



## Performance Comparison of YOLOv8 and DETR in White Blood Cell Detection

Andrian Rakhmatsyah<sup>1\*</sup>, Maman Abdurrohman<sup>2</sup>, Bayu Erfianto<sup>3</sup>, Delita Prihatni<sup>4</sup>

<sup>1,2,3</sup>School of Computing, Universitas Telkom, Bandung, Indonesia

<sup>4</sup>Department of Clinical Pathology, Faculty of Medicine, Universitas Padjajaran, Bandung, Indonesia

<sup>1</sup>kangandrian@telkomuniversity.ac.id, <sup>2</sup>abdurrohman@telkomuniversity.ac.id, <sup>3</sup>erfianto@telkomuniversity.ac.id,

<sup>4</sup>delitapri@yahoo.com

### Abstract

*Automated detection and classification of white blood cells (WBCs) from microscopic images play a vital role in supporting the diagnosis of hematological diseases. Accurate and robust object detection algorithms are essential for handling interclass similarities and imbalanced datasets. This study aims to evaluate and compare the performance of two modern object detection algorithms—Detection Transformer (DeTR) and YOLOv8—in performing multiclass WBC classification using public datasets from various sources with diverse visual characteristics. Five experimental scenarios were designed based on varying class distributions and data augmentation techniques, including horizontal/vertical flipping and random rotation. Both methods were trained and evaluated on the same dataset partitions, and their performances were assessed using the following standard metrics: precision, recall, and F1-score for each WBC class. The results show that YOLOv8 consistently achieved superior and more stable performance across all scenarios, with average F1-scores close to 1.00 even in augmented and imbalanced conditions. In contrast, DeTR performed competitively in balanced scenarios but showed lower consistency, particularly in classes such as Neutrophil and Monocyte. Data augmentation positively affected both models, although the gains were more prominent in YOLOv8. This study highlights the strong potential of YOLOv8 in real-time WBC classification tasks and presents DeTR as a viable yet less-optimized approach for this application. These findings contribute to the advancement of medical image-based object detection and offer valuable insights into the selection of appropriate algorithms for hematological image analysis.*

**Keywords:** data augmentation; DeTR; hematology; medical image analysis; object detection; transformer; white blood cell classification; YOLOv8

*How to Cite:* A. Rakhmatsyah, M. Abdurrohman, B. Erfianto, and D. Prihatni, "Performance Comparison of YOLOv8 and DETR in White Blood Cell Detection", J. RESTI (Rekayasa Sist. Teknol. Inf.), vol. 9, no. 4, pp. 923 - 934, Aug. 2025.  
*Permalink/DOI:* <https://doi.org/10.29207/resti.v9i4.6864>

Received: June 30, 2025

Accepted: August 13, 2025

Available Online: August 23, 2025

*This is an open-access article under the CC BY 4.0 License  
Published by Ikatan Ahli Informatika Indonesia*

### 1. Introduction

Object detection is one of the rapidly developing fields in digital image processing and artificial intelligence [1]. This technology has a wide range of applications, including security surveillance, autonomous vehicles, and medical analysis. In the medical field, particularly in microscopic analysis of white blood cells (WBC), object detection is a crucial aspect in assisting the diagnosis of diseases such as leukemia, infections, and other hematological disorders. With the growing need for rapid and accurate diagnoses, the development of more precise detection methods is becoming increasingly crucial. However, image-based WBC detection still faces several challenges. Variations in the shape, size, and color of WBCs under different health conditions often make it difficult for detection models

to generalize patterns effectively [2], [3]. In addition, limited and non-diverse datasets can lead to *overfitting*, resulting in decreased model performance when applied to new data [4]. Therefore, innovative approaches are necessary to enhance detection accuracy by leveraging data augmentation techniques and more advanced artificial intelligence-based detection models.

Various *Computer Vision*-based solutions have been developed to detect and classify WBCs in microscopic images. Several primary methods have been employed, including color-based segmentation methods that utilize RGB or HSV color channels, as well as thresholding methods to distinguish objects from the background [5]. However, these methods are less effective at handling overlapping cells. The following methods utilize morphology-based segmentation and the Watershed

algorithm [6]. These methods employ morphological operations and watershed segmentation to separate WBCs from the background image. While providing better results than thresholding, these methods still face limitations when dealing with the variations in size and shape of WBCs. CNN-based methods such as VGGNet, ResNet, and Inception have been applied in WBC detection and classification [7] - [10]. CNNs are capable of extracting complex features but require a large amount of training data to avoid overfitting. YOLO-based and Faster R-CNN detection, such as with YOLOv3 and Faster R-CNN, have been used in WBC detection due to their real-time detection capabilities and high precision [11]. Various YOLO variants have shown promising results in WBC classification.

For example, YOLOv2 has been applied in combination with the *Bag-of-Features* method to classify WBC types from microscopic images [12]. At the same time, YOLOv4 has been utilized for multi-class WBC classification, demonstrating high accuracy and efficiency in detection time [13]. In recent developments, YOLOv8 has attracted widespread attention due to its ability to efficiently and accurately detect small objects. Its implementation has been tested in various domains, ranging from medical images such as lung disease classification [14], and cancer diagnosis from medical images [15], to melanoma segmentation and dental radiography [16]. In addition, the development of detection algorithms based on YOLOv8 has been carried out in variants such as DC-YOLOv8 for small object detection using camera sensors [17] and an improved YOLOv8 for comprehensive detection of blood cells from microscopic images [18]. In [19], CST-YOLO is introduced, a combined approach that combines YOLOv7 and CNN-Swin Transformer for high-accuracy blood cell detection. This approach emphasizes the importance of integrating spatial features from the Transformer in detecting complex microscopic objects.

Nevertheless, the explicit application of YOLOv8 in the microscopic analysis of WBCs, especially for multi-classification tasks based on diverse public datasets, remains very limited. This indicates a relevant research opportunity to evaluate the effectiveness and robustness of this model in the context of hematology. However, this model has limitations in handling small objects or overlapping cells [20]. The use of Attention Mechanism and Transformer has begun to be applied in medical image analysis, such as with the Vision Transformer (ViT)-based approach [21]. However, the use of a Detection Transformer (DeTR) for WBC detection is still very limited in the scientific literature. Some studies have indeed developed DETR variants such as Deformable DETR [22] or improved DETR for WBC detection [1]. Nevertheless, most of these studies are more focused on architectural modifications, leaving the performance exploration of the original DETR [22] in image-based WBC detection and classification tasks

as a largely unexplored area. A study by [23], proposed an improved DETR-based WBC detection network and showed promising results; however, this approach has not yet thoroughly examined the original DETR architecture [24] on diverse public datasets. Although several of the methods mentioned above have been implemented, challenges remain in improving precision, recall, and f1-score, especially in multi-class WBC classification using publicly available datasets on the Internet.

This study focuses on evaluating the performance of DeTR and YOLOv8 in the multi-classification task of WBCs (Basophils, Eosinophils, Lymphocytes, Monocytes, and Neutrophils) using publicly available datasets found on the Internet. The selection of these two methods was based on the consideration that YOLOv8 is a CNN-based object detection method that was developed since YOLO was first introduced, and was officially released in early 2023, right when this research began. Specifically, this research aims to assess the extent to which DeTR and YOLOv8 can recognize and classify various types of WBCs from public datasets. It analyzes the impact of dataset augmentation on the performance of both models in terms of improving precision, recall, and F1-scores. It compares the effectiveness of DeTR and YOLOv8 in multi-classification to determine the most optimal method for image-based WBC detection.

This approach involves several key steps, including collecting public datasets from various sources that contain WBC images with multi-class labels. Then, dataset augmentation is conducted using rotation and flipping techniques to increase the variety of training data.

The next step involves training the DeTR and YOLOv8 models using the augmented dataset. Finally, the performance of the models is evaluated using precision, recall, and f1-score metrics, along with a comparative analysis between the two models. The selection of DeTR and YOLOv8 is based on several key advantages, including DeTR's capability to understand the spatial relationships between objects. DeTR utilizes a transformer architecture that can recognize spatial relationships between pixels in an image, making it well-suited for detecting overlapping cells. Furthermore, DeTR does not rely on anchor boxes. Instead, DeTR uses a set-based loss approach, which is more flexible compared to traditional detection models that depend on anchor boxes [22]. YOLOv8's efficiency in real-time detection offers improvements in speed and accuracy, making it an ideal choice for medical applications that require rapid detection [15], [17]. YOLOv8 can work with small datasets through transfer learning, which allows model accuracy to be improved with limited training data [15]. The contributions of this research to object detection with WBC multi-classification include exploring DeTR and YOLOv8 for the WBC multi-classification task, which has not been extensively addressed in previous studies; the

application of optimal dataset augmentation to improve model accuracy in detecting WBCs from public datasets [25] - [27]; and comprehensive analysis of precision, recall, and f1-score to evaluate the effectiveness of the proposed combination of methods. Thus, this study is expected to make a significant contribution to optimizing the selection of object detection methods and the development of object detection technology in medical analysis while also recommending future exploration of dataset variability. Specifically, the aim is to enhance the accuracy of WBC detection and classification to support faster and more accurate diagnosis of hematological diseases.

