International Journal of Future Engineering Innovations

Comparative Performance Evaluation of Five Machine Learning Models for Diagnosing Blood Diseases from Microscopic Images

Zaid Rajih Mohammed [1*], Mohammed Hayder Kadhim [2]
[1] Faculty of Medical Sciences, Jabir ibn Hayyan University for Medical and Pharmaceutical Sciences, Najaf, Iraq
[2] The College of Medicine, Jabir ibn Hayyan University for Medical and Pharmaceutical Sciences, Najaf, Iraq

* Corresponding Author: **Zaid Rajih Mohammed**

Article Info

ISSN (online): 3049-1215
Volume: 03
Issue: 06
Received: 02-09-2026
Accepted: 03-10-2026
Published: 01-11-2026
Page No: 54-64

Abstract
Machine-learning models for blood-smear diagnosis routinely report accuracies above 95%, but most studies split images, not patients, into training and test sets, so cells of the same person appear on both sides. We compare five model families under one leakage-free protocol: Random Forest and a support vector machine on 135 handcrafted colour, texture and shape features, a custom convolutional network, transfer-learning CNNs (EfficientNet-B0 and ResNet-50) and a vision transformer (DeiT-Small). Two public tasks are used: acute lymphoblastic leukemia versus normal B-cells (C-NMC 2019, 12,418 cells, 93 subjects) and parasitized versus uninfected red cells (NIH malaria, 27,558 cells, 200 patients). Splits are disjoint by subject; a pixel-hash audit of C-NMC found 110 identical images filed under different patient IDs, which we merged. Every model is trained with three seeds and evaluated by AUC, accuracy, sensitivity, specificity, F1 and balanced accuracy, with DeLong and McNemar tests and patient-level bootstrap intervals. On C-NMC, ResNet-50 (AUC 0.9141 ± 0.0168) and DeiT-Small (0.9066 ± 0.0199) lead and do not differ significantly (DeLong p = 0.190 after Holm correction); the handcrafted-feature models reach 0.867–0.876. On malaria all models exceed AUC 0.98; DeiT-Small is highest (0.9922 ± 0.0008), but Random Forest (0.9896 ± 0.0003) makes as few errors at a validated threshold. Splitting by image instead of by patient inflates the C-NMC AUC of EfficientNet-B0 from 0.880 to 0.974 and of Random Forest from 0.866 to 0.931, close to the values reported in recent literature. In malaria, 79% of the cells missed by all six models come from a single patient. Leakage-free protocols, not larger models, are what separate realistic from optimistic results.

DOI: https://doi.org/10.54660/IJFEI.2026.3.6.54-64

Keywords: blood smear, leukemia, malaria, deep learning, vision transformer, random forest, data leakage, patient-level validation

1. Introduction

1.1. Background

Microscopic examination of stained blood smears remains the reference method for diagnosing leukemia and malaria in many hospitals. It is slow, depends on the experience of the microscopist and is hard to scale where specialists are few. Automated classification of single cells promises a fast second reader. Acute lymphoblastic leukemia (ALL), the most common childhood cancer, requires distinguishing immature lymphoblasts from normal B-lymphocytes that look very similar; malaria diagnosis requires detecting Plasmodium parasites inside red blood cells.

1.2. Research Gap

Recent studies (Table 1) report accuracies of 95–99% and AUC above 0.99 on the same two public datasets used here [1-4, 8-15]. Most do not state how images were split, and those that do split images rather than patients [5, 14]. Because cells of one patient share staining, illumination and morphology, an image-level split lets a model recognise the patient instead of the disease. Leakage of this kind inflates accuracy by up to 18 points in histopathology [16], ultrasound [17] and neuroimaging [18], but its size has not been measured on blood-smear benchmarks. Comparisons between handcrafted features, CNNs and vision transformers are also usually run with one seed and without significance tests, so small differences are reported as rankings.

Table 1: Recent related works (2024–2026). Split: train/test split level as stated by the authors; n/r = not reported.

