Explainable Deep Learning Methodologies for Comprehensive White Blood Cell Analysis

M.Tech Thesis

by

Adit Srivastava

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

Explainable Deep Learning Methodologies for Comprehensive White Blood Cell Analysis

A THESIS

Submitted in partial fulfillment of the

requirements for the award of the degree

of

Master of Technology

by

Adit Srivastava

2302101002

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY INDORE

May 2025

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled **Explainable Deep Learning Methodologies for Comprehensive White Blood Cell Analysis** in the partial fulfillment of the requirements for the award of the degree of **Master of Technology** and submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Indore,** is an authentic record of my own work carried out during the period from July 2023 to July 2025 under the supervision of Dr. Puneet Gupta, Indian Institute of Technology Indore, India.

The matter presented in this thesis has not been submitted by me for the award of any other degree of this or any other institute.

Adit Srivastava
15/05/2025

Signature of the Student with Date

(Adit Srivastava)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

16/05/2025

Signature of Thesis Supervisor with Date

(Dr. Puneet Gupta)

Adit Srivastava has successfully given his M.Tech. Oral Examination held on **30th April, 2025**.

Signature(s) of Supervisor(s) of M.Tech. thesis

Date:16/05/2025

Subhra Mazumdar

Signature of Chairman, PG Oral Board

Date: 18.05.2025

Signature of HoD

Date: 18-May-2025

ACKNOWLEDGEMENT

I want to take this moment to sincerely thank everyone who has supported me along this journey, making it both joyful and rewarding. Their cooperation has been invaluable to my scientific progress and personal development, and I am deeply grateful for their assistance.

Above all, I extend my heartfelt gratitude to my supervisor, Dr. Puneet Gupta, whose persistent support, wise counsel, and steadfast guidance have been crucial throughout this endeavor. His mentorship has provided me with the direction and confidence needed to navigate the challenges of my research. This work would not have been possible without his expertise and dedication, for which I am immensely appreciative. I am equally thankful to my lab mates, whose collaboration, feedback, and camaraderie have greatly enriched the quality of my research. Their helpful suggestions and considerate assistance have made this experience both memorable and intellectually rewarding. I am also deeply grateful to Dr. Ranveer Singh, Head of the Department of Computer Science and Engineering, for his invaluable guidance and supervision, which have significantly contributed to the success of this project. Special recognition goes to the Director of the Indian Institute of Technology Indore, Prof. Suhas Joshi, for fostering an excellent learning environment and providing me with the opportunity to develop my research skills. My experience has been greatly enhanced by the resources and support offered by this institution.

Last but not least, I want to express my sincere gratitude to my parents for their constant love, support, and encouragement. Throughout my life, their confidence in my skills has continuously given me strength and inspired me. Their steadfast presence has enabled me to pursue my aspirations, and I will always remain grateful for their sacrifices and dedication. To everyone who has contributed to this work in any capacity, from their guidance and advice to their encouragement and patience, thank you. Each individual has played a crucial role in making this journey a success. I am truly grateful for your support during this significant chapter of my academic life.

Adit Srivastava

Dedicated to My Family

List of Publications

Publications from Thesis

International Conferences

C1. **Adit Srivastava**, Aravind Ramagiri, Puneet Gupta, and Vivek Gupta. "SANGAM: Synergizing Local and Global Analysis for Simultaneous WBC Classification and Segmentation". In: International Conference on Pattern Recognition. Springer. 2025, pp. 154–169.

ABSTRACT

The analysis of white blood cells (WBCs) is a critical aspect of health monitoring and diagnosis, providing valuable information on a patient's immune health. Pathologists typically follow a systematic approach to this task, involving three sequential steps: localizing WBCs, analyzing their morphological attributes, and classifying them based on these features. Despite the interdependence of these processes, existing literature often fails to address their synergy comprehensively. Most current systems focus on individual tasks, such as segmentation or classification, without integrating these steps in a way that enhances their mutual strengths. Additionally, these systems rarely provide transparent explanations for their decisions, which are crucial for practical applications where interpretability and trust in automated systems are paramount. Deep learning models, in particular, are frequently criticized for their opacity, offering minimal insights into the rationale behind their predictions. Another significant limitation of existing methods is their lack of versatility. There is a growing demand for adaptable systems that can be fine-tuned on datasets with limited ground truth annotations or even none for specific tasks, while maintaining consistent effectiveness across various tasks.

The proposed system, introduced in this thesis, addresses these challenges comprehensively. Results from experiments on benchmark datasets show that the system performs better than current WBC segmentation, classification, and morphological attribute prediction methods. It employs a novel hybrid architecture that combines transformers for modeling long-range dependencies with convolutional neural networks for capturing intricate local details. This integration enables precise segmentation, providing structural cues for reliable morphological attribute prediction, which subsequently guides WBC classification. By seamlessly linking these tasks, the system emulates the decision-making process of pathologists, enhancing both performance and interpretability. Addressing the limitations of traditional deep learning approaches, the system is both effective and adaptable, making it well-suited for real-world healthcare applications where transparency is essential.

Contents

List of Figures

List of Tables

Acronyms

Acc	Accuracy
CA	Cross Attention
CEL	Cross Entropy Loss
CSD	Cross Scale Decoder
CNNs	Convolutional Neural Networks
DETR	Detection Transformer
DL	Deep Learning
DSC	Mean Dice Similarity Coefficient
F-m	F-measure
FSD	Feature Synergy Decoder
IoU	Mean Intersection over Union
MAttrP	Morphological Attribute Prediction
MAtts	Morphological Attributes
PCI	Progressive Contextual Integration
PFA	Progressive Feature Aggregation
Pre	Precision
Rec	Recall
Spec	Specificity
STR	Spatial Texture Refinement
SupConLoss	Supervised Contrastive Loss
SOTA	State of the Art
Swin T	Swin Transformer

SVM	Support Vector Machine
ViT	Vision Transformer
WBCs	White Blood Cells

Chapter 1

Introduction

White blood cells (WBCs) are crucial immune system components found in blood [2]. They are classified into five categories, each of which plays a crucial role: lymphocytes, eosinophils, monocytes, neutrophils, and basophils. [3]. Neutrophils are the primary defense against infections caused by pathogens and bacteria [4]. When monocytes develop into macrophages, they play a crucial role in eliminating cellular debris. Eosinophils are essential in the fight against parasite infections and allergic reactions, whereas lymphocytes are in charge of locating and destroying malignant and virus-infected cells [5]. By releasing histamine, basophils help control allergic reactions [6]. Given their crucial role in diagnosing a variety of conditions, including leukemia, lymphoma, and infections, analyzing WBCs is of great clinical significance. This underscores the urgent need for systems that combine segmentation, classification, and morphological interpretation of WBCs to enable accurate and efficient diagnostics.

In the past, analyzing blood smears involved pathologists manually examining cells under a microscope. Although this approach was precise, it was laborious and prone to mistakes [7]. To address these issues, early computer-aided systems were developed, relying on handcrafted features and traditional machine learning approaches [8, 9]. These systems typically broke down the process into steps like pre-processing, feature extraction, segmentation, and classification [10, 11]. However, they had significant limitations, such as heavy dependence on manually designed features and the risk of errors compounding at each stage, which often affected the accuracy of the final results.

Recent years have seen a revolution in WBC analysis with deep learning (DL) [12, 13, 14, 15], greatly minimizing the need for manual feature extraction. Systems like ResNet, MobileNet [16], and DenseNet [17] have excelled in segmentation and classification tasks by capturing detailed local features [18]. However, these convolutional neural networks (CNNs) based systems often face limitations in understanding broader contextual information, which is vital for interpreting the overall structure of cells [19]. To address this gap, Transformer-based architectures such as Swin Transformer (Swin T) [2], Vision Transformer (ViT) [20], and Detection Transformer (DETR) [21] have emerged. These systems leverage their ability to capture long-range relationships [22] for a more holistic feature representation. Nevertheless, even these advanced systems may struggle with capturing subtle details in intricate regions like the nucleus and cytoplasm [20]. To overcome these challenges, hybrid systems that integrate the localized precision of CNNs with the global contextual understanding of Transformers have gained attention [5]. Many such systems use decoders that stack multi-level features [5], incorporating convolutions to retain local context. By integrating attention mechanisms during feature merging, these systems can highlight relevant features, improving the alignment of morphological details with broader contextual information [23], and optimizing the synergy between CNNs and Transformers. Some DL systems [24] also perform Morphological Attribute Prediction (MAttrP) alongside segmentation or classification tasks. However, these attributes are rarely leveraged to guide or improve the core tasks, leaving the systems less transparent and less reliable for clinical use [1, 25]. Typically, segmentation, classification, and MAttrP are treated as separate tasks or are only weakly integrated, failing to utilize interpretable attributes to enhance system performance in a clinically meaningful way. This lack of explainability is a significant drawback, as DL systems often operate as opaque systems, providing limited insights into the reasoning behind their predictions [26]. In clinical settings, this opacity undermines trust, especially when the system's outputs do not align with the diagnostic logic used by experts [1, 25]. Although some systems offer visual tools like heatmaps, they rarely capture the detailed reasoning that pathologists apply [27]. Pathologists, for instance, examine WBCs in blood smear slides by analyzing their morphology, considering attributes like shape, size, color, and texture, alongside nuclear features [1] to make accurate classifications [4].

It is crucial to realise that MAttrP, segmentation, and classification are all interrelated processes. Similar to how pathologists find cells on a blood smear slide, segmentation identifies the regions of interest and provides MAttrP with essential structural information. Fine features like vacuoles, which resemble blobs, and larger ones like cell and nucleus size, which call for a more comprehensive contextual understanding, are examples of these interpretable qualities. This emphasises how systems that combine CNNs and Transformers are required in order to take advantage of both localised and global properties. Adding characteristics that pathologists frequently employ, like cell size, cytoplasmic makeup, and nucleus shape, can greatly increase classification accuracy [1]. Additionally, insights gained from classification can further refine segmentation, especially in complex cases like basophils, where the cytoplasm and nucleus overlap [1]. Despite the obvious synergy between these tasks, using interpretable features to improve classification and hence segmentation is still underexplored. A combined approach that integrates all these elements holds great potential for achieving more accurate and clinically reliable WBC analysis. Moreover, a significant limitation in many datasets [7, 28] is that they often provide segmentation ground truth, classification labels, or both, but typically lack ground truth for WBC morphological attributes (MAtts) [1]. This highlights the need for versatile systems that can be fine-tuned using classification labels, segmentation labels, or a combination of both, while performing well across all aspects of WBC analysis. Furthermore, incorporating MAttrP improves the system's interpretability, offering valuable insights into the reasoning behind classification decisions. This transparency helps build trust with clinical pathologists, strengthening their confidence in the system's results.

This study introduces a novel, *explainable framework inspired by the diagnostic strategies of pathologists*. The framework integrates segmentation, MAttrP, and classification into a unified pipeline for WBC analysis. The following summarizes this thesis's main contributions.

- **Comprehensive Framework**: A unified approach for WBC analysis is developed, combining segmentation, MoAP, and classification in a manner that mirrors the analytical process of experienced pathologists.

- **Innovative Decoder Architecture**: A novel decoder design is presented that blends the global context modeling capabilities of transformers with the intricate feature learning strengths of CNNs. This approach enables the seamless fusion of precise low-level details with abstract high-level representations.

- **Synergistic Workflow**: The framework establishes an interconnected workflow where segmentation outputs provide structural information for MAttrP. These predicted attributes subsequently guide the classification process, and the classification results, in turn, contribute to refining the segmentation masks.

- **Versatile and Adaptable System**: The proposed system is designed for universal adaptability, allowing fine-tuning across diverse WBC datasets. It demonstrates superior performance in segmentation, classification, and attribute prediction, even when datasets are annotated for only one of these tasks. By delivering interpretable insights, the framework enhances clinical reliability and facilitates its application in real-world diagnostic scenarios.

The remainder of this thesis is structured as follows: Chapter 2 covers the review of the literature, Chapter 3 details the proposed systems, Chapter 4 presents the experimental findings, and Chapter 5 concludes the research.

Chapter 2

Literature Survey

WBC analysis involves three key tasks: segmentation, classification, and MAttrP. Among these, segmentation and classification have been extensively studied, whereas MAttrP has received relatively limited attention.

2.0.1 WBC Segmentation

2-class and 3-class segmentation are the two main categories of WBC segmentation (refer to Fig. 2.1). Everything else is considered the background in 2-class segmentation, which only identifies the nucleus. On the other hand, 3-class segmentation divides the backdrop, cytoplasm, and nucleus into different groups.

Traditional techniques, including thresholding, morphological operations, non-local filtering, and clustering [18, 19, 29, 30], have demonstrated notable success in 2-class segmentation. However, these systems often perform poorly under diverse imaging conditions [10], necessitating a shift towards DL based approaches. DL systems, particularly CNNs and Transformers, have shown significant promise. For instance, [31] integrates CNNs with attention mechanisms, while [3] combines traditional systems with U-Net architectures. Although there is limited focus on 3-class segmentation, notable contributions include [32] and [5]. [32] employs CNNs and U-Net, while [5] introduces an integrated segmentation-classification framework that stacks high-level features and low-level and in the decoder.