## 2. Methods

This section outlines the research design, the source of the dataset, and the stages of data acquisition and preparation before its use in the model training and evaluation process. The objective of this study is to evaluate the performance of two object detection algorithms, specifically Detection Transformer (DeTR) and YOLOv8, in classifying white blood cells (WBCs) from microscopic images. Figure 1 presents an overview of the research's general design.

### 2.1 Data Collection and Preprocessing

To achieve the research objectives, a dataset comprising microscopic images was sourced from several public and open repositories, specifically from [28], Kaggle [29], BCCD [30], and Raabin [31], each containing label annotations for various types of WBC from healthy individuals (normal).

The datasets obtained from these sources were randomly selected and subsequently combined into a single collection, which was then divided into training, evaluation, and testing datasets. The Roboflow tool facilitated the annotation process, including label conversion to bounding box format, image augmentation, and the division of data into training, validation, and testing datasets. Figure 2 illustrates some examples of the datasets utilized. The image sizes are at 640 x 640 pixels. This dataset was further processed and adjusted to align with the input formats required by the methods employed, namely DeTR and YOLOv8. The COCO format was utilized for DeTR, while the YOLO format was employed for detection using the YOLOv8 method.

In the scenarios of Model 1 and Model 2, the annotated dataset is employed for each respective method, as illustrated in Figure 1. In this context, the term "model" refers to the experimental scenario being conducted. Model 2 utilizes the dataset from Model 1, supplemented with additional random samples from the dataset sources while ensuring a balanced distribution of datasets across classes. Conversely, Models 3, 4, and 5 employ the annotated dataset from Model 2, with further data augmentation applied to the training dataset. Data augmentation techniques such as flipping and rotation were applied exclusively to the training

dataset to enhance model generalization and mitigate overfitting. Validation and test datasets remained unaltered to ensure that performance metrics accurately reflect real-world model performance on unseen, original data distributions. Table 1 provides a detailed description of the dataset scenarios utilized, which correspond to Figure 1. The distribution of the dataset is presented in Table 2.

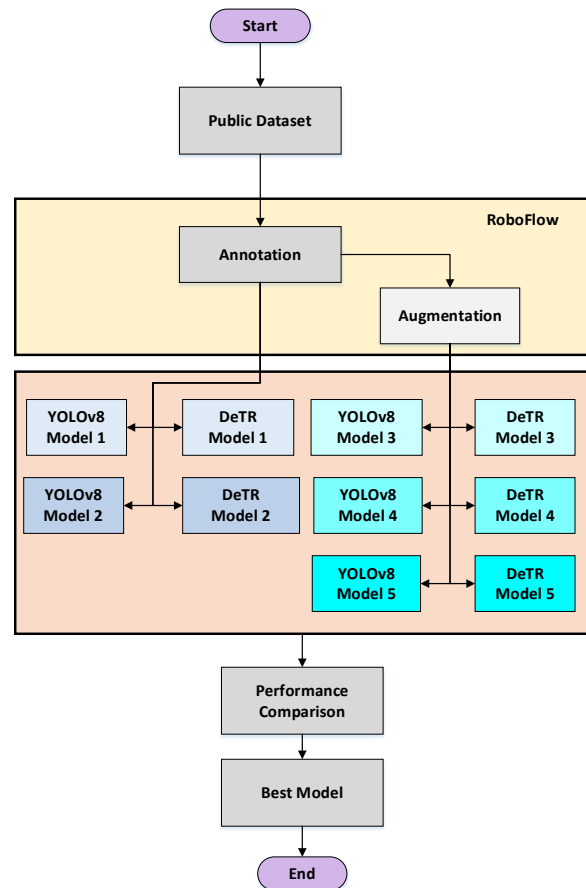


Figure 1. Research Desain

For Model 2, the dataset is allocated with 80% for training, 10% for validation, and 10% for testing. In the case of Model 3, the training dataset constitutes 90%, while both validation and testing datasets comprise 5% each. For Models 4 and 5, the training dataset is 92%, while the validation and testing datasets each represent 4%.

Table 1. Experimental Dataset Model

| Scene   | Dataset                                                                          |
|---------|----------------------------------------------------------------------------------|
| Model 1 | Randomly distributed dataset for each class without augmentation                 |
| Model 2 | Balanced dataset with more samples than Model 1, no augmentation                 |
| Model 3 | Model 2 dataset with randomly horizontal/vertical flipping augmentation          |
| Model 4 | Model 2 dataset with random rotation (-15° and +15°) augmentation                |
| Model 5 | Model 2 dataset with combined flipping and rotation (-15° and +15°) augmentation |

Augmentation was carried out randomly using RoboFlow. The dataset distribution is presented in Table 2. The *Trn* column is for training data, the *Val*

column is for validation data, and the *Tst* column is for testing data. Table 2 shows that the augmentation data contains zeros in Models 1 and 2, both in the training, validation, and test datasets. Zeros indicate that neither Model 1 nor Model 2 includes any augmentation data.

Similarly, Models 3, 4, and 5 contain zeros in both the validation and test datasets. Meanwhile, the training dataset contains a number of randomly generated data points from the augmentation technique.

Table 2. Dataset

| WBC Type     | Model 1 |     |     | Model 2 |     |     | Model 3 |     |     | Model 4 |     |     | Model 5 |     |     |
|--------------|---------|-----|-----|---------|-----|-----|---------|-----|-----|---------|-----|-----|---------|-----|-----|
|              | Trn     | Val | Tst | Trn     | Val | Tst | Trn     | Val | Tst | Trn     | Val | Tst | Trn     | Val | Tst |
| Basophil     | 238     | 42  | 29  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  |
| Eosinophil   | 343     | 58  | 37  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  |
| Lymphocyte   | 303     | 56  | 34  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  |
| Monocyte     | 287     | 51  | 34  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  |
| Neutrophil   | 535     | 88  | 50  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  | 505     | 63  | 63  |
| Augmentation | 0       | 0   | 0   | 0       | 0   | 0   | 3321    | 0   | 0   | 4966    | 0   | 0   | 5026    | 0   | 0   |
| Total        | 1706    | 295 | 184 | 2523    | 315 | 315 | 5846    | 315 | 315 | 7491    | 315 | 315 | 7551    | 315 | 315 |

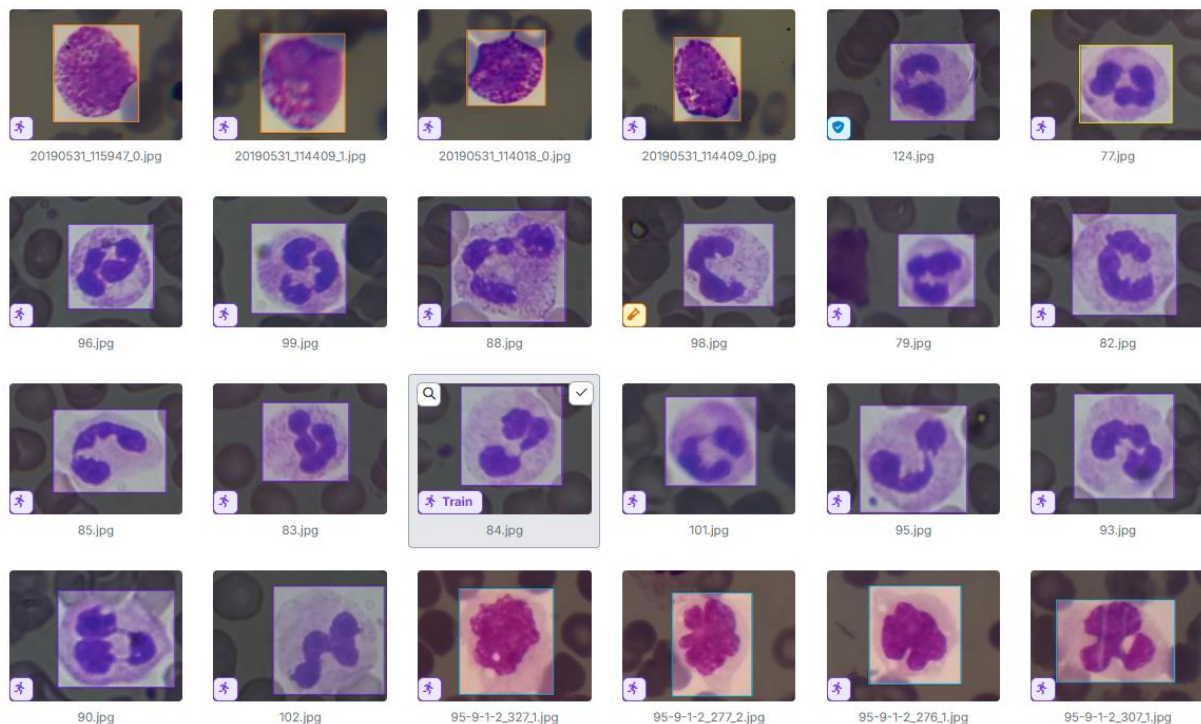


Figure 2. Training dataset samples resulting from annotation and labeling with Roboflow

## 2.2 Experimentation and Evaluation

In this study, YOLOv8 and DeTR were implemented using the PyTorch 2.2 framework, Ultralytics YOLOv8n, and Transformers 4.35.0, with Python 3.10 as the programming language. The workstation utilized an NVIDIA RTX 4070 124B graphics processing unit (GPU), a Windows 11 operating system, and an Intel Core i7-14700K processor paired with 128GB of DDR5 memory.