#	Study	Task / data	Method	Result	Split
1	Mohan *et al.* (2025)	ALL vs normal B-cells; C-NMC, 10,661 images	DenseNet-121 + ResNet-34 deep features fused, classified by SVM/KNN/RF/AdaBoost/BPN	Acc 92.5%; F1 93.1%; AUC 0.975; t-test and Wilcoxon $p<0.05$	n/r
2	Jammal *et al.* (2025)	ALL vs benign; C-NMC 15,135 images / 118 patients; ALL Image 3,256 images / 89 patients	CNN features + ViT contextual encoding + fuzzy reasoning; Grad-CAM, SHAP, IG, LIME	Acc 0.9715 C-NMC (SGD), 1.0000 ALL Image; 0.5002 on C-NMC with Nadam	n/r
3	Jammal (2026)	ALL vs normal; ALL Image; C-NMC	Meta-Conformer-XAI: CNN + ViT dual-attention fusion, meta-learned path controller	Acc 0.9636 C-NMC, 0.9924 ALL Image; AUC > 0.99 on both	n/r
4	Hendrawan *et al.* (2025)	ALL vs healthy cells; C-NMC 2019 (size n/r)	Pretrained ResNet-50 features + ViT self-attention (hybrid)	Acc 98% (abstract seen via search index only; verify from full text)	n/r
5	Ashikuzzaman *et al.* (2026)	Leukemia vs normal; Kaggle leukemia dataset, 10,700 images	14 pretrained CNNs (EfficientNet-B1/B3/B5, ResNet, Xception, DenseNet, VGG); Grad-CAM, LIME; McNemar	Acc up to 96%; F1 up to 97%; AUC up to 0.987 (mean of 10 seeds)	image
6	Manalı *et al.* (2025)	Blood cancer (WBC) classification; Public Kaggle WBC dataset (size n/r)	13 models: 4 CNNs, 5 ViTs, 4 hybrids under identical settings	Best F1 98.76% (MobileViT-XS); DenseNet121 98.65%; ViT-Base 87.36%	n/r
7	Kasim *et al.* (2025)	Healthy vs lymphoblast vs myeloblast; ALL-IDB + Munich AML Morphology	VGG16 / InceptionV3 / ResNet50 features + RF, SVM, XGBoost, MLP	Best Acc 88% (InceptionV3 + SVM); ResNet50 overfitted	n/r
8	Boit *et al.* (2024)	Malaria: parasitized vs uninfected; NIH malaria, 27,558 images	EDRI hybrid multi-architecture deep model	Acc 97.68%	n/r
9	Mujahid *et al.* (2024)	Malaria: infected vs uninfected; RBC smear images (n/r in abstract)	EfficientNet vs other pretrained models; k-fold CV	Acc 97.57%	n/r
10	Benachour *et al.* (2025)	Malaria: parasitized vs uninfected; 27,558 images (24,958 train + 2,600 test)	Five CNN variants (basic to augmented, VGG16 transfer learning)	Simple CNN Acc > 99%, beating deeper models	n/r
11	Shahin *et al.* (2025)	Malaria: parasitized vs uninfected; Thin-smear dataset, 27,558 images	ResNet50 / VGG16 / DenseNet-201 features + PCA + SVM and LSTM, majority voting	Acc 96.47%; Sens 96.03%; Spec 96.90%; F1 96.45%	n/r
12	Reddy *et al.* (2025)	Malaria: parasitized vs uninfected; Thin-smear cell images (n/r in abstract)	Ensemble of VGG16, ResNet50V2, DenseNet201, VGG19 vs custom CNN and CNN-SVM	Ensemble Acc 97.93%; custom CNN 97.20%; CNN-SVM 82.47%	n/r
13	Gasmi *et al.* (2026)	Malaria: infected vs uninfected; NIH malaria 27,558 images (+ NM-AIST 3,544)	Track A: CNN features + calibrated classic classifiers; Track B: end-to-end CNNs; weighted pairwise ensembles	Acc 96.35% on held-out data (ensemble)	n/r
14	Dehbi *et al.* (2026)	Malaria: parasitized vs uninfected; NIH/NLM, 27,558 images	Two custom CNNs + score-level ensemble; validation-only F2 threshold tuning	Acc 95.26%; Sens 97.05%; AUC 0.9876 (CI 0.9842–0.9910)	image
15	Özdemir *et al.* (2025)	Malaria: parasitized vs uninfected; Blood smear images (n/r in abstract)	ViT-L/16 + EfficientNet-B0 feature fusion, SHAP selection, boosting ensemble	Acc 95.61%	n/r
16	Nogueira *et al.* (2026)	Data leakage (oral cancer histopathology); Public oral cancer histopathology dataset	11 CNN and Transformer models, leakage-free vs two leakage scenarios	Sample-related leakage inflated Acc by up to 18% (mean +14.3%)	image vs leakage-free
17	Wang (2026)	Patient-level validation (breast ultrasound); BUS-UCLM, 264 images / 36 patients	7 CNNs + ViT + ResNet18-ViT fusion; patient-level 5-fold CV vs image-level	Patient-level best AUC 0.67; image-level split substantially inflated results	patient
18	Young *et al.* (2025)	Data leakage (review); 44 studies	PRISMA-ScR scoping review with a three-tier leakage-risk framework	Subject-wise splits 66–90% Acc vs 95–99% for high-leakage studies	patient vs leaky splits compared

