JJEE

Volume 12 | Number 3 | September 2026 | Pages 559-575

Jordan Journal of Electrical Engineering

ISSN (print): 2409-9600, ISSN (online): 2409-9619

Homepage: jjee.ttu.edu.jo

Automated Extraction and Size-based Filtering of High-quality RBC Images from Blood Smears in Thailand for Morphological Analysis

B. Angmanee[1], S. Wanram[2], A. Thedsakhulwong[3*]

[1, 3] Department of Physics, Faculty of Science, Ubon Ratchathani University, Thailand
E-mail: amorn.t@ubu.ac.th
[2] Department of Pathology, College of Medicine and Public Health, Ubon Ratchathani University, Thailand

Received: Jul 01, 2025 *Revised: Feb 03, 2026* *Accepted: Feb 25, 2026* *Available online: Aug 15, 2026*

***Abstract*—** This study presents a semi-automated and reproducible pipeline for extracting high-quality, morphology-preserved single RBC images from peripheral blood smears of Thai patients. The dataset includes six clinically confirmed hematological conditions: Iron Deficiency Anemia, Thalassemia Trait, Hb H Disease, Thalassemia Hb E Disease, Symptomatic Thalassemia Hb E Disease, and Homozygous Hb E. Whole-slide images were acquired using an Aperio AT2 scanner at 40× magnification and processed through a Python-based workflow integrating mean shift filtering, adaptive thresholding, contour-based segmentation, and edge-contact rejection to exclude incomplete or touching cells. Isolated RBCs were filtered by size and overlaid onto standardized black backgrounds ranging from 32×32 to 1024×1024 pixels. The proposed pipeline generated a curated dataset of 12,229 single-cell RBC images, with the 128×128 pixel group identified as the most stable and representative resolution across all disease categories. Baseline CNN experiments demonstrate that the extracted images are learnable and structurally consistent, confirming dataset usability without claiming optimized or clinically deployable performance. By providing a region-specific RBC dataset derived from a Thai cohort with a high prevalence of hemoglobinopathies, together with a transparent extraction framework and baseline CNN validation, this work establishes a solid foundation for future CNN-based RBC morphology analysis and population-specific digital hematopathology research.

***Keywords*—** Red blood cell segmentation; Blood smear analysis; Image preprocessing; Medical image dataset.

1. INTRODUCTION

The morphological examination of red blood cells (RBCs) under a microscope remains a cornerstone of hematological diagnostics. Variations in RBC shape and size—such as microcytosis, target cells, elliptocytes, or irregular biconcave structures—serve as morphological hallmarks of blood disorders including thalassemia and iron deficiency anemia [1–2]. In Southeast Asia, particularly Northeastern Thailand, these conditions are highly prevalent due to the regional genetic background and high carrier rates of hemoglobinopathies [3–6]. According to the Thai Ministry of Public Health, approximately 30–40% of the population are thalassemia carriers, with Hb E and α-thalassemia being the most common forms, emphasizing the need for reliable and low-cost diagnostic tools. While molecular testing has improved precision, morphological assessment of peripheral blood smears remains a critical screening step in routine practice, especially in low-resource laboratories. Recent advances in digital pathology further emphasize the need for precise RBC morphology to support computational analysis. Recent advances in artificial intelligence, especially convolutional

* Corresponding author
DOI: https://doi.org/10.5455/jjee.204-1751358445

neural networks (CNNs), have enabled automatic classification of hematological images with high accuracy [9–12]. Nevertheless, CNN performance depends heavily on the quality, scale, and diversity of available datasets.

Most existing public RBC datasets suffer from several limitations: (i) they are either synthetic or limited to small, curated subsets, (ii) they often lack standardized backgrounds or consistent resolution, and (iii) they rarely represent real patient diversity from specific populations such as Thai or Southeast Asian cohorts [20]. These limitations restrict the ability of CNN models to learn fine-grained morphological variations that are essential for accurate RBC classification. Furthermore, models trained on non-representative datasets often fail to transfer effectively to real-world clinical samples, resulting in reduced diagnostic reliability. Consequently, the absence of a clinically grounded and ethnically relevant dataset remains a major obstacle to developing generalizable diagnostic models. In this study, CNN experiments are conducted solely for dataset validation and baseline morphology learning, rather than for clinical diagnosis or decision-making.

To address this gap, this study develops a semi-automated image processing pipeline for generating large-scale, morphology-preserved single-cell RBC images from high-resolution blood smear slides of Thai patients. Unlike prior pipelines that depend on manual cropping or fixed-coordinate extraction, the proposed approach employs contour-based segmentation with touching-cell rejection and size-standardized overlay, ensuring consistent image quality across diverse hematological conditions. This pipeline also incorporates adaptive preprocessing steps that minimize staining variability and slide-to-slide artifacts, further improving feature consistency.

As a result, the extracted images maintain high morphological fidelity, making them suitable for downstream tasks such as CNN training, clustering, and quantitative morphometric analysis.

The dataset produced from this pipe-line includes six clinically verified classes—Iron Deficiency Anemia, Thalassemia Trait, Hb H Disease, Thalassemia Hb E Disease, Thalassemia Hb E Disease (Symptomatic), and Homozygous Hb E—providing one of the most comprehensive RBC image resources derived from a real Thai cohort. In summary, the key contributions of this study are as follows:

1. Dataset creation: generation of over 12,000 single-cell RBC images from Aperio AT12 scans of clinically confirmed Thai patients across six major blood disorders.
2. Reproducible segmentation pipeline: development of a Python-based workflow integrating mean-shift filtering, Otsu thresholding, contour extraction, and size-based overlay with touching-cell rejection.
3. Standardized morphological quality: all cells are morphology-preserved, centered, and resolution-consistent to ensure compatibility with downstream architectures.
4. Clinical and regional relevance: establishment of the first RBC dataset designed specifically for the Northeastern Thai population, providing a valuable reference for digital hematopathology and AI-driven thalassemia research.

At a high level, the proposed workflow (Fig. 1) begins with high-resolution whole-slide scanning, followed by a structured sequence of preprocessing, segmentation, and morphology-based filtering steps designed to isolate high-quality single-cell RBC images. The resulting dataset supports robust and reproducible model training, helping bridge the gap between traditional hematology and modern AI-based diagnostics for population-specific applications.

2. MATERIALS AND METHODS

2.1. Sample Preparation and Slide Acquisition

Peripheral blood smear samples were collected from six Thai patients, each clinically diagnosed with one of the following hematological conditions: Iron Deficiency Anemia (IDA), Thalassemia Trait (TT), Hb H Disease (HbH), Thalassemia Hb E Disease (HbE), Thalassemia Hb E Disease (Symptomatic) (HbE Sx), and Homozygous Hb E (Homo HbE). For each diagnostic category, a single patient contributed one stained peripheral blood smear slide, resulting in a total of six slides, each representing a distinct disease class. This one-patient-per-class protocol was implemented to ensure clearly defined diagnostic boundaries during dataset construction and to establish a transparent link between clinical confirmation and image-level ground truth. All diagnoses were verified using a multi-step laboratory workflow, including Complete Blood Count (CBC), High Performance Liquid Chromatography (HPLC), and hemoglobin electrophoresis. These diagnostic methods were performed according to standard clinical procedures at the collaborating medical laboratory. The final diagnosis for each patient was jointly confirmed by a board-certified pathologist and a hematologist, who conducted independent reviews of laboratory reports and smear morphology prior to inclusion in the study. Whole blood samples were collected in EDTA anticoagulant tubes and prepared using a manual smear technique to preserve cellular morphology. Each smear was stained using the standard Wright–Giemsa protocol, ensuring consistent chromatic contrast and morphology across all diagnostic classes. The stained slides were digitized using an Aperio AT2 whole-slide scanner operating in brightfield, high-resolution mode at 40× magnification, yielding images at 0.1658 μm/pixel [13]. The resulting SVS-format whole-slide images served as the input for all subsequent image processing steps. This study was approved by the Ubon Ratchathani University Research Ethics Committee under protocol code UBU–REC–77/2568, with approval granted on 12 March 2025. All samples were anonymized prior to analysis, and the study was conducted in compliance with institutional and national research ethics guidelines.