The training process was carried out for 100 epochs for each method. For YOLOv8, a learning rate of 0.0001 was used, weight decay = 0.0001, batch size = 4, and backbone learning rate = 0.00001. In contrast, the DeTR experiments utilized a batch size of 16, the AdamW optimizer, a learning rate of 0.0001, a weight decay of 0.0001, and a backbone learning rate of 0.00001. Implementation was conducted using the GPU to accelerate the process. After training, both methods were evaluated to ensure that the models could

accurately detect and classify white blood cells (WBC). To verify the reliability of the methods, testing was conducted using a new dataset distinct from the datasets employed in the previous training and evaluation stages.

## 3. Results and Discussions

The performance metrics used in this study are *precision*, *recall*, and *F1-Score* [33], [34]. Precision is a metric used to measure the proportion of optimistic predictions that are correct compared to the total number of positive predictions made by the model. In other words, precision indicates how accurately the model identifies positive classes without misclassifying negative objects as positive. *Recall* measures the proportion of positive entities successfully identified by the model out of all actual positive entities. This metric measures the model's ability to detect all positive entities without missing those that should be detected.

Meanwhile, the F1-Score is the harmonic mean of precision and recall, providing a balance between these two metrics. The F1-Score is used to take both precision and recall into account simultaneously, especially in cases of class imbalance. The formulas used are as shown in Equations 1 - 3.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (1)$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \quad (2)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Increasing precision often reduces recall and vice versa. The F1-Score is used to evaluate overall model performance when both precision and recall are equally important. In the detection of abnormal WBC types, high recall is important to ensure that no abnormal cells are missed. In contrast, high precision ensures that the cells identified as abnormal are truly abnormal. The F1-Score provides an overall picture of the model's performance.

### 3. Results and Discussions

This study conducts a comparative analysis of the performance of YOLOv8 and *Detection Transformer*

(DETR) in the recognition and classification of five types of white blood cells (WBCs) using publicly available datasets obtained from various sources. An augmentation approach was employed to evaluate its impact on the performance of both methodologies.

#### 3.1 YOLOv8 Performance

The experimental procedure commenced with the application of the YOLOv8 method to train the dataset, followed by an evaluation using the validation dataset, and subsequently, testing was conducted with the test dataset, as illustrated in Table 2. The various model scenarios are delineated in Table 1. The outcomes of the tests across all scenarios are presented in Table 3. In this context, 'Prec' is precision, 'Rec' is recall, and 'F1' refers to the F1-score.

Table 3 and Figure 3 shows the performance of YOLOv8 across five dataset scenarios for each WBC type. In Model 1, where the data distribution is unbalanced, and no augmentation techniques were used to increase the training data, both the Eosinophil and Lymphocyte classes had an F1-score of less than 0.95. The F1-scores for the Monocyte and Neutrophil classes were the lowest of all, at 0.87.

Table 2. YOLOv8 Performance

| WBC Type   | Model 1 |      |      | Model 2 |      |      | Model 3 |      |      | Model 4 |      |      | Model 5 |      |      |
|------------|---------|------|------|---------|------|------|---------|------|------|---------|------|------|---------|------|------|
|            | Prec    | Rec  | F1   | Prec    | Rec  | F1   | Prec    | Rec  | F1   | Prec    | Rec  | F1   | Prec    | Rec  | F1   |
| Basophil   | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 |
| Eosinophil | 0.88    | 0.97 | 0.92 | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 | 1.00    | 0.97 | 0.98 | 1.00    | 1.00 | 1.00 |
| Lymphocyte | 1.00    | 0.88 | 0.94 | 1.00    | 1.00 | 1.00 | 1.00    | 0.97 | 0.98 | 0.97    | 0.98 | 0.98 | 1.00    | 0.98 | 0.99 |
| Monocyte   | 0.79    | 0.97 | 0.87 | 1.00    | 1.00 | 1.00 | 0.97    | 1.00 | 0.98 | 0.95    | 0.97 | 0.96 | 0.98    | 1.00 | 0.99 |
| Neutrophil | 0.95    | 0.80 | 0.87 | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 | 1.00    | 1.00 | 1.00 |



Figure 3. YOLOv8 Performance Metrics per WBC Type across Models

In Model 2, with a scenario of balanced dataset size and without augmentation, all metrics—Precision, Recall, and F1-score—reached a perfect value of 1.00. This shows that balancing data distribution and increasing the amount of training data immediately yielded maximum impact. In Model 3, with the same dataset size as Model 2 and the addition of basic augmentation techniques such as horizontal and vertical flipping, almost all F1-scores remained at 1.00, with only Lymphocyte and Monocyte dropping slightly to 0.98. This indicates that flipping does not significantly reduce performance and helps improve robustness.

In Model 4, with the same dataset size as Model 2 and the addition of rotation augmentation techniques  $\pm 15^\circ$ , the F1-score for the Monocyte class dropped slightly to less than 0.98. At the same time, Lymphocytes and Eosinophils had F1-scores of 0.98. The slight decrease may occur because rotation changes the nucleus orientation too much due to the effect of complex augmentation. In Model 5, with the same dataset size as Model 2 but employing a combination of flipping and rotation, along with a  $\pm 15^\circ$  augmentation technique, all performance metrics returned to near-perfect values, specifically greater than 0.98. This shows that combining augmentation techniques tends to rebalance and enhance robustness, particularly for Monocyte. Of the five existing classes, Basophil consistently showed perfect results in all models, indicating that the class is easily identifiable. For Eosinophil, the F1-Score increased from 0.92 to 1.00, although it dropped slightly in Model 4 due to rotation. The Lymphocyte class was most sensitive to rotational augmentation, as indicated by its F1-score dropping to 0.98 in Models 3 and 4, although it rose again in Model 5 to 0.99. The same

trend occurred with Monocyte, whose F1-score started at 0.87 and gradually increased to a maximum of 0.99 in Model 5. The Neutrophil class maintained a consistently high and stable F1-score, indicating that the augmentation techniques had little impact. When viewed in terms of the impact of adding basic flipping and rotation augmentation techniques, Model 3 had almost no adverse effects, as it was able to add variety without disrupting core cell patterns. In Model 4, however, the rotation augmentation technique may slightly affect the shape of the cell nuclei, particularly for Lymphocytes and Monocytes. Meanwhile, the combination of both augmentation techniques rebalanced the results and provided better flexibility. Figure 4 illustrates the trend of YOLOv8's F1-score across various scenarios of data augmentation for each WBC type. Starting from Model 2 onward, the F1-scores remain near-perfect for all classes, indicating highly balanced performance in both precision and recall across all experimental settings.

### 3.2 DETR Performance

Similar to the experiments using YOLOv8, the experiments using DETR were also conducted using the datasets shown in Table 2, following the scenarios presented in Table 1. The performance results from tests using DETR are presented in Table 4.

Table 4 and Figure 5 show DETR's performance in five dataset scenarios for each WBC type. The performance metrics used include precision, recall, and F1-score. Figure 5 illustrates the variation in DETR's classification stability across different augmentation strategies and cell types.

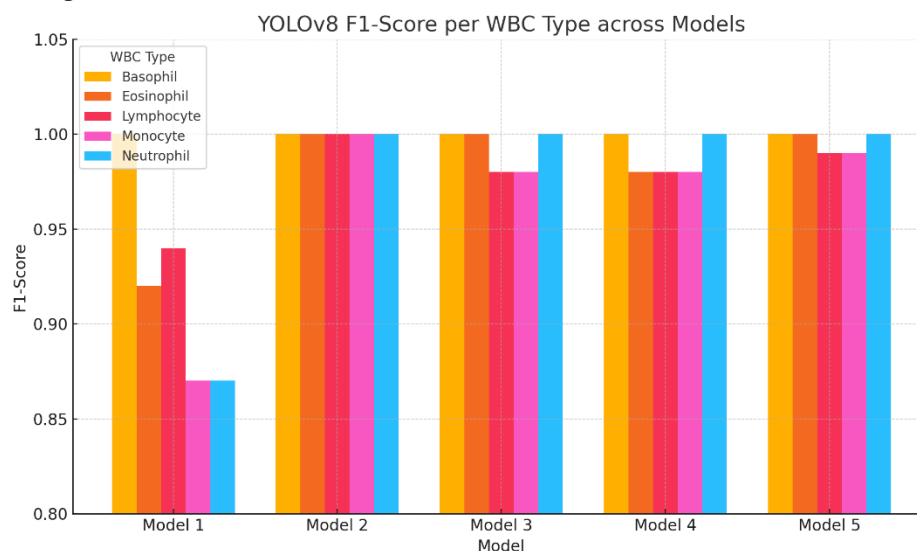


Figure 4. F1-Score per WBC Type for YOLOv8 across Experimental Scenarios

In Model 1, where the dataset quantity distribution is random, and no augmentation techniques are applied, performance is very low across all classes. This is indicated by F1-scores less than 0.6 for all classes. Figure 4 illustrates the F1-score values for each WBC

type. The Neutrophil class has an F1-score of only 0.28, indicating that Model 1 is vulnerable to imbalanced data. This strengthens the point that DETR does not perform well on small and imbalanced datasets.

