (left) and labeled mask (right).

ACKNOWLEDGMENT

The Embedded System and IC Design (ESID) Research Laboratory, situated within the Department of Electrical Engineering at the Faculty of Engineering, Chulalongkorn University, has provided support for this research.

REFERENCES

[1] S. P. Hunger, and C.G. Mullighan, "Acute lymphoblastic leukemia in children," New England Journal of Medicine, vol.373, no.16, 2015, pp.1541-1552.

[2] R. D. Labati, V. Piuri, and F. Scotti, "ALL-IDB: The acute lymphoblastic leukemia image database for image processing," 2011 18th IEEE International Conference on Image Processing, Brussels, Belgium, 2011, pp.2045-2048.

[3] N. M. Deshpande, S. Gite, B. Pradhan, K. Kotecha, and A. Alamri, "Improved Otsu and Kapur approach for white blood cells segmentation based on LenTLBO optimization for the detection of Leukemia," Mathematical Biosciences and Engineering, vol.19, no.2, 2022, pp.1970-2001.

[4] Y. Li, R. Zhu, L. Mi, Y. Cao, and D. Yao, "Segmentation of white blood cell from acute lymphoblastic leukemia images using dual-threshold method,", Computational and Mathematical Methods in Medicine, vol.2016

[5] T. Tran, O. H. Kwon, K. R. Kwon, S. H. Lee, and K. W. Kang, "Blood cell image segmentation using deep learning semantic segmentation," 2018 IEEE International Conference on Electronics and Communication Engineering (ICECE), Xi'an, China, 2018, pp. 13-16

[6] H. Chen, X. Qi, L. Yu, and P. A. Heng, "DCAN: deep contiur-aware networks for accurate gland segmentation," Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2016, pp.2487-2496

[7] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M. and Luo, P., "SegFormer: Simple and efficient design for semantic segmentation with transformers". Advances in Neural Information Processing Systems, vol.34, 2021, pp.12077-12090.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., "An image is worth 16x16 words: Transformers for image recognition at scale". arXiv preprint arXiv:2010.11929, 2020.

[9] Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C., and Dosovitskiy, A., "Do vision transformers see like convolutional neural networks?". arXiv preprint arXiv:2108.08810, 2021.

[10] Lin, T., Maire, M., Belongie, S., Bouredev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollar, P., "Microsoft coco: common objects in context". arXiv preprint arXiv:1405.0312, 2014.