International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Deep Learning-Based Smart AI System for Blood Cell Classification and Anemia Detection

Vishrutha Rani S N, Manushree N*, Manisha K Sakre, Manasi N, Disha Jain

Department of Information Science and Engineering, Jawaharlal Nehru New College of Engineering, Shivamogga, Karnataka, India - 577201

Abstract - Anemia affects an estimated 1.92 billion people worldwide, yet automated hematology analyzers remain unaffordable in many low- and middle-income settings, and manual peripheral blood smear (PBS) examination is slow and varies between observers. This paper presents HemaVision, an open-access, dual-stream multimodal research prototype that combines PBS photomicrographs with tabular Complete Blood Count (CBC) indices for preliminary anemia screening. The framework integrates a YOLOv8n detector for blood cell enumeration (89.14% mAP@0.50); a MobileNetV3-Small smear classifier trained on 811 perceptual-hash-audited image clusters to prevent cross-split contamination (75.19% accuracy, 0.8167 ROC-AUC); a Random Forest CBC classifier evaluated under a circular target leakage audit, in which supplying hemoglobin merely memorizes the WHO definition (99.51% accuracy) while a hemoglobin-free model reaches 83.56% accuracy and 0.8974 ROC-AUC; and two fusion strategies, a 12-dimensional Stacking Ensemble (99.44% accuracy) and a Gated Multimodal Unit (GMU) with missing-modality gating (97.50% accuracy). The system is delivered as a FastAPI service with Streamlit and React interfaces and an automated PDF reporting engine. Because the image and tabular datasets lack patient-level linkage, fusion metrics were obtained on synthetically paired data and reflect algorithmic behavior rather than clinical efficacy. HemaVision is an investigational research prototype and not a cleared medical device.

Key Words: anemia screening, multimodal deep learning, peripheral blood smear, complete blood count, gated multimodal unit, target leakage audit.

1. INTRODUCTION

Anemia is a widespread global public health disorder characterized by a quantitative deficiency in circulating erythrocytes or an abnormally low concentration of intracellular hemoglobin, thereby impairing systemic physiological oxygen transport to peripheral tissues [1], [2]. According to epidemiological assessments published by the World Health Organization (WHO) and the Global Burden of Disease study, anemia afflicts approximately 1.92 billion individuals globally (roughly 24.3% of the human population), with the highest morbidity concentrated among preschool children, pregnant women, and economically marginalized populations in low- and middle-income countries (LMICs) [1]. Chronic or unmanaged anemia precipitates cognitive impairment, diminished working capacity, maternal mortality, and elevated cardiovascular risks [2].

Clinical screening and diagnosis in modern healthcare systems rely primarily on two diagnostic modalities:

- *Automated Complete Blood Count (CBC):* Conducted using multi-channel automated hematology analyzers utilizing electrical impedance (Coulter principle), laser flow cytometry, and spectrophotometry. Although rapid and highly quantitative, automated analyzers carry capital costs ranging from $15,000 to $80,000 USD, require continuous cold-chain chemical reagents, demand constant recalibration, and depend on uninterrupted electrical infrastructure, rendering them inaccessible in remote and resource-constrained health centers [3], [4].
- *Peripheral Blood Smear (PBS) Microscopy:* The standard morphological examination method wherein a thin capillary or venous blood film is smeared on a glass slide, stained with Romanowsky-type dyes (Wright-Giemsa or Leishman), and visually inspected under an optical light microscope at 400× to 1000× oil-immersion magnification. Cytopathologists visually evaluate erythrocyte cellularity, variations in size (anisocytosis), variations in shape (poikilocytosis), central pallor expansion (hypochromia), and white blood cell differential counts. However, manual PBS examination requires 15 to 30 minutes per slide, is susceptible to optical fatigue, and demonstrates inter-observer classification divergence exceeding 15% to 20% under high clinical workloads [4], [5].

Artificial intelligence (AI), particularly computer vision (CV) and supervised machine learning (ML), presents unprecedented opportunities to automate and scale hematological triage [5], [6]. Deep convolutional neural networks (CNNs) extract hierarchical spatial representations directly from digital photomicrographs, while gradient-boosted decision trees and multi-layer perceptrons model complex nonlinear interactions among tabular laboratory parameters [6], [7].

Despite substantial literature, existing automated hematology systems remain almost universally unimodal:

- *Vision-Only Systems:* Evaluate erythrocyte crops or microscopic fields of view (FOVs) to detect sickle cells, malaria trophozoites, or morphological abnormalities [8]–[10], but cannot quantify systemic hemoglobin concentration or cellular volume distributions.
- *Tabular-Only Systems:* Ingest CBC panel values to classify hematological diseases [11]–[13], but lack the capacity to inspect visual cytomorphology, detect parasitological inclusions, or identify slide-level cellular artifacts.

Combining microscopic image cytology with tabular laboratory parameters within a single unified screening platform represents a compelling yet underexplored paradigm [14], [15]. However, the development of multimodal clinical AI systems introduces critical methodological pitfalls that are frequently obscured in contemporary literature:

- *Data Contamination:* Public microscopic datasets frequently contain uncurated near-duplicate frames and overlapping fields of view from the same patient slide,

which artificially inflate performance metrics when randomly split [16].

- *Circular Target Leakage:* Public tabular datasets often derive target diagnostic labels directly via rigid hemoglobin threshold cutoffs (e.g., <12.0 g/dL for females, <13.0 g/dL for males). Supplying raw hemoglobin as an input feature allows decision trees to trivially memorize the labeling rule rather than learning pathological disease patterns [17].
- *Unlinked Data Fusion:* High multimodal accuracies are routinely reported without disclosing that the underlying image and laboratory datasets originate from completely separate, unlinked cohorts [14], [18].

To address these challenges, this paper presents HemaVision, an open-access, dual-stream multimodal machine learning framework designed for preliminary anemia screening. The framework accepts a digital blood smear photomicrograph, a set of tabular CBC parameters, or both, dynamically adapting its inference pipeline across missing modalities.