2.2. Overview of Image Processing Pipeline

The image extraction pipeline was implemented in Python 3.10 and designed to operate on standard computing hardware while maintaining reproducibility and transparency across processing steps. The workflow integrates multiple stages of whole-slide imaging, preprocessing, segmentation, and morphology-based filtering to produce standardized single-cell RBC images suitable for downstream analysis (Fig. 1). All digitized whole-slide images, generated using the Aperio AT2 scanner [13], were processed using OpenCV 4.8.1 for image manipulation, with support from NumPy and PIL for array operations and image handling [16–17]. The pipeline consists of four major phases:

1. Whole-slide acquisition and ROI selection: High-resolution SVS files scanned at 40× magnification (0.1658 μm/pixel) were imported using the OpenSlide library [14]. Selected regions of interest (ROIs), typically measuring 4000–5000 pixels per side, were extracted for processing.

2. Preprocessing: Pyramid mean shift filtering was applied to each ROI to enhance cellular boundaries while minimizing background noise. This step improves robustness against staining variability and supports effective thresholding [17].
3. Segmentation and candidate extraction: Grayscale conversion followed by Otsu's thresholding [8] was used to generate binary masks for contour detection. External contours representing potential RBC candidates were isolated and evaluated for morphology completeness, including boundary contact checks to eliminate touching or partially segmented cells.
4. Standardized overlay and results organization: non-touching cells were extracted using a mask-based cropping strategy and overlaid onto black square canvases based on their maximal dimension. This produced consistently centered, morphology-preserved images across multiple predefined size groups. The results were saved into structured folders to facilitate downstream CNN training and reproducibility.

Figure 1 provides a high-level illustration of these steps, showing the flow from whole-slide acquisition to standardized image output. This modular design allows each processing block to be independently modified or extended—for example, by integrating stain-normalization modules or advanced segmentation models—without altering the core workflow. The complete step-by-step procedure of the proposed RBC extraction pipeline is summarized in Algorithm 1.

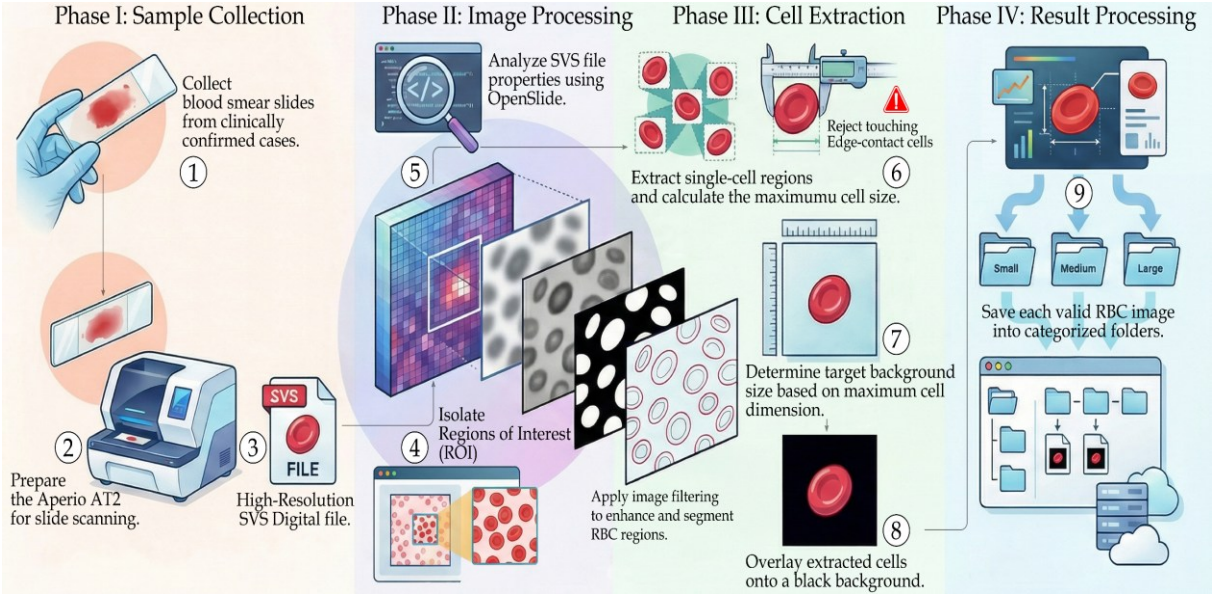

Fig. 1. Overview of the proposed RBC extraction pipeline, illustrating sample collection, whole-slide imaging, preprocessing, contour-based segmentation with touching-cell rejection, and size-based overlay for standardized single-cell RBC image generation.

2.3. Preprocessing and Segmentation

Each input image is first resized to 100% scale to preserve its original resolution. The image then undergoes pyramid mean shift filtering, a nonlinear smoothing technique that enhances the homogeneity of image regions while preserving edges. This step is particularly useful in blood smear images where variation in staining intensity or background noise may interfere with thresholding. Following filtering, the image is converted to grayscale, and Otsu's thresholding [8] is applied to automatically generate a binary mask that separates the foreground (cells) from the background.

$$Mask\ (x,y) = \begin{cases} 255, & if\ I(x,y) < T_{Otsu} \\ 0, & otherwise \end{cases} \quad (1)$$

This binarization approach is adaptive and minimizes intra-class variance, making it well-suited for processing a diverse range of smear conditions. The resulting binary mask is then processed to extract the outermost boundaries of potential RBC candidates using a standard external contour detection technique [17-18]. The use of external contour detection ensures that each candidate region corresponds to a contiguous foreground object, providing a well-defined basis for subsequent mask generation. At this stage, each detected contour is treated as an independent RBC candidate and converted into a binary mask that preserves the exact cell boundary. This contour-derived mask representation is essential for reliable downstream processing, as it enables precise differentiation between isolated cells, overlapping cells, and incomplete structures located near image borders. By operating at the mask level rather than relying solely on fixed-coordinate regions, the proposed segmentation strategy establishes a robust foundation for touching-cell rejection and morphology-preserving single-cell extraction described in the following sections.