In Model 2, there is a substantial improvement across all classes. The F1-score values are  $\geq 0.85$ , and there is a significant increase in the Neutrophil class from 0.28 to 0.87. This shows that balanced dataset distribution is

crucial for DETR's performance. In Model 3, where the dataset is balanced but basic augmentation techniques, such as flipping, are added, the performance tends to stagnate or decline slightly.

Table 3. DETR Performance

| WBC Type   | Model 1 |      |      | Model 2 |      |      | Model 3 |      |      | Model 4 |      |      | Model 5 |      |      |
|------------|---------|------|------|---------|------|------|---------|------|------|---------|------|------|---------|------|------|
|            | Prec    | Rec  | F1   | Prec    | Rec  | F1   | Prec    | Rec  | F1   | Prec    | Rec  | F1   | Prec    | Rec  | F1   |
| Basophil   | 0.50    | 0.52 | 0.51 | 1.00    | 0.97 | 0.98 | 0.98    | 1.00 | 0.99 | 1.00    | 0.94 | 0.97 | 1.00    | 1.00 | 1.00 |
| Eosinophil | 0.37    | 0.60 | 0.46 | 0.90    | 0.86 | 0.88 | 0.76    | 0.89 | 0.82 | 0.69    | 0.87 | 0.77 | 0.71    | 0.92 | 0.80 |
| Lymphocyte | 0.65    | 0.50 | 0.57 | 0.97    | 0.97 | 0.97 | 0.94    | 1.00 | 0.97 | 0.98    | 1.00 | 0.99 | 0.94    | 1.00 | 0.97 |
| Monocyte   | 0.43    | 0.47 | 0.45 | 0.76    | 0.95 | 0.85 | 0.66    | 0.97 | 0.78 | 0.72    | 0.94 | 0.81 | 0.85    | 0.87 | 0.86 |
| Neutrophil | 0.69    | 0.18 | 0.28 | 0.98    | 0.78 | 0.87 | 1.00    | 0.26 | 0.41 | 1.00    | 0.40 | 0.57 | 0.97    | 0.49 | 0.65 |

Model 3, as shown by the F1-score from Figure 6, indicates a decrease in monocytes from 0.85 to 0.78. There was also a very drastic drop for the Neutrophil class, down to 0.41. The flipping augmentation technique used to increase the dataset size disrupts the spatial relationships learned by DETR, particularly for Neutrophils. The spatial context here refers to the relative positional relationships between parts of the image. This includes the layout of objects or object parts within the space, such as the shape of the cell nucleus, the direction of orientation, and the distribution of granules. DETR uses a transformer encoder-decoder

that operates based on self-attention. A self-attention-based model demonstrates that DETR not only examines local features but also learns the global spatial relationships among all patches in the image [22]. In the case of WBC detection, DETR will understand the position of the neutrophil nuclear lobes relative to the surrounding granules and the alignment of the oval shape of monocytes. If this structure changes drastically, for example, due to flipping or rotation, the spatial patterns previously learned will become invalid [34], [35]. Likewise, self-attention will struggle to construct accurate representations.

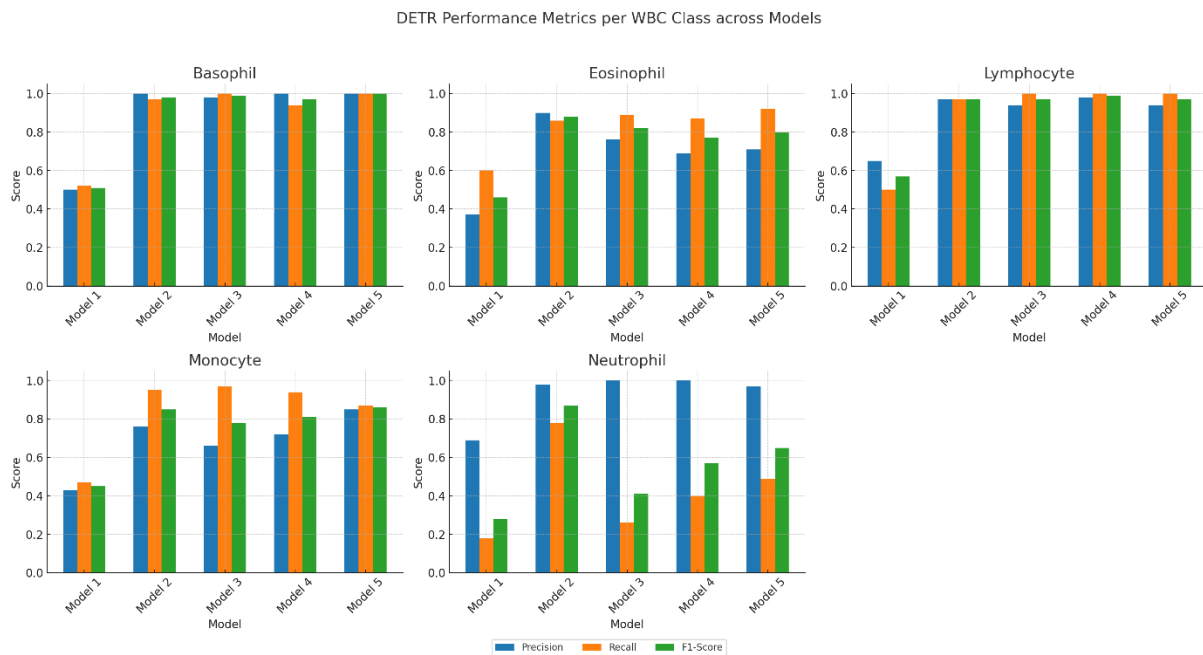


Figure 5. DETR Performance Metrics per WBC Type across Models

In contrast, YOLOv8 focuses more on local features, such as grain texture or intensity patterns. YOLOv8, being a CNN-based model, is more tolerant of positional or spatial changes due to its translation-equivariant nature, which is inherent in the convolution. The addition of flipping or rotation augmentation techniques to increase the dataset size can help with generalization without compromising the learned patterns.

The Basophil class, as in YOLOv8, is a class that has consistently increased since Model 2 and remains

stable, with an F1-score value of greater than 0.97. The performance is relatively stable, with a slight decrease occurring when the rotation augmentation technique is added to the training dataset, specifically in Model 4, where an F1-score of 0.77 is observed for the Eosinophil class due to the application of rotation augmentation, which altered the global arrangement of eosinophilic granules and nucleus orientation, leading to higher feature confusion. The Lymphocyte class is the most stable across all models, with F1-scores of  $\geq 0.97$ . The Monocyte class is sensitive to augmentation, dropping to an F1-score of 0.78 in Model 3, likely

caused by horizontal and vertical flipping that mirrored the oval or horseshoe-shaped nucleus, creating misleading patterns for DETR's global attention. Conversely, in Model 5, which combined flipping and rotation, the F1-score for Monocytes recovered, suggests that diverse combined augmentations helped the model learn robust spatial variations. Meanwhile, the Neutrophil class is the most unstable among the models.

The impact of the flipping augmentation technique to increase the training dataset is detrimental to the neutrophil and monocyte classes. Meanwhile, the rotation augmentation technique can decrease performance in Model 4. The combination of flipping and rotation augmentation techniques partially restores

the performance decline. However, it is still below the performance shown in Model 2, where the dataset used did not involve any augmentation techniques.

The performance demonstrated in the application of DETR for WBC detection and classification is highly dependent on data that is balanced and sufficiently large. Geometric augmentation can potentially degrade performance, especially for cells with distinctive nuclear shapes and orientations, such as Monocytes and Neutrophils. Model 3 and Model 4 exhibit a sharp performance decline, indicating that DETR is highly sensitive to changes in spatial context. Meanwhile, Model 5 shows that augmentation can be beneficial when combined carefully, but the results are still below the Model without augmentation (Model 2).