1.1 Contributions

The principal contributions of this work are as follows:

- *End-to-End Multimodal Framework:* We design, train, and deploy an asynchronous screening system combining an anchor-free YOLOv8n cell enumerator, a MobileNetV3-Small cytomorphology classifier, a tabular Random Forest panel, and two multimodal fusion mechanisms (Stacking Ensemble and Gated Multimodal Unit).
- *Perceptual Hash Near-Duplicate Audit and Cluster Remediation:* We execute an exhaustive difference-hash (dHash) and deep cosine similarity audit across 837 blood smear images, discovering 27 candidate duplicate pairs and 14 cross-split contaminated pairs in standard baselines. We construct 811 atomic, indivisible slide clusters, showing that cluster-stratified retraining yields a defensible performance of 75.19% accuracy and 0.8167 ROC-AUC.
- *Empirical Circular Target Leakage Audit:* We formally identify and audit the target leakage mechanism inherent in CBC machine learning models. We demonstrate that while standard models achieve 99.51% to 100.0% accuracy through hemoglobin memorization, a genuine leakage-aware screening model strictly excluding hemoglobin achieves 83.56% accuracy and 0.8974 ROC-AUC by redistributing decision splits across erythrocyte cytometric features (MCV, MCH, MCHC).
- *Multi-Paradigm Multimodal Integration:* We implement and compare a 12-dimensional decision-level Stacking Ensemble against a deep feature-level Gated Multimodal Unit (GMU) equipped with explicit missing-modality binary masks, evaluating both across held-out synthetic test cohorts.
- *Deployable Full-Stack Diagnostic Prototype:* We engineer an asynchronous FastAPI REST service with cached singleton inference, a Python Streamlit laboratory dashboard, a React 19 / TanStack enterprise single-page interface, and an automated clinical PDF report generator (FPDF2) with persistent regulatory disclaimers.

1.2 Scope and Intended Use

HemaVision is an investigational research prototype developed exclusively for scientific exploration, algorithmic benchmarking, and educational triage research. It has not undergone prospective multi-center clinical trials, lacks 510(k)/De Novo clearance from the U.S. FDA, and holds no CE-mark certification under European Union Medical Device Regulation (EU MDR 2017/745). It must not be deployed as a standalone diagnostic instrument, used to prescribe medication or blood transfusions, or substituted for standard laboratory pathology workflows. All diagnostic probabilities represent algorithmic estimates; definitive clinical diagnosis requires standard quantitative venous hemoglobin testing confirmed by a licensed pathologist.

1.3 Organization of the Paper

Section 2 provides a literature review and taxonomy of related work. Section 3 details the datasets, preprocessing, model architectures, mathematical formulations, algorithmic workflows, and software implementation across backend, frontend, database, and reporting layers. Section 4 presents the experimental setup, evaluation metrics, held-out test results, and runtime profiles. Section 5 analyzes the findings, target leakage dynamics, and fusion behaviors. Section 6 discusses limitations and future research directions and concludes the paper.

2. LITERATURE REVIEW

The application of machine learning to hematology spans multiple disparate domains: computer vision for microscopic cytology, machine learning for complete blood count interpretation, multi-omics biomarker discovery, and deep multimodal data integration. As illustrated in Fig. 1, the surveyed literature is categorized into four primary paradigms: (i) image-based cytology and cell detection, (ii) tabular machine learning on hematological indices, (iii) deep learning for disease classification, and (iv) multimodal integration studies.

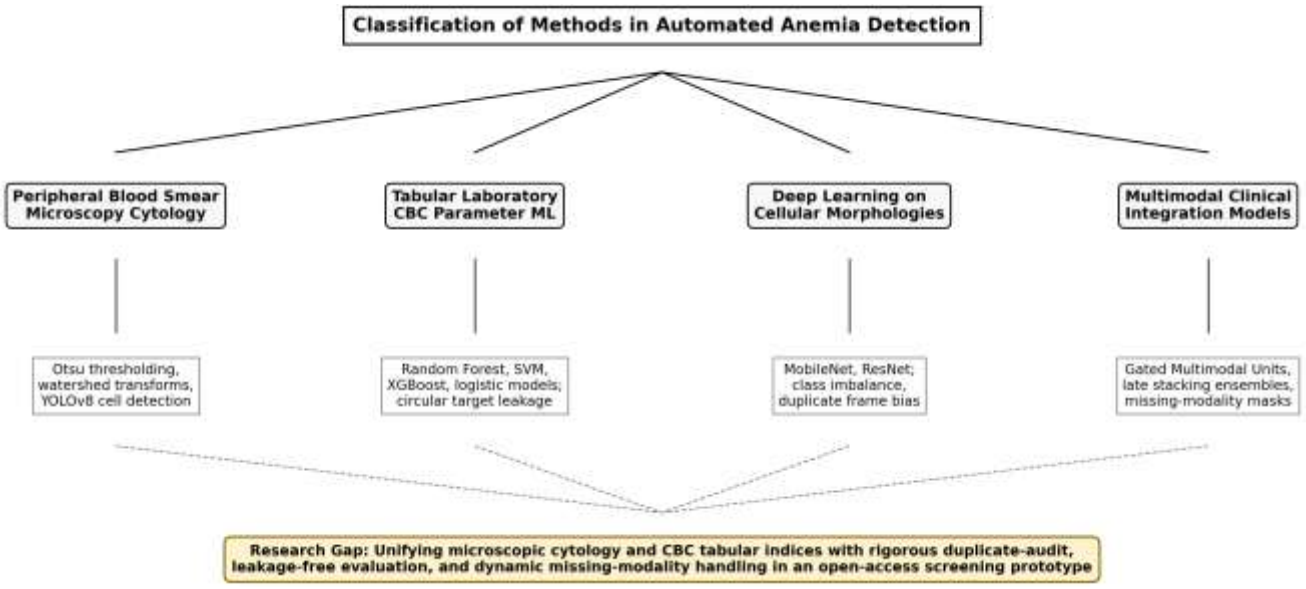


Fig -1: Classification of Methods Used in Automated Anemia Detection and Screening

2.1 Image-Based Cytology and Cell Detection

Automated segmentation and counting of blood cells from digital microscopy is a foundational task in digital pathology. Shenggan [8] curated the open-access Blood Cell Count and Detection (BCCD) dataset, providing 364 microscopic annotations of RBCs, WBCs, and platelets. Early attempts utilized traditional morphological image processing, including Otsu thresholding, watershed transforms, and circular Hough transforms, but exhibited extreme vulnerability to non-uniform illumination and overlapping erythrocyte rouleaux [9].

With the advent of deep object detection, single-stage anchor-free architectures such as YOLO (You Only Look Once) established dominant benchmarks. Jocher et al. [10] developed YOLOv8, introducing an anchor-free split-head architecture and Mosaic augmentations that achieved superior spatial efficiency on dense object arrays. However, applying object detection to blood smears introduces specific domain challenges: severe class imbalance (erythrocytes outnumber platelets and leukocytes by up to 100:1) and chromatic channel ordering inconsistencies (RGB vs. BGR conversions) that can invert cellular color representations and cause false leukocyte classifications [10].

2.2 Tabular Machine Learning on Hematological Indices

Tabular analysis of complete blood counts represents the most common machine learning application in hematology. Sanadhya et al. [11] evaluated multi-layer perceptrons, Support Vector Machines (SVM), and Random Forests on clinical blood count parameters, reporting diagnostic accuracies exceeding 98%. Similarly, AlSalemi et al. [12] implemented ensemble models to differentiate iron deficiency anemia from β-thalassemia trait using red cell cytometric indices.