2.4. Touching Cell Rejection

To isolate only complete and well-formed RBCs, a binary mask is generated for each detected contour. The corresponding cropped region is examined for edge contact, which is a common indicator of either partially segmented or overlapping cells. The system detects edge contact by scanning the top, bottom, left, and right borders of the binary crop. If any nonzero pixel is found on the boundary, the cell is deemed to be touching or incomplete and is excluded from further processing. This step is essential to maintain dataset quality and prevent distorted or merged cells from contaminating the training set. All rejected touching cells are still saved separately into a dedicated folder for potential future inspection or post-filtering analysis. Touching or partially segmented cells introduce ambiguity in single-cell morphology, as overlapping boundaries distort shape features and may lead CNN models to learn non-representative patterns. To avoid this issue, edge-touching candidates are conservatively rejected to preserve instance-level morphological integrity. Rejected cells are retained in a separate folder to support future cell separation or instance segmentation studies.

$$\exists\ x \in \{0, W-1\}\ or\ y \in \{0, H-1\}\ \ such\ that\ Mask(x,y) = 0 \quad (2)$$

where W and H denote the width and height of the cropped binary mask. Cells touching any image edge were excluded and stored separately for review. As illustrated in Fig. 1, touching-cell rejection is applied immediately after contour extraction to prevent incomplete or overlapping cells from entering the final dataset.

2.5. Cell Extraction and Size Classification

Following segmentation and touching-cell rejection, each remaining RBC candidate was isolated using a mask-based cropping strategy designed to preserve complete cellular morphology and eliminate background artifacts. For every detected contour, the corresponding binary mask was applied to the original ROI using a bitwise AND operation, ensuring that only the true cellular region was retained. This approach contrasts sharply with traditional coordinate-based cropping, which frequently captures surrounding debris, partial cells, or background noise, thereby reducing morphological clarity. As illustrated in Fig. 2,

mask-based extraction yielded substantially cleaner and more clinically meaningful images, addressing a known limitation in existing RBC datasets that rely on bounding-box or fixed-window extraction methods. To support downstream machine learning applications, particularly, all extracted cells were standardized using a size-based overlay system. After cropping, the width (w) and height (h) of each cell were measured, and the maximal dimension was defined as $L = \max(w, h)$. This value determined the target canvas size (S) to which each cell would be aligned. Cells were placed at the exact center of a square black canvas of predefined resolution, ensuring uniform positioning, consistent scaling, and morphology preservation. This design also prevents unintended geometric distortions that may occur with arbitrary resizing or stretching, thereby improving the stability and interpretability of CNN training. The image size classification system employed in this study is summarized in Table 1. The 128×128 pixel class was selected as the standard reference resolution, as it provided an optimal trade-off between morphological detail and computational efficiency during CNN training. Smaller size classes (e.g., 32×32) typically contain platelets or small debris, while larger classes (≥256×256) often represent close-paired RBCs or aggregated structures that do not correspond to isolated single-cell morphology.

Table 1. RBC image size grouping based on maximum cell dimension.

Cell size range [pixels]	Overlay image size [pixels]	Description
L ≤ 16	32 × 32	Very small RBCs
16 < L ≤ 32	32 × 32	Small RBCs
32 < L ≤ 128	128 × 128	Standard size for CNN training
128 < L ≤ 256	256 × 256	Large RBCs
256 < L ≤ 512	512 × 512	Very large RBCs
512 < L ≤ 1024	1024 × 1024	Extra-large RBCs
L > 1024	Excluded (oversized)	Discarded due to extreme size

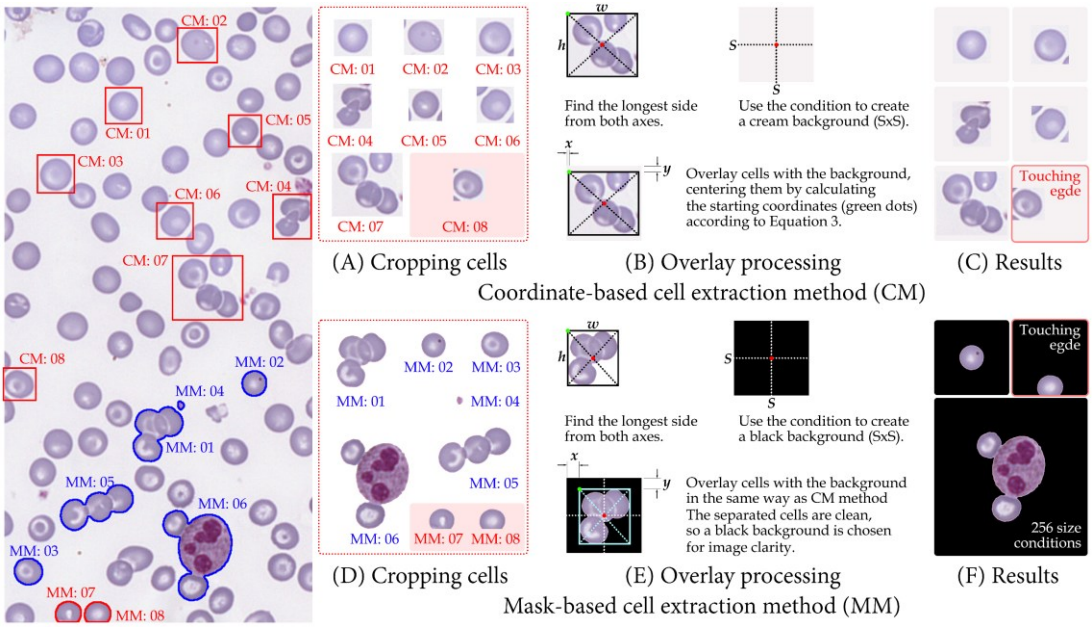

Fig. 2. Comparison between coordinate-based (CM) and mask-based (MM) single-cell extraction methods. (A–C) CM: cells are cropped using coordinate-based regions, followed by size-based overlay, which may include touching or partially cropped cells. (D–F) MM: cells are extracted using binary masks, enabling cleaner separation, explicit touching-cell rejection, and standardized overlay on a black background. This comparison illustrates how mask-based extraction reduces boundary ambiguity and background interference, thereby preserving true single-cell morphology for robust downstream CNN-based analysis.

Oversized regions (>1024 px) were excluded due to their lack of relevance for single-cell analysis. This structured, size-aware extraction and overlay process produced a high-quality, morphology-consistent image dataset across six clinically verified hematological conditions. By ensuring uniform resolution, central alignment, and minimal background interference, the proposed method directly addresses several limitations identified in previous RBC datasets, such as inconsistent image boundaries, mixed object types, and variability in spatial scaling. Figure 2 compares coordinate-based and mask-based cell extraction methods, highlighting the inherent limitations of coordinate cropping and the advantages of mask-based extraction for clean single-cell isolation and effective touching-cell rejection. In contrast to coordinate-based approaches, the mask-based method provides explicit control over cell boundaries, enabling incomplete or edge-touching cells to be reliably identified and excluded prior to overlay. This capability is particularly important in dense smear regions, where coordinate cropping may inadvertently include overlapping cells or truncated structures. By enforcing mask-level integrity checks together with size-based standardization, the proposed approach minimizes label noise and preserves true single-cell morphology. This comparison demonstrates that mask-based extraction consistently yields cleaner and more interpretable single-cell images than coordinate-based cropping. As a result, the extracted dataset is well-suited for robust CNN feature learning by reducing background interference, preventing geometric distortion, and ensuring that all samples correspond to complete and isolated RBCs. Consequently, the dataset provides a reliable foundation for reproducible machine learning experiments and downstream morphological analysis in digital hematopathology.