It is crucial to recognize that the performance of existing WBC segmentation systems remains limited due to their inability to fully harness the synergy between CNNs and Transformers. For instance, [5] identifies weaknesses in decoder design, where features are merely stacked without leveraging intelligent merging techniques. Incorporating attention mechanisms to selectively emphasize relevant features holds significant potential for improving the efficacy of these hybrid systems [23]. Furthermore, the use of classification insights to enhance segmentation accuracy is still underdeveloped, highlighting an important area for future research.

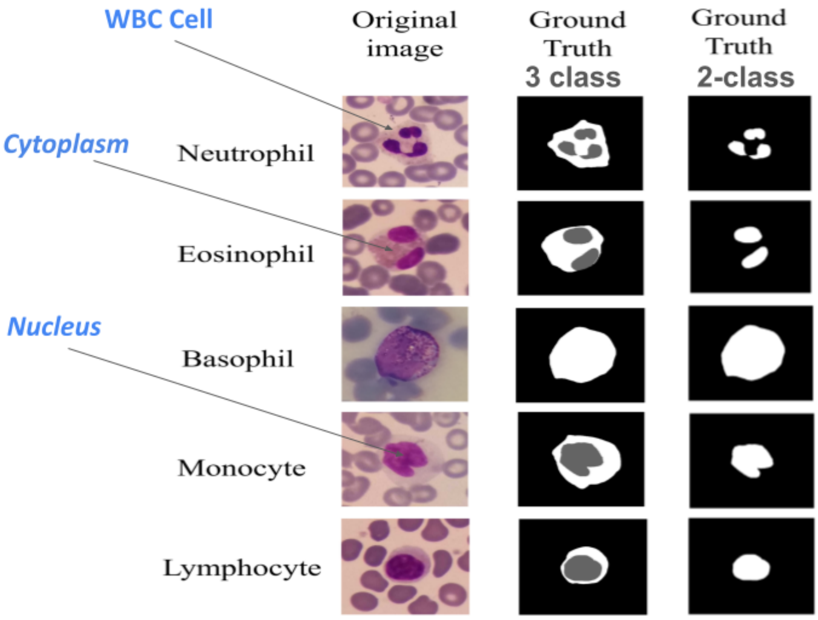

Figure 2.1: WBC Segmentation: 2-class (nucleus vs. background) and 3-class (nucleus, cytoplasm, background), alongside the categorization of different WBC types based on distinct features such as nucleus size, cytoplasm texture, and staining patterns.

2.0.2 WBC Classification

DL has revolutionized the classification of WBC (refer to Fig. 2.1 for a visual depiction of the distinct appearances of various WBC classes) by replacing traditional techniques with more efficient and accurate models. CNN-based approaches [33, 34, 17, 35] have significantly improved classification accuracy by processing entire cell images. However, CNNs' inherent limitation in capturing global contextual information has driven the adoption of Transformer-based architectures, such as the ViT [20], which excels at modeling long-range dependencies. The integration of segmentation into classification pipelines has further enhanced results. For example, [36] combines DeepLabv3+ for segmentation with AlexNet for classification, while [3] integrates U-Net with ResNet. However, resizing segmented regions in these systems often blurs boundaries, obscuring critical morphological details essential for clinical analysis. Addressing this issue, [5] proposes an approach that preserves key features during segmentation, ensuring granularity is retained.

Despite these advancements, existing systems fail to replicate the diagnostic approach of pathologists [1]. Pathologists typically localize WBCs within an image, examine their MAtts, and then proceed to classification [37, 23, 1]. This process enhances interpretability and decision-making, fostering trust in the analysis. While some systems incorporate segmentation to aid classification [5, 36, 3], they primarily use segmentation for WBC localization and structural cues rather than explicitly predicting MAtts. MAtts [1] range from fine details like vacuoles, which require local feature learning, to broader features such as cell shape and nucleus size, necessitating global context. Current hybrid systems [5] capable of both local and global learning should optimize segmentation to guide MAttrP before performing classification. By simulating the diagnostic behavior of pathologists, these systems could significantly improve clinical applicability and reliability.

2.0.3 WBC Morphological Attribute Prediction

The prediction of MAtts, such as nucleus size, nucleus-cytoplasm ratio, cytoplasmic texture, etc. [1] (refer to Fig 2.2 & 2.3) has received less attention than segmentation and classification.

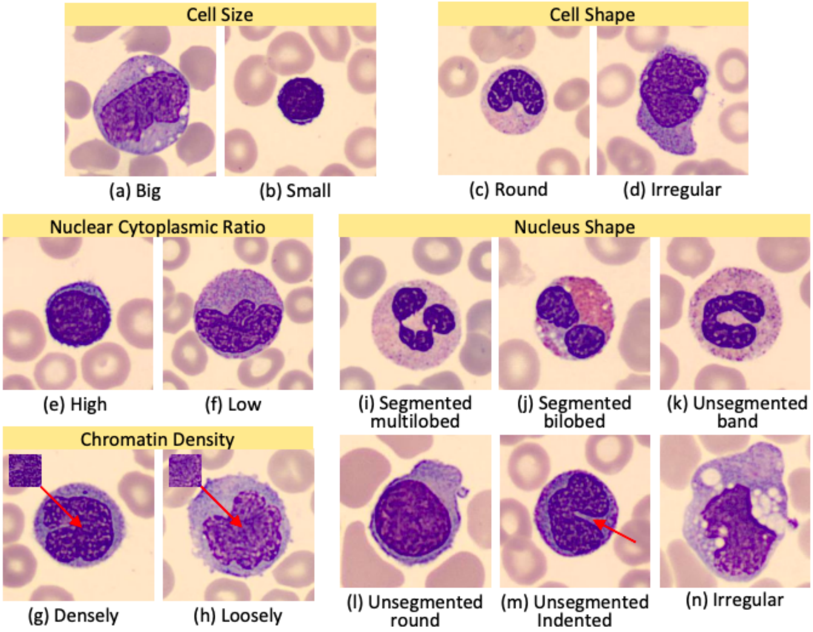

Figure 2.2: Explanatory characteristics of WBCs [1], including their dimensions, structural form, nucleus-to-cytoplasm ratio, nuclear contour, and chromatin density.

These attributes are crucial for clinical diagnostics; however, traditional methods relying on manual measurements or semi-automated systems are often time-consuming and subjective. Recent advancements in DL have enabled automated systems for predicting MAtts, but a major limitation of these systems is their lack of interpretability, which undermines clinical reliability. Techniques like Grad-CAM and LIME aim to improve interpretability but fall short of replicating the explicit analytical reasoning used by pathologists, who rely on MAtts such as nuclear shape, cytoplasmic granularity, cell size, etc [1, 25]. For instance, [1] utilized various image encoders, including CNNs (e.g., ResNet, ShuffleNet, DenseNet) and Transformers (e.g., Swin T, ViT), to predict attributes. However, their approach did not integrate segmentation or classification tasks. Similarly, [24] focused on enhancing interpretability by combining segmentation, classification, and MAttrP within a single framework. Their system, based on a DETR model, employed three separate heads for each task: segmentation, classification, and attribute prediction. Despite these innovations, their design lacked effective synergy between segmentation, classification, and MAttrP, falling short of replicating the reasoning pathologists use in clinical diagnostics.

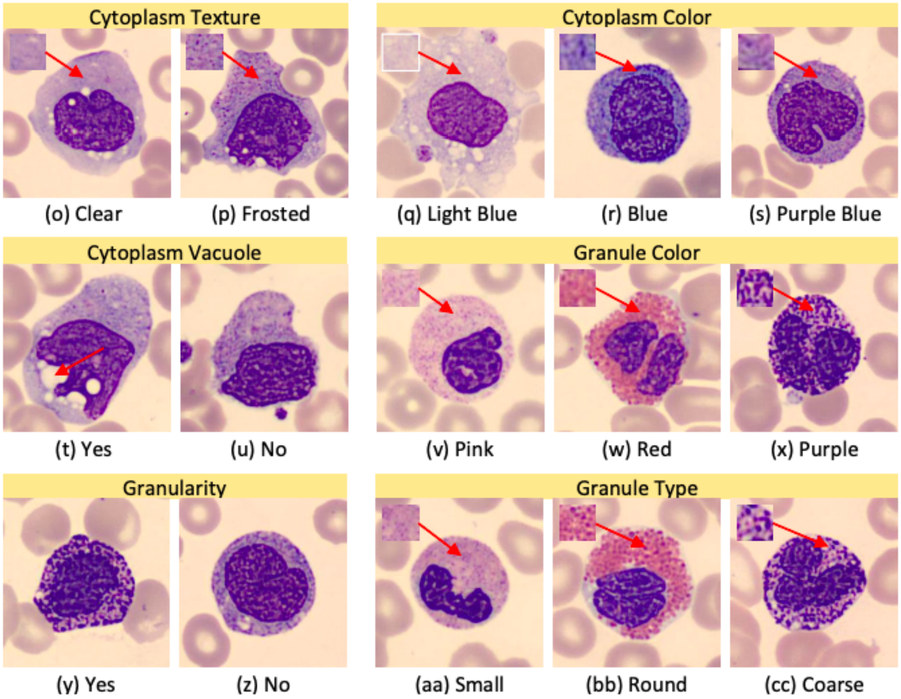

Figure 2.3: Explanatory Attributes of WBCs [1] such as the texture and color of the cytoplasm, presence of vacuoles, characteristics of granules, including their color, type, and overall granularity.

Future advancements in MAttrP should prioritize better task integration and interpretability. Systems must align predictions with clear, observable characteristics to simulate the analytical processes of pathologists [1, 23]. This approach could significantly improve clinical reliability and foster greater trust in automated diagnostic tools.

2.0.4 Synergy between WBC Segmentation, Classification, and Morphological Attribute Prediction

Limited research has delved into establishing a cohesive integration of segmentation, classification, and MAttrP within a unified framework. While some systems utilize segmentation to aid classification [3, 36] or employ classification to refine segmentation [5], the explicit inclusion of MAttrP as a guiding factor remains unexplored.

To emulate the holistic reasoning process of pathologists [1], future systems need to create stronger task interconnections. This involves leveraging segmentation to inform MAttrP, using MAttrP to improve classification, and applying classification insights to enhance segmentation. By fostering seamless integration among these tasks, automated systems can simulate the synergistic diagnostic approach of pathologists, paving the way for interpretable and more reliable clinical tools.

The tables below provide a concise summary of the referenced works.

Systems	**Methodology Adopted**
[18, 19, 29, 30]	Primarily use image processing methods such as morphological operations, non-local filtering, clustering, and gray-level thresholding, for 2-class segmentation.
[31, 32]	Combines CNN architectures with attention mechanisms to perform 2-class segmentation.
[32]	Uses CNN, U-Net, and SegNet for 3-class segmentation.
[24]	Employs Transformer with sparse attention for segmentation, classification, and explanation.

Table 2.1: Literature Review Summary: Part 1.

Systems	**Methodology Adopted**
[3]	Uses U-Net for 2-class segmentation and ResNet for classification.
[10]	Extracts shape and color features for classification using Support Vector Machine (SVM), but traditional systems struggle to adapt to different microscope settings and image variations.
[33, 34, 17, 35]	The whole image is input into various CNN models for classification however, CNNs often underperform as they overlook global information essential for accuracy.
[20]	Uses Deep ViT for classification to address CNN limitations.
[36]	Employs DeepLabv3+ for 2-class segmentation and then uses AlexNet for classification.
[18]	SqueezeNet is used for classification after 2-class segmentation using a non-local average filter for thresholding.
[24]	Uses DETR-based models for classification, attribute extraction, and segmentation, but lacks explainability for guiding these tasks.
[1]	Utilizes deep CNNs (e.g., VGG16, ResNet, DenseNet) for MAttrP but lacks global context modeling and does not incorporate explainability to guide segmentation or classification.

Table 2.2: Literature Review Summary: Part 2.

Chapter 3

Methodology

This chapter presents the proposed methodology, which comprises two works, with the second being an extension of the first. The initial work focuses on integrating segmentation and classification, aiming to design a system that effectively combines the strengths of CNNs and Transformers. The second work builds upon this foundation by incorporating explainability and emulating pathologist-like reasoning, while further enhancing the decoder design introduced in the first work. Together, these efforts maximize the potential of the CNN-Transformer combination.

3.1 Proposed Work I: SWASTIC (Synergistic WBC Segmentation and Classification Integrated Computational System)

This section introduces our proposed system, SWASTIC, which is structured into three key stages. The initial stage integrates a Transformer encoder with a Feature Synergy Decoder (FSD) to carry out WBC segmentation. The FSD is instrumental in harmonizing CNN features within the Transformer-based architecture. In the second stage, the classification

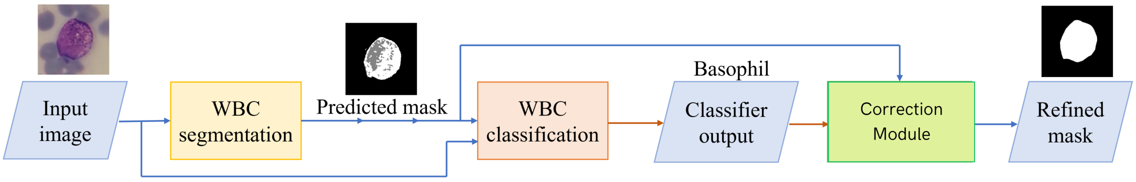

Figure 3.1: Flow graph of proposed system I: SWASTIC.

network, based on the Swin T, processes both the segmented regions and the original input image to identify various WBC types. This segmentation information enables the classification network to focus on essential WBC regions, such as nuclei and cytoplasm. Recognizing that the initial segmentation may not always be accurate, the final stage adjusts and improves the segmentation outcomes by leveraging insights from the WBC classification. The detailed flow of SWASTIC is illustrated in Fig. 3.1.