1.3. Contributions

1. A controlled comparison of five model families (Random Forest, SVM, a custom CNN, transfer-learning CNNs and a vision transformer) on two diseases, with identical preprocessing, patient-disjoint splits, three seeds and significance tests.
2. A measurement of image-level leakage on C-NMC and NIH malaria, and a duplicate-image audit of C-NMC that found identical cells filed under different patient IDs.
3. A per-patient error analysis showing that the residual malaria errors are concentrated in a few patients, and an efficiency comparison on a 4 GB consumer GPU.

2. Related Work

2.1. Leukemia (C-NMC)

Mohan *et al*. [1] fused DenseNet-121 and ResNet-34 features and classified them with SVM, k-NN and Random Forest, reporting 92.5% accuracy and AUC 0.975 on the C-NMC training subset. Jammal *et al*. combined CNN features with a vision transformer and fuzzy reasoning (97.15% accuracy) [2], and later a meta-learned CNN–transformer hybrid with AUC above 0.99 [3]. Hendrawan *et al*. [4] joined a ResNet-50 backbone with a ViT. Ashikuzzaman *et al*. [5] compared 14 transfer-learning CNNs with ten seeds and McNemar tests on a Kaggle leukemia set, with an explicitly image-level split. For other leukemia data, Manalı *et al*. [6] compared CNNs, ViTs and hybrids under identical settings and found plain ViTs weakest, and Kasim *et al*. [7] reported a more modest 88% for deep features with SVM on small multiclass sets.

2.2. Malaria (NIH)

On the NIH set, reported accuracies range from 95.3% to above 99%: a multi-architecture hybrid [8], EfficientNet [9], a small custom CNN that matched deeper transfer-learning models [10], deep features with SVM and LSTM [11], CNN ensembles [12], two-track classic-ML and CNN ensembles [13], and ViT–EfficientNet-B0 fusion [15]. Dehbi *et al*. [14] tuned a sensitivity-oriented threshold on validation data only and stated that their image-level split could not guarantee patient independence.

2.3. Leakage in Medical Image Evaluation

Nogueira and Gomes [16] showed that sample-level leakage raised the accuracy of eleven CNNs and transformers on oral cancer images by 14 points on average; Wang [17] found that only patient-level validation gave reliable rankings in breast ultrasound; and a review of Alzheimer studies [18] reported 66–90% accuracy for subject-wise splits against 95–99% for leaky ones. Table 1 summarises the works.

3. Materials and Methods

3.1. Datasets and Leakage-Free Splits

C-NMC 2019 [19] contains segmented cells from bone-marrow smears labelled ALL or normal; we used the labelled training and preliminary-test parts (12,528 images, 101 patient IDs). The NIH malaria set [20] contains 27,558 segmented red cells from 200 patients (150 infected, 50 healthy); the patient is encoded in each file name. A SHA-1 hash of the decoded pixels of every image revealed 110 images in C-NMC that occur under two different patient IDs; 12 IDs share identical cells and were merged with a union–find step into fewer subjects (93 in total), keeping one copy of each image. No duplicates were found in the malaria set. Subjects were split 70/15/15 into training, validation and test parts, stratified by subject class, with seed 2026. An automatic check confirms that no subject and no pixel hash occurs in two parts (fingerprint of splits.json: 2c8efd575bae). Healthy malaria cells also come from infected patients' slides, so a patient-level split separates slides, not only diseases. Table 2 gives the sizes.

Table 2: Datasets after duplicate removal and the patient-disjoint split.

Dataset	Images	Dup. removed	Subjects	Train / val / test images	Train / val / test subjects	Positive (test)
C-NMC	12,418	110	93	8,092 / 2,948 / 1,378	65 / 14 / 14	61.3%
NIH malaria	27,558	0	200	17,759 / 5,420 / 4,379	140 / 30 / 30	52.8%

3.2. Preprocessing

Every image is resized to 224 × 224 pixels, enhanced once by contrast-limited adaptive histogram equalisation (CLAHE) on the L channel of the LAB colour space (clip limit 2.0, 8 × 8 tiles) [21] and smoothed with a 3 × 3 median filter. The same images feed all models. During network training, flips, 90° rotations and ±10% brightness and saturation jitter are applied on the GPU.