2.6. Image Output and Folder Structuring

After size-based cropping and standardized overlay processing, all resulting RBC images were exported into a hierarchical directory structure designed to maximize reproducibility, traceability, and ease of use for downstream machine learning applications. Each processed cell was saved into a size-specific subdirectory—for example, "128 size", "256 size", or "touching"—corresponding to its assigned overlay dimension or exclusion category. This organizational structure ensures that images can be efficiently accessed for model training, cross-size comparison, or statistical analysis without requiring additional manual sorting. To support full transparency and quality verification, each ROI produced a set of auxiliary images documenting every major processing step, including: the original ROI (full-color), the mean-shift filtered image, grayscale-converted version, Otsu-thresholded binary mask, the final contour mask, and the contour overlay visualization. These multi-stage visual records enable researchers to trace the derivation of each extracted RBC image and assess segmentation accuracy on a per-ROI basis. This form of traceability is critical for digital hematopathology, where subtle segmentation errors can propagate into CNN training and impact classification performance. The dataset directory follows a structured and modular design inspired by FAIR data principles. Each file is named using a standardized convention that captures relevant metadata such as size class, ROI index, and segmentation status. This guarantees that users can reliably reproduce experiments, re-train models, or re-label samples without ambiguity. Furthermore, the clear separation of high-quality RBCs, touching cells, and oversized/invalid regions provides a foundation for future work involving multi-class segmentation, noise modeling, or instance separation algorithms. This structured organization

enables direct and reproducible use of the dataset for CNN training, evaluation, and future extensions.

Algorithm 1. RBC Extraction using Size-Based Overlay Filtering	
Input	ROI image I from stained blood smear slide; size thresholds [16, 32, 128, 256, 512, 1024] maximum size S_max = 1024 px
Output	Folder-sorted RBC images F[i] for CNN use; grouped by size and type
1:	Initialize output folders and counters
2:	Apply **pyrMeanShiftFiltering** to image I to enhance object boundaries
3:	Convert to grayscale and apply **Otsu's thresholding** to obtain binary mask
4:	Use cv2.findContours to extract external contours from the binary mask
5:	For each contour c in contours:
6:	Create cropped mask Mc and original image crop around c
7:	**If any pixel in Mc touches image boundary**, classify as **Touching**, and skip Otherwise, calculate $L = max(H, W)$ of cropped region
8:	If L > 1024, classify as **Oversize**, skip
9:	Else, determine target canvas size S:
10:	If L ≤ 16 → S = 32
11:	If 16 < L ≤ 32 → S = 32
12:	If 32 < L ≤ 128 → S = 128
13:	If 128 < L ≤ 256 → S = 256
14:	If 256 < L ≤ 512 → S = 512
15:	If 512 < L ≤ 1024 → S = 1024
16:	Create black canvas of size S×S, paste cropped cell image at center
17:	Save resulting overlaid image to corresponding folder

2.7. Software and Libraries

The image processing pipeline was implemented in Python 3.10 using widely adopted open-source libraries to ensure reproducibility and seamless deployment. OpenCV 4.8.1 served as the primary engine for image manipulation, supporting mean-shift filtering, thresholding, contour detection, and image I/O operations [18]. NumPy 1.26 enabled efficient array manipulation and numerical computation, while Scikit-learn 1.3 was used for optional color profiling and K-means clustering to support future dataset extensions. To generate standardized overlays, the pipeline used the cvzone.overlayPNG function, precisely centering each cropped cell on a uniform black background with proper alpha blending to maintain consistent visual presentation for downstream machine learning tasks. Supporting libraries such as PIL (Pillow) were employed for auxiliary image handling and format conversion. All scripts were executed on a Windows 11 system with an Intel Core i5 CPU and an NVIDIA GTX/RTX GPU, demonstrating that the pipeline can operate efficiently on standard laboratory hardware and can be readily reproduced by other researchers. The selection of widely adopted open-source libraries ensures that the proposed pipeline can be reproduced without reliance on proprietary software or specialized computing environments. The use of OpenCV and NumPy enables efficient pixel-level operations and contour-based processing, while optional Scikit-learn components provide flexibility for future dataset extension. Importantly, the modular software design allows individual processing stages to be modified or replaced without affecting the overall workflow. Together, these design choices facilitate

reproducible implementation and seamless integration with the CNN-based experiments presented in the subsequent sections.

3. RESULTS AND DISCUSSION

3.1. Qualitative Visualization of Segmentation and Overlay

To demonstrate the end-to-end performance of the proposed RBC extraction pipeline, Fig. 3 presents a step-by-step visualization of the segmentation, filtering, and overlay process. The workflow begins with SVS file extraction (A), where a high-resolution whole-slide image is scanned and a ROI is selected for analysis (B). This ROI undergoes preprocessing beginning with pyramid mean shift filtering (C) to reduce background noise and enhance cell boundaries, followed by grayscale conversion (D) and Otsu's thresholding (E) to generate a binary mask for contour detection. Connected regions are identified (F), cell contours are extracted (G), and each candidate undergoes boundary-based evaluation (H) to measure dimensions and remove edge-touching or oversized cells. Valid cells are then cropped and overlaid onto a black square canvas based on size classification (I), ensuring centered and normalized RBCs across output sizes ranging from 32 to 1024 pixels.