3.1.1 SWASTIC: Segmentation Module

This section outlines our segmentation framework, which combines CNN and Transformer capabilities to achieve precise WBC segmentation. Our approach performs 3-class segmentation, enabling straightforward adaptation to 2-class segmentation. It features a Transformer encoder for feature extraction and an FSD that employs Spatial Texture Refinement (STR) and Progressive Feature Aggregation (PFA) modules. The STR enhances local texture details using convolutional methods, while the PFA integrates multiple-level features. Refer to Fig. 3.2 for a diagrammatic representation of the module. The following subsections provide details.

3.1.1.1 Transformer Encoder

The Transformer encoder is based on a pyramid architecture inspired by PVTv2 [38], using convolutional operations in place of traditional positional encoding. For an input image I

of dimensions $H \times W \times 3$, the encoder produces hierarchical features $\{F_i \mid i = 1, \ldots, 4\}$. Each feature F_i has the resolution:

$$\text{Resolution of } F_i = \left[\frac{H}{2^{k-1}}, \frac{W}{2^{k-1}}, D_i\right],$$

where $k = \{3, 4, 5, 6\}$ corresponds to feature levels, enabling multiscale feature extraction.

3.1.1.2 Feature Synergy Decoder (FSD)

The FSD integrates local and global information through the STR and PFA modules, enhancing the segmentation by preserving fine details and capturing broader contextual information.

3.1.1.2.1 Spatial Texture Refinement (STR) The STR module refines local details in the feature maps using convolution operations. For an input feature F_i at level i with channel dimension D_i, the refinement process is described by:

$$F_i^{\text{STR}} = \text{ReLU}\left(\text{Conv}_{2D}\left(\text{ReLU}\left(\text{Conv}_{2D}(F_i)\right)\right)\right),$$

where Conv_{2D} is a 2D convolution preserving spatial resolution while adjusting channel dimensions. The STR module outputs enhanced features F_i^{STR} for further processing.

3.1.1.2.2 Progressive Feature Aggregation (PFA) The PFA module combines refined features across consecutive levels to merge local details and global context. Let F_i^{STR} and F_{i-1}^{STR} be the refined features from levels i and $i-1$, respectively. These features are concatenated and processed as:

$$F_i^{\text{PFA}} = \text{Conv}_{2D}\left(\text{Concat}\left(F_i^{\text{STR}}, F_{i-1}^{\text{STR}}\right)\right),$$

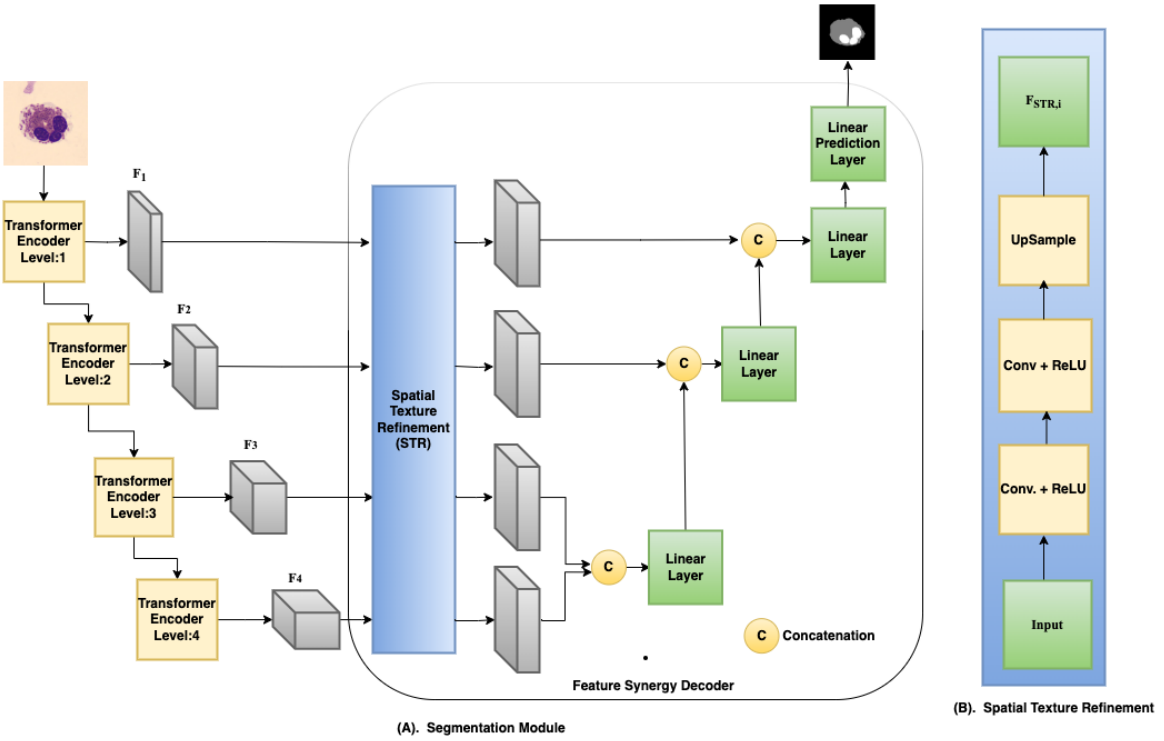

Figure 3.2: (A) Segmentation Module of SWASTIC. (B) Spatial Texture Refinement block.

where Concat denotes concatenation along the channel dimension. The resulting feature F_i^{PFA} maintains the channel size D_i, ensuring consistency for subsequent operations. By combining the outputs of STR and PFA, the FSD effectively unifies local and global insights, leveraging CNNs for texture detail and Transformers for contextual understanding.

3.1.2 SWASTIC: Classification Module

Research has shown that focusing on important WBC areas, such as the cytoplasm and nucleus, leads to more accurate classification [18]. Thus, we isolate and emphasise the crucial WBC structures using the segmentation result from the previous step. This procedure, called Target Area (TA) extraction, modifies the input image so that key WBC parts are highlighted while other regions are suppressed. The processed image is then passed to the Swin T [39] for classification into one of five WBC categories. The Swin T is well-suited for this task

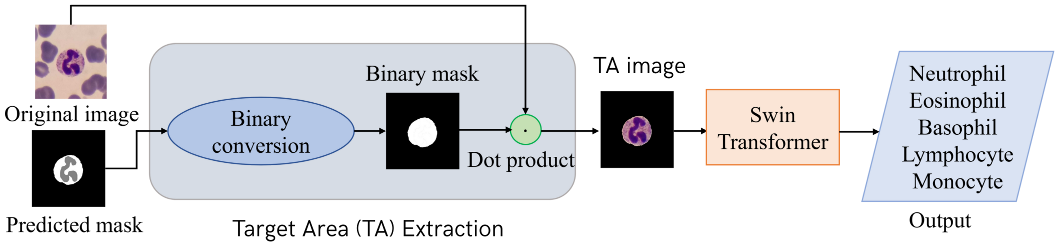

Figure 3.3: Classification Module of SWASTIC.

due to its hierarchical feature extraction and its ability to preserve global context. Fig. 3.3 illustrates the workflow of our classification framework. Detailed steps are outlined below.

3.1.2.1 Target Area (TA) Extraction

In this step, we isolate the nucleus and cytoplasm, which are the relevant WBC structures for classification, while eliminating the background. A binary mask $T(a, b)$ is created to identify the WBC structures, with relevant regions assigned a value of 1, and non-relevant regions set to 0. This binary mask is derived from the segmentation mask $S(a, b)$ from Section 3.1.1 and is defined as:

$$T(a,b) = \begin{cases} 0, & \text{when } S(a,b) = 0 \\ 1, & \text{when } S(a,b) \in \{128, 255\} \end{cases}$$

Where $S(a, b) = 128$ and $S(a, b) = 255$ correspond to the cytoplasm and nucleus areas, respectively, and $S(a, b) = 0$ corresponds to the background.

Using this binary mask T, the input image $I(a, b)$ is modified to emphasize the relevant WBC parts, while suppressing the background pixels. The transformed image, $A(a, b)$, is calculated as:

$$A(a, b) = I(a, b) \odot T(a, b),$$

where $\odot$ denotes pixel-wise multiplication. This operation ensures that the classification model focuses on the significant WBC structures.

3.1.2.2 Swin Transformer-Based Classification

The processed image $A(a,b)$ is provided as input to the Swin T [39], which classifies it into one of the five WBC types. Leveraging its hierarchical feature extraction and shifted window mechanisms, the Swin T is very useful for examining complex WBC structures, including the cytoplasm and nucleus, because it is excellent at capturing both global and detailed contextual characteristics.

3.1.3 SWASTIC: Correction Module

A feedback-based correction strategy is employed to address inaccuracies in the initial segmentation by utilizing classification outcomes. This method proves especially effective for basophils, where dense, darkly stained cytoplasmic granules frequently obscure the lobed nucleus, leading to misclassification [28, 1]. By incorporating the classifier's reliable predictions (refer to Table 4.5), ambiguous regions are reassigned with greater accuracy, as illustrated in Fig. 3.1.

3.1.4 Loss Functions Utilized in SWASTIC Training

The Proposed System I employs module-specific loss functions to optimize each task. For segmentation (refer to Section 3.1.1), we use the Dice loss, defined as

$$\mathcal{L}_{\text{Overlap}} = 1 - \frac{2\sum_{j=1}^{M} \hat{y}_j \, y_j}{\sum_{j=1}^{M} \hat{y}_j + \sum_{j=1}^{M} y_j + \delta}, \tag{3.1}$$

where M represents the total number of pixels, $\hat{y}_j \in [0,1]$ denotes the predicted probability for the j-th pixel, $y_j \in \{0,1\}$ indicates the corresponding ground-truth label, and δ is a small constant added to prevent division by zero. Minimizing $\mathcal{L}_{\text{Overlap}}$ increases the alignment between the ground truth and the predicted output, effectively reducing both false negative and false positive rates. For the classification module (Section 3.1.2), we minimize the multiclass cross-entropy loss (CEL):

$$\mathcal{L}_{\text{CEL}} = -\frac{1}{M}\sum_{k=1}^{M}\sum_{l=1}^{K} t_{k,l}\,\log(\hat{t}_{k,l}), \tag{3.2}$$

where M represents the total number of samples, K is the number of categories, $t_{k,l}$ indicates the one-hot encoded ground-truth label for the k-th sample and the l-th class, and $\hat{t}_{k,l}$ is the model's predicted probability for the same. This loss penalizes discrepancies between predicted and actual class distributions, ensuring that the classifier learns to assign high probability to the correct labels.

3.2 Proposed Work II: LEUCOSIGHT (WBC Insights)

This work builds upon the previous SWASTIC framework, with a primary focus on incorporating interpretability to guide decision-making, simulate clinician reasoning, and enhance trustworthiness for real-world deployment. A significant limitation of the prior work was the lack of explainability and insufficient attention to how WBC-related attributes could be integrated with segmentation and classification tasks. This work seeks to address these shortcomings.

The architectural design of the system is inspired by the behavior of pathologists when diagnosing histopathology slides. Typically, they first localize the WBCs, then examine their morphological features, and based on this, classify the WBCs. Following a similar reasoning, our segmentation module first localizes the WBCs and provides structural cues to the MAttrP module. The design of the segmentation module advances the CNN and transformer combination used in prior work. It incorporates attention mechanisms to integrate features across different hierarchies while utilizing convolution operations to capture local learning, as attention is typically focused on global learning.