However, a fundamental methodological flaw permeates numerous published CBC studies: circular target leakage [17]. In clinical practice, the World Health Organization defines anemia mathematically as venous hemoglobin falling below specific physiological thresholds (<12.0 g/dL in non-pregnant adult females, <13.0 g/dL in adult males) [1]. Public machine learning repositories frequently generate the binary classification target (Result) by applying this exact threshold equation to the Hemoglobin column. Models trained with raw hemoglobin as an input feature do not discover latent biological disease markers; instead, decision trees simply isolate the threshold boundary, yielding artificially perfect metrics (99% to 100%) that fail to provide real-world screening utility when hemoglobin is unavailable [17].

2.3 Deep Learning for Erythrocyte Cytomorphology

Convolutional neural networks have been widely applied to classify blood smear fields into pathological categories. Howard et al. [6] introduced MobileNetV3, utilizing hardware-aware Neural Architecture Search (NAS), hard-swish activation functions, and Squeeze-and-Excitation (SE) channel attention modules to achieve high representational capacity with low computational footprint.

In hematology, researchers have applied MobileNet, ResNet, and VGG backbones to the public AneRBC dataset [19] to categorize smears as normal or anemic. Nonetheless, published baselines suffer from two critical defects: (1) specificity collapse, in which extreme class imbalance (typically 5:1 to 6:1 favoring positive cases) causes unweighted cross-entropy loss to drive predictions entirely toward the majority class, yielding high overall accuracy alongside catastrophic false-positive rates (specificity as low as 5.0%) [16]; and (2) optical near-duplicate contamination, in which digital microscopy repositories frequently extract multiple image crops from identical glass slides or adjacent focal planes, so that near-identical frames leak across training and testing partitions and yield over-optimistic evaluation metrics [16].

2.4 Multimodal Clinical Integration

Multimodal learning integrates heterogeneous data sources—such as imaging, structured laboratory records, and clinical text—to enhance diagnostic robustness [14], [15]. Arevalo et al. [14] proposed the Gated Multimodal Unit (GMU), which employs an internal sigmoid gating mechanism to dynamically regulate the intermediate feature contributions of visual and textual modalities. Kumar et al. [15] applied multimodal transformers to pulmonary disease detection by combining radiographs and clinical metadata. Ballard et al. [18] systematically reviewed multi-omics and multimodal integration, emphasizing that intermediate feature-level fusion and late decision-level ensembles generally outperform naive early concatenation by preventing high-dimensional image tensors from overwhelming low-dimensional tabular features.

2.5 Comparative Summary and Research Gap

Table -1 summarizes principal studies in automated hematological analysis. No existing open-access system integrates deep visual cytology with tabular CBC parameters into a unified, crash-safe screening application while transparently exposing near-duplicate slide leakage, circular hemoglobin target leakage, and multimodal dataset unlinking.

Table -1: Summary of Related Work in Automated Hematology and Anemia Detection

Ref. / Study	Modality, architecture and task	Key result	Identified limitations
[8] Shenggan (2019)	Microscopy; Faster R-CNN; RBC, WBC and platelet detection	mAP ~0.80	Small cohort (364 images); severe class imbalance.
[9] Acevedo et al. (2020)	Microscopy; VGG-16; leukocyte subtypes	96.2% accuracy	Requires manual single-cell pre-cropping; no CBC.
[11] Sanadhya et al. (2022)	Tabular CBC; Random Forest; anemia screening	99.1% accuracy	Severe circular target leakage via raw hemoglobin.
[12] AlSalemi et al. (2023)	Tabular CBC; SVM / XGBoost; anemia differentiation	95.4% accuracy	Single-center cohort; no visual cytology.
[19] AneRBC Baseline	Microscopy; MobileNetV2; anemia screening	80.6% accuracy	Specificity collapsed to 5.0%; unpurged duplicate slides.
Ours: HemaVision	**Smear (YOLOv8 + MobileNet) + CBC; Stacking and GMU; anemia screening and cell enumeration**	**Stacking 99.44%, GMU 97.50%, leak-free CBC 83.56%**	**Evaluated on synthetic multimodal combinations; public data unlinked.**

3. METHODOLOGY

3.1 System Overview

HemaVision operates as an asynchronous, dual-stream diagnostic pipeline capable of processing multimodal or single-modality inputs, as shown in Fig. 2. The framework accepts a digital blood smear photomicrograph, a set of 5 tabular CBC parameters, or both, validates data integrity, executes modality-specific deep feature extraction, and routes predictions through a late Stacking Ensemble or an experimental deep Gated Multimodal Unit (GMU).

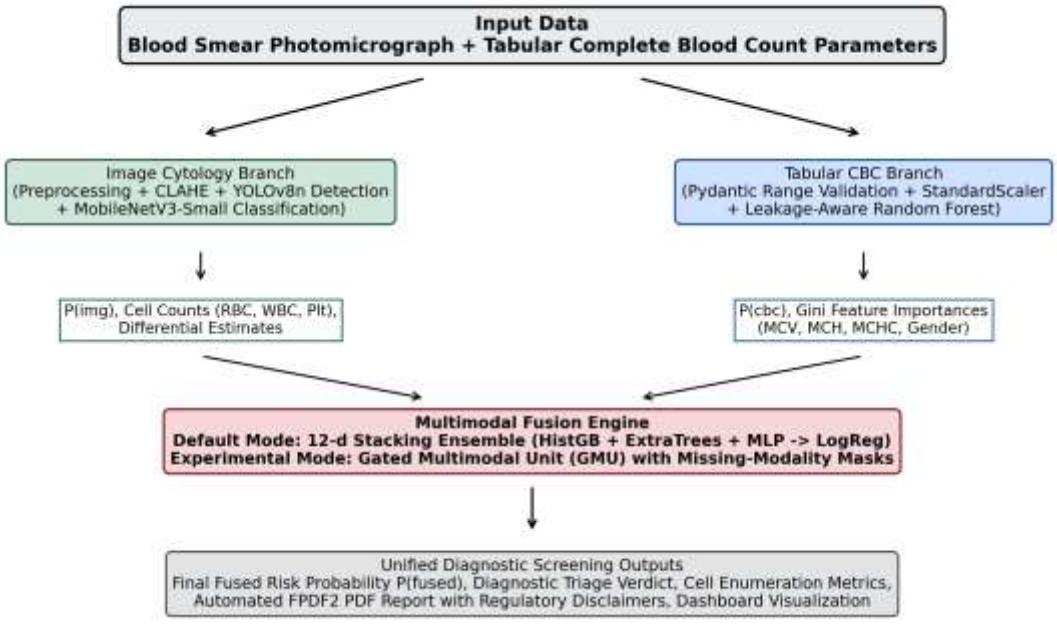


Fig -2: Block Diagram of the Proposed Multimodal Screening Framework

3.2 Dataset Provenance and Cohorts

The empirical framework is constructed from four open-access repositories, whose partitions are summarized in Table -2:

- *BCCD Cell Detection Dataset:* Curated by Shenggan [8], containing 364 peripheral blood smear images with 4,888 annotated bounding boxes (4,155 RBCs, 372 WBCs, 361 platelets). To counteract platelet under-representation, three-fold geometric augmentations (flips, composite rotations) expanded the training split to 674 images.
- *AneRBC Anemia Microscopy Dataset:* Curated by Mendeley Data (DOI: 10.17632/hms3sjzt7f.1) [19], providing 738 blood smear photomicrographs categorized into Anemic and Normal smears.
- *Chula-RBC-12 Dataset:* Curated by Chulalongkorn University, contributing 121 specialized erythrocyte

disease photomicrographs (covering iron deficiency, thalassemia, hereditary spherocytosis, elliptocytosis, and megaloblastic anemia) merged into the anemia image pool.

