# Early Diagnoses of Acute Lymphoblastic Leukemia Using YOLOv8 and YOLOv11 Deep Learning Models

Alaa Awad, Mohamed Hegazy, Salah A. Aly Faculty of Computing and Data Science, Badya University October 6th City, Giza, Egypt

Abstract-Thousands of individuals succumb annually to leukemia alone. This study explores the application of image processing and deep learning techniques for detecting Acute Lymphoblastic Leukemia (ALL), a severe form of blood cancer responsible for numerous annual fatalities. As artificial intelligence technologies advance, the research investigates the reliability of these methods in real-world scenarios. The study focuses on recent developments in ALL detection, particularly using the latest YOLO series models, to distinguish between malignant and benign white blood cells and to identify different stages of ALL, including early stages. Additionally, the models are capable of detecting hematogones, which are often misclassified as ALL. By utilizing advanced deep learning models like YOLOv8 and YOLOv11, the study achieves high accuracy rates reaching 98.8%, demonstrating the effectiveness of these algorithms across multiple datasets and various real-world situations.

Index Terms—Lymphoblastic Leukemia, YOLOv8 and Yolov11 Deep Learning Models

# I. INTRODUCTION

Today, we live in a technology-driven era where computer science is harnessed to mimic human intelligence, enabling more accurate and faster decision-making. As a result, many researchers have tackled the challenge of leukemia detection through Artificial Intelligence, employing various deep learning methodologies such as MobileNetV2 [1], attention mechanisms [2], and YOLO [3]. Several datasets have been used in different studies, including the ALL image dataset [4] and the CNMC-2019 dataset [5].

Most research to date has relied on single-cell datasets to train AI models. However, in real-world applications, models need to handle multi-cell images and still maintain high accuracy. This paper aims to bridge this gap by training the proposed model on multi-cell samples.

In this study, we used image processing techniques such as segmentation to prepare the dataset. Additionally, we applied transfer learning and fine-tuning techniques on models like YOLOv11 [6] and YOLOv8 [7] achieving results above 98% accuracy.

The contributions of this research are summarized as follows:

- 1) Up to our knowledge, this is the first work to utilize YOLOv11 for ALL blood cancer detection.
- 2) The integration of two datasets to improve generalization across different sample types.

- The classification of white blood cells as malignant or benign is addressed
- 4) Successful detection of hematogones, which are often misclassified as ALL.
- A comparative analysis of our results with previous work in the field.

# II. RELATED WORK

Hosseini et al. [8] aimed to detect B-cell acute lymphoblastic leukemia (B-ALL) and its subtypes using a deep CNN. Talaat et al. [9] applied an attention mechanism to detect and classify leukemia cells.

Yan [10] worked with the single-cell dataset CNMC-2019 [5] to classify normal and cancerous white blood cells using three models: YOLOv4, YOLOv8, and a CNN. Data augmentation was applied to the CNN and YOLOv4 models. The CNN model, featuring convolutional layers, max-pooling layers, and ReLU activation, achieved 93% accuracy, while YOLOv4 and YOLOv8 surpassed 95

Devi et al. [11] combined custom-designed and pretrained CNN architectures to detect ALL in the augmented ALL image dataset [4]. The custom CNN extracted hierarchical features, while VGG-19 extracted high-level features and performed classification, achieving 97.85% accuracy. In contrast, Khosrosereshki [12] used image processing and a Fuzzy Rule-Based inference system for this task.

Rahmani et al. [13] utilized the C-NMC 2019 dataset, applying preprocessing techniques such as grayscaling and masking, followed by feature extraction via transfer learning using VGG19, ResNet50, ResNet101, ResNet152, Efficient-NetB3, DenseNet-121, and DenseNet-201. Feature selection employed Random Forest, Genetic Algorithms, and Binary Ant Colony Optimization. The classification, done through a multilayer perceptron, achieved slightly above 90% accuracy.

Kumar et al. [14] focused on classifying different blood cancers, including ALL and Multiple Myeloma, in white blood cells. After preprocessing and augmentation, feature selection was done using SelectKBest. Their model comprised two blocks with convolutional and max-pooling layers, followed by fully connected and classification layers, achieving 97.2% accuracy.