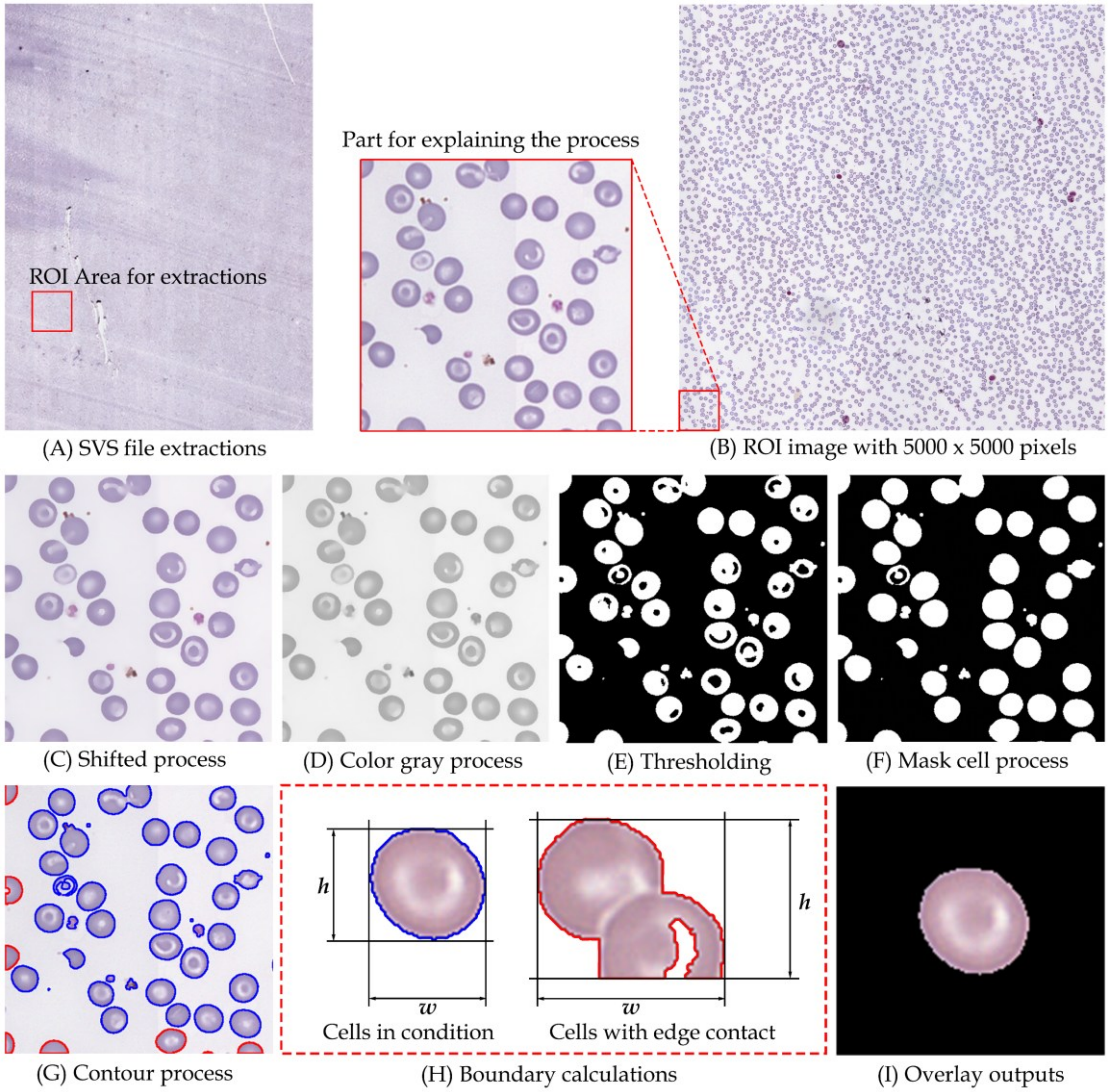

Fig. 3. End-to-end visualization of RBC segmentation, filtering, and size-based overlay across representative conditions.

3.2. Quantitative Cell Extraction Statistics

To evaluate the effectiveness of the proposed pipeline across multiple pathological conditions, we extracted and categorized individual RBC images from blood smear samples belonging to six diagnostic groups: Iron Deficiency Anemia (IDA), Thalassemia Trait (TT, Trait), Hb H Disease (HbH), Thalassemia Hb E Disease (HbE), Thalassemia Hb E Disease with symptoms (HbE (Sx)), and Homozygous E (Homo HbE). Table 2 summarizes the number of single-cell images obtained from each condition, categorized into size-based overlay groups and filtered categories (e.g., touching, oversized). Each cell was classified into one of the following categories based on its maximal dimension $L = max(w, h)$: 16, 32, 128, 256, 512, 1024 pixels, or "oversize" (excluded). Cells detected near image borders were rejected and stored as "touching". A representative portion of the dataset—particularly images sized 128×128—was considered optimal for CNN training due to their balance between resolution and morphological clarity. Table 2 demonstrates that the 128×128 group constituted the largest portion of usable data across all disease conditions, especially in Homo HbE and HbE (Sx), where segmentation yielded over 3,000 images in this category. Conversely, very small cells (≤16 px) and oversized cells (>1024 px) were rare and excluded. Touching cells accounted for ~5–10% of all detections, highlighting the necessity of edge-contact filtering in dense smear regions. This observation confirms the effectiveness of the pipeline's size-based filtering strategy in isolating morphologically meaningful RBCs. This quantitative analysis affirms that the pipeline is capable of producing thousands of high-quality, size-standardized single-cell images directly from real-world clinical samples. This confirms the robustness of the size filter and its suitability as the target resolution.

Table 2. Number of extracted RBC images per disease and size group, including excluded categories.

Sample	Touching Result	16 sizes	32 sizes	128 sizes	256 sizes	512 sizes	1024 sizes	Over size
IDA	35	853	211	912	2	1	0	0
TT	59	40	298	2393	16	0	0	0
HbH	68	22	232	1599	185	7	0	0
HbE	71	96	415	3244	132	0	0	0
Hb E (Sx)	49	408	723	674	173	19	0	0
Homo HbE	66	150	291	3407	18	0	0	0

3.3. Quantitative Evaluation of RBC Extraction and Size-Based Filtering

This section quantitatively evaluates the effectiveness and robustness of the proposed RBC extraction pipeline across six hematological conditions. Rather than reporting raw image counts alone, the analysis focuses on extraction yield, size distribution stability, and rejection behavior to assess the suitability of the generated dataset for downstream machine learning applications. Table 2 summarizes the number of extracted objects categorized by size group and exclusion type. Across all disease classes, the pipeline consistently produced a high proportion of isolated single-cell RBC images, with the 128×128 pixel group emerging as the dominant and most reliable category. In total, 12,229 images were assigned to this group, accounting for more than 80% of all valid extractions. This consistency across disease types supports the selection of 128×128 pixels as the standard reference resolution for CNN-based analysis. The proportion of edge-touching cells ranged from approximately 5% to 10% depending on the smear density and pathological condition. Higher rejection rates were

observed in samples with dense cellular distribution, such as Hb E (Sx) and Homo HbE, reflecting increased cell crowding and overlap. Importantly, the boundary-contact rejection strategy effectively prevented incomplete or partially segmented cells from entering the training dataset, thereby reducing label noise and preserving morphological integrity. Very small objects ($\leqslant$16 pixels) and 32-pixel candidates were primarily associated with smear debris, platelet fragments, or staining artifacts rather than true RBC morphology. These categories were therefore excluded from the primary dataset. Larger size groups (256×256 pixels and above) occurred less frequently and typically represented touching or closely paired RBCs, which are unsuitable for single-cell classification without further instance separation. Objects exceeding 1024 pixels were rare and consistently excluded as oversized clusters. Overall, the quantitative distribution of extracted images demonstrates that the proposed pipeline achieves both selectivity and stability across diverse pathological conditions. By concentrating the majority of valid outputs into a single, morphology-preserving size class while systematically rejecting ambiguous or low-quality regions, the pipeline produces a dataset that is well-aligned with the requirements of reproducible and robust CNN training.

3.4. Morphological Group Analysis and Rationale for Data Inclusion

Beyond size-based categorization, the extracted objects were further analyzed according to their morphological characteristics and structural completeness. Representative examples of each morphological group are illustrated in Fig. 4, providing visual context for the classification and exclusion criteria applied in this study. This analysis establishes a principled rationale for data inclusion and exclusion decisions, particularly in the context of training CNNs, where label noise and structural ambiguity can significantly degrade model performance. The 128×128 pixel group represents the primary category of interest, as it predominantly contains isolated, well-formed single RBCs with clearly preserved morphology. Cells in this group typically exhibit intact boundaries, uniform staining, and minimal background interference. These properties are essential for CNN-based feature learning, as they allow models to focus on intrinsic morphological patterns rather than artifacts introduced by segmentation errors or inconsistent scaling. Occasional isolated WBCs were also observed in this group; however, their distinct morphology allows for future multi-class or exclusion-based handling without compromising dataset integrity. The 256×256 pixel group mainly consists of touching or closely paired RBCs. Although these objects retain substantial morphological information, their overlapping boundaries introduce ambiguity in instance-level labeling. Including such samples without explicit cell separation may lead CNNs to learn mixed or distorted features that do not correspond to true single-cell morphology. Consequently, these images were excluded from the primary training set but retained for potential future research on instance segmentation, ellipse fitting, or cell separation algorithms. Objects assigned to the 32×32 pixel group were primarily identified as platelets, platelet aggregates, or small debris. While these elements are not suitable for RBC morphology analysis, they may hold value in broader hematological studies. To prevent false-positive learning and class contamination, this group was excluded from the RBC dataset but preserved as a separate category for potential downstream applications. The $\leqslant$16 pixel group contained very small fragments resulting from staining noise, smear artifacts, or residual debris. These objects lack diagnostic or morphological relevance and were therefore discarded entirely. The touching result category, defined by edge-contact detection, includes partially

visible or truncated cells located at ROI boundaries. Such incomplete representations can introduce severe label noise and bias CNNs toward learning artificial edge patterns. Excluding these samples ensures that the dataset contains only structurally complete cells, thereby improving training stability and generalization. Finally, 512×512, 1024×1024, and oversized groups represent large clusters or stain blobs where individual cell boundaries are indistinguishable, even under manual inspection. These objects were excluded due to their incompatibility with single-cell analysis.