3.3. Handcrafted Features

For Random Forest and SVM, 135 features are extracted per image: 16-bin histograms of the H, S, V, L, a and b channels (96) and the mean, standard deviation and skewness of R, G and B (9); grey-level co-occurrence features [22] at distances 1 and 3 (12) and a uniform local-binary-pattern histogram [23] (10); and eight shape descriptors of the stained region segmented by Otsu's threshold on saturation. Features are mapped to a normal distribution by a rank-based quantile transform fitted on the training images; this scaler was chosen on validation AUC over standardisation and a robust scaler (Table 8) because several histogram bins have extreme values after standardisation (46 features in malaria and 40 in C-NMC with $|z| > 10$). Bins that are constant in the training part (17 and 27) have no effect.

3.4. Models

Random Forest [24] uses 500 trees with class-balanced weights and fixed settings. The SVM [25] uses an RBF kernel with $C \in \{1, 10, 100\}$ and $\gamma \in \{\text{scale}, 0.01, 0.001\}$ chosen on validation AUC; its score is the decision function. The custom CNN has a stride-2 stem followed by four convolution blocks (32–64–128–256 filters, two 3 × 3 convolutions with batch normalisation each) and global pooling. EfficientNet-B0 [26], ResNet-50 [27] and DeiT-Small/16 [28] are ImageNet-pretrained models from the timm library [29] with a new two-class head; all layers are fine-tuned. We count the two transfer-learning CNNs as one family, giving five families and six models (Table 3).

Table 3: Models compared. Parameters of the networks from the trained models.

Model	Family	Input	Parameters	Configuration
Random Forest	Classic ML	135 features	—	500 trees, balanced class weights
SVM	Classic ML	135 features	—	RBF, C × γ grid on validation
Custom CNN	CNN from scratch	224 × 224 RGB	1.2 M	stride-2 stem + 4 conv blocks
EfficientNet-B0	Transfer-learning CNN	224 × 224 RGB	4.0 M	ImageNet, fine-tuned
ResNet-50	Transfer-learning CNN	224 × 224 RGB	23.5 M	ImageNet, fine-tuned
DeiT-Small	Vision transformer	224 × 224 RGB	21.7 M	ImageNet, patch 16, fine-tuned

3.5. Training

Networks are trained with AdamW [30] (learning rate 1e-4, 1e-3 for the custom CNN, weight decay 0.05) and a cosine schedule for at most 20 epochs on C-NMC and 10 on malaria, with cross-entropy weighted by inverse class frequency and early stopping on validation AUC (patience 5). Batch sizes are 64 (custom CNN), 32 (EfficientNet-B0), 32 (DeiT-Small) and 16 (ResNet-50). All runs use full fp32 precision on an NVIDIA GTX 1650 SUPER (4 GB), because mixed precision was 2.5–5 times slower on this card without tensor cores. Each model is trained with seeds 0, 1 and 2; the SVM is deterministic and gives the same model for every seed. Training can stop and resume from a per-epoch checkpoint; resumed runs are not bit-identical to uninterrupted ones, so replication rests on the three seeds, and their wall-clock training time is excluded from the time table. Random Forest is fitted on all CPU cores, the SVM on one; handcrafted features are extracted on one core.

3.6. Evaluation

Disease is the positive class. We report accuracy, sensitivity, specificity, F1, balanced accuracy and AUC on the test part, as mean ± standard deviation over seeds, at two thresholds: 0.5, and the Youden threshold [31] chosen on the validation part of each run and then applied unchanged to the test part. Pairs of models are compared on seed 0 with the DeLong test for correlated AUCs [32] and the exact McNemar test [33] on the validation-threshold decisions, with Holm correction [34] over the 15 pairs of each task. 95% confidence intervals are obtained by resampling test patients (2000 resamples), because cells of one patient are correlated. Fig. 8 uses Grad-CAM [35] for ResNet-50 and attention rollout [36] for DeiT-Small.

4. Results

4.1. Overall Performance

Table 4: C-NMC test results, mean ± sd over three seeds (SVM: deterministic, —). First row of each model: threshold 0.5; second row (val. thr.): Youden threshold chosen on validation. Metrics other than AUC in %.