Saikia et al. [15] introduced VCaps-Net, a fine-tuned VGG16 combined with a capsule network for ALL detection. Using the ALL-IDB1 dataset [16] and a private dataset, VCaps-Net maintained spatial relationships in images through capsule vectors, avoiding the loss often caused by MaxPooling, and achieved 98.64% accuracy.

# III. DATASETS AND DATA COLLECTIONS

Available datasets are divided into two types: single-cell and multi-cell datasets. Single-cell datasets typically contain images with a single white blood cell per image, whereas multi-cell datasets depict multiple cells within each sample. Since multi-cell datasets better represent real-life scenarios when working with blood cells, we chose to focus on them. The two datasets selected for this study are the Acute Lymphoblastic Leukemia (ALL) image dataset from Kaggle [4] and ALL-IDB1 [16], both of which contain multiple white blood cells per sample.

Table I summarizes the statistics of the ALL image dataset, which contains 3,256 images in total, divided into four categories: Benign, Early, Pre, and Pro. The benign class includes Hematogones, a condition where lymphoid cells accumulate in a pattern similar to ALL but are non-cancerous and generally harmless. The dataset consists of 504 benign images and 2,752 malignant cells, further categorized into 985 early-stage samples, 963 pre-stage samples, and 804 pro-phase samples.

TABLE I: Sample distribution per class for ALL image dataset

Class	Samples Per Class
Benign	504
Early	985
Pre	963
Pro	804
Total	3256

On the other hand, Table II is affiliated with ALL-IDB1 dataset includes 108 images in total divided into 59 normal blood samples and 49 cancerous ones. This balance between normal and cancerous samples is crucial for the model to effectively learn the distinguishing features of ALL cells. We

TABLE II: Sample distribution per class for ALL-IDB1 dataset

Class	Samples Per Class
Normal	59
Cancer	49
Total	108

decided to merge the normal cells from ALL-IDB1 with the benign cells from the ALL dataset into a single category called Normal. Similarly, we combined the Early, Pre, and Pro classes from the ALL dataset with the Cancer class from ALL-IDB1 into one category called Cancer. As a result, we focused on two classes: Normal and Cancer. This approach exposes the models to different datasets and various shapes of blast cells, allowing for more practical detection and classification.

# IV. MODELS AND METHODOLOGIES

The implementation is divided into several phases, as illustrated in Fig. 1. The first phase involves data preparation,

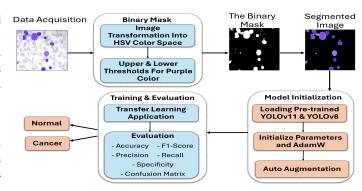


Fig. 1: The implementation process including the data preparation and models training and evaluation

where image segmentation techniques are applied to isolate the relevant elements. Next, the pretrained YOLOv11s and YOLOv8 models are loaded. These models enable data augmentation and experimentation with various optimizers and learning rates. In the final phase, the models are trained to fine-tune the pretrained weights for the specific task at hand.

**Dataset Preparation** The dataset underwent preprocessing to enhance model performance by removing irrelevant elements like different backgrounds and unrelated blood components. Image segmentation was applied using OpenCV, converting images to the HSV color space and creating a binary mask to isolate white blood cells. To improve robustness and mitigate overfitting, various augmentation techniques were implemented, including geometric transformations, mosaic augmentation, random erasing, and randaugment for diverse data variations.

**Image Classification:** The detection of blast cells can be approached in various ways, with image classification being one of the most common. Numerous deep learning architectures have been developed to support this task, with Convolutional Neural Networks (CNNs) being the most widely used for image and video datasets. Models like VGG, AlexNet, and GoogleNet (Inception) [17] are all based on CNNs. In this paper, we focus on two versions of YOLO: YOLOv8 and YOLOv11.

To train and optimize model performance, we employed two key techniques: transfer learning and fine-tuning.

First, transfer learning was applied using a pretrained YOLOv8 model, imported after installing the Ultralytics package, and then training it on our custom segmented dataset. Various tests were conducted using different optimizers and hyperparameters, but the final results were based on 100 epochs of training, using the AdamW optimizer with a learning rate of 0.000714.