- *Master Complete Blood Count Dataset:* Ingestion and harmonization of the Kaggle Anemia Dataset (1,421 records) and clinical complete blood count records (diagnosed_cbc_data_v4.csv, 1,281 records) into data/anemia/processed/master_cbc_dataset.csv. The combined dataset contains 2,695 patient records (1,429 Anemic [53.0%], 1,266 Normal [47.0%]), covering Gender, Hemoglobin, MCV, MCH, and MCHC.

Table -2: Dataset Provenance, Modalities, and Partition Sizes Across Tasks

Dataset / modality	Total units	Train	Val	Test
BCCD Cell Detection *Smear microscopy*	364 images / 4,888 cells	256 (674 augmented)	52 images	56 images
AneRBC + Chula-RBC-12 *Smear microscopy*	837 images (811 clusters)	567 clusters (583 img.)	121 clusters (125 img.)	123 clusters (129 img.)
Kaggle Anemia Cohort *Tabular laboratory*	1,421 patient records	994 records	213 records	214 records
Master Consolidated CBC *Tabular laboratory*	2,695 patient records	1,617 records	269 records	809 records

3.3 Near-Duplicate Audit and Atomic Cluster Partitioning

To resolve data contamination, an automated perceptual difference-hashing (dHash) and deep feature cosine similarity audit was executed across all 837 microscopy images:

- *Identification:* Images with dHash Hamming distance $d_H \leq 4$ and deep feature cosine similarity ≥ 0.985 were extracted as candidate near-duplicates.
- *Audit Findings:* Exactly 27 candidate pairs were identified. Of these, 25 pairs were exact duplicates (pixel mean squared error MSE $< 10^{-4}$), 2 pairs were likely sequential same-slide microscope fields, and 14 pairs crossed training, validation, and test boundaries in unclustered baselines. Exactly 1 pair (151.jpg [Anemic] vs. 152.jpg [Normal]) possessed conflicting diagnostic labels ($d_H = 4$, cosine similarity = 0.9961); to protect class purity, these images were not merged.
- *Connected Component Clustering:* Connected component graph partitioning grouped all linked duplicates into 811 atomic, indivisible cluster units. Stratified group splitting was performed at the cluster level, guaranteeing exactly 0 cross-split cluster leakage:

567 (Train) + 121 (Val) + 123 (Test) = 811 Total Clusters (100.0% reconciled)

3.4 Image Cytology Pipeline (YOLOv8 and MobileNetV3)

Submitted blood smears undergo optical standardization in src/preprocess.py: (1) application of a 3×3 Gaussian smoothing kernel to attenuate CMOS sensor noise; (2) conversion from RGB to CIE L*a*b* color space; and (3) extraction of the luminance (L*) channel and application of Contrast Limited Adaptive Histogram Equalization (CLAHE) with clip limit C = 2.0 and grid size 8×8 tiles:

$$L_{\text{clahe}} = \text{CLAHE}(L^*, \text{clipLimit} = 2.0,\ \text{tileGrid} = (8,8)) \quad (1)$$

Channel recombination standardizes arrays to 640×640 for YOLOv8 or 224×224 for MobileNetV3.

Cell detection employs an anchor-free single-stage YOLOv8n detector (best.pt, 3.01M parameters, 8.1 GFLOPs). Class mappings are strictly defined as 0: Platelet, 1: RBC, 2: WBC. The pipeline resolves an RGB-to-BGR channel inversion bug that previously caused red erythrocytes to appear blue, triggering false leukocyte detections. For high-resolution digital smears (>640×640 pixels) or low-density fields, tiled inference partitions the image into overlapping 320×320 patches with a 20% spatial overlap stride, followed by global Non-Maximum Suppression (NMS) at an IoU threshold of 0.45:

$$\text{IoU}(B_i, B_j) = \frac{\text{Area}(B_i \cap B_j)}{\text{Area}(B_i \cup B_j)} \leq 0.45 \quad (2)$$

Standardized 224×224×3 tensors are normalized via ImageNet channel parameters (μ = [0.485, 0.456, 0.406], σ = [0.229, 0.224, 0.225]):

$$x'_c = \frac{x_c - \mu_c}{\sigma_c} \quad (3)$$

The tensor is processed through the MobileNetV3-Small backbone to generate a 576-dimensional pooled embedding f_{img}, projected through a linear classification head:

$$P_{\text{img}} = \frac{\exp(z_1)}{\exp(z_0)+\exp(z_1)}, \quad z = W_{\text{head}} f_{\text{img}} + b_{\text{head}} \quad (4)$$

Classification is determined via a Youden-calibrated threshold [20], $\tau^* = 0.5007$: $y_{\text{img}} = I(P_{\text{img}} \geq 0.5007)$.

3.5 Tabular CBC Pipeline and Circular Target Leakage

The tabular pipeline processes 5 clinical parameters: Gender, Hemoglobin, MCV, MCH, and MCHC. Features are standardized via StandardScaler: $z_i = (x_i - \mu_i) / \sigma_i$. Classification is performed by a balanced Random Forest ensemble (N = 100 trees).