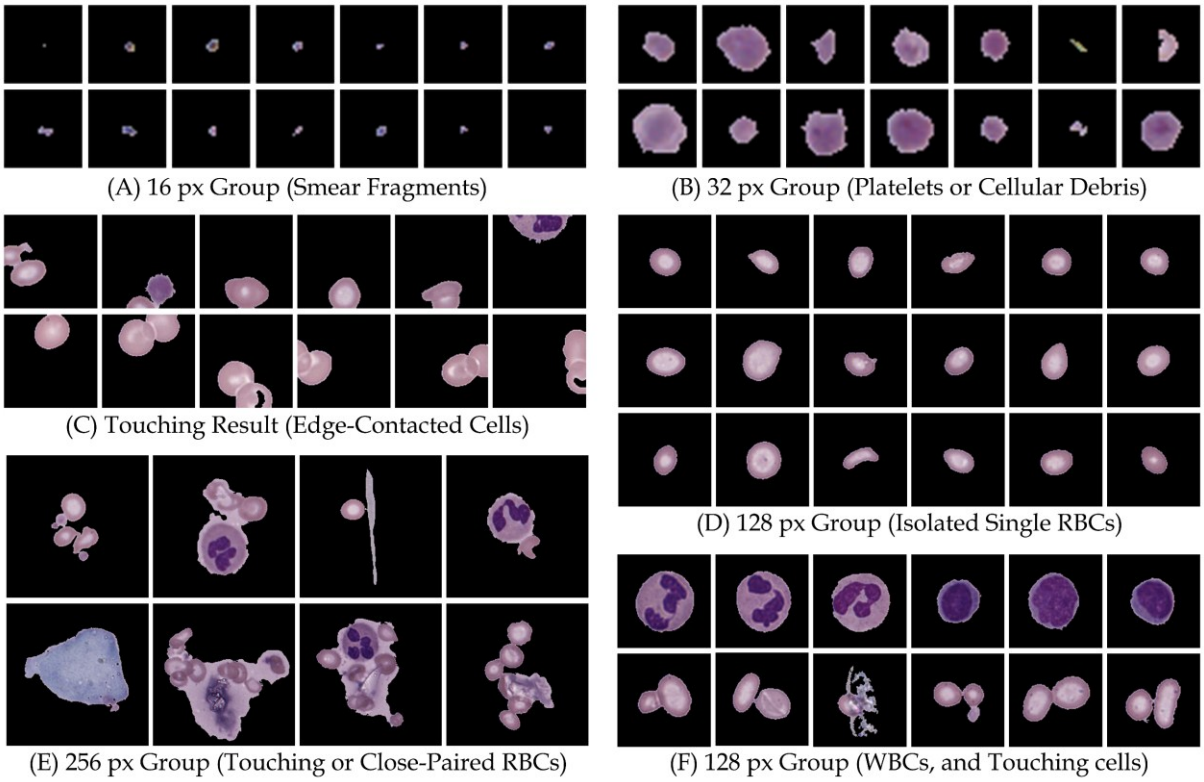

Fig. 4. Representative examples from each morphological group: Touching, 16, 32, 128, and 256.

Collectively, this morphology-aware filtering strategy ensures that the final dataset emphasizes structural completeness, scale consistency, and diagnostic relevance. By explicitly excluding ambiguous or noisy samples, the proposed pipeline produces a curated dataset optimized for robust machine learning training while maintaining traceability for future algorithmic extensions. Although labels are assigned at the disease level, each category reflects dominant and recurring RBC morphological patterns associated with the corresponding hematological condition. Therefore, disease-level labeling is appropriate for baseline morphology learning, where the objective is to capture consistent structural traits rather than to perform definitive diagnosis.

3.5. Baseline CNN Evaluation for Dataset Usability

To demonstrate the practical usability of the extracted RBC dataset for CNN-based analysis, a baseline morphological classification experiment was conducted. This evaluation aims to verify that the proposed extraction pipeline produces morphology-preserved images that are learnable by standard CNN architectures, rather than to establish an optimized or clinically deployable model. All experiments used standardized 128×128 pixel RBC images,

which represent the most stable and dominant size group identified in previous analyses. Two baseline CNN architectures—a 4-layer CNN and an 8-layer CNN—were trained for 100 epochs using a fixed learning rate of 1.00×10^{-6} to emphasize training stability. Figure 5 presents the training and validation accuracy curves for both models. The 4-layer CNN (Fig. 5a) shows gradual and stable convergence with a small gap between training and validation accuracy, indicating that the dataset contains sufficient and consistent morphological information for effective feature learning. In contrast, the deeper 8-layer CNN (Fig. 5b) achieves higher training accuracy but exhibits larger fluctuations in validation performance, reflecting increased sensitivity to model complexity and training configuration. Importantly, both models successfully learned meaningful morphological patterns without training collapse or severe instability. These results confirm that the extracted dataset is compatible with CNN-based learning. At the same time, the observed differences between model depths highlight that further optimization of architecture design, regularization, and training strategies is necessary for task-specific applications. Overall, this baseline evaluation serves as a proof of dataset usability, demonstrating that the proposed pipeline provides a reliable foundation for future CNN-based morphological classification studies.