Second, YOLOv11s, the latest version in the YOLO series, was also trained on our custom dataset. We experimented with the small version of the model to observe the performance on different versions of YOLO. This gave us the chance to understand the enhancements in the new version of the model.

**Performance Metrics:** Accuracy is an overall indicator on how well the model performs taking into consideration the number of correctly identified samples out of all the given samples. This is represented by the summation of true positives and true negatives divided by the total number of examples consisting of True Positive(TP), True Negative(TN), False Positive (FP) and False Negative (FN) as expressed in Equation 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Presented in Equation 2, the sample precision which is identified by the ratio of correctly classified instances to the total number of classified instances.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, or Sensitivity, is calculated as the ratio of correctly identified instances to the total number of instances, as described in Equation 3.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Another significant metric that contributed to our results is f1 score. It is obtained by calculating the harmonic mean of precision and recall, as illustrated in Equation 4.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$
 (4)

In addition to the previous indicators, we calculated the specificity using the formula in Equation 5.It refers to the proportion of correctly identified negative instances among all actual negative cases. It reflects the model's ability to accurately classify instances from the opposite disease classes.

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

# V. EXPERIMENTAL RESULTS

In this section, the performance of our different trained models is evaluated using the different metrics mentioned in the previous section. Starting with YOLOv11s were trained for 100 epochs, which achieved 97.4% training accuracy and 98.8% testing accuracy, while YOLOv11s has achieved a slightly higher testing accuracy of 98.8%.

The accuracy graph for the nano version of YOLOv11 shown in Fig. 2b demonstrates improvement in the accuracy's progress with some fluctuations at the beginning. These variations decrease gradually as the number of epochs increases until the graph curve becomes more stable. It can be seen that the training and validation losses were declining steadily as the training advanced in Figures. 2a and 6a.

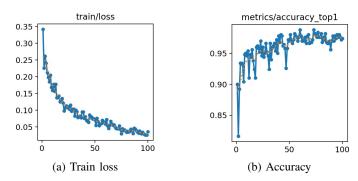


Fig. 2: Performance metrics for YOLOv11s. (a) Train loss, (b) Accuracy, and (c) Validation loss.

The confusion matrix in Fig. 3 offers valuable insights into the YOLOv11s model's performance, highlighting which classes are accurately detected and where errors occur. This analysis helps identify areas for improvement to enhance the model's effectiveness. The matrix indicates that the model achieved high accuracy in detecting cancer across all stages, though it did misclassify 0.06% of healthy white blood cells as cancerous. Overall, the analysis validates the model's strong performance while pinpointing specific issues that require optimization.

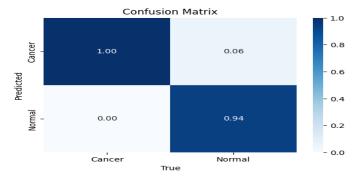
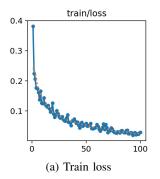
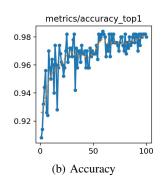


Fig. 3: Normalized confusion matrix for YOLOv11s.

Next, we analyze the results from YOLOv8 shown in figures 4a, 4b and 6b in which the accuracy on the training dataset was 98%, and it peaked at 98.4 when evaluated on the testing dataset. Compared to YOLOv11, YOLOv8's small version achieves a slightly lower accuracy. The accuracy, training and validation loss graphs of YOLOv8 shown below follow similar patterns as those of YOLOv11 with the later having more stable curves.





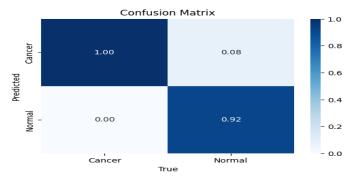
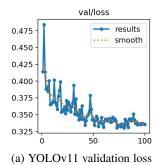


Fig. 5: Confusion matrix for YOLOv8s.



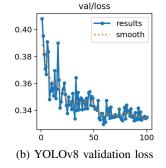


Fig. 6: The validation losses for both (a) YOLOv11 and (b) YOLOv8.