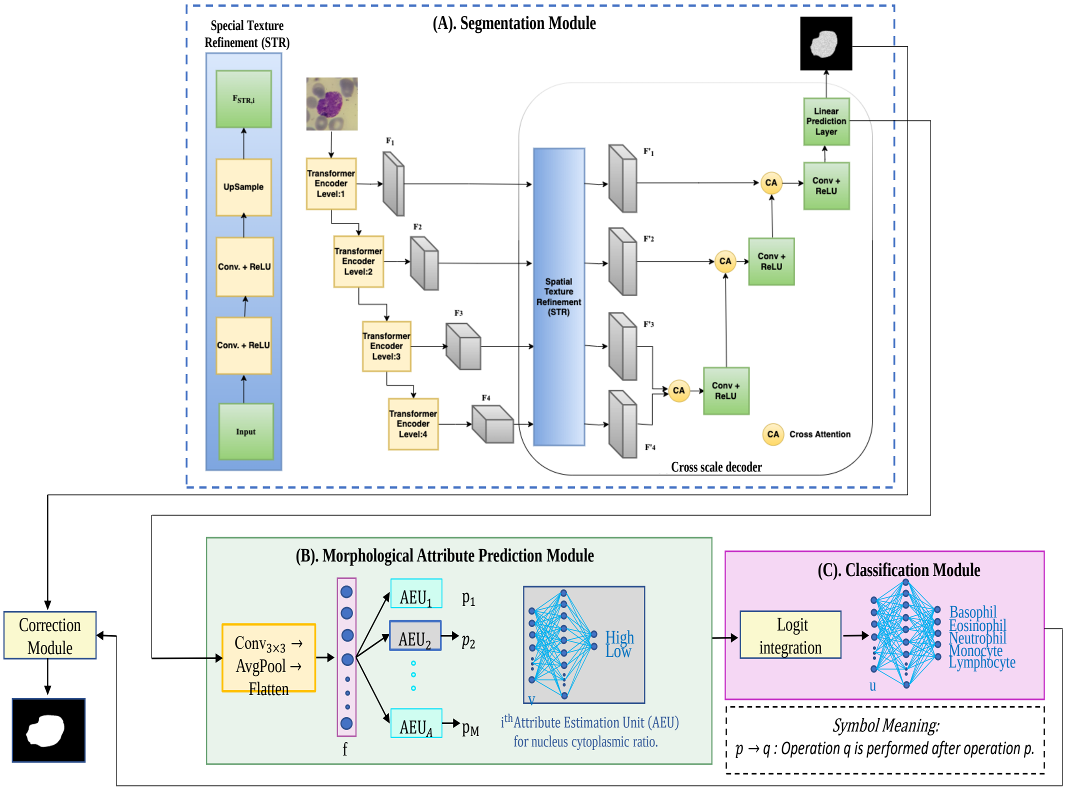
(A). Segmentation Module
Special Texture Refinement (STR)
$F_{STR,i}$
UpSample
Conv. + ReLU
Conv. + ReLU
Input
Transformer Encoder Level:1
Transformer Encoder Level:2
Transformer Encoder Level:3
Transformer Encoder Level:4
F_1
F2
F3
F4
Spatial Texture Refinement (STR)
F'1
F'2
F'3
F'4
CA
Conv + ReLU
Linear Prediction Layer
CA Cross Attention
Cross scale decoder
(B). Morphological Attribute Prediction Module
Conv$_{3\times3}$ → AvgPool → Flatten
f
AEU$_1$
AEU$_2$
AEU$_A$
p$_1$
p_2
p$_M$
v
High
Low
ithAttribute Estimation Unit (AEU) for nucleus cytoplasmic ratio.
(C). Classification Module
Logit integration
u
Basophil
Eosinophil
Neutrophil
Monocyte
Lymphocyte
Symbol Meaning:
$p \rightarrow q$: Operation q is performed after operation p.
Correction Module

Figure 3.4: Proposed Work II: LEUCOSIGHT. (A) Segmentation Module, (B) Morphological Attribute Prediction Module, (C) Classification Module.

This approach enhances efficacy. Moreover, the MAttrP module learns relevant WBC attributes, and its output is used for WBC classification, simulating clinical behavior. Finally, the classification output is leveraged to correct segmentation errors. In this way, a proper synergy between all WBC analysis tasks is established. Fig. 3.4 illustrates the workflow of the LEUCOSIGHT system. The following subsections will delve deeper into the system's design.

3.2.1 LEUCOSIGHT: WBC Segmentation Module

This segmentation approach integrates CNNs and transformers to achieve the desired 3-class segmentation of WBCs, which can be effortlessly adapted to derive a 2-class segmentation. The ViT encoder, influenced by PVTv2 [38], extracts features at multiple resolutions. These features are decoded using a cross-scale decoder (CSD), which consists of two main stages: STR and progressive contextual integration (PCI).

3.2.1.1 Transformer Encoder

Given an input image I of size $H \times W \times 3$, the transformer encoder generates feature maps at different scales:

$$F_l \in \mathbb{R}^{\frac{H}{2^l} \times \frac{W}{2^l} \times C_l}, \quad l \in \{2, 3, 4, 5\}.$$

Here, l indicates the feature level, with higher levels capturing more abstract spatial patterns and contextual information. Convolutional layers replace positional encoding to retain local details effectively.

3.2.1.2 Cross-Scale Decoder

By utilising the advantages of transformers and CNNs, the CSD efficiently integrates contextual and intricate learning strategies while processing the feature maps in two phases, STR and PCI.

3.2.1.2.1 Spatial Texture Refinement: Each feature map F_l undergoes refinement through a series of convolution operations, enabling each pixel to interact with its neighboring pixels. This process preserves fine details within the feature map. The refined features are then upsampled to match the highest resolution:

$$F'_l = \text{Upsample}\Big(\text{Conv}_{1\times1}(\text{Conv}_{1\times1}(F_l))\Big), \quad F'_l \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_5}. \tag{3.3}$$

The refined feature maps F'_2, F'_3, F'_4, F'_5 are subsequently passed to the PCI stage.

3.2.1.2.2 Progressive Contextual Integration: The PCI module incrementally integrates features from various levels, beginning with lower resolutions and progressing towards higher resolutions. It ensures that high-level features are guided by low-level features through attention, thereby capturing contextual information. Convolutional layers are then applied on top, helping retain fine details. By combining cross-attention (CA) [40] with convolutional operations, the module preserves local features while simultaneously incorporating global context. The following outlines the detailed steps.

1. **Integration of F'_5 and F'_4:**

   - Compute the cross-attention map:

   $$A_{54} = \text{CA}(F'_5, F'_4, F'_4), \tag{3.4}$$

   where the key and value are both represented by F'_4 and the query is F'_5.

   - Update F'_5:

   $$F_{54} = \gamma_{54} \cdot A_{54} + F'_5, \tag{3.5}$$

   where γ_{54} is a trainable weight.

   - Apply convolution operations:

   $$F'_{54} = \text{Conv}_{5\times5}(\text{Conv}_{3\times3}(F_{54})). \tag{3.6}$$

2. **Integration of F'_{54} and F'_3:**

- Compute the cross-attention map:

$$A_{543} = \text{CA}(F'_{54}, F'_3, F'_3). \tag{3.7}$$

- Update F'_{54}:

$$F_{543} = \gamma_{543} \cdot A_{543} + F'_{54}, \tag{3.8}$$

  where γ_{543} is a trainable weight.

- Apply convolution operations:

$$F'_{543} = \text{Conv}_{5\times5}(\text{Conv}_{3\times3}(F_{543})). \tag{3.9}$$

3. **Integration of F'_{543} and F'_2:**

- Compute the cross-attention map:

$$A_{5432} = \text{CA}(F'_{543}, F'_2, F'_2). \tag{3.10}$$

- Update F'_{543}:

$$F_{5432} = \gamma_{5432} \cdot A_{5432} + F'_{543}, \tag{3.11}$$

  where γ_{5432} is a trainable weight.

- Apply convolution operations:

$$F'_{5432} = \text{Conv}_{5\times5}(\text{Conv}_{3\times3}(F_{5432})). \tag{3.12}$$

3.2.1.2.3 Final Output: The final feature map F'_{5432} is transformed through a 1×1 convolution layer to adjust its channel dimensions:

$$F_{\text{out}} = \text{Conv}_{1\times1}(F'_{5432}), \quad F_{\text{out}} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_{\text{out}}} \tag{3.13}$$

After that, this output undergoes scaling to match the resolution of the original input:

$$S_{\text{output}} = \text{Upsample}(F_{\text{out}}). \tag{3.14}$$

The resulting segmentation map is ready for further analysis and classification.

3.2.2 LEUCOSIGHT: Morphological Attributes Prediction Module

The module for morphological attribute prediction leverages the segmentation results $\mathbf{S}_{\text{output}}$, produced by the CSD (refer to Section 3.2.1.2), to estimate M distinct attributes. The following steps outline the process:

3.2.2.1 Transformation of Segmentation Features

The segmentation map $\mathbf{S}_{\text{final}}$ with dimensions $H \times W \times C_s$ undergoes transformation to prepare for attribute estimation:

$$\mathbf{G} = \text{Conv}_{3\times3}(\mathbf{S}_{\text{final}}), \quad \mathbf{G} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C_t}. \tag{3.15}$$

3.2.2.2 Extraction of Global Descriptors

To summarize spatial information into global features, adaptive average pooling is applied:

$$\mathbf{R} = \text{AdaptiveAvgPool}(\mathbf{G}), \quad \mathbf{R} \in \mathbb{R}^{1\times1\times C_t}. \tag{3.16}$$

This operation creates a compact representation of the segmentation features. Flattening this pooled tensor yields a feature vector:

$$\mathbf{f} = \text{Flatten}(\mathbf{R}), \quad \mathbf{f} \in \mathbb{R}^{C_t}. \tag{3.17}$$

3.2.2.3 Attribute Estimation Unit

The feature vector $\mathbf{f}$ is processed by a series of independent units, each designed for predicting a specific attribute. Each unit comprises:

- **Intermediate Layer:** A dense layer with N_ℓ neurons, followed by a ReLU activation:

$$\mathbf{h}_i = \text{ReLU}(\mathbf{W}_{\ell,i}\mathbf{f} + \mathbf{b}_{\ell,i}), \quad \mathbf{h}_i \in \mathbb{R}^{N_\ell}. \tag{3.18}$$

  Here, $\mathbf{W}_{\ell,i}$ is a weight matrix of size $N_\ell \times C_t$, and $\mathbf{b}_{\ell,i}$ is a bias vector of length N_ℓ.

- **Output Layer:** A dense layer with N_i output neurons, where N_i is the number of classes for the i-th attribute. A softmax activation is used to generate probabilities:

$$\mathbf{p}_i = \text{Softmax}(\mathbf{W}_{o,i}\mathbf{h}_i + \mathbf{b}_{o,i}), \quad \mathbf{p}_i \in \mathbb{R}^{N_i}. \tag{3.19}$$

  Here, $\mathbf{W}_{o,i}$ is a weight matrix of dimensions $N_i \times N_\ell$, and $\mathbf{b}_{o,i}$ is a bias vector of length N_i.

3.2.2.4 Consolidation of Predictions

The predictions from all M attributes are aggregated into a set:

$$\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M\}. \tag{3.20}$$

3.2.3 LEUCOSIGHT: WBC Classification Module

The WBC classification component integrates the attribute logits predicted by the attribute estimation units (Section 3.2.2) to predict the WBC type. This approach ensures that the morphological characteristics of WBCs serve as the basis of classification judgements.

3.2.3.1 Logit Integration

Logits generated by each attribute estimation unit are concatenated into a single feature vector:

$$\mathbf{u} = \text{Concat}(\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M), \quad \mathbf{u} \in \mathbb{R}^{\sum_{i=1}^{M} N_i}, \tag{3.21}$$

where $\mathbf{p}_i \in \mathbb{R}^{N_i}$ is the output logits for the i-th attribute, and N_i is the number of classes associated with it.

3.2.3.2 Classification Block

The fully connected WBC classification block receives the concatenated feature vector $\mathbf{u}$ and includes the following:

- **Hidden Layer:** A dense layer with N_h neurons, activated using ReLU:

  $$\mathbf{h}_{\text{cls}} = \text{ReLU}(\mathbf{W}_h \mathbf{u} + \mathbf{b}_h), \quad \mathbf{h}_{\text{cls}} \in \mathbb{R}^{N_h}. \tag{3.22}$$

  Here, $\mathbf{W}_h$ is a weight matrix of size $N_h \times (\sum_{i=1}^{M} N_i)$, and $\mathbf{b}_h$ is a bias vector of length N_h.

- **Output Layer:** A dense layer with C neurons to estimate probabilities for C WBC classes:

  $$\mathbf{y}_{\text{wbc}} = \text{Softmax}(\mathbf{W}_c \mathbf{h}_{\text{cls}} + \mathbf{b}_c), \quad \mathbf{y}_{\text{wbc}} \in \mathbb{R}^{C}. \tag{3.23}$$

  Here, $\mathbf{W}_c$ is a weight matrix of size $C \times N_h$, and $\mathbf{b}_c$ is a bias vector of length C.

3.2.3.3 Final Decision

The predicted WBC type is obtained by selecting the class with the highest probability:

$$\hat{y}_{\text{cls}} = \arg\max(\mathbf{y}_{\text{wbc}}), \tag{3.24}$$

where $\hat{y}_{\text{cls}}$ represents the final classification output.

3.2.4 LEUCOSIGHT: WBC Correction Module

The initial segmentation process described in Section 3.2.1 may result in inaccuracies, particularly for basophils. These errors are addressed through a correction module. Distinct nuclei and cytoplasm characterize basophils, but their dense, dark-staining cytoplasmic granules often obscure the lobed nucleus, creating a blended appearance [28, 1]. To mitigate this challenge, segmentation inaccuracies are refined by incorporating insights from WBC classification outcomes. As demonstrated in Table 4.6, the classification model effectively distinguishes basophils due to their distinctive morphological features [1]. The system uses morphology-based classification to accurately identify basophils, reassigning any misclassified regions based on the classification results. This approach ensures a more precise depiction of cellular structures. Fig. 3.4 shows an example of this refinement procedure.