Model	AUC	Acc.	Sens.	Spec.	F1	Bal. acc.
Random Forest	0.8672 ± 0.0010	81.25 ± 0.04	90.57 ± 0.45	66.48 ± 0.78	85.56 ± 0.05	78.53 ± 0.17
val. thr.	0.8672 ± 0.0010	81.11 ± 0.11	90.77 ± 1.85	65.79 ± 3.20	85.49 ± 0.19	78.28 ± 0.68
SVM	0.8757 ± —	79.97 ± —	84.14 ± —	73.36 ± —	83.75 ± —	78.75 ± —
val. thr.	0.8757 ± —	80.70 ± —	85.56 ± —	72.98 ± —	84.46 ± —	79.27 ± —
Custom CNN	0.8367 ± 0.0048	65.46 ± 6.53	51.91 ± 16.43	86.93 ± 9.50	63.83 ± 10.78	69.42 ± 3.70
val. thr.	0.8367 ± 0.0048	77.99 ± 0.59	86.59 ± 2.19	64.35 ± 2.16	82.82 ± 0.72	75.47 ± 0.31
EfficientNet-B0	0.8779 ± 0.0156	76.42 ± 4.76	70.73 ± 11.83	85.43 ± 6.88	78.27 ± 6.08	78.08 ± 2.78
val. thr.	0.8779 ± 0.0156	78.37 ± 1.55	91.95 ± 0.43	56.85 ± 4.69	83.92 ± 0.91	74.40 ± 2.13
ResNet-50	0.9141 ± 0.0168	82.85 ± 2.87	81.81 ± 6.16	84.49 ± 2.38	85.32 ± 3.03	83.15 ± 1.91
val. thr.	0.9141 ± 0.0168	82.54 ± 1.71	92.70 ± 2.01	66.42 ± 4.79	86.69 ± 1.23	79.56 ± 2.15
DeiT-Small	0.9066 ± 0.0199	83.89 ± 3.05	85.64 ± 1.71	81.11 ± 5.18	86.71 ± 2.40	83.38 ± 3.44
val. thr.	0.9066 ± 0.0199	82.41 ± 2.68	93.73 ± 1.14	64.48 ± 8.19	86.76 ± 1.64	79.10 ± 3.69

Table 5: NIH malaria test results, mean ± sd over three seeds (SVM: deterministic, —). First row of each model: threshold 0.5; second row (val. thr.): Youden threshold chosen on validation. Metrics other than AUC in %.

Model	AUC	Acc.	Sens.	Spec.	F1	Bal. acc.
Random Forest	0.9896 ± 0.0003	95.41 ± 0.10	92.46 ± 0.16	98.71 ± 0.10	95.51 ± 0.10	95.59 ± 0.10
val. thr.	0.9896 ± 0.0003	95.48 ± 0.09	92.62 ± 0.36	98.68 ± 0.22	95.58 ± 0.10	95.65 ± 0.08
SVM	0.9804 ± —	95.23 ± —	92.30 ± —	98.50 ± —	95.33 ± —	95.40 ± —
val. thr.	0.9804 ± —	95.18 ± —	92.13 ± —	98.60 ± —	95.28 ± —	95.36 ± —
Custom CNN	0.9813 ± 0.0034	94.75 ± 0.11	90.86 ± 0.28	99.10 ± 0.07	94.81 ± 0.12	94.98 ± 0.10
val. thr.	0.9813 ± 0.0034	94.95 ± 0.23	91.55 ± 0.70	98.74 ± 0.39	95.03 ± 0.25	95.15 ± 0.21
EfficientNet-B0	0.9849 ± 0.0031	93.56 ± 1.25	89.23 ± 2.38	98.40 ± 0.08	93.59 ± 1.32	93.82 ± 1.18
val. thr.	0.9849 ± 0.0031	92.43 ± 1.64	86.50 ± 3.18	99.08 ± 0.21	92.33 ± 1.80	92.79 ± 1.54
ResNet-50	0.9856 ± 0.0043	95.14 ± 0.46	92.12 ± 0.91	98.52 ± 0.07	95.24 ± 0.48	95.32 ± 0.44
val. thr.	0.9856 ± 0.0043	94.97 ± 0.70	91.77 ± 1.59	98.55 ± 0.37	95.06 ± 0.74	95.16 ± 0.65
DeiT-Small	0.9922 ± 0.0008	95.80 ± 0.33	93.05 ± 0.56	98.87 ± 0.12	95.90 ± 0.33	95.96 ± 0.32
val. thr.	0.9922 ± 0.0008	95.92 ± 0.56	93.40 ± 1.08	98.74 ± 0.34	96.03 ± 0.57	96.07 ± 0.53

On C-NMC (Table 4) the two large pretrained networks lead: ResNet-50 reaches AUC 0.9141 ± 0.0168 and DeiT-Small 0.9066 ± 0.0199, followed by EfficientNet-B0 (0.878), the SVM (0.876), Random Forest (0.867) and the custom CNN (0.837). AUC is stable across seeds, but accuracy at 0.5 is not: the custom CNN varies by ±6.53 points and EfficientNet-B0 by ±4.76, because their scores are poorly calibrated around 0.5. The validation threshold removes most of this variation (custom CNN ±0.59).