The Circular Target Leakage Phenomenon. In hematological datasets, the target label y is derived directly from the WHO diagnostic definition:

$$y = \mathbb{I}(\ (\text{Gender} = 0 \wedge \text{Hb} < 12.0) \vee (\text{Gender} = 1 \wedge \text{Hb} < 13.0)) \quad (5)$$

Supplying Hemoglobin allows decision trees to isolate this rule, resulting in Gini feature importances where hemoglobin accounts for 83.32% of all decision splits. The Leakage-Aware Screening Model restricts features to:

$$x_{\text{leakage-free}} = [\text{Gender}, \text{MCV}, \text{MCH}, \text{MCHC}] \quad (6)$$

3.6 Multimodal Integration Architectures

Production Stacking Ensemble (Default Mode). This mode constructs a 12-dimensional engineered multimodal vector:

$$v_{\text{fusion}} = [\ P_{\text{img}}, C_{\text{img}}, E_{\text{img}}, P_{\text{cbc}}, \text{Hb}, \text{MCV}, \text{MCH}, \text{MCHC}, \text{Gender}, \text{MCV/Hb}, |P_{\text{img}} - P_{\text{cbc}}|, P_{\text{img}} P_{\text{cbc}}] \quad (7)$$

where C_{img} is visual confidence and E_{img} is prediction entropy. Base estimators include HistGradientBoosting (150 trees), ExtraTrees (100 trees), and MLP (64-32 neurons). The meta-learner is a calibrated regularized Logistic Regression:

$$P_{\text{fused}} = \sigma\left(w_0 + \sum_m w_m \ \text{logit}(\tilde{P}_m)\right), \quad \text{logit}(p) = \ln\frac{p}{1-p} \tag{8}$$

Experimental Gated Multimodal Unit (GMU Mode). The deep GMU projects image and tabular features into a common latent dimension d = 128:

- *Visual latent projection:* $x_{img} = W_{proj}\, f_{img} + b_{proj} \in \mathbb{R}^{128}$.
- *Tabular latent projection:* $x_{cbc} = \text{Swish}(\text{LN}(W_2 \cdot \text{Swish}(\text{LN}(W_1\, x_{cbc} + b_1)) + b_2)) \in \mathbb{R}^{128}$.
- *Candidate modality embeddings:* $h_{img} = \tanh(W_{img}\, x_{img} + b_{img})$, $h_{cbc} = \tanh(W_{cbc}\, x_{cbc} + b_{cbc})$.
- *Dynamic multimodal gating:* $z = \sigma(W_z\, [h_{img}, h_{cbc}] + b_z) \in (0, 1)^{128}$.
- *Mask-aware latent fusion:* let $m_{img}, m_{cbc} \in \{0, 1\}$ represent binary presence masks:

$$h_{\text{fused}} = z^* \odot h_{\text{img}} + (1 - z^*) \odot h_{\text{cbc}} \tag{9}$$

where z* is clamped to 1 when CBC is missing, and to 0 when the smear image is missing. The classification head computes the final screening probability via a sigmoid.

3.7 Core System Algorithms

The execution logic is formalized in Algorithms 1, 2, and 3.

Algorithm 1: Microscopic Smear Cytology Pipeline

Input: Raw blood smear image I_raw, threshold θ_conf
Output: Cell counts {N_rbc, N_wbc, N_plt}, P_img, I_annot
Validate dimensions of I_raw; raise format error if invalid
I_smooth ← GaussianBlur(I_raw, kernel = (3, 3))
I_lab ← ConvertColor(I_smooth, RGB_to_LAB)
L_clahe ← ApplyCLAHE(I_lab.L, clipLimit = 2.0, grid = (8, 8))
I_prep ← ConvertColor(Merge(L_clahe, a, b), LAB_to_RGB)
if Width(I_raw) > 640 or Height(I_raw) > 640 then
  B ← TiledYOLOv8(I_prep, patch = (320, 320), overlap = 0.2)
else
  B ← YOLOv8n_Forward(I_prep, conf = θ_conf, iou = 0.45)
end if
N_plt ← Count(B, 0); N_rbc ← Count(B, 1); N_wbc ← Count(B, 2)
I_annot ← DrawBoundingBoxes(I_raw, B)
I_tensor ← ImageNetNormalize(Resize(I_raw, (224, 224)))
z_img ← MobileNetV3_Forward(I_tensor)
P_img ← Softmax(z_img)[1]
return {N_rbc, N_wbc, N_plt}, P_img, I_annot

Algorithm 2: Complete Blood Count Processing and Leakage-Aware ML

Input: Dict D_cbc = {Gender, Hb, MCV, MCH, MCHC}, flag M_leak
Output: Tabular anemia probability P_cbc, feature importances Imp
Validate ranges: Gender in {0, 1}, Hb in [1, 25], MCV in [30, 150]
if M_leak = FALSE then
  x_raw ← [Gender, MCV, MCH, MCHC]
  scaler ← LoadScaler("models/scaler_leakage_free.joblib")
  model ← LoadModel("models/anemia_leakage_free.joblib")
else
  x_raw ← [Gender, Hb, MCV, MCH, MCHC]
  scaler ← LoadScaler("models/scaler.joblib")
  model ← LoadModel("models/anemia_model.joblib")
end if
x_scaled ← scaler.transform(x_raw)
P_cbc ← model.predict_proba(x_scaled)[1]
Imp ← model.feature_importances_
return P_cbc, Imp

Algorithm 3: Mask-Aware Gated Multimodal Unit (GMU)

Input: Tensor X_img ∈ ℝ^(3×224×224), vector x_cbc ∈ ℝ^5, masks m_img, m_cbc
Output: Screening probability P_fused, gate summary z_scalar
if m_img = 1 then
  f_img ← MobileNetV3_ExtractEmbedding(X_img) // 576-d
  x_i ← W_proj · f_img + b_proj; h_i ← tanh(W_img · x_i + b_i)
else h_i ← 0_d; end if
if m_cbc = 1 then
  x_scaled ← GMU_StandardScaler(x_cbc)
  x_c ← GMU_TabularMLP(x_scaled); h_c ← tanh(W_cbc · x_c + b_c)
else h_c ← 0_d; end if
if m_img = 1 and m_cbc = 1 then
  z ← sigmoid (W_z · [h_i, h_c] + b_z)
else if m_img = 1 and m_cbc = 0 then z ← 1_d
else if m_img = 0 and m_cbc = 1 then z ← 0_d
else return 0.0, 0.0; end if
h_fused ← (z ⊙ h_i) + ((1 − z) ⊙ h_c)
P_fused ← sigmoid(GMU_ClassificationHead(h_fused))
return P_fused, Mean(z)

3.8 System Implementation

HemaVision is engineered as a decoupled software suite comprising four modular layers: an asynchronous FastAPI backend service, an SQLite database persistence tier, two user interfaces, and an automated verification test suite. The end-to-end prediction flow is shown in Fig. 3.

3.8.1 Backend Service Architecture (FastAPI)

The backend service (backend/main.py) exposes an asynchronous RESTful API running via Uvicorn. Key specifications include: (a) a singleton inference cache, in which models are loaded into system memory once during application startup, preventing repetitive disk I/O latency; (b) asynchronous task workers, in which diagnostic reports are compiled in the background using FastAPI BackgroundTasks, decoupling PDF generation from HTTP response cycles; and (c) REST endpoints: POST /api/analyze ingests multipart form requests; GET /api/history returns paginated analysis summaries; GET /api/history/{id} returns full serialized records; GET /api/reports/{filename} streams generated clinical PDF reports; and GET /api/health provides hardware status and checkpoint integrity verification.