3.2.5 Loss Functions Utilized in LEUCOSIGHT Training

The proposed system II uses specialized loss functions tailored to each module to achieve optimal performance. The segmentation module (Section 3.2.1) leverages Dice Loss to reduce overlap inaccuracies and ensure accurate segmentation results. For the MAttrP module (Section 3.2.2), CEL addresses prediction errors, while Supervised Contrastive Loss (SupConLoss) [41] improves feature learning by enhancing intra-class consistency and inter-class differentiation in the attribute estimation units (Section 3.2.2.3). SupConLoss is defined as:

$$\mathcal{L}_{\text{SupCon}} = \sum_{x \in S} \frac{-1}{|Q(x)|} \sum_{q \in Q(x)} \log \frac{\exp\left(\frac{\mathbf{f}_x \cdot \mathbf{f}_q}{\theta}\right)}{\sum_{r \in R(x)} \exp\left(\frac{\mathbf{f}_x \cdot \mathbf{f}_r}{\theta}\right)}, \tag{3.25}$$

Where: S is the set of all samples, $Q(x)$ is the set of positive samples for anchor x, $R(x)$ is the set of all samples except x, $\mathbf{f}_x$ represents the feature representation of the anchor sample

x, $\mathbf{f}_q$ represents the feature representation of the positive sample q and θ is the temperature parameter. This loss increases inter-class differentiation by pushing away features from different classes, while improving intra-class consistency by bringing together characteristics of the same class in the embedding space. This dual emphasis on feature alignment and separation makes it well-suited for attribute prediction tasks. The WBC classification module (Section 3.2.3) also relies on CEL to refine classification accuracy. Modules can be trained individually or in combination, depending on the availability of ground truth datasets, as detailed in Section 4.3.

Chapter 4

Results

4.1 Experimental Settings

Our proposed work I, SWASTIC, carries out both WBC segmentation and classification, whereas our proposed work II, LEUCOSIGHT, expands its capabilities to incorporate MAttrP in addition to WBC segmentation and classification. Consequently, our experiments leverage datasets that provide detailed ground-truth labels for these tasks. The study utilized three publicly accessible WBC benchmark datasets: Raabin, LISC, and WBCAtt (see Table 4.1). Two distinct segmentation tasks were performed. In the 3-class segmentation task, the ground-truth annotations were adjusted to categorize the background, nuclei, and cytoplasm, corresponding to black, gray, and white regions, respectively. Conversely, the 2-class segmentation task simplified the ground-truth annotations by combining the background with the cytoplasm as black and representing nuclei as white. The performance of WBC segmentation was assessed using metrics such as the mean Dice Similarity Coefficient (DSC), mean Intersection over Union (IoU), and accuracy (Acc). The evaluation of WBC classification and MAttrP employed several metrics, including precision (Pre), specificity (Spec), accuracy, recall (Rec), and the F-measure (F-m). These abbreviations are consistently referenced across Sections 4.4, 4.5, 4.6, and 4.7. Furthermore, it is crucial to highlight that our system's performance was benchmarked against state-of-the-art (SOTA) methods, ensuring identical training and testing conditions for a fair comparison.

Table 4.1: Datasets utilized in the study.

Dataset	Description
WBCAtt	• 10,298 images (1024 x 716 px) • Cells: Monocytes (1,420), Lymphocytes (1,214), Basophils (1,218), Eosinophils (3,117), Neutrophils (3,392) • Annotated with 10 attributes: Granule type/color, Cytoplasm vacuole/color/texture, Chromatin density, NC ratio, Nucleus shape, Cell shape/size
Raabin	• 1,145 images (575x575 px) • Cells: Neutrophils (242), Eosinophils (201), Basophils (218), Lymphocytes (242), Monocytes (242) • Masks: Background (black), Cytoplasm (white), Nucleus (grey) • Classification (Train/Test split): Monocytes (561/234), Eosinophils (744/322), Neutrophils (6,231/2,660) Lymphocytes (2,427/1,034), Basophils (212/89)
LISC	• 242 images (720x576 px) • Cells: Monocytes (48), Eosinophils (39), Neutrophils (50), Lymphocytes (52), Basophils (53) • Masks: Background (black), Cytoplasm (grey), Nucleus (light grey)

4.2 Implementation Settings

The experimental setup leveraged PyTorch and was executed on a system equipped with an NVIDIA GeForce RTX 3080 GPU, an Intel Core i7-10700K processor, and 32 GB of RAM. Both proposed approaches were optimized using the AdamW optimizer [42] with an initial learning rate of 0.0001. The training process spanned 200 epochs with a batch size of 32. Additional information on training and evaluation protocols can be found in Section 4.3.

4.3 Training and Testing Settings

Proposed System I

All images were standardized by resizing them to a resolution of 224×224 pixels as part of the preprocessing step. For the experiments detailed in Sections 4.4 and 4.5, the following workflow was employed: on the Raabin dataset, the segmentation module was trained using 912 images with Dice loss and tested on 233 images. Subsequently, with the segmentation weights frozen, the classification module was trained on 10,175 images using CEL and evaluated on 4,339 held-out samples. The same approach was applied to the LISC dataset, where the data was randomly divided into 20% for testing and 80% for training. Freezing the segmentation network ensured that classification learning relied exclusively on features extracted from accurately segmented regions.

Proposed System II

The images were initially scaled to a dimension of 224×224 pixels. For the experiments described in Sections 4.4, 4.5, and 4.6, we proceeded as follows: On the Raabin dataset, the segmentation network was trained on 912 images using the Dice loss and evaluated on 233 test images. With the segmentation weights then frozen, the MAttrP and classification networks were jointly trained on 10,175 images using CEL for the WBC class labels, and their performance was measured on 4,339 held-out samples. For the LISC dataset, we randomly split the data into 20% for testing and 80% for training. Before being frozen, the segmentation network was trained using dice loss; subsequently, the MAttrP and classification networks were trained together, again using CEL for the WBC labels, and evaluated on the reserved test set. In the WBCAtt dataset, we followed its predefined train–test partitions. The segmentation and MAttrP networks were trained simultaneously using a combination of SupConLoss and CEL for the attribute labels, then frozen. Finally, the classification network was trained and assessed on the designated test set using CEL for the WBC class labels.

In Section 4.7, we further investigated cross-dataset generalization by training on the WBCAtt dataset and fine-tuning on Raabin, as well as training on WBCAtt and LISC interchangeably. The same training and evaluation protocols mentioned above were consistently applied throughout these experiments.

4.4 Performance Evaluation on WBC Segmentation

This subsection compares the efficacy of our suggested WBC segmentation modules incorporated into LEUCOSIGHT and SWASTIC with SOTA methods. Table 4.2 provides a comprehensive quantitative analysis, showcasing the enhancements achieved. Additionally, the qualitative performance of the proposed approaches is illustrated in Fig. 4.2 and 4.1. As evidenced in Table 4.2, our proposed system consistently outperforms current SOTA approaches across both 2-class and 3-class segmentation tasks. The remarkable performance in 2-class segmentation arises from the synergistic strengths of Transformers and CNNs. While CNNs excel in capturing intricate local structural details, Transformers are adept at modeling broader contextual dependencies. Conversely, traditional image processing techniques, as discussed in [19, 29, 30], often struggle to generalize effectively across a variety of microscopic conditions and image variations [10], resulting in limited utility. Similarly, CNN-based systems such as Mask R-CNN, SqueezeNet, and U-Net++ [43, 44, 18, 3, 31], although proficient in extracting localized features, cannot account for long-range dependencies, which compromises their segmentation accuracy.

Table 4.2: Comparative Analysis of WBC Segmentation Performance: Proposed Works vs. SOTA methods.

2-class segmentation

System	**Raabin Dataset**			**System**	**LISC Dataset**		
	DSC	**IOU**	**Acc**		**DSC**	**IOU**	**Acc**
[43]	0.9725	0.9462	-	-	-	-	-
[31]	0.9620	0.9283	-	-	-	-	-
[44]	0.9203	0.8512	-		-	-	-
[29]	0.9535	0.9114	-	[19]	0.8982	0.8760	0.9590
[10]	0.9668	0.9372	-	[30]	0.9029	0.8980	0.9793
[3]	0.9472	0.9218	0.9893	[18]	0.8920	0.8973	0.9684
PWI	**0.9819**	**0.9657**	**0.9960**	**PWI**	**0.9342**	**0.9083**	**0.9887**
PWII	**0.9897**	**0.9761**	**0.9985**	**PWII**	**0.9510**	**0.9243**	**0.9937**

3-class segmentation

System	**Raabin Dataset**			**System**	**LISC Dataset**		
	DSC	**IOU**	**Acc**		**DSC**	**IOU**	**Acc**
SNet [32]	0.8290	0.8271	0.9425	SNet [32]	0.7689	0.7513	0.9408
UNet [32]	0.8443	0.8400	0.9679	UNet [32]	0.7875	0.7692	0.9588
CNN [32]	0.7818	0.7737	0.9092	CNN [32]	0.7183	0.7110	0.9190
PWIWCM	0.8932	0.8561	0.9812	PWIWCM	0.8120	0.7474	0.9972
PWI	**0.9385**	**0.9211**	**0.9915**	**PWI**	**0.8590**	**0.8005**	**0.9981**
PWIIWCM	0.9315	0.9265	0.9912	PWIIWCM	0.8547	0.8094	0.9930
PWII	**0.9510**	**0.9348**	**0.9967**	**PWII**	**0.8813**	**0.8265**	**0.9994**

‘SNet: SegNet’, ‘PWI: Proposed Work I (SWASTIC)’, ‘PWII: Proposed Work II (LEUCOSIGHT)’, ‘PWIWCM: Proposed Work I without correction module’, ‘PWIIWCM: Proposed Work II without correction module’, Hybrid approaches, including our earlier system SWASTIC, achieve notable improvements by combining CNNs for extracting localized features with Transformers for capturing global context. However, SWASTIC's reliance on conventional feature-merging techniques, where multi-level features are aggregated and processed through convolution, limits its ability to harmonize morphological details with higher-level context. In contrast, LEUCOSIGHT adopts an attention-based feature-merging mechanism, which prioritizes relevant features while seamlessly aligning intricate details with global patterns. This design enhances the interaction between CNNs and Transformers, yielding superior segmentation results. In the domain of 3-class segmentation, which remains less explored in existing literature, our system demonstrates a clear edge over other methods. Architectures such as U-Net, SegNet, and similar CNN-based systems [32] face challenges in addressing long-range dependencies. While SWASTIC partially mitigates these limitations, its traditional feature-merging strategy restricts its ability to fully harness the CNN-Transformer combination. LEUCOSIGHT, with its attention-driven feature-merging approach, effectively bridges this gap by aligning morphological details with comprehensive contextual representations, leading to enhanced segmentation accuracy.

A standout feature of our work is the integration of WBC classification, which significantly bolsters segmentation performance. To evaluate this impact, we compared our methods against baseline systems, PWIWCM and PWIIWCM, which perform segmentation independently of classification insights. As shown in Table 4.2, both SWASTIC and LEUCOSIGHT consistently outperform these baselines, underscoring the pivotal role of classification-driven refinement in improving segmentation outcomes. This enhancement is visually evident in Fig. 4.2, where LEUCOSIGHT effectively corrects segmentation errors in the basophil class during 3-class segmentation, demonstrating the effectiveness of classification-based adjustments.

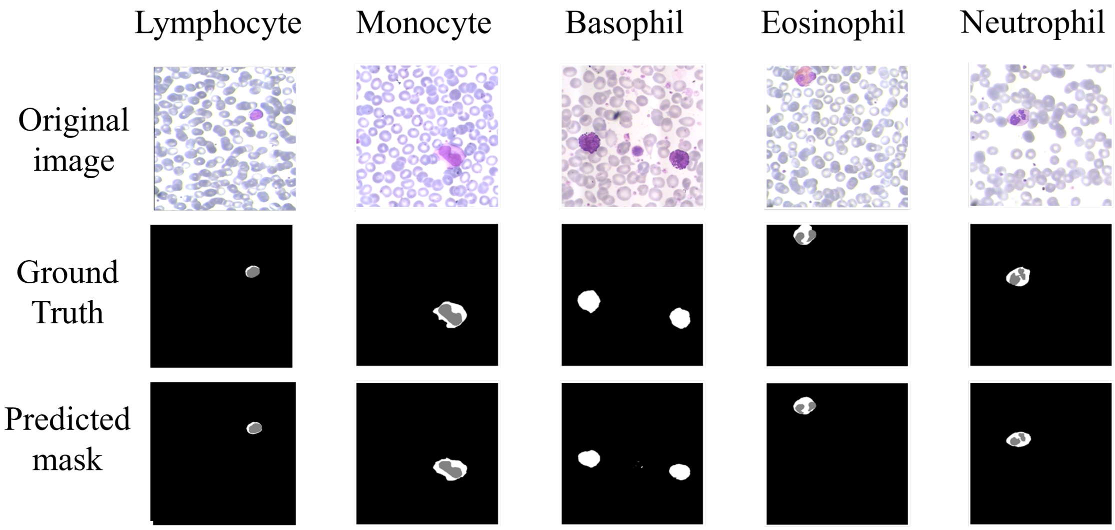

Figure 4.1: Qualitative outcomes of SWASTIC for 3-class segmentation.