On malaria (Table 5) every model exceeds AUC 0.98. DeiT-Small is highest (0.9922 ± 0.0008, accuracy 95.80%), followed closely by Random Forest (0.990, 95.41%) and ResNet-50 (0.986). Sensitivity is lower than specificity for every model (93.05% vs 98.87% for DeiT-Small); Section 4.6 shows where the missed cells come from. Fig. 2 shows the ROC curves and Fig. 3 the confusion matrices.

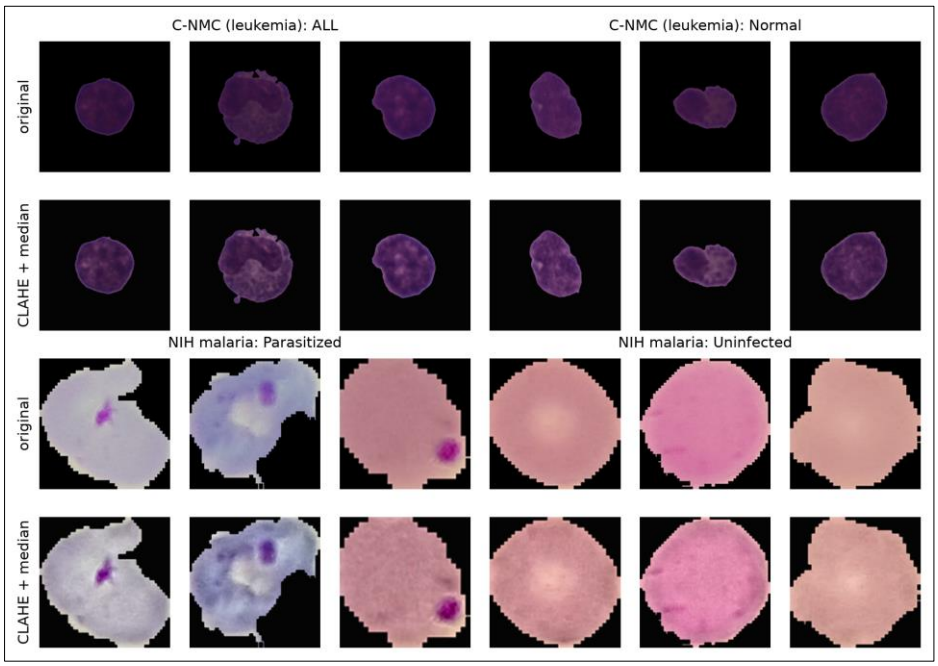


Fig 1: Example test cells of each class before and after CLAHE and median filtering.

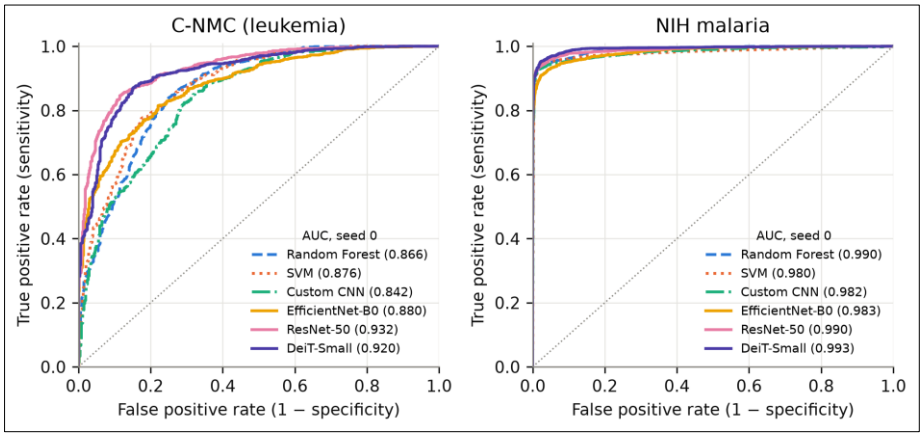


Fig 2: ROC curves of the six models on the test sets (seed 0); dashed and dotted lines: handcrafted-feature models.