3.8.2 Frontend Implementations

Streamlit Clinical Dashboard (frontend/app.py): A Python-native clinical interface operating on port 8501. It provides a dual-column layout for simultaneous microscopic image drag-and-drop and CBC parameter entry. A sidebar radio toggle allows clinicians to switch between the Production Stacking Ensemble and the Experimental GMU, dynamically displaying context-specific regulatory warning banners, color-coded YOLOv8 cell detection bounding boxes, leukocyte differential pie charts, and risk probability gauges.

React 19 / TanStack Web Application (ui/): An enterprise single-page application built with React 19, TypeScript, Vite 8, TanStack Router, TanStack Query, and Tailwind CSS 4. Configured in high-contrast slate/navy diagnostic styling (#0f172a to #1e1b4b) optimized for clinical monitors. It provides client-side form validation, asynchronous query state management, interactive bounding-box toggles, and direct links to generated PDF reports.

3.8.3 Database Persistence Layer

Persistence is managed via SQLAlchemy ORM connected to an embedded SQLite database (data/blood_analyzer.db). The analyses table schema records the analysis GUID (BRA-

YYYYMMDDHHMMSS-XXXXXX), timestamp, cell counts, serialized WBC differential distributions, single-modality probabilities, raw CBC inputs, final fused screening verdicts, and report paths.

3.8.4 Automated Clinical Reporting Engine (FPDF2)

The automated reporting engine (src/report_generator.py) generates multi-page clinical PDF summaries using fpdf2. The document layout conforms to standardized clinical documentation norms: Document Control Header, Patient CBC Summary Table with directional abnormality flags, Cytological Photomicrograph with bounding boxes, Cellular Enumeration Table, Matplotlib WBC Differential Pie Chart, Multimodal Risk Assessment Card, and a bold red-bordered Regulatory Disclaimer Banner.

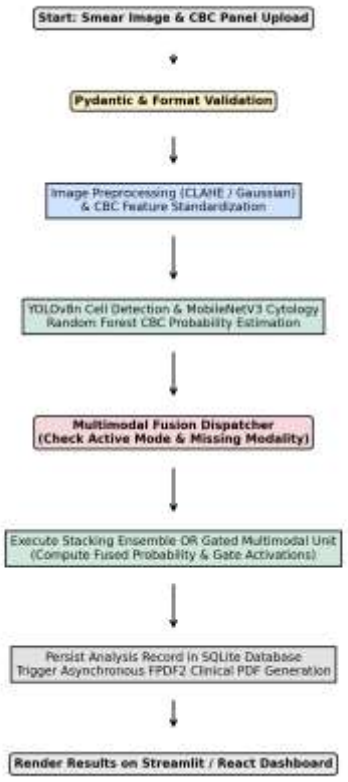


Fig -3: Flowchart of Prediction Process

4. RESULTS

4.1 Experimental Setup

All evaluations were executed on isolated, held-out test partitions. The YOLOv8 detector was evaluated on 56 held-out test images from the BCCD repository. The MobileNetV3 smear classifier was evaluated across the 123 held-out atomic clusters (129 physical images) established during the perceptual hash audit. The CBC tabular models were evaluated on 809 held-out patient records from the Master CBC dataset. Multimodal fusion models were evaluated across 400 synthetically paired test instances generated from held-out splits. In the tables below, Acc., Sens., Spec., and AUC denote accuracy, sensitivity, specificity, and ROC-AUC, respectively.

4.2 Smear Cytomorphology Classification

Table -3 compares the successive iterations of the MobileNetV3 smear classifier.

Table -3: Comparative Evaluation of Smear Cytomorphology Classifiers (MobileNetV3)

Model / cohort	Acc.	Sens.	Spec.	F1	AUC
Old Baseline (Unweighted) *Unpurged 108 test images (AneRBC)*	80.56%	97.73%	5.00%	0.8912	0.7477
Baseline Re-evaluation *Purged 128 test images (AneRBC)*	81.25%	83.33%	70.00%	0.8824	0.8752
Weighted Retraining *Purged 128 test images (AneRBC)*	75.78%	75.93%	75.00%	0.8410	0.8500
Cluster-Remediated (Active) ***Remediated 129 test images (811 clusters)***	**75.19%**	**77.27%**	**63.16%**	**0.8416**	**0.8167**

4.3 CBC Models and Target Leakage Audit

Table -4 reports the tabular CBC models with and without hemoglobin as an input feature.

Table -4: Evaluation of Complete Blood Count Models and Target Leakage Audit

Model / cohort	Acc.	Sens.	Spec.	F1	AUC
Standard RF (Kaggle Split) *Gender, Hb, MCV, MCH, MCHC*	100.0%	100.0%	100.0%	1.0000	1.0000
Master Dataset RF (Target Leak) *Gender, Hb, MCV, MCH, MCHC*	99.51%	99.53%	99.47%	0.9953	0.9999
Leakage-Aware Model (Active) ***Gender, MCV, MCH, MCHC (no Hb)***	**83.56%**	**89.28%**	**77.11%**	**0.8521**	**0.8974**

4.4 Multimodal Fusion Performance

Table -5 benchmarks the two fusion engines on synthetically paired test data, and Fig. 4 compares the metrics across model tiers.

Table -5: Comparative Benchmark of Multimodal Fusion Engines (Evaluated on Synthetic Combinations)

Model / cohort	Acc.	Sens.	Spec.	F1	AUC
Production Stacking Ensemble *Held-out 360 synthetic pairs*	99.44%	99.44%	99.44%	0.9944	0.9999
Probability Averaging Baseline *Held-out 400 synthetic pairs*	97.75%	98.00%	97.50%	0.9776	0.9922
Experimental GMU Dual-Stream *Held-out 400 synthetic pairs*	97.50%	97.00%	98.00%	0.9749	0.9978
GMU CBC-Only Mode (m_img = 0) *Held-out 400 synthetic pairs*	97.00%	98.00%	96.00%	0.9703	0.9980
GMU Image-Only Mode (m_cbc = 0) *Held-out 400 synthetic pairs*	69.00%	71.00%	67.00%	0.6961	0.7495

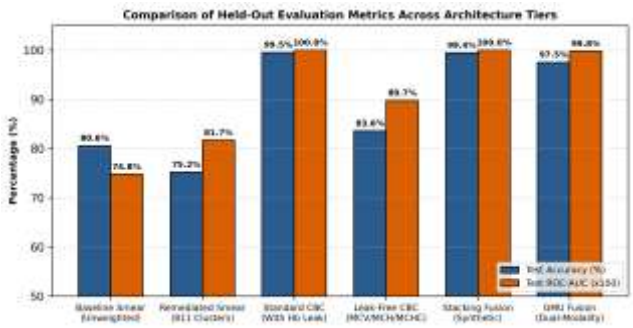


Fig -4: Metric Comparison Across Model Tiers

4.5 Cell Detection and Runtime Profile

YOLOv8n cell enumeration achieved 89.14% mAP@0.50 (91.4% RBC, 92.8% WBC, 83.2% Platelets); sample detections are shown in Fig. 5. Tiled inference eliminated border boundary truncations. Warm latency on an Intel i7 CPU averaged 560.1 ms (YOLO: 184.2 ms, MobileNet: 42.8 ms, CBC: 8.4 ms, Stacking: 12.1 ms, PDF: 312.6 ms in background). All 42 automated QA test suites passed with 100% success.