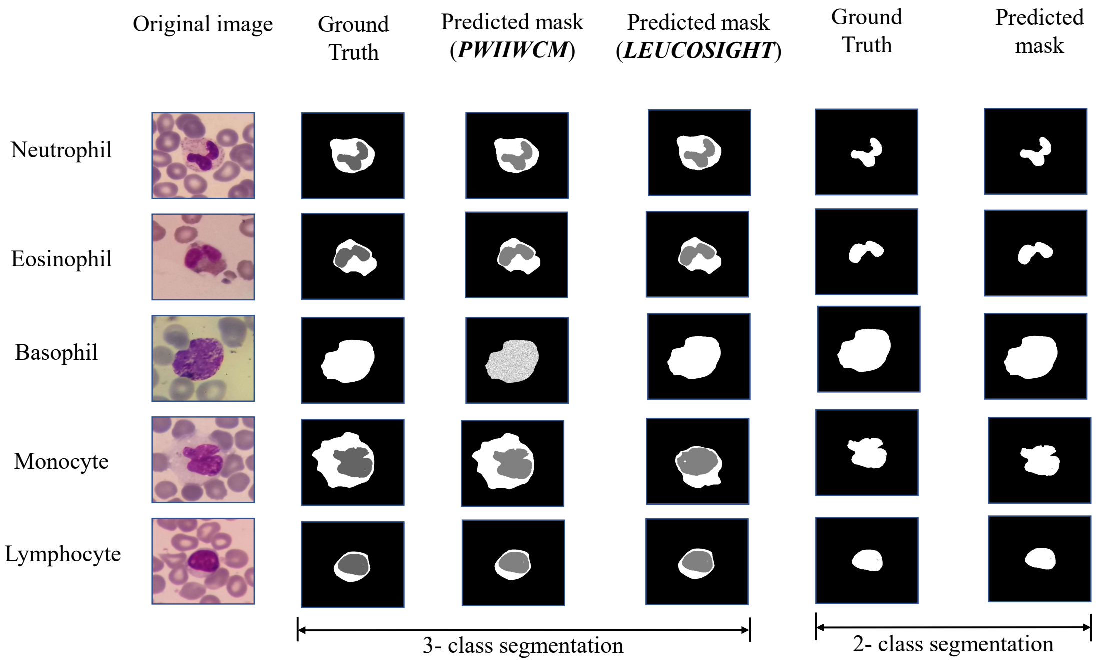

Figure 4.2: Qualitative outcomes of LEUCOSIGHT for both 2-class and 3-class segmentation.

4.5 Performance Evaluation on WBC Classification

We assess the effectiveness of our proposed modules for WBC classification, integrated into SWASTIC and LEUCOSIGHT, by comparing them with existing SOTA methods, as summarized in Table.4.3 and 4.4. The results highlight the substantial improvements achieved by our systems over prior approaches.

Existing methods, such as those outlined in [10], combine image features with SVM. However, these approaches often fail to generalize effectively across varied datasets and conditions. Similarly, CNN-based systems [33, 34, 17], including architectures like DenseNet, ShuffleNet, ResNet-50, ConvNeXt, GoogLeNet, and MobileNet, focus on WBC image classification but cannot capture the global contextual information essential for precise predictions. Transformer-based models address this limitation by introducing mechanisms to analyze broader contexts. Models like ViT and its derivatives [20, 24] combine attention mechanisms with convolutional layers to enhance classification accuracy. The Swin-T further advances this approach by employing hierarchical feature extraction, effectively capturing both local and global contexts [39]. Additionally, integrated systems such as those proposed in [18, 15] highlight the critical role of segmentation as a supporting task in improving classification outcomes. Hybrid systems like SWASTIC, an earlier contribution, combine CNNs and Transformers to generate detailed segmentation masks that inform classification. However, SWASTIC relies on traditional feature stacking techniques, which limit its ability to preserve nuanced local contextual information. In contrast, our extended system, LEUCOSIGHT, adopts a hierarchical feature-merging strategy using cross-attention mechanisms, complemented by convolutional layers to retain local details. This approach eliminates the limitations of simple feature stacking, ensuring an optimal integration of global and local information.

Table 4.3: Comparative Analysis of WBC Classification Performance: Proposed Works vs. SOTA methods.

System	**Raabin Dataset**				**System**	**LISC Dataset**			
	Acc	**Pre**	**Rec**	**F-m**		**Acc**	**Pre**	**Rec**	**F-m**
VGG16	0.8901	0.8503	0.8602	0.8554	VGG16	0.8652	0.8401	0.8453	0.8426
Tavakoli Algo. [10]	0.9473	-	-	-	Tavakoli Algo. [10]	0.9224	-	-	-
DenseNet	0.9105	0.8754	0.8806	0.8778	DenseNet	0.8853	0.8605	0.8653	0.8627
ResNet50	0.9357	0.9152	0.9203	0.9178	ResNet50	0.9106	0.8953	0.9002	0.8977
[33]	0.9525	0.9043	0.9342	0.9189	[33]	0.9324	0.9295	0.9272	0.9284
ShuffleNet	0.9258	0.9004	0.9053	0.9027	ShuffleNet	0.9005	0.8806	0.8852	0.8826
[34]	0.9283	-	-	-	[34]	0.8754	-	-	-
Swin T.	0.9703	0.9704	0.9705	0.9706	Swin T.	0.9407	0.9305	0.9353	0.9327
ViT	0.9405	0.9304	0.9352	0.9328	ViT	0.9206	0.9102	0.9154	0.9127
[20]	0.9556	0.9403	0.9452	0.9428	[17]	0.9745	0.9712	0.9645	0.9678
ConvNeXT	0.9707	0.9603	0.9654	0.9628	ConvNeXT	0.9609	0.9504	0.9553	0.9527
MobileNet	0.9026	0.9089	0.9101	0.9094	MobileNet	0.8862	0.8889	0.8901	0.8894
[18]	0.9665	0.9604	0.9553	0.9578	[18]	0.9686	0.9425	0.9493	0.9458
PWI	**0.9854**	**0.9705**	**0.9834**	**0.9768**	**PWI**	**0.9836**	**0.9835**	**0.9804**	**0.9818**
PWIIWAPM	0.9821	0.9669	0.9721	0.9694	PWIIWAPM	0.9794	0.9737	0.9728	0.9727
PWII	**0.9913**	**0.9756**	**0.9852**	**0.9803**	**PWII**	**0.9894**	**0.9847**	**0.9828**	**0.9837**

‘PWI: Proposed Work I (SWASTIC)’, ‘PWII: Proposed Work II (LEUCOSIGHT)’, ‘PWIIWAPM: Proposed Work II without Attribute Prediction module’

Beyond merely segmenting regions of interest, the segmentation outputs in LEUCOSIGHT provide structural cues that aid in learning MAtts, effectively emulating the diagnostic reasoning of pathologists. It is also evident that segmentation alone does not significantly enhance classification, as shown by the suboptimal performance of the baseline PWIIWAPM system in Table 4.3. Instead, utilizing segmentation outputs to predict key MAtts crucial to

WBC classification leads to markedly improved results.

Table 4.4: Comparative Analysis of WBC Segmentation Performance: Proposed Works vs. SOTA methods.

System	**WBCAtt Dataset**			
	Acc	**Pre**	**Rec**	**F-m**
VGG16	0.8804	0.8703	0.8752	0.8728
Tavakoli Algo. [10]	0.8806	0.8853	0.8814	0.8832
DenseNet	0.9007	0.8904	0.8925	0.8916
ResNet50	0.9409	0.9206	0.9253	0.9227
Faster-RCNN	0.9337	0.9375	0.9332	0.9354
ShuffleNet	0.9109	0.9003	0.9025	0.9014
GoogLeNet	0.8932	0.8853	0.8894	0.8876
Swin T.	0.9609	0.9406	0.9502	0.9448
ViT	0.9308	0.9205	0.9284	0.9246
CUSS-Net [15]	0.9637	0.9623	0.9626	0.9628
ConvNeXT	0.9653	0.9557	0.9604	0.9578
MobileNet	0.8991	0.9015	0.9039	0.9036
HemaX [24]	0.9624	0.9545	0.9573	0.9558
PWII	**0.9896**	**0.9839**	**0.9864**	**0.9851**

'PWII: Proposed Work II (LEUCOSIGHT)'

For the Raabin and LISC datasets, where ground truth labels for MAtts are unavailable, our system still demonstrates superior performance. During training, the MAttrP module, which operates jointly with the classification module while keeping the segmentation module frozen, learns latent morphological features by leveraging structural information from segmentation outputs and minimizing classification loss. Despite the absence of explicit morphology labels, the MAttrP module effectively captures meaningful representations as

Table 4.5: The experimental results for individual WBC classification classes provide a detailed view of SWASTIC overall classification performance after 3-class segmentation as summarized in Table 4.3.

Class	Raabin				LISC			
	Acc	**Pre**	**Rec**	**F-m**	**Acc**	**Pre**	**Rec**	**F-m**
M	0.990	0.906	0.914	0.910	0.983	0.917	1.000	0.957
L	0.989	0.977	0.979	0.978	1.000	1.000	1.000	1.000
B	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
E	0.996	0.966	0.984	0.975	0.983	1.000	0.857	0.923
N	0.994	0.997	0.993	0.995	1.000	1.000	1.000	1.000
Overall	**0.985**	**0.970**	**0.983**	**0.972**	**0.983**	**0.983**	**0.980**	**0.979**

M: Monocyte, L: Lymphocyte, B: Basophil, E: Eosinophil, N: Neutrophil

part of the end-to-end framework. On the WBCAtt dataset, the segmentation module indirectly gains valuable feature representations through weak supervision provided by the MAttrP head, which relies on spatial patterns. In both scenarios, proxy supervision derived from gradients of upstream or downstream tasks compensates for the lack of direct supervision, allowing robust feature learning and consistently strong performance.

4.6 Performance Evaluation on WBC Morphological Attributes

We evaluate the performance of our system, LEUCOSIGHT, by benchmarking it against SOTA methods. The attribute prediction framework, illustrated in Fig. 4.3, incorporates a variety of encoders, including CNN-based models (ResNet, DenseNet, MobileNet, Shuf-

Table 4.6: The experimental results for individual WBC classification classes provide a detailed view of LEUCOSIGHT overall classification performance as summarized in Table 4.3.

Class	**Raabin**				**LISC**			
	Acc	**Pre**	**Rec**	**F-m**	**Acc**	**Pre**	**Rec**	**F-m**
M	0.982	0.945	0.960	0.952	0.988	0.915	0.970	0.942
L	0.989	0.975	0.975	0.975	0.991	0.980	0.980	0.980
B	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
E	0.993	0.961	0.980	0.970	0.987	0.930	0.950	0.940
N	0.992	0.996	0.990	0.993	0.998	0.998	0.998	0.998
Overall	**0.991**	**0.976**	**0.985**	**0.980**	**0.989**	**0.985**	**0.983**	**0.984**

M: Monocyte, L: Lymphocyte, B: Basophil, E: Eosinophil, N: Neutrophil

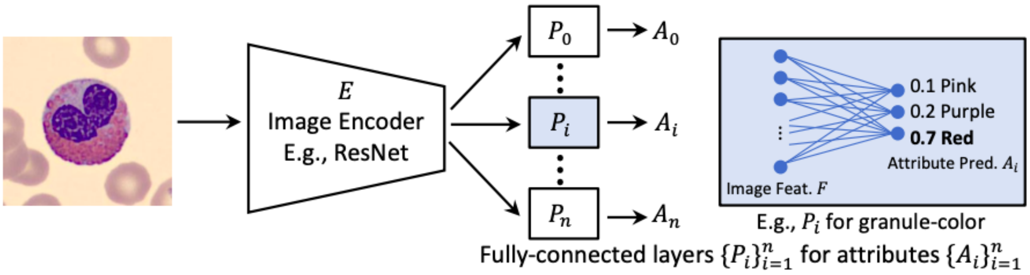

Figure 4.3: Attribute Prediction System [1].

fleNet, ConvNeXt, and VGG) and Transformer-based models (ViT and Swin-T). The feature vectors extracted by these encoders are processed through attribute prediction blocks for classification. Tables 4.8 and 4.9 offer performance comparisons across classification metrics, whereas Table 4.7 presents aggregated results averaged across attribute classes. LEUCOSIGHT combines the advantages of Transformers and CNNs to overcome the drawbacks of conventional techniques. While CNNs such as ResNet, DenseNet, and ConvNeXt effectively capture local features, they struggle with global context modeling. Conversely, Transformer models, such as ViT and Swin-T, excel at capturing global dependencies but often lack detailed local feature preservation. By integrating these complementary capabilities, LEUCOSIGHT enables both global and local feature extraction, enhancing segmentation and facilitating accurate attribute prediction. This integration empowers the MAttrP module to achieve superior performance, establishing LEUCOSIGHT as a robust and reliable solution.

Table 4.7: Summary of the average of each classification metric computed across all attributes for each System, as presented in Table 4.8 and Table 4.9 on the WBCAtt Dataset.