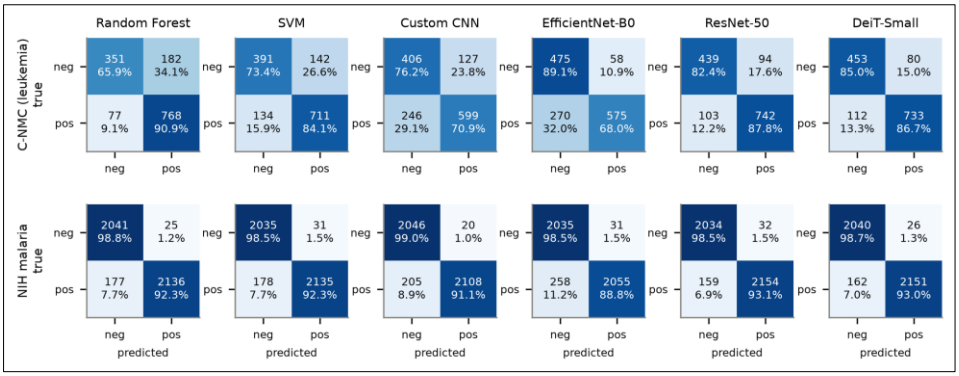


Fig 3: Confusion matrices on the test sets (seed 0, threshold 0.5); percentages per true class.

4.2. Statistical Comparison

Table 6: Pairwise tests on seed 0, Holm-corrected over the 15 pairs of each task (selected pairs; all pairs in out/table_delong.csv and out/table_mcnemar.csv). McNemar: cells only A / only B classifies correctly at the validation thresholds.

Task	A vs B	AUC A / B	DeLong p	McNemar A / B	McNemar p
C-NMC	ResNet-50 vs DeiT-Small	0.932 / 0.920	0.190	80 / 73	1.000
C-NMC	EfficientNet-B0 vs ResNet-50	0.880 / 0.932	< 0.001	59 / 150	< 0.001
C-NMC	SVM vs ResNet-50	0.876 / 0.932	< 0.001	109 / 161	0.020
C-NMC	Random Forest vs ResNet-50	0.866 / 0.932	< 0.001	102 / 148	0.043
C-NMC	Random Forest vs SVM	0.866 / 0.876	0.405	95 / 89	1.000
C-NMC	SVM vs EfficientNet-B0	0.876 / 0.880	0.610	164 / 125	0.126
C-NMC	Custom CNN vs EfficientNet-B0	0.842 / 0.880	< 0.001	121 / 110	1.000
NIH malaria	ResNet-50 vs DeiT-Small	0.990 / 0.993	0.003	57 / 70	1.000
NIH malaria	Random Forest vs DeiT-Small	0.990 / 0.993	0.004	58 / 79	0.632
NIH malaria	Random Forest vs ResNet-50	0.990 / 0.990	1.000	39 / 47	1.000
NIH malaria	Random Forest vs SVM	0.990 / 0.980	< 0.001	33 / 24	1.000
NIH malaria	SVM vs DeiT-Small	0.980 / 0.993	< 0.001	58 / 88	0.145
NIH malaria	EfficientNet-B0 vs DeiT-Small	0.983 / 0.993	< 0.001	31 / 157	< 0.001
NIH malaria	Custom CNN vs EfficientNet-B0	0.982 / 0.983	1.000	145 / 50	< 0.001

On C-NMC (Table 6) ResNet-50 and DeiT-Small do not differ (DeLong p = 0.190, McNemar p = 1.000), and both have higher AUC than every other model. At the validation threshold ResNet-50 also makes significantly fewer errors than Random Forest (p = 0.043), the SVM (p = 0.020) and EfficientNet-B0 (p < 0.001). Random Forest and the SVM do not differ (DeLong p = 0.405). On malaria DeiT-Small has a significantly higher AUC than ResNet-50 (p = 0.003) and Random Forest (p = 0.004), but the differences are small: 0.003 over ResNet-50 on seed 0 (0.007 on the seed mean) and 0.003 over Random Forest (0.003). At the validation thresholds Random Forest, the SVM, ResNet-50 and DeiT-Small make statistically indistinguishable numbers of errors (McNemar p ≥ 0.14); only EfficientNet-B0 is worse (p < 0.001 against DeiT-Small).

Table 7: 95% confidence intervals from 2000 patient-level bootstrap resamples (seed 0, threshold 0.5; % except AUC).