# VI. CONCLUSION AND COMPARISON

In conclusion, the integration of AI in the medical field is a massive step in the advancement of the health system and services provided to patients. This study was able to detect the presence of ALL in blood even at early stages using YOLOv11 and YOLOv8. The performance of YOLOv11 proved to be slightly better than that of YOLOv8 achieving better accuracies which can be significant in cancer diagnosis. Table III demonstrates a comparison between our findings and some of the previous studies.

### REFERENCES

 M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," 2019. [Online]. Available: https://arxiv.org/abs/1801.04381

TABLE III: Comparison of Different Approaches for Detecting Acute Lymphoblastic Leukemia (ALL)

Study	Methodology	Accuracy	Dataset
Yan [10]	YOLOv4,	CNN: 93%,	CNMC-2019 [5]
	YOLOv8, and	YOLOv4:	
	CNN	> 95%,	
		YOLOv8:	
		> 95%	
Devi et al. [11]	Custom + pre-	97.85%	ALL dataset [4]
	trained CNN		
Saikia et al. [15]	VCaps-Net	98.64%	ALL-IDB1 [16]
Kumar et al. [14]	Custom CNN	97.2%	Custom dataset
Our study	YOLOv11s and	YOLOv11s:	ALL-IDB1 [16]
	YOLOv8s	98.8%,	+ ALL dataset
		YOLOv8s:	[4]
		98.4%	

- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762
- [3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [4] 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," *Int. J. Intell. Syst.*, 2021.
- [5] S. Mourya, S. Kant, P. Kumar, A. Gupta, and R. Gupta, "All challenge dataset of isbi 2019 (c-nmc 2019) (version 1)," 2019. [Online]. Available: https://doi.org/10.7937/tcia.2019.dc64i46r
- [6] Ultralytics, "Yolov11 key features," 2024. [Online]. Available: https://docs.ultralytics.com/models/yolo11/#key-features
- [7] R. Varghese and S. M., "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), 2024.
- [8] A. Hosseini et al., "A mobile application based on efficient lightweight cnn model for classification of b-all cancer from non-cancerous cells: A design and implementation study," *Informatics in Medicine Unlocked*, vol. 39, 2023.
- [9] F. M. Talaat and S. A. Gamel, "A2m-leuk: attention-augmented algorithm for blood cancer detection in children," *Neural Computing and Applications*, vol. 35, no. 24, 2023.
- [10] E. Yan, "Detection of acute myeloid leukemia using deep learning models based systems," in *IFMBE Proceedings*, 2024, pp. 421–431.
- [11] J. R. Devi, P. S. Kadiyala, S. Lavu, N. Kasturi, and L. Kosuri, "Enhancing acute lymphoblastic leukemia classification with a rapid and effective cnn model," 2024.
- [12] M. A. Khosrosereshki and M. B. Menhaj, "A fuzzy based classifier for diagnosis of acute lymphoblastic leukemia using blood smear image processing," in 2017 5th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), 2017.
- [13] A. M. Rahmani et al., "A diagnostic model for acute lymphoblastic leukemia using metaheuristics and deep learning methods," 2024. [Online]. Available: https://arxiv.org/abs/2406.18568
- [14] D. Kumar, N. Jain, A. Khurana, S. Mittal, S. C. Satapathy, R. Senkerik, and J. D. Hemanth, "Automatic detection of white blood cancer from bone marrow microscopic images using convolutional neural networks," *IEEE Access*, vol. 8, 2020.
- [15] R. Saikia, A. Sarma, K. M. Singh, and S. S. Devi, "Vcaps-net: Fine-tuned vgg16 with capsule network for acute lymphoblastic leukemia detection on a diverse dataset," in 2024 6th International Conference on Energy, Power and Environment (ICEPE), 2024.
- [16] A. Genovese, V. Piuri, K. N. Plataniotis, and F. Scotti, "Dl4all: Multi-task cross-dataset transfer learning for acute lymphoblastic leukemia detection," *IEEE Access*, vol. 11, pp. 65 222–65 237, 2023.
- [17] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2014. [Online]. Available: https://arxiv.org/abs/1409.4842