System	**Avg Pre.**	**Avg Rec.**	**Avg F-m.**	**Avg Acc.**	**Avg Spec.**
VGG16 [45]	89.462	89.305	89.306	90.137	90.964
ResNet50 [46]	90.651	90.256	90.435	90.923	91.766
DenseNet [47]	91.316	90.997	91.143	91.558	92.239
ShuffleNet [48]	91.249	90.965	91.088	91.533	92.277
ViT-Base [49]	90.517	90.581	90.483	90.980	91.716
ConvNeXt [50]	91.192	90.677	90.924	91.251	91.976
Swin T [39]	91.580	91.206	91.385	91.736	92.403
HemaX [24]	92.831	92.067	92.447	92.922	93.587
LEUCOSIGHT	**97.479**	**97.642**	**97.564**	**98.081**	**98.252**

4.7 Cross-Dataset Generalization

This section presents the cross-dataset generalization results from Table 4.10, showcasing significant performance enhancements on the Raabin and WBCAtt datasets compared to the discussion in Section 4.5. These improvements stem from the system's initial training on datasets containing at least two label types, often including classification labels, while

Table 4.8: Classification metrics for the first set of attributes from the WBCAtt Dataset, evaluated using various encoders in the Attribute Prediction System.

Attribute	**Metric**	**System**							
		VGG16	**ResNet50**	**DenseNet**	**ShuffleNet**	**ViT**	**ConvNeXt**	**Swin T**	**LEUCOSIGHT**
Cell Size	Pre.	83.452	84.217	84.839	85.093	85.612	85.903	86.214	97.642
	Rec.	83.448	83.691	84.718	84.931	85.411	85.865	86.123	97.582
	F-m.	83.450	83.953	84.779	84.961	85.509	85.884	86.168	97.611
	Acc.	85.031	85.512	86.003	86.217	86.512	86.793	87.145	98.025
	Spec.	86.503	87.017	87.305	87.508	87.824	87.993	88.337	98.232
Cell Shape	Pre.	89.085	90.736	91.019	91.215	91.534	91.801	92.007	98.218
	Rec.	90.132	90.648	91.217	91.406	91.718	91.905	92.213	98.116
	F-m.	89.609	90.684	91.123	91.309	91.611	91.853	92.105	98.153
	Acc.	91.015	91.507	91.902	92.014	92.305	92.487	92.718	98.508
	Spec.	92.018	92.512	92.704	92.915	93.109	93.284	93.508	98.713
Nuclear Cytoplasmic Ratio	Pre.	96.781	97.472	97.158	97.268	96.879	96.807	97.009	98.901
	Rec.	95.122	95.305	96.505	96.705	95.915	95.655	96.205	98.752
	F-m.	95.932	96.384	96.824	96.976	96.382	96.223	96.593	98.822
	Acc.	96.504	96.885	97.256	97.405	96.753	96.604	97.102	99.001
	Spec.	97.803	98.102	98.356	98.405	97.903	97.855	98.205	99.501
Chromatin Density	Pre.	83.973	84.553	86.103	86.203	84.733	85.563	85.803	95.502
	Rec.	86.713	88.523	86.753	86.903	84.613	85.873	86.303	96.753
	F-m.	85.323	86.373	86.423	86.553	84.653	85.703	86.053	96.102
	Acc.	85.803	86.403	86.703	86.903	85.003	85.903	86.403	97.002
	Spec.	86.503	87.003	87.503	87.603	86.203	86.703	87.103	97.503
Cytoplasm Vacuole	Pre.	91.243	92.713	93.503	93.753	90.623	92.843	93.103	96.653
	Rec.	85.963	87.083	89.153	89.303	90.683	88.043	88.903	96.503
	F-m.	88.363	89.573	91.203	91.403	90.633	90.263	90.963	96.583
	Acc.	89.003	90.503	91.403	91.603	91.003	90.303	91.203	97.003
	Spec.	90.503	91.703	92.203	92.403	91.803	91.403	92.003	98.003

Table 4.9: Classification metrics for the second set of attributes from the WBCAtt Dataset, evaluated using various encoders in the Attribute Prediction System.

Attribute	**Metric**	**System**							
		VGG16	**ResNet50**	**DenseNet**	**ShuffleNet**	**ViT-Base**	**ConvNeXt**	**Swin**	**LEUCOSIGHT**
Cytoplasm Texture	Pre.	92.291	93.226	93.805	92.953	92.867	93.935	94.507	96.503
	Rec.	95.215	95.957	95.406	94.805	95.180	95.008	95.604	97.203
	F-m.	93.612	94.492	94.606	93.871	93.946	94.445	95.048	96.847
	Acc.	94.801	95.209	95.202	94.603	94.905	95.101	95.502	97.004
	Spec.	95.601	96.007	96.108	95.904	95.802	96.202	96.503	97.504
Nucleus Shape	Pre.	74.883	77.083	79.803	80.003	77.213	78.513	79.303	96.673
	Rec.	74.183	75.703	78.903	79.103	75.883	77.683	78.503	96.853
	F-m.	74.533	76.373	79.353	79.553	76.053	78.083	78.893	96.763
	Acc.	75.503	77.003	79.403	79.703	76.503	78.003	79.003	97.003
	Spec.	76.003	77.503	79.903	80.203	77.103	78.503	79.503	97.503
Cytoplasm Color	Pre.	84.194	88.248	88.607	87.804	87.624	88.058	89.401	96.003
	Rec.	83.937	88.063	88.804	88.106	88.041	88.287	89.603	95.802
	F-m.	84.067	88.151	88.706	87.958	87.832	88.166	89.503	95.902
	Acc.	85.002	88.503	89.006	88.307	88.205	88.408	89.704	96.205
	Spec.	86.002	88.903	89.507	89.108	88.609	88.809	90.005	96.408
Granule Type	Pre.	99.281	99.364	99.484	99.402	99.323	99.496	99.523	99.801
	Rec.	99.426	99.529	99.565	99.502	99.456	99.641	99.604	99.854
	F-m.	99.354	99.446	99.524	99.456	99.392	99.562	99.563	99.826
	Acc.	99.503	99.607	99.623	99.581	99.557	99.704	99.665	99.902
	Spec.	99.602	99.705	99.683	99.621	99.653	99.754	99.703	99.953
Granule Color	Pre.	98.724	98.888	98.842	98.805	98.786	99.004	98.926	98.903
	Rec.	98.914	98.963	98.946	98.889	98.874	99.123	99.005	99.002
	F-m.	98.819	98.925	98.892	98.843	98.824	99.063	98.967	98.954
	Acc.	99.006	99.107	99.081	99.002	99.058	99.205	99.105	99.152
	Spec.	99.107	99.207	99.127	99.102	99.154	99.256	99.168	99.206

Table 4.10: Cross-Dataset Generalization Classification Results.

<table>
<tr><th colspan="6">Trained on WBCAtt, Fine-Tuned on Raabin (Results on Raabin)</th></tr>
<tr><th></th><th>Acc</th><th>Pre</th><th>Rec</th><th>F-m</th><th>Spec</th></tr>
<tr><td><b>PWII</b></td><td>0.9953</td><td>0.9796</td><td>0.9892</td><td>0.9846</td><td>0.9930</td></tr>
<tr><th colspan="6">Trained on Raabin, Fine-Tuned on WBCAtt (Results on WBCAtt)</th></tr>
<tr><th></th><th>Acc</th><th>Pre</th><th>Rec</th><th>F-m</th><th>Spec</th></tr>
<tr><td><b>PWII</b></td><td>0.9936</td><td>0.9845</td><td>0.9894</td><td>0.9867</td><td>0.9925</td></tr>
<tr><th colspan="6">Trained on WBCAtt, Fine-Tuned on LISC (Results on LISC)</th></tr>
<tr><th></th><th>Acc</th><th>Pre</th><th>Rec</th><th>F-m</th><th>Spec</th></tr>
<tr><td><b>PWII</b></td><td>0.9563</td><td>0.9495</td><td>0.9549</td><td>0.9521</td><td>0.9602</td></tr>
</table>

‘PWII: Proposed Work II (LEUCOSIGHT)’

leveraging weak supervision to address missing annotations like segmentation and MAtts during joint training. Subsequent training phases, which introduced previously unavailable labels, further refined the system’s performance.

On the Raabin dataset, classification accuracy improved, and predictions of attributes for 100 sampled images aligned with expected WBC class characteristics, despite the absence of ground truth annotations. Similarly, training on WBCAtt not only enhanced classification accuracy but also produced visually correct segmentation masks for 100 sampled WBC images, with precise localization of WBC regions, even without segmentation annotations. This demonstrates the system’s adaptability and robustness. The similarity in feature distributions between Raabin and WBCAtt facilitated effective fine-tuning, enabling the system to generate segmentation masks for WBCAtt and predict MAtts for Raabin, despite the lack of corresponding annotations. However, the results highlight a limitation: when datasets with highly divergent feature distributions are used, fine-tuning may yield less substantial improvements, as observed in the case of training on WBCAtt and fine-tuning on LISC. Addressing this limitation is suggested for future research.

4.8 Ablation Analysis

4.8.1 Impact of choosing 3-class segmentation over 2-class segmentation

Our first proposed system, SWASTIC, begins with a 3-class segmentation as its first stage, followed by the classification of WBCs. To examine the importance of this 3-class segmentation, we designed an alternative system named SW2CS, derived from SWASTIC. The main difference between SWASTIC and SW2CS is that SW2CS uses a 2-class segmentation approach instead of a 3-class segmentation. Table 4.11 displays the outcomes of the WBC classification comparison between SWASTIC and SW2CS. The findings reveal that SWASTIC achieves significantly better performance than SW2CS. This improvement arises because SW2CS relies solely on nuclear information for classification, which increases the likelihood of misclassification. For instance, similarities in nuclear shapes between certain WBC types, such as eosinophils and neutrophils or monocytes and lymphocytes, can be confusing. In contrast, SWASTIC uses 3-class segmentation, integrating additional cytoplasmic features, improving the ability to accurately classify WBCs.

Likewise, in our second proposed system, LEUCOSIGHT, which also employs 3-class segmentation followed by MAttrP and WBC classification, we observed a similar trend. To further validate the role of 3-class segmentation, we developed a counterpart system, LW2CS, based on LEUCOSIGHT. The WBC classification outcomes for LW2CS and LEUCOSIGHT are presented in Table 4.12. The results confirm that LEUCOSIGHT outperforms LW2CS. The limitation of LW2CS lies in its reliance solely on nuclear regions to

Table 4.11: Performance comparison between SWASTIC and SW2CS on the Raabin dataset.

Class	**SW2CS**					**SWASTIC**				
	Acc	**Pre**	**Rec**	**F1S**	**Spe**	**Acc**	**Pre**	**Rec**	**F1S**	**Spe**
M	0.987	0.894	0.893	0.888	0.994	0.990	0.906	0.914	0.910	0.994
L	0.985	0.965	0.972	0.968	0.989	0.989	0.977	0.979	0.978	0.992
B	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
E	0.973	0.784	0.879	0.829	0.981	0.996	0.966	0.984	0.975	0.997
N	0.968	0.984	0.964	0.974	0.974	0.994	0.997	0.993	0.995	0.995
Overall	0.956	0.925	0.936	0.931	0.986	0.985	0.971	0.983	0.971	0.996

M: Monocyte, L: Lymphocyte, B: Basophil, E: Eosinophil, N: Neutrophil
‘SW2CS: SWASTIC performing 2-class segmentation instead of 3-class’

guide MAttrP, capturing features specific to the nucleus alone. On the other hand, LEUCOSIGHT’s 3-class segmentation approach includes the cytoplasm as a distinct class, enabling the extraction of essential cytoplasmic attributes such as vacuoles, cytoplasmic texture, nucleus-to-cytoplasm ratio, etc. These attributes are critical for accurately distinguishing between WBC types, leading to LEUCOSIGHT’s superior performance.

4.8.2 Impact of different components in the segmentation module of LEUCOSIGHT

This section delves into the design of the segmentation module within LEUCOSIGHT, focusing on its implementation across the Raabin and LISC datasets, particularly emphasizing the CSD. As shown in Table 4.13, the segmentation module based solely on the STR block (SMSTR) exhibits the lowest performance. This is attributed to its bottom-up concatenation

Table 4.12: Performance comparison between LEUCOSIGHT and LW2CS on the Raabin dataset.

Class	LW2CS					LEUCOSIGHT				
	Acc	**Pre**	**Rec**	**F1S**	**Spe**	**Acc**	**Pre**	**Rec**	**F1S**	**Spe**
M	0.977	0.964	0.973	0.968	0.979	0.982	0.945	0.960	0.952	0.994
L	0.958	0.959	0.952	0.955	0.969	0.989	0.975	0.979	0.980	0.996
B	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
E	0.939	0.941	0.928	0.934	0.931	0.993	0.961	0.980	0.970	0.997
N	0.944	0.939	0.954	0.946	0.947	0.992	0.996	0.990	0.993	0.996
Overall	0.963	0.961	0.961	0.960	0.965	0.991	0.976	0.985	0.980	0.997

M: Monocyte, L: Lymphocyte, B: Basophil, E: Eosinophil, N: Neutrophil
'LW2CS: LEUCOSIGHT performing 2-class segmentation instead of 3-class'

strategy, which hinders the effective integration of hierarchical features. Introducing cross-attention significantly improves performance by leveraging features with the most global context as queries, while the encoder's high-level outputs serve as keys and values. This mechanism aligns global and semantic information effectively but falls short in capturing finer local details, as it prioritizes enhancing the global context. The impact of this limitation is reflected in the performance of SMSTR-CA.