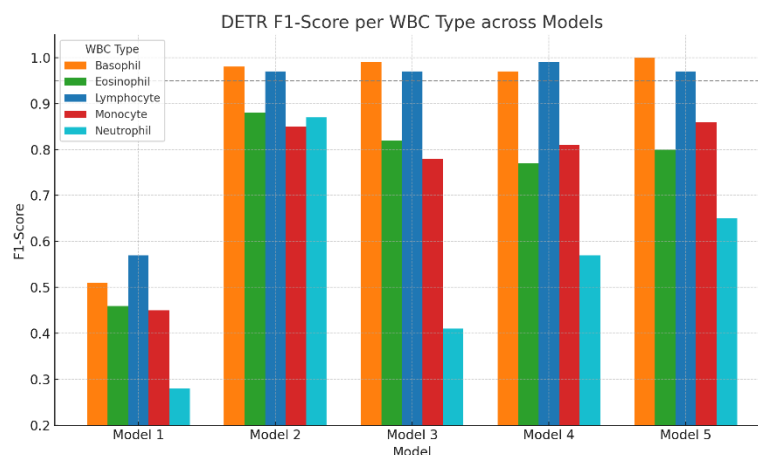


Figure 6. F1-Score per WBC Type for DETR across Experimental Scenarios

### 3.3 Performance Comparison of YOLOv8 and DETR

In general, YOLOv8 demonstrates more stable, higher, and more consistent performance compared to DETR across all metrics—Precision, Recall, and F1-Score—even from Model 1. While DETR shows a sharp improvement starting from Model 2 onward, it still experiences significant fluctuations, especially in classes such as Neutrophil and Monocyte.

The robustness of YOLOv8 is evident in its ability to maintain high scores even when the initial data in Model 1 was imbalanced. On the other hand, DETR is more susceptible to data imbalance, as indicated by its low and unstable scores in Model 1, as well as large fluctuations even after augmentation techniques have been applied.

#### 3.3.1 Comparative Analysis of YOLOv8 and DETR Performance by Class.

In the Basophil class, YOLOv8 maintains perfect Precision, Recall, and F1-scores across all models. In contrast, DETR achieves stability only in Model 2. YOLOv8 is much more robust for the Basophil class. For the Eosinophil class, DETR experiences decreases and fluctuations in all metrics, especially after Model 2, indicating sensitivity to the augmentation techniques

used. Meanwhile, YOLOv8 consistently maintains high scores starting from Model 2.

In the Lymphocyte class, both models demonstrate good performance after Model 2. However, YOLOv8 shows greater stability even in Model 1, while DETR only becomes consistent after Model 2. For the Monocyte class, YOLOv8 is consistent and shows improvement across all metrics, whereas DETR exhibits fluctuating performance, especially in the Recall and F1-score. The Neutrophil class poses a significant challenge for DETR, as the F1-score in Model 1 is only 0.28 and increases slowly, yet never reaches complete consistency. Although Neutrophils had the highest number of samples at Model 1 and the F1-score remained low (0.28), this was not due to limited data but rather caused by high intra-class morphological variation, including the lobulated nuclear shape and inconsistent granule distribution. Moreover, the absence of augmentation techniques such as flipping and rotation made the model highly sensitive to nucleus orientation. The global spatial dependency in DETR's self-attention mechanism further contributed to performance drops, as it struggled to capture invariant features under different spatial configurations. These factors collectively resulted in low recall and overall F1-score for Neutrophils despite sufficient data quantity.

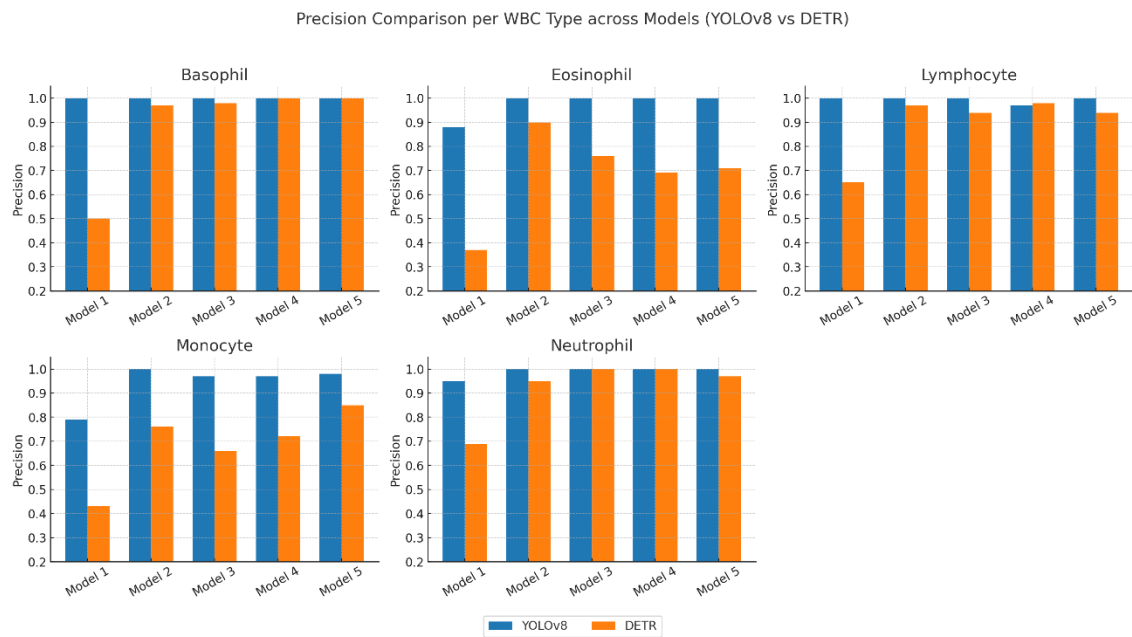


Figure 7. Precision Comparison per WBC Type across Models (YOLOv8 vs DETR)

Figure 7 illustrates the precision trends of YOLOv8 and DETR across all experimental scenarios for each WBC type. YOLOv8 maintains high and consistent precision for all classes from Model 2 onwards. DETR shows

more fluctuations, particularly in classes such as Monocyte and Neutrophil, highlighting its greater sensitivity to class imbalance and the need for augmentation.

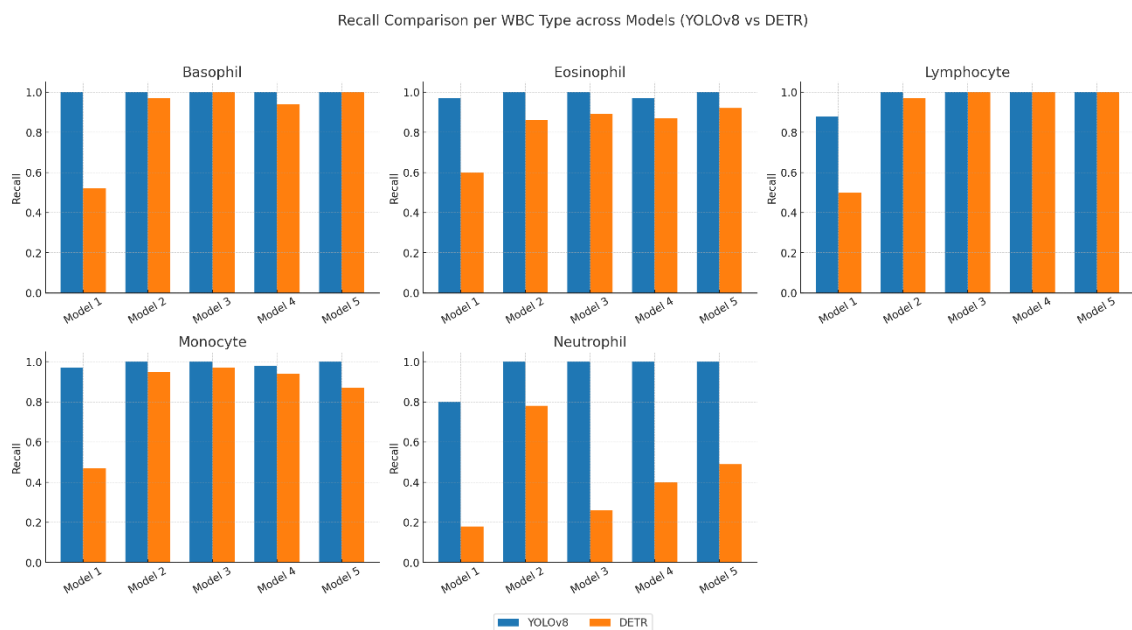


Figure 8. Recall Comparison per WBC Type across Models (YOLOv8 vs DETR)

Figure 8 illustrates the recall performance of YOLOv8 and DETR in all experimental scenarios for each WBC type. YOLOv8 quickly stabilizes at near-perfect recall values starting from Model 2. DETR, although improving with augmentation, remains inconsistent in detecting Neutrophils and Monocytes, indicating a greater dependency on balanced and enriched training data.