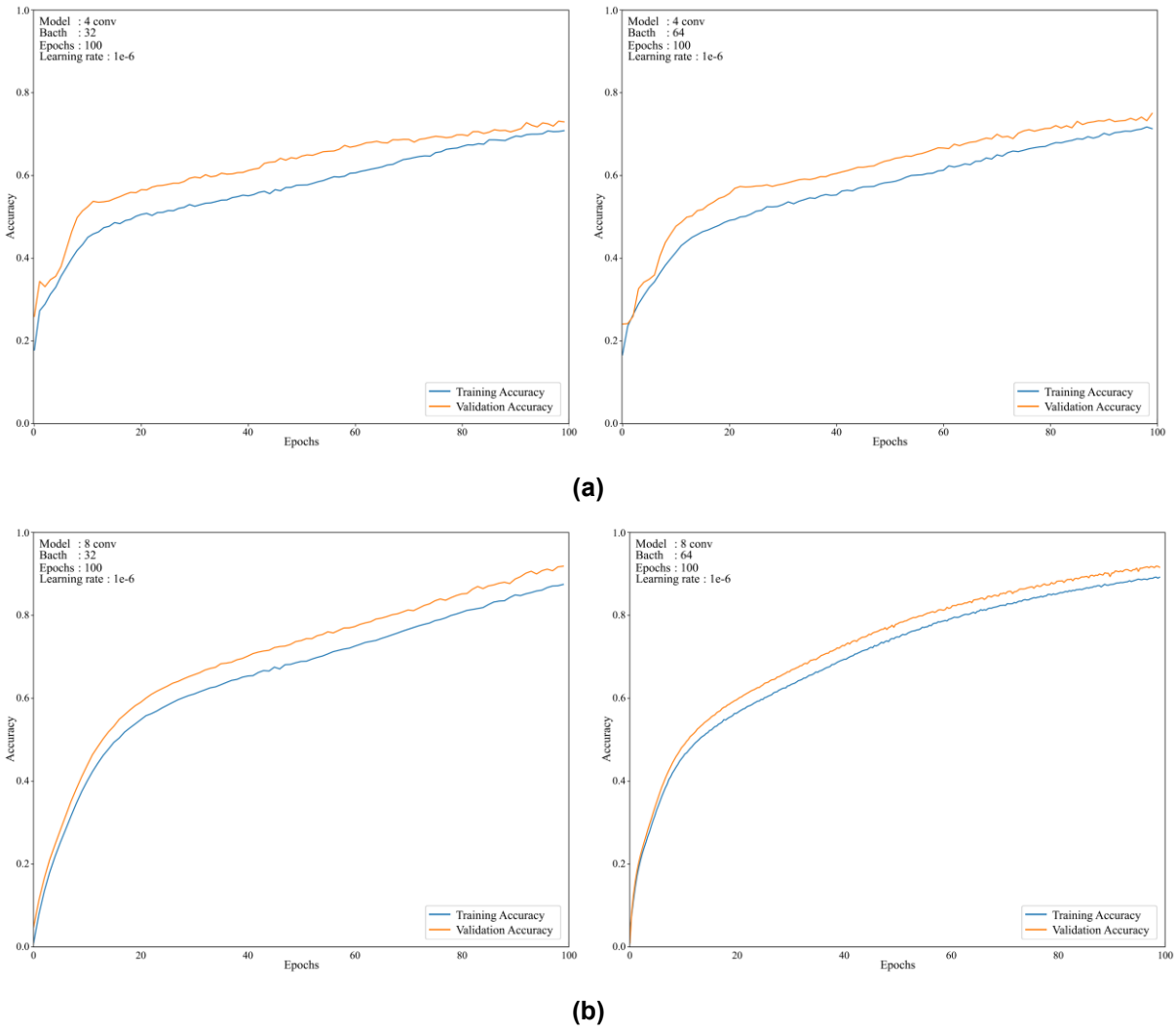

Fig. 5. Training and validation accuracy of baseline CNN models demonstrating dataset learnability across 100 epochs using a learning rate of 1.00×10^{-6}: a) Results from the 4-layer CNN model; and b) results from the 8-layer CNN model.

4. DISCUSSIONS

4.1. Baseline CNN Evaluation for Dataset Usability

This study focuses on the development of a reproducible pipeline for extracting morphology-preserved single-cell RBC images from real clinical blood smear slides. Rather than proposing a fully optimized classification system, the primary objective is to validate that the extracted dataset is structurally consistent, learnable, and suitable for downstream CNN–based analysis. The quantitative extraction results demonstrate that the proposed pipeline consistently produces a large proportion of isolated single RBC images within a standardized resolution range. The dominance of the 128×128 pixel group across all disease categories indicates that the size-based filtering strategy effectively captures complete RBC morphology while suppressing artifacts, debris, and ambiguous cell clusters. This consistency is critical for CNN training, as uncontrolled variation in scale and background content can lead to unstable feature learning and reduced generalization. The baseline CNN evaluation further supports the usability of the dataset. As shown in Fig. 5, both shallow and deeper CNN architectures were able to learn meaningful morphological patterns from the extracted images without training collapse or severe instability. The stable convergence observed in the 4-layer CNN suggests that the dataset contains sufficiently strong and consistent morphological signals. In contrast, the increased fluctuation observed in the 8-layer CNN highlights the sensitivity of deeper architectures to model complexity, training configuration, and dataset scale. Importantly, these findings should not be interpreted as an assessment of optimal classification performance. Instead, they demonstrate that the proposed extraction pipeline produces data that are compatible with CNN-based learning and capable of supporting further methodological development. The observed performance differences between model depths emphasize that classification accuracy is influenced not only by data quality but also by architectural design, regularization strategies, and hyperparameter selection, all of which require task-specific optimization. Overall, the results confirm that the extracted dataset provides a reliable foundation for future research involving advanced CNN architectures, fine-grained morphological classification, and quantitative shape analysis. By decoupling dataset validation from model optimization, this study establishes a transparent and reproducible baseline that can be extended in subsequent work.

4.2. Comparison with Prior RBC Datasets and Morphological Studies

Previous studies on RBC image analysis have demonstrated the potential of image processing and deep learning techniques for hematological classification. However, many existing datasets and pipelines differ substantially in terms of data source, scale, and clinical grounding, which limits direct comparison and generalizability. Several publicly available RBC datasets are constructed from manually cropped images, synthetic data, or highly curated microscope fields. While these datasets are useful for algorithm development, they often lack standardized background control, consistent spatial resolution, or explicit handling of touching and incomplete cells. As a result, CNN models trained on such data may inadvertently learn background patterns, scale-specific artifacts, or annotation biases rather than intrinsic RBC morphology. In contrast, the present study emphasizes automated extraction from whole-slide images (WSIs) derived from real clinical blood smears. By incorporating contour-based segmentation, edge-contact rejection, and size-standardized

overlay, the proposed pipeline explicitly addresses common sources of noise in RBC datasets. This design choice prioritizes morphological completeness and structural consistency over raw dataset size, which is particularly important for reproducible CNN training. From a population perspective, most existing RBC datasets are derived from non-Southeast Asian cohorts and rarely reflect the high prevalence of Hb E–related disorders. The dataset presented in this study is specifically constructed from Thai patients with clinically confirmed diagnoses, providing a regionally relevant resource that captures morphology patterns underrepresented in existing benchmarks. This distinction is critical for developing models intended for deployment in endemic regions, where morphology distributions may differ substantially from those in commonly used datasets. It is important to note that this study does not aim to outperform prior classification systems or establish state-of-the-art accuracy. Instead, its contribution lies in providing a transparent and reproducible dataset generation framework, along with baseline evidence of CNN learnability. By clearly separating dataset validation from model optimization, this work complements existing algorithm-focused studies and establishes a foundation for future comparative and multi-center research.