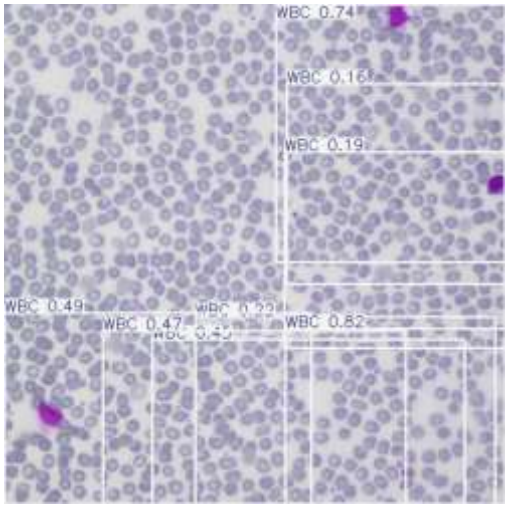


Fig -5: YOLOv8 Cell Detection Annotations

5. DISCUSSION

5.1 Why Image Specificity Collapsed and How Clustering Remediated It

The baseline MobileNetV3 classifier exhibited an acute specificity collapse, correctly identifying only 1 out of 20 normal test smears (specificity = 5.00%). This failure stemmed directly from severe class imbalance (5.7:1 ratio favoring positive cases) combined with unweighted cross-entropy loss and a default $\tau = 0.50$ decision threshold. Under gradient descent, the network minimized loss by classifying nearly every input as anemic.

Furthermore, our perceptual hash audit revealed that 14 near-duplicate image pairs spanned training and test splits, artificially inflating baseline accuracy. By constructing 811 atomic clusters and retraining with inverse class weights, specificity improved from 5.00% to 63.16%, with sensitivity stabilizing at 77.27% and ROC-AUC at 0.8167. This demonstrates that rigorous duplicate auditing and threshold calibration via Youden's index ($\tau^* = 0.5007$) are mandatory when evaluating medical microscopy datasets.

5.2 Biological Foundations of the Leakage-Free CBC Model

When Hemoglobin is supplied to a Random Forest classifier, it accounts for 83.32% of all tree splits. This confirms that the model simply memorizes the World Health Organization threshold definition used to assign the dataset target labels [1], [17].

When Hemoglobin was strictly removed, the Random Forest model achieved 83.56% accuracy and 0.8974 ROC-AUC on 809 held-out patients. Feature importance was redistributed across erythrocyte cytometric parameters: MCV (38.63%), MCH (27.77%), MCHC (24.08%), and Gender (9.52%). This performance reflects fundamental hematological biology: in iron deficiency anemia and thalassemia, impaired hemoglobin synthesis restricts erythrocyte cellular expansion, producing microcytosis (low MCV) and hypochromia (low MCH and MCHC) [1], [2]. These morphological indices serve as valid physiological surrogates for screening anemia risk when hemoglobin cannot be directly measured. However, because normocytic anemias (such as aplastic anemia or hemolytic anemia) present with normal MCV and MCH, a tabular screening model lacking hemoglobin cannot detect all anemia etiologies.

5.3 Multimodal Gating Dynamics in the GMU

In the deep Gated Multimodal Unit, the mean gate activation across 400 held-out synthetic test instances settled at $\bar{z} = 0.4499 \pm 0.0821$. This indicates that the network learned an approximately balanced weighting between visual smear cytomorphology ($z \approx 0.45$) and tabular laboratory parameters ($1 - z \approx 0.55$).

When evaluated under forced single-modality masking, clamping $m_{img} = 0$, $m_{cbc} = 1$ ($z = 0$) yielded 97.00% accuracy and 0.9980 ROC-AUC, reflecting the strong discriminatory power of tabular CBC parameters. Clamping $m_{img} = 1$, $m_{cbc} = 0$ ($z = 1$) yielded 69.00% accuracy and 0.7495 ROC-AUC, aligning with the performance of standalone image classification on noisy crops.

These gating behaviors demonstrate that the GMU architecture dynamically down-weights noisy visual features when informative laboratory indices are available, while preserving inference capability when modalities are missing.

5.4 The Methodological Reality of Synthetic Data Fusion

Crucially, the high accuracies recorded for both the Stacking Ensemble (99.44%) and the GMU (97.50%) were derived from synthetically paired combinations of unlinked images and CBC records. Because open-access repositories do not link blood smear photomicrographs to corresponding clinical laboratory records, synthetic pairing was executed under class-consistent criteria.

While these experiments validate tensor operations, gradient convergence, and missing-modality gating, they do not establish clinical diagnostic efficacy on real human patients. In a prospective clinical cohort where blood smears and CBC panels are extracted from the same venipuncture draw, inter-modality discordance, staining variability, and atypical cell morphologies may degrade performance. Reporting synthetic multimodal accuracy as clinical patient efficacy is scientifically indefensible; researchers must transparently distinguish algorithmic validation on synthetic benchmarks from prospective clinical trials.

5.5 Clinical Safety and Regulatory Boundaries

HemaVision is engineered strictly as a clinical decision-support research prototype. It does not possess FDA 510(k) clearance or CE-mark certification under EU MDR 2017/745. The provisional 0.50 decision threshold, risk gauge visualizations, and automated PDF summaries are intended to assist educational and scientific inquiry. Incorrect model predictions could lead to misdiagnoses or inappropriate clinical interventions if accepted without physician oversight. Therefore, all model outputs must be independently reviewed and verified by a licensed clinical pathologist alongside gold-standard automated hematology laboratory testing.

6. CONCLUSIONS

6.1 Limitations

1) Unlinked Multimodal Cohorts: Blood smear images and tabular CBC records originate from separate public datasets with zero patient-level linkage. All multimodal fusion metrics reflect synthetic pairing.

2) Absence of Patient Identifiers in Smear Repositories: While grouping 837 images into 811 atomic clusters resolved field-of-view duplicate leakage, the underlying repositories lack patient-level identifiers. Multiple slides could still originate from the same patient.

3) Missing WBC Subtype Model: The deep leukocyte classifier checkpoint (wbc_classifier_sklearn.pth) is absent from the repository; the system transparently defaults to WHO physiological reference distributions for leukocyte differentials.