Figure 9 compares the F1 scores of YOLOv8 and DETR across the five experimental models for each type of white blood cell. YOLOv8 consistently achieves near-perfect performance from Model 2 onwards in all classes, demonstrating strong robustness and generalization. In contrast, DETR exhibits greater variability, particularly in the Neutrophil and Monocyte classes, and relies more on data augmentation to stabilize performance.

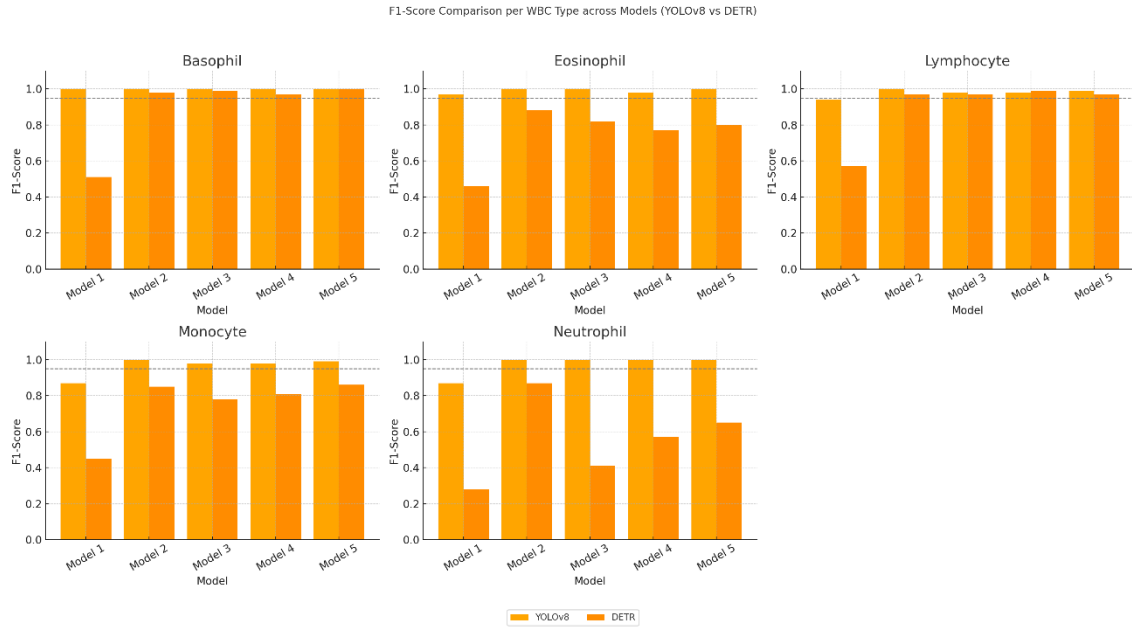


Figure 9. F1-Score Comparison per WBC Type across Models (YOLOv8 vs DETR)

### 3.3.2 Comparative Performance Analysis of YOLOv8 and DETR by Model

Figure 10 shows the average F1-score of YOLOv8 and DETR across five experimental scenarios. YOLOv8 maintains a high and consistent F1-score performance even from Model 1, which demonstrates resilience to class imbalance. In contrast, DETR only shows significant improvement after Model 2, with ongoing instability in subsequent models, highlighting its greater reliance on additional data.

To demonstrate the performance gap between the two methods for the scenario, the difference in the F1-score between the two methods is used, as shown in Equation

4. The difference in F1-Score is made to provide a sharper and more quantitative comparative evaluation of classification or detection performance within the same experimental context. As shown in Figure 11.

$$\text{Difference F1-Score} = \left( \frac{\sum_{i=1}^5 F1_{YOLOv8,i}}{5} \right) - \left( \frac{\sum_{i=1}^5 F1_{DETR,i}}{5} \right) \quad (4)$$

Where  $i$  is the index for each WBC class (Basophil, Eosinophil, Lymphocyte, Monocyte, Neutrophil), the average is calculated per Model. The F1-score comes from the detection experiments of each respective method.

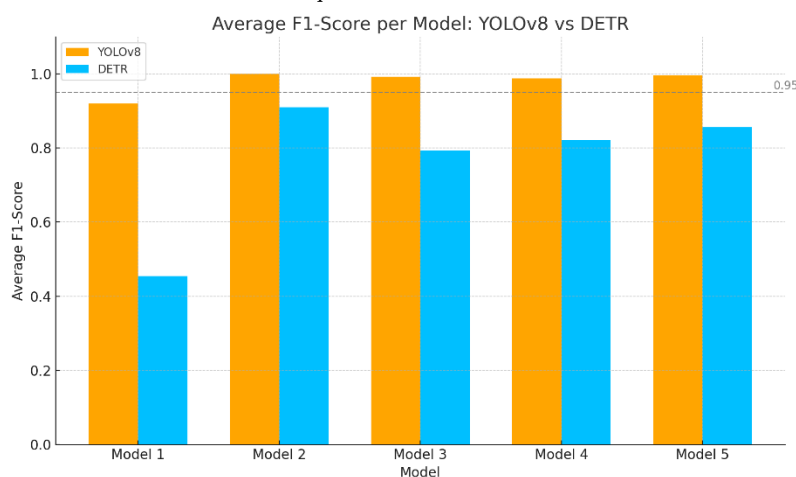


Figure 10. Average F1-Score per Model (YOLOv8 vs DETR)

The comparison of performance between YOLOv8 and DETR based on the difference in F1-scores for each Model, as shown in Figure 11, provides a clear illustration of the level of robustness and generalization of both methods to variations in data distribution and

augmentation techniques. Figure 11 also shows that YOLOv8 consistently outperforms in all scenarios, with the most significant difference observed in Model 1, which is the condition where class distribution is highly imbalanced and no augmentation has been applied. In

this Model, YOLOv8 demonstrates a significantly higher F1-score compared to DETR, indicating that YOLOv8 has superior capability in handling

imbalanced data and detecting all types of WBCs evenly, including minor classes such as basophils [37].

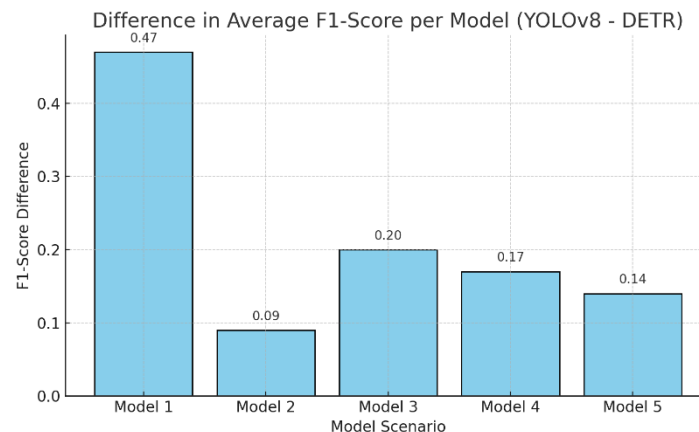


Figure 11. Average F1-Score Difference per Model (YOLOv8 vs DETR)

As balancing and augmentation were applied to Model 2 through Model 5, DETR showed significant performance improvements. Nevertheless, the F1-score gap between the two methods remained consistent, although it narrowed slightly. Indicates that while DETR greatly benefits from the augmentation process, its performance still does not match the stability and consistency demonstrated by YOLOv8. This further reinforces that YOLOv8 is more robust, not only to imbalanced data but also to distribution changes resulting from data augmentation.

In the study [32] on multilabel classification performance metrics, it is emphasized that the F1-score, especially in macro or per-class form, is very useful for evaluating a model's ability to generalize equally to all classes without bias toward the majority class. This study stresses the importance of considering label distribution when interpreting the F1-score, particularly in medical classification scenarios with complex data distributions. In this context, the high and even performance demonstrated by YOLOv8 shows its capability to maintain balanced detection across all classes, including those that are clinically crucial but underrepresented.

Furthermore, in [33], it is stated that the F1-score is a highly informative metric for measuring the trade-off between precision and recall, which is crucial in real-world object detection, especially in the medical field where prediction errors can have significant consequences. In the context of WBC detection from microscopic images, an imbalance between false positives and false negatives can lead to incorrect diagnoses. Therefore, the significant difference in F1-scores between YOLOv8 and DETR in the initial models provides strong evidence that YOLOv8 is capable of maintaining a more balanced and reliable detection performance even before augmentation is applied. At the same time, DETR shows a greater dependence on data modification to achieve similar performance.

Overall, analysis of F1-score differences per Model shows that YOLOv8 has advantages in terms of robustness against uneven data distribution, performance consistency across classes, and resistance to fluctuations caused by augmentation. While DETR demonstrates improved performance with augmentation, its performance remains unbalanced between classes and is not as strong as YOLOv8 in the initial condition without augmentation. This makes YOLOv8 a more reliable choice for implementing image-based medical object detection systems in complex real-world environments.