To overcome this drawback, convolution operations are applied to the feature grid after the attention mechanism. These operations enhance the module's ability to capture local details by enabling pixel-level interactions with neighboring regions, thus improving local feature representation. Combining attention mechanisms with convolution operations results in markedly better performance, as demonstrated by SMSTR-CAWC. The fusion of convolutional techniques, characteristic of CNNs, with the attention mechanisms from transformers

Table 4.13: Study of different components in Segmentation Module of LEUCOSIGHT.

System	**Raabin Dataset**			**LISC Dataset**		
	DSC	**IoU**	**Acc**	**DSC**	**IoU**	**Acc**
2-Class Segmentation						
SMSTR	0.9617	0.9343	0.9812	0.8923	0.8835	0.9721
SMSTR-CA	0.9753	0.9514	0.9916	0.9317	0.9145	0.9865
SMSTR-CAWC	**0.9902**	**0.9755**	**0.9989**	**0.9503**	**0.9247**	**0.9941**
3-Class Segmentation						
SMSTR	0.8819	0.8625	0.9537	0.8317	0.8241	0.9453
SMSTR-CA	0.9123	0.8931	0.9724	0.8512	0.8437	0.9643
SMSTR-CAWC	**0.9325**	**0.9270**	**0.9907**	**0.8550**	**0.8102**	**0.9927**

‘SMSTR: Segmentation Module with Spatial Texture Refinement’, ‘SMSTR-CA: SMSTR with Cross Attention’, ‘SMSTR-CAWC: SMFR-CA with Convolution operation’

creates a robust synergy, significantly boosting the effectiveness of the segmentation module.

Chapter 5

Conclusion

This study presents an innovative system for WBC analysis, drawing inspiration from diagnostic methods traditionally employed by pathologists. Unlike earlier systems, which either used segmentation exclusively to aid classification or implemented segmentation, MAttrP, and classification simultaneously through multi-head outputs without fully leveraging the synergy between these tasks, our approach capitalizes on their interdependence. In clinical practice, pathologists typically identify WBCs on histopathology slides, examine their morphological attributes in detail, and then classify them based on these features. Mimicking this systematic workflow, our proposed system, LEUCOSIGHT, begins by segmenting WBCs from microscopic images, accurately localizing them, and extracting critical structural information. These structural cues are subsequently utilized to predict MAttrs, which serve as a foundation for precise classification.

A significant innovation of this work lies in its emphasis on explainability, an area where conventional DL systems often fall short due to their opaque decision-making nature. In contrast, LEUCOSIGHT provides classification decisions accompanied by morphology-based explanations, reflecting the logical, interpretable decision-making process pathologists employ when analyzing WBCs. The system employs a hybrid architecture that synergistically combines CNNs and transformers. CNNs excel at capturing fine-grained, localized structural features, while transformers are adept at modeling global contextual relationships and dependencies. By integrating these strengths, the system achieves superior segmentation accuracy through precise identification and delineation of structural boundaries. This approach represents a significant improvement over previous methods, such as our earlier work, SWASTIC, which relied on simpler hierarchical feature concatenation. The advanced fusion strategy used in LEUCOSIGHT ensures a more effective combination of features, leading to enhanced overall performance. Furthermore, the system demonstrates remarkable adaptability, capable of being fine-tuned for diverse WBC datasets. Even when initially trained on datasets lacking certain ground truths, it excels across various tasks and conditions. This flexibility makes it particularly well-suited for use in resource-limited environments or with datasets featuring sparse annotations.

Future enhancements will focus on improving the system's robustness to effectively generalize across multiple datasets. Addressing variability in data distributions will ensure consistent performance across diverse scenarios, paving the way for broader applications in WBC diagnostics and analysis.

Bibliography

[1] Satoshi Tsutsui, Winnie Pang, and Bihan Wen. Wbcatt: A white blood cell dataset annotated with detailed morphological attributes. *Advances in Neural Information Processing Systems*, 36:50796–50824, 2023.

[2] Hüseyin Üzen and Hüseyin Fırat. A hybrid approach based on multipath swin transformer and convmixer for white blood cells classification. *Health Information Science and Systems*, 12(1):33, 2024.

[3] Jose Luis Diaz Resendiz, Volodymyr Ponomaryov, Rogelio Reyes Reyes, and Sergiy Sadovnychiy. Explainable cad system for classification of acute lymphoblastic leukemia based on a robust white blood cell segmentation. *Cancers*, 15(13):3376, 2023.

[4] Nisha Ramesh, Bryan Dangott, Mohammed E Salama, and Tolga Tasdizen. Isolation and two-step classification of normal white blood cells in peripheral blood smears. *Journal of pathology informatics*, 3(1):13, 2012.

[5] Adit Srivastava, Aravind Ramagiri, Puneet Gupta, and Vivek Gupta. Sangam: Synergizing local and global analysis for simultaneous wbc classification and segmentation. In *International Conference on Pattern Recognition*, pages 154–169. Springer, 2025.

[6] Adnan Haider, Muhammad Arsalan, Young Won Lee, and Kang Ryoung Park. Deep features aggregation-based joint segmentation of cytoplasm and nuclei in white blood cells. *IEEE Journal of Biomedical and Health Informatics*, 26(8):3685–3696, 2022.

[7] Seyed Hamid Rezatofighi and Hamid Soltanian-Zadeh. Automatic recognition of five types of white blood cells in peripheral blood. *Computerized Medical Imaging and Graphics*, 35(4):333–343, 2011.

[8] Pilar Gómez-Gil, Manuel Ramírez-Cortés, Jesús González-Bernal, Ángel García Pedrero, César I Prieto-Castro, Daniel Valencia, Rubén Lobato, and José E Alonso. A feature extraction method based on morphological operators for automatic classification of leukocytes. In *2008 Seventh Mexican International Conference on Artificial Intelligence*, pages 227–232. IEEE, 2008.

[9] CS Hinge, AG Ambekar, and SS Kulkarni. Classification of rbc and wbc in peripheral blood smear using knn. *Indian Journal of Research*, 2(1):2250–1991, 2013.

[10] Sajad Tavakoli, Ali Ghaffari, Zahra Mousavi Kouzehkanan, and Reshad Hosseini. New segmentation and feature extraction algorithm for classification of white blood cells in peripheral smear images. *Scientific Reports*, 11(1):19428, 2021.

[11] Adnan Khashman. Ibcis: Intelligent blood cell identification system. *Progress in Natural Science*, 18(10):1309–1314, 2008.

[12] Ahmed Ismail Shahin, Yanhui Guo, Khalid Mohamed Amin, and Amr A Sharawi. White blood cells identification system based on convolutional deep neural learning networks. *Computer methods and programs in biomedicine*, 168:69–80, 2019.

[13] Xufeng Yao, Kai Sun, Xixi Bu, Congyi Zhao, and Yu Jin. Classification of white blood cells using weighted optimized deformable convolutional neural networks. *Artificial Cells, Nanomedicine, and Biotechnology*, 49(1):147–155, 2021.

[14] Rabiah Al-Qudah and Ching Y Suen. Improving blood cells classification in peripheral blood smears using enhanced incremental training. *Computers in Biology and Medicine*, 131:104265, 2021.

[15] Xiaogen Zhou, Zhiqiang Li, Yuyang Xue, Shun Chen, Meijuan Zheng, Cong Chen, Yue Yu, Xingqing Nie, Xingtao Lin, Luoyan Wang, et al. Cuss-net: a cascaded unsupervised-based strategy and supervised network for biomedical image diagnosis and segmentation. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2444–2455, 2023.

[16] Bairaboina Sai Sambasiva Rao and Battula Srinivasa Rao. An effective wbc segmentation and classification using mobilenetv3–shufflenetv2 based deep learning framework. *IEEE Access*, 11:27739–27748, 2023.

[17] Hua Chen, Juan Liu, Chunbing Hua, Jing Feng, Baochuan Pang, Dehua Cao, and Cheng Li. Accurate classification of white blood cells by coupling pre-trained resnet and densenet with scam mechanism. *BMC bioinformatics*, 23(1):282, 2022.

[18] S Ratheesh and A Ajisha Breethi. Deep learning based non-local k-best renyi entropy for classification of white blood cell subtypes. *Biomedical Signal Processing and Control*, 90:105812, 2024.

[19] Jamal Ferdosi Bilkis. Unified approach for white blood cell segmentation, feature extraction, and counting using max-tree data structure. *International Journal of Advanced Computer Science and Applications*, 11(9), 2020.

[20] Rufus Rubin, SM Anzar, Alavikunhu Panthakkan, and Wathiq Mansoor. Transforming healthcare: Raabin white blood cell classification with deep vision transformer. In *2023 6th International Conference on Signal Processing and Information Security (ICSPIS)*, pages 212–217. IEEE, 2023.

[21] Bing Leng, Chunqing Wang, Min Leng, Mingfeng Ge, and Wenfei Dong. Deep learning detection network for peripheral blood leukocytes based on improved detection transformer. *Biomedical Signal Processing and Control*, 82:104518, 2023.

[22] Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

[23] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

[24] Aditya Shankar Pal, Debojyoti Biswas, Joy Mahapatra, Debasis Banerjee, Prantar Chakrabarti, Alejandro F Frangi, and Utpal Garain. Pathologist-like explanations unveiled: An explainable deep learning system for white blood cell classification. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024.

[25] Gautam Rajendrakumar Gare, Andrew Schoenling, Vipin Philip, Hai V Tran, P de-Boisblanc Bennett, Ricardo Luis Rodriguez, and John Michael Galeotti. Dense pixel-labeling for reverse-transfer and diagnostic learning on lung ultrasound for covid-19 and pneumonia detection. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1406–1410. IEEE, 2021.

[26] Atul Rawal, James McCoy, Danda B Rawat, Brian M Sadler, and Robert St Amant. Recent advances in trustworthy explainable artificial intelligence: Status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*, 3(6):852–866, 2021.

[27] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9, 2020.

[28] Zahra Mousavi Kouzehkanan, Sepehr Saghari, Eslam Tavakoli, Peyman Rostami, Mohammadjavad Abaszadeh, Farzaneh Mirzadeh, Esmaeil Shahabi Satlsar, Maryam Gheidishahran, Fatemeh Gorgi, Saeed Mohammadi, et al. Raabin-wbc: a large free access dataset of white blood cells from normal peripheral blood. *bioRxiv*, pages 2021–05, 2021.

[29] Zahra Mousavi Kouzehkanan, Sajad Tavakoli, and Arezoo Alipanah. Easy-gt: Open-source software to facilitate making the ground truth for white blood cells nucleus. *arXiv e-prints*, pages arXiv–2101, 2021.

[30] S Sapna and A Renuka. Computer-aided system for leukocyte nucleus segmentation and leukocyte classification based on nucleus characteristics. *International Journal of Computers and Applications*, 42(6):622–633, 2020.

[31] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[32] Şeyma Nur Özcan, Tansel Uyar, and Gökay Karayeğen. Comprehensive data analysis of white blood cells with classification and segmentation by using deep learning approaches. *Cytometry Part A*, 2024.

[33] Lei Jiang, Chang Tang, and Hua Zhou. White blood cell classification via a discriminative region detection assisted feature aggregation network. *Biomedical Optics Express*, 13(10):5246–5260, 2022.

[34] Siraj Khan, Muhammad Sajjad, Naveed Abbas, José Escorcia-Gutierrez, Margarita Gamarra, and Khan Muhammad. Efficient leukocytes detection and classification in microscopic blood images using convolutional neural network coupled with a dual attention network. *Computers in Biology and Medicine*, page 108146, 2024.

[35] Saba Saleem, Javeria Amin, Muhammad Sharif, Muhammad Almas Anjum, Muhammad Iqbal, and Shui-Hua Wang. A deep network designed for segmentation and classification of leukemia using fusion of the transfer learning models. *Complex & Intelligent Systems*, pages 1–16, 2021.

[36] M Roy Reena and PM Ameer. Localization and recognition of leukocytes in peripheral blood: A deep learning approach. *Computers in Biology and Medicine*, 126:104034, 2020.

[37] Jimut Bahan Pal, Aniket Bhattacharyea, Debasis Banerjee, and Br Tamal Maharaj. Advancing instance segmentation and wbc classification in peripheral blood smear through domain adaptation: A study on pbc and the novel rv-pbs datasets. *Expert Systems with Applications*, 249:123660, 2024.

[38] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[41] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

[42] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[43] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[44] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[46] Brett Koonce. Resnet 50. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72. Springer, 2021.

[47] Chaoning Zhang, Philipp Benz, Dawit Mureja Argaw, Seokju Lee, Junsik Kim, Francois Rameau, Jean-Charles Bazin, and In So Kweon. Resnet or densenet? introducing dense shortcuts to resnet. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3550–3559, 2021.

[48] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018.

[49] Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 589–598, 2021.

[50] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023.