4) Staining and Optical Acquisition Variance: The image classification models have not been evaluated across slides prepared with varying Wright-Giemsa vs. Leishman staining protocols, varying illumination color temperatures, or smartphone-attached optical lenses.

5) Lack of Prospective Clinical Validation: The entire framework has been evaluated retrospectively on historical benchmarks without prospective validation in active clinical triage workflows.

6.2 Future Work

1) Prospective Clinically Linked Cohort Collection: Partner with teaching hospitals under Institutional Review Board (IRB) approval to collect 1,000+ patient samples with simultaneously paired venipuncture draws of digital blood smear photomicrographs and 5-part automated hematology analyzer CBC panels.

2) True Multimodal GMU Retraining: Train the Gated Multimodal Unit on genuine patient-paired cohorts to evaluate true clinical diagnostic lift over unimodal baselines.

3) Multi-Class Anemia Etiology Subtyping: Expand the classification head beyond binary screening to differentiate specific anemia etiologies, including Iron Deficiency Anemia, β-Thalassemia trait, Megaloblastic Anemia, Sideroblastic Anemia, and Aplastic Anemia.

4) Quantized Mobile Edge Deployment: Quantize YOLOv8n and MobileNetV3-Small to 8-bit integers (INT8) for edge deployment on low-cost Android smartphones paired with 3D-printed optical microscope eyepiece adapters.

6.3 Concluding Remarks

This paper presented HemaVision, an open-access multimodal machine learning framework for preliminary anemia screening integrating peripheral blood smear photomicrographs and complete blood count parameters. By combining an anchor-free YOLOv8n cell enumerator, a MobileNetV3-Small cytomorphology classifier, a tabular Random Forest panel, and two multimodal fusion mechanisms (a 12-dimensional Stacking Ensemble and an experimental Gated Multimodal Unit), the framework demonstrates the feasibility of unified hematological screening.

Through rigorous empirical audits, we uncovered and remediated critical methodological pitfalls common in medical AI:

- An exhaustive difference-hash and cosine similarity audit of 837 microscopy images identified 27 candidate duplicate pairs and 14 cross-split contaminated pairs in standard baselines. Reorganizing the dataset into 811 atomic clusters resolved data leakage, establishing a validated smear classification performance of 75.19% accuracy and 0.8167 ROC-AUC.
- An empirical target leakage audit revealed that standard CBC models achieve 99.51% to 100.0% accuracy through trivial memorization of the WHO hemoglobin threshold. A genuine screening model strictly omitting hemoglobin achieved 83.56% accuracy and 0.8974 ROC-AUC by learning biological surrogate markers in erythrocyte cytometric parameters (MCV, MCH, MCHC).
- Multi-paradigm multimodal integration demonstrated high synthetic test performance (99.44% Stacking Ensemble, 97.50% GMU), while mask-aware gating provided crash-safe operation across missing modalities.

Crucially, we emphasize that high multimodal performance on synthetically paired datasets does not establish clinical diagnostic efficacy on human patients. HemaVision is an investigational research prototype, providing a transparent, scientifically validated baseline for future prospective clinical research in AI-assisted hematology.

REFERENCES

1. World Health Organization, Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity, Vitamin and Mineral Nutrition Information System, Geneva: World Health Organization, 2011 (WHO/NMH/NHD/MNM/11.1).
2. N. J. Kassebaum, R. Jasrasaria, M. Naghavi, et al., "A systematic analysis of global anemia burden from 1990 to 2010," Blood, vol. 123, no. 5, pp. 615–624, 2014, doi: 10.1182/blood-2013-06-508325.
3. J. A. Brittin and G. Brecher, "Instrumentation and automation in hematology," Prog. Hematol., vol. 7, pp. 299–340, 1971.
4. B. J. Bain, Blood Cells: A Practical Guide, 5th ed. Chichester, UK: Wiley-Blackwell, 2015.
5. G. L. Kumar and D. L. Bailey, Education Guide: Immunohistochemical Staining Methods, 6th ed. Carpinteria, CA: Dako North America, 2013.
6. A. Howard, M. Sandler, G. Chu, et al., "Searching for MobileNetV3," in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), 2019, pp. 1314–1324, doi: 10.1109/ICCV.2019.00140.
7. L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
8. Shenggan, "BCCD (Blood Cell Count and Detection) Dataset," MIT License, 2019. [Online]. Available: https://github.com/Shenggan/BCCD_Dataset
9. A. Acevedo, A. Merino, L. Alférez, et al., "A dataset of microscopic peripheral blood cell images for development of

automatic recognition systems," Data in Brief, vol. 30, Art. no. 105474, 2020, doi: 10.1016/j.dib.2020.105474.

10. G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," Version 8.1.0, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
11. S. Sanadhya, B. P. Sharma, and R. K. Gupta, "Machine learning approaches for anemia detection using complete blood count parameters," in Proc. Int. Conf. Comput. Intell. Commun. Netw. (CICN), 2022, pp. 210–215, doi: 10.1109/CICN56167.2022.10008321.
12. R. AlSalemi, H. H. Al-Jubouri, and M. A. Al-Mallah, "Automated classification of microcytic anemias based on routine complete blood cell count parameters using machine learning algorithms," Comput. Biol. Med., vol. 154, Art. no. 106584, 2023, doi: 10.1016/j.compbiomed.2023.106584.
13. F. Pedregosa, G. Varoquaux, A. Gramfort, et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.
14. J. Arevalo, T. Solorio, M. Montes-y-Gómez, and F. A. González, "Gated Multimodal Units for multimodal fusion," in ICLR 2017 Workshop Track, Toulon, France, 2017, pp. 1–9.
15. S. Kumar, S. Sharma, and K. T. Megra, "Transformer enabled multi-modal medical diagnosis for tuberculosis classification," J. Big Data, vol. 12, Art. no. 5, 2025, doi: 10.1186/s40537-024-01054-w.
16. L. Alzubaidi, M. A. Fadhel, O. Al-Shamma, et al., "Deep learning models for classification of red blood cells in microscopy images to aid in sickle cell anemia diagnosis," Diagnostics, vol. 10, no. 10, Art. no. 775, 2020, doi: 10.3390/diagnostics10100775.
17. S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," ACM Trans. Knowl. Discov. Data, vol. 6, no. 4, Art. no. 15, 2012, doi: 10.1145/2382577.2382579.
18. J. Ballard, Z. Wang, W. Li, et al., "Deep learning-based approaches for multi-omics data integration and analysis," BioData Min., vol. 17, Art. no. 38, 2024, doi: 10.1186/s13040-024-00391-z.
19. Mendeley Data, "AneRBC: Peripheral Blood Smear Dataset for Anemia Classification," Mendeley Data, V1, 2020, doi: 10.17632/hms3sjzt7f.1.
20. W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950, doi: 10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3.