### 3.4 Limitations and Future Work

Although the metrics of precision, recall, and F1-score provide an overall picture of model performance, both as a whole and per class, they are all limited in their ability to describe object detection capabilities comprehensively. F1-score is indeed very appropriate for illustrating the balance between accuracy and completeness in classification, especially in imbalanced datasets [32], [33]. However, this study has not yet analyzed aspects such as spatial localization, bounding box accuracy, or the precision of object positioning. Therefore, the interpretation of robustness in this study is from a classification perspective, specifically in terms of distinguishing between classes, and does not yet include the spatial integration of detection and classification.

In this study, the dataset used originates from various public sources as described in section 2.1, which are then compiled into a single collection and randomly distributed. Each source has distinct visual characteristics, including staining types, slide quality, resolution, and annotation styles. This results in a diverse combined dataset, thus representing intra-domain variation within the scope of training and testing data. Although the experiments were designed to evaluate the influence of data quantity and augmentation variation on model performance, this

approach also remains limited in comprehensively measuring robustness. The basic augmentation techniques used, can simulate a small portion of visual variation but do not fully reflect what might be encountered in clinical practice, such as variations in lighting or changes in image resolution. Thus, the robustness evaluated in this study still refers to in-domain robustness and does not yet include out-of-distribution robustness, as all the compiled data from various sources are used together in both the training and testing processes. Although using varying amounts of training data for each Model has successfully tested the sensitivity of performance to data size, this characteristic still does not include evaluation of the Model's generalization to truly novel data that has never been seen or is not represented within the training distribution. Future work that can be carried out includes incorporating out-of-distribution testing scenarios to measure the Model's ability to generalize across domains more comprehensively.

Furthermore, although the public datasets used in this study come from several different sources, no explicit separation or separate evaluation has been conducted based on the domain of the datasets. This is a limitation because it is unknown to what extent the Model's performance is affected by such domain variation and whether the Model tends to overfit to a single type of domain or can generalize to others. Robustness is important because systems used in real-world settings must handle variations in how tasks are performed across different labs or hospitals [36], [37], [38]. The results of the per-class analysis in this study indicate an imbalance in the detection of certain classes, such as neutrophils and monocytes, which may be related to differences in visual morphology stemming from domain variation or staining types.

Finally, although model-wise analyses were conducted to assess the impact of balancing and augmentation on model performance, this study has not yet included analyses of model stability under repeated testing or visualization methods based on interpretability, such as attention maps, which could explain why the Model misdetects or misclassifies objects. In a clinical context, where system reliability must be traceable, it is imperative, mainly when used as a diagnostic aid or as an automatic pre-screening system.

Considering these limitations, the robustness evaluation presented in this study provides a comprehensive preliminary analysis from the perspective of data distribution-based classification and augmentation. Nevertheless, it does not yet fully encompass the Model's resilience to spatial variations, data domains, and the interpretability of model errors within the complex context of medical applications.

Based on the experimental results, several WBC classes consistently show lower or unstable performance, particularly in the DETR model, including Neutrophils, Monocytes, and Eosinophils. This is a significant

concern in the development of image-based medical classification systems, which often have complex morphology and tend to exhibit overlapping visual features, making it challenging for models to distinguish them accurately. In particular, for Neutrophils, although they are the most clinically dominant type of WBC [39], performance drops are observed with DETR in some scenarios. It indicates a limitation of the architecture in capturing distinctive local features in cell images. Therefore, the development of detection models that are more sensitive to subtle visual features and local details becomes a priority for further advancement. One future direction is the exploration of transformer-based architectures with more adaptive spatial attention, such as the Swin Transformer, which has proven to be more effective in capturing multi-scale spatial context [40], [41], [42]. Implementing models with more flexible local-global attention capabilities may mitigate misclassification.

Additionally, the performance of YOLOv8 on the Monocyte and Lymphocyte classes, which also shows a relative decline—though not as pronounced as with DETR—suggests that even modern CNN architectures still have limitations in differentiating cells with similar nuclear textures and shapes. Therefore, future development may include strategies for incorporating contextual information or domain-specific features, such as multimodal fusion with numerical data (absolute counts or WBC ratios) into the learning system. This strategy enables the Model to classify not only based on visual morphology but also through clinically relevant quantitative indicators. Future work should also include retraining based on active learning, that is, selectively adding examples from underperforming classes based on predictions with high uncertainty or input samples considered ambiguous, uncertain, or close to the decision boundary. In cases where the Model lacks high confidence in classifying those samples into a single class, such samples are considered informative and valuable for relabeling or reuse in training to enhance performance. Finally, considering that some misclassifications are suspected to arise due to visual similarity between cells, it is crucial to integrate interpretability methods into future development.

By combining approaches such as architectural improvement, multimodal data integration, and training techniques that are more sensitive to class imbalance, future work is expected to produce classification systems that are more robust, generalizable, and—most importantly—reliable in real clinical contexts, where accuracy for each cell class has diagnostic implications that cannot be underestimated.

#### 4. Conclusions

This study shows that YOLOv8 consistently outperforms DETR in terms of F1-score, precision, and recall while also being more stable across classes and

scenarios. YOLOv8 demonstrates a positive response to augmentation and balancing and has proven to be more adaptive to changes in the amount of training data. On the other hand, DETR shows more varied performance and is more sensitive to differences in data distribution. Although YOLOv8's performance is generally more stable and consistent across all scenarios, experimental results also show that DETR managed to achieve performance metrics (F1-score) above 95%, especially in balanced and augmented data scenarios and in several classes such as Basophil and Lymphocyte (starting from Model 2). It indicates that the DETR architecture still possesses strong capabilities for detecting and classifying small objects such as WBCs, particularly when provided with representative data distribution and proper training. With high F1-score metrics achieved in those classes, DETR proves that this transformer-based approach is not only promising in theory but also practically competitive, even without the need for complex architectural modifications. Therefore, achieving performance above 95% reinforces the argument that DETR deserves further exploration, especially as a foundation for the development of more structured and interpretable end-to-end detection systems in the medical domain. In this context, DETR methods are not the result of synthesis, such as Deformable DETR, Improved DETR, or even integrating DETR with backbones like the vision transformer (Vision Transformer and Swin Transformer). This study also contributes by implementing DETR in the context of WBC classification—an area that has been rarely explored to date. It positions the research as a relevant starting point for understanding DETR's potential in the medical domain, especially for tasks involving the detection of small objects with similar morphology. Thus, although YOLOv8 excelled in terms of performance metrics in this experiment, DETR still holds long-term potential worth further exploration through architecture optimization, integration with augmentation techniques, and domain-specific adaptation in the future.

### Acknowledgements

This work was supported by Penelitian DRTPM 2023/Penelitian Disertasi Doktor” based on Grand-in-id and Agreement No. KWR4.090/LIT07./PPM-LIT/2023 and 993/SP2H/RT-MONO/LL4/2023.