4.3. Limitations and Future Work

Despite the strengths of the proposed extraction pipeline and dataset, several limitations should be acknowledged to place the findings in proper context and guide future research directions. First, the dataset is constructed from a limited number of patients, with one clinically confirmed case representing each disease category. While this design ensures clear diagnostic ground truth and avoids label ambiguity, it does not capture the full inter-patient variability commonly observed in clinical practice. Consequently, the current dataset should be regarded as a morphology-preserved reference set rather than a population-scale benchmark. Second, the baseline CNN evaluation presented in this study is intentionally conservative. The models were not extensively optimized in terms of architecture depth, data augmentation, or regularization strategies. As a result, the reported training and validation behaviors should not be interpreted as upper-bound performance. Instead, they serve to demonstrate dataset learnability and structural consistency. Future work should explore advanced CNN architectures, systematic hyperparameter tuning, and cross-validation protocols to improve classification robustness. Third, the current annotations are limited to disease-level grouping and size-based filtering. Fine-grained morphological labels—such as target cells, elliptocytes, or anisocytosis—were not explicitly annotated. Incorporating expert-driven or semi-automated morphological labeling would enable more detailed phenotype analysis and support multi-label or hierarchical classification tasks. In addition, although the pipeline effectively rejects touching and incomplete cells, densely packed RBC regions remain challenging for single-cell analysis. Future extensions may integrate instance segmentation, ellipse fitting, or deep-learning-based separation methods to recover usable cells from complex clusters. Future work will also focus on expanding the dataset to include additional patients, institutions, and staining variations, thereby improving generalizability. The modular design of the proposed pipeline allows seamless integration of stain normalization, color augmentation, and domain adaptation techniques without restructuring the core workflow. Overall, these limitations highlight opportunities for methodological extension rather than weaknesses of the proposed approach. By providing a transparent extraction framework and a validated baseline dataset, this study establishes a solid foundation for

subsequent research in CNN-based RBC morphology analysis and population-specific hematological modeling.

5. CONCLUSIONS

This study presents a semi-automated and reproducible pipeline for extracting high-quality, morphology-preserved single RBC images from whole-slide blood smear scans of Thai patients. By integrating contour-based segmentation, edge-contact rejection, and size-standardized overlay, the proposed method effectively isolates complete single-cell RBC images while suppressing artifacts, debris, and ambiguous cell clusters commonly encountered in real clinical smears. Quantitative analysis demonstrates that the pipeline consistently produces a dominant and stable group of 128×128 pixel RBC images across six hematological conditions. This standardized resolution preserves essential morphological information while remaining computationally efficient, making it well suited for CNN–based analysis. Baseline CNN experiments further confirm that the extracted dataset is learnable by standard CNN architectures, providing clear evidence of dataset usability without claiming optimized or clinically deployable performance. Rather than focusing on classification accuracy, the contribution of this work lies in establishing a transparent dataset generation framework and validating its compatibility with deep learning workflows. The dataset reflects morphology patterns from a clinically confirmed Thai cohort, addressing a gap in region-specific RBC resources that are underrepresented in existing benchmarks. The modular design of the proposed pipeline enables straightforward extension to larger cohorts, multi-center data, and advanced modeling techniques, including deeper CNNs, instance segmentation, and quantitative shape analysis. As such, this work provides a solid foundation for future research in population-specific digital hematopathology and CNN-based RBC morphology studies. Importantly, the proposed framework emphasizes dataset quality and reproducibility over task-specific model optimization, ensuring that subsequent studies can build upon a well-defined and transparent data foundation. By decoupling dataset construction from classification performance claims, this work promotes fair comparison, reuse, and extension across different modeling strategies. The presented pipeline and dataset therefore serve not only as a technical contribution, but also as an enabling resource for future methodological and clinical research in digital hematopathology.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Ubon Ratchathani University (protocol code UBU–REC–77/2568 and date of approval: 12 March 2025).

Informed Consent Statement: Informed consent was waived due to the use of de-identified blood smear images, which were obtained from Surasak Wanram, College of Medicine and Public Health, Ubon Ratchathani University. The images contained no personally identifiable information, and all data usage complied with institutional policies.

Data Availability Statement: The data supporting the reported results of this study are not publicly available due to privacy and ethical restrictions related to patient confidentiality. However, deidentified datasets may be made available from the corresponding author upon reasonable request and with appropriate ethical approval.

Acknowledgments: The authors would like to acknowledge the Faculty of Science and College of Medicine and Public Health, Ubon Ratchathani University, for providing administrative and technical support. We also thank the hematology laboratory staff for their assistance in sample preparation and provision of blood smear slides.

REFERENCES

[1] G. d'Onofrio, G. Zini, *Morphology of Blood Disorders*, 2nd edition, New Jersey: Wiley-Blackwell, 2014.

[2] J. Ford, "Red blood cell morphology," *International Journal of Laboratory Hematology*, vol. 35, no. 3, pp. 351–357, 2013, doi: 10.1111/ijlh.12082.

[3] S. Tariq et al., "Diagnosis of thalassemia and iron deficiency anemia using confocal and atomic force microscopy," *Laser Physics Letters*, vol. 14, no. 11, p. 115703, 2017, doi: 10.1088/1612-202X/aa8bca.

[4] P. Wasi et al., "Thalassemia in Thailand," *Annals of the New York Academy of Sciences*, vol. 344, pp. 352–363, 1980, doi: 10.1111/j.1749-6632.1980.tb33674.x.

[5] N. Teawtrakul et al., "The impact and disease burden of thalassemia in Thailand," *The Journal of the Medical Association of Thailand*, vol. 95, no. 7, pp. 211–216, 2012.

[6] T. Munkongdee et al., "Update in laboratory diagnosis of thalassemia," *Frontiers in Molecular Biosciences*, vol. 7, pp. 74–86, 2020, doi: 10.3389/fmolb.2020.00074.

[7] J. Sharif et al., "Red blood cell segmentation using masking and watershed algorithm: A preliminary study," International Conference on Biomedical Engineering, 2012, pp. 258–262, doi: 10.1109/ICoBE.2012.6179016.

[8] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979, doi: 10.1109/TSMC.1979.4310076.

[9] F. Chollet, *Deep Learning with Python*, 2nd edition, New York: Manning Publications, 2022.

[10] G. Litjens al., "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017, doi: 10.1016/j.media.2017.07.005.

[11] G. Huang, Z. Liu, L. Maaten, K. Weinberger, "Densely connected convolutional networks," IEEE Conference on Computer Vision and Pattern Recognition, 2017, doi: 10.1109/CVPR.2017.243.

[12] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," IEEE Conference on Computer Vision and Pattern Recognition, 2016, doi: 10.1109/CVPR.2016.90.

[13] *Leica Biosystems Aperio AT2 Digital Pathology Scanner User Guide*, Nussloch: Leica Biosystems, 2018.

[14] *OpenSlide Library Documentation*, Version 3.4.1, OpenSlide.org, 2015.

[15] ASUS, *ASUS TUF Gaming F15 FX506LHB User Manual*, Taipei: ASUS, 2020.

[16] NVIDIA Corporation, *CUDA Toolkit Documentation*, Version 11.4. Santa Clara: NVIDIA, 2021.

[17] A. Geron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd edition, California: O'Reilly Media, 2019.

[18] J. Solem, *Programming Computer Vision with Python,* California: Creative Commons, 2012.

[19] H. Lim, M. Mashor, R. Hassan, "White blood cell segmentation for acute leukemia bone marrow images," International Conference on Biomedical Engineering, 2012, doi: 10.1109/ICoBE.2012.6179038.

[20] R. Hegd et al., "Peripheral blood smear analysis using image processing approach for diagnostic purposes: a review," *Biocybernetics and Biomedical Engineering*, vol. 38, no. 3, pp. 467–480, 2018, doi: 10.1016/j.bbe.2018.03.002.