### References

- [1] M. Li et al., “Peripheral Blood Leukocyte Detection Based on an Improved Detection Transformer Algorithm,” *Sensors*, vol. 23, no. 16, 2023, doi: 10.3390/s23167226.
- [2] X. Zheng, P. Tang, L. Ai, D. Liu, Y. Zhang, and B. Wang, “White blood cell detection using saliency detection and CenterNet: A two-stage approach,” *J Biophotonics*, vol. 16, no. 3, p. e202200174, Mar. 2023, doi: 10.1002/JBIO.202200174.
- [3] R. Biswal, P. K. Mallick, A. R. Panda, G. S. Chae, and A. Mishra, “White Blood Cell Classification Using Pre-Trained Deep Neural Networks and Transfer Learning,” *2023 1st International Conference on Circuits, Power, and Intelligent Systems, CCPIS 2023*, 2023, doi: 10.1109/CCPIS59145.2023.10291642.
- [4] Y. Jiang et al., “Automatic recognition of white blood cell images with memory efficient superpixel metric GNN: SMGNN,” *Mathematical Biosciences and Engineering* 2024 2:2163, vol. 21, no. 2, pp. 2163–2188, 2024, doi: 10.3934/MBE.2024095.
- [5] “Automated Identification and Classification of White Blood Cells (Leukocytes) in Digital Microscopic Images.” Accessed: Feb. 26, 2025. [Online]. Available: <https://www.ijcaonline.org/specialissues/rtippr/number2/977-100/>
- [6] C. Shyalika and P. Kumara, “Morphological Analysis and Diagnosis of Human Leukemic Cells Using Image Processing: A Review,” *2022 2nd International Conference on Computer, Control and Robotics, ICCCR 2022*, pp. 251–257, 2022, doi: 10.1109/ICCCR54399.2022.9790102.
- [7] J. Zhao, M. Zhang, Z. Zhou, J. Chu, and F. Cao, “Automatic detection and classification of leukocytes using convolutional neural networks,” *Med Biol Eng Comput*, vol. 55, no. 8, pp. 1287–1301, Aug. 2017, doi: 10.1007/S11517-016-1590-X.
- [8] H. Kutlu, E. Avcı, and F. Özyurt, “White blood cells detection and classification based on regional convolutional neural networks,” *Med Hypotheses*, vol. 135, p. 109472, Feb. 2020, doi: 10.1016/J.MEHY.2019.109472.
- [9] S. Khan, M. Sajjad, N. Abbas, J. Escorcia-Gutierrez, M. Gamarra, and K. Muhammad, “Efficient leukocytes detection and classification in microscopic blood images using convolutional neural network coupled with a dual attention network,” *Comput Biol Med*, vol. 174, p. 108146, May 2024, doi: 10.1016/J.COMPBIOMED.2024.108146.
- [10] V. Shalini and K. S. A. Viji, “AUTOMATIC DETECTION AND CLASSIFICATION OF LEUKEMIA FROM BLOOD SMEAR IMAGE USING SENET CONVOLUTIONAL NEURAL NETWORK,” *J Theor Appl Inf Technol*, vol. 101, no. 15, pp. 6065 – 6075, 2023, [Online]. Available: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85172690043&partnerID=40&md5=9b0a734b7dbb1a7ac70c996d8b9f3512>
- [11] B. Liu, “Blood Cell Count and Detection Method Based on YOLO,” *Highlights in Science, Engineering and Technology*, vol. 27, pp. 594–599, Dec. 2022, doi: 10.54097/hset.v27i.3822.
- [12] M. Sharif et al., “Recognition of different types of leukocytes using YOLOv2 and optimized bag-of-features,” *IEEE Access*, vol. 8, pp. 167448–167459, 2020, doi: 10.1109/ACCESS.2020.3021660.
- [13] L. D. Trong, T. P. Thanh, H. P. Manh, and D. N. Minh, “Yolov4 in White Blood Cell Classification,” pp. 387–400, 2023, doi: 10.1007/978-981-19-6004-8\_31.
- [14] N. H. Harun and N. Fared, “An Experiment on Lung Disease Classification using YOLOv8,” *Applied Mathematics and Computational Intelligence (AMCI)*, vol. 13, no. 3, pp. 176–185, Oct. 2024, doi: 10.58915/AMCI.V13I3.626.
- [15] N. Palanivel, S. Deivanai, G. Lakshmi Priya, B. Sindhuja, and M. Shamrin Millet, “The Art of YOLOv8 Algorithm in Cancer Diagnosis using Medical Imaging,” *2023 International Conference on System, Computation, Automation and Networking, ICSCAN 2023*, 2023, doi: 10.1109/ICSCAN58655.2023.10395046.
- [16] S. Albahli, “A Robust YOLOv8-Based Framework for Real-Time Melanoma Detection and Segmentation with Multi-Dataset Training,” *Diagnostics* 2025, Vol. 15, Page 691, vol. 15, no. 6, p. 691, Mar. 2025, doi: 10.3390/DIAGNOSTICS15060691.
- [17] H. Lou et al., “DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor,” *Electronics* 2023, Vol. 12, Page 2323, vol. 12, no. 10, p. 2323, May 2023, doi: 10.3390/ELECTRONICS12102323.
- [18] W. Jianjing, X. Baolin, Z. Shurong, and Z. Xin, “Complete Blood Cell Detection Based on Improved YOLOv8,” *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, pp. 148–152, 2024, doi: 10.1109/ICSESS62520.2024.10719382.

- [19] M. Kang, C.-M. Ting, F. F. Ting, and R. Phan, "CST-YOLO: A Novel Method for Blood Cell Detection Based on Improved YOLOv7 and CNN-Swin Transformer," 2023.
- [20] Y. Guo and M. Zhang, "Blood Cell Detection Method Based on Improved YOLOv5," *IEEE Access*, vol. 11, pp. 67987–67995, 2023, doi: 10.1109/ACCESS.2023.3290905.
- [21] C. Matsoukas, J. F. Haslum, M. Sorkhei, M. Söderberg, and K. Smith, "Pretrained ViTs Yield Versatile Representations For Medical Images," Mar. 2023, Accessed: Feb. 26, 2025. [Online]. Available: <https://arxiv.org/abs/2303.07034v3>
- [22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12346 LNCS, pp. 213–229, 2020, doi: 10.1007/978-3-030-58452-8\_13/FIGURES/8.
- [23] B. Leng, C. Wang, M. Leng, M. Ge, and W. Dong, "Deep learning detection network for peripheral blood leukocytes based on improved detection transformer," *Biomed Signal Process Control*, vol. 82, p. 104518, Apr. 2023, doi: 10.1016/J.BSPC.2022.104518.
- [24] X. Zhu et al., "Deformable DETR: Deformable Transformers for End-to-End Object Detection," Oct. 2020, Accessed: Jun. 09, 2023. [Online]. Available: <https://arxiv.org/abs/2010.04159v4>
- [25] N. Bayat, D. D. Davey, M. Coathup, and J. H. Park, "White Blood Cell Classification Using Multi-Attention Data Augmentation and Regularization," *Big Data and Cognitive Computing 2022*, Vol. 6, Page 122, vol. 6, no. 4, p. 122, Oct. 2022, doi: 10.3390/BDCC6040122.
- [26] Y. Deng and H. Li, "Deep learning for few-shot white blood cell image classification and feature learning," *Comput Methods Biomech Biomed Eng Imaging Vis*, vol. 11, no. 6, pp. 2081–2091, Nov. 2023, doi: 10.1080/21681163.2023.2219341.
- [27] H. Nozaka et al., "The Effect of Data Augmentation in Deep Learning Approach for Peripheral Blood Leukocyte Recognition," *Stud Health Technol Inform*, vol. 290, pp. 273–277, Jun. 2022, doi: 10.3233/SHTI220077.
- [28] A. Acevedo, A. Merino, S. Alf  rez,   . Molina, L. Bold  , and J. Rodellar, "A dataset for microscopic peripheral blood cell images for development of automatic recognition systems," vol. 1, 2020, doi: 10.17632/SNKD93BNJR.1.
- [29] "Blood Cell Images." Accessed: Jun. 17, 2025. [Online]. Available: <https://www.kaggle.com/code/plavak10/blood-cell-images/input>
- [30] "BCCD Dataset." Accessed: Jul. 09, 2025. [Online]. Available: <https://public.roboflow.com/object-detection/bccd>
- [31] "Home - Raabin Health Database." Accessed: Dec. 29, 2023. [Online]. Available: <https://raabindata.com/>
- [32] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences (Switzerland)*, vol. 14, no. 21, p. 9863, Nov. 2024, doi: 10.3390/AP14219863/S1.
- [33] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," *International Conference on Systems, Signals, and Image Processing*, vol. 2020-July, pp. 237–242, Jul. 2020, doi: 10.1109/TWSSIP48289.2020.9145130.
- [34] M. Nourelahi, L. Kotthoff, P. Chen, and A. Nguyen, "How explainable are adversarially-robust CNNs?," May 2022, Accessed: Jun. 29, 2025. [Online]. Available: <https://arxiv.org/pdf/2205.13042>
- [35] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Oct. 2020, Accessed: Jun. 08, 2023. [Online]. Available: <https://arxiv.org/abs/2010.11929v2>
- [36] Z. Su, J. Guo, X. Yang, Q. Wang, F. Coenen, and K. Huang, "Navigating Distribution Shifts in Medical Image Analysis: A Survey," Nov. 2024, Accessed: Jun. 24, 2025. [Online]. Available: <https://arxiv.org/pdf/2411.05824>
- [37] M. Sikaroudi, M. Hosseini, S. Rahnamayan, and H. R. Tizhoosh, "ALFA - Leveraging All Levels of Feature Abstraction for Enhancing the Generalization of Histopathology Image Classification Across Unseen Hospitals," *Proceedings - 2023 IEEE/CVF International Conference on Computer Vision Workshops, ICCVW 2023*, pp. 2656–2665, 2023, doi: 10.1109/ICCVW60793.2023.00281.
- [38] S. Jiang, Z. Wu, H. Yang, K. Xiang, W. Ding, and Z. S. Chen, "A prior knowledge-guided distributionally robust optimization-based adversarial training strategy for medical image classification," *Inf Sci (N Y)*, vol. 673, p. 120705, Jul. 2024, doi: 10.1016/J.INS.2024.120705.
- [39] C. Schultz, "The neutrophil," *General Internal Medicine and Clinical Innovations*, vol. 5, no. 3, 2020, doi: 10.15761/GIMCI.1000198.
- [40] I. Pacal, "A novel Swin transformer approach utilizing residual multi-layer perceptron for diagnosing brain tumors in MRI images," *International Journal of Machine Learning and Cybernetics*, vol. 15, no. 9, pp. 3579–3597, Sep. 2024, doi: 10.1007/S13042-024-02110-W/METRICS.
- [41] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.
- [42] T. Y. Kim, A. Niaz, J. S. Choi, and K. N. Choi, "One-Stage Detection Model Based on Swin Transformer," *IEEE Access*, vol. 12, pp. 60960–60972, 2024, doi: 10.1109/ACCESS.2024.3393152.