Task	Model	AUC	Accuracy	Sensitivity	Specificity
C-NMC	Random Forest	0.7685–0.9229	70.23–89.56	79.56–97.11	48.83–74.89
	SVM	0.7887–0.9240	72.36–85.91	73.76–92.49	58.66–80.28
	Custom CNN	0.7182–0.9171	63.61–80.53	57.97–81.23	58.64–85.14
	EfficientNet-B0	0.7857–0.9450	64.58–87.34	52.92–85.59	80.60–93.02
	ResNet-50	0.8704–0.9671	79.24–90.57	78.20–94.68	71.64–88.49
	DeiT-Small	0.8493–0.9620	80.36–90.88	78.80–94.76	73.04–90.03
NIH malaria	Random Forest	0.9788–0.9964	90.97–98.19	83.97–97.54	98.16–99.33
	SVM	0.9576–0.9946	91.02–97.98	84.02–97.34	97.89–99.08
	Custom CNN	0.9580–0.9968	89.91–98.15	82.08–96.97	98.56–99.46
	EfficientNet-B0	0.9606–0.9968	86.94–97.71	76.86–96.81	97.81–99.13
	ResNet-50	0.9773–0.9980	91.54–98.27	85.20–97.95	97.77–99.07
	DeiT-Small	0.9845–0.9985	90.88–98.57	83.65–98.37	98.22–99.19

The patient-level intervals (Table 7) are wide on C-NMC, where the test part has only 14 subjects: the AUC interval of ResNet-50 (0.8704–0.9671) overlaps those of all other models. Statistical tests on individual cells therefore overstate the certainty of the C-NMC ranking, and any claim of a single best model on this dataset requires more test patients.

4.3. Handcrafted-Feature Scaling

Table 8: Feature scaler for the SVM, chosen on validation AUC (C and γ re-selected for each scaler). Test AUC is shown for transparency and was not used for the choice.

Task	Scaler	C	γ	Val. AUC	Test AUC
C-NMC	Standardisation	10	0.01	0.9115	0.8876
	Robust (5–95%, clipped)	10	scale	0.9106	0.8926
	Quantile (normal)	1	0.01	0.9181	0.8757
NIH malaria	Standardisation	100	0.001	0.9803	0.9155
	Robust (5–95%, clipped)	10	0.01	0.9829	0.9524
	Quantile (normal)	10	scale	0.9858	0.9804

With standardised features the SVM reached only 0.9155 test AUC on malaria, far below Random Forest on the same features; the quantile transform, which had the best validation AUC in both tasks, raised it to 0.9804 (Table 8). On C-NMC the same choice lowered the test AUC from 0.8876 to 0.8757; we keep it because choosing a scaler per task on test results would itself be leakage.

4.4. Training and Inference Cost

Fitting Random Forest takes seconds, but extracting the handcrafted features takes 3.1 minutes for the C-NMC training set and 19.8 minutes for malaria on one core, which also limits their inference to 15–43 images per second (Table 9). The custom CNN is the fastest model at inference (662 images/s on malaria) and DeiT-Small the slowest network (142 images/s) and the longest to train (50.1 minutes on malaria). Fig. 5 relates AUC to throughput; Fig. 4 shows the training curves

Table 9: Training time per run (mean over the seeds that were not resumed; ResNet-50 on C-NMC from 2 seeds, DeiT-Small on C-NMC from 2 seeds) and inference cost. Networks: GTX 1650 SUPER. RF and SVM: feature extraction on one CPU core (parallelisable) is included in the total and in the inference cost; the RF fit uses all cores, the SVM fit one.

Task	Model	Feature extr. (min)	Fit (min)	Total (min)	Images/s	Latency (ms)
C-NMC	Random Forest	3.1	0.0	3.1	43	75.0
	SVM	3.1	0.5	3.6	42	23.8
	Custom CNN	—	7.1	7.1	303	1.4
	EfficientNet-B0	—	22.3	22.3	298	7.6
	ResNet-50	—	25.2	25.2	239	6.5
	DeiT-Small	—	30.1	30.1	175	7.6
NIH malaria	Random Forest	19.8	0.1	19.9	15	118.3
	SVM	19.8	1.3	21.1	15	67.4
	Custom CNN	—	9.4	9.4	662	1.4
	EfficientNet-B0	—	28.7	28.7	423	7.8
	ResNet-50	—	43.0	43.0	224	6.5
	DeiT-Small	—	50.1	50.1	142	7.5

4.5. Image-Level Leakage and Ablation (Single Runs)

Splitting the same images by image instead of by patient raised the C-NMC AUC of EfficientNet-B0 from 0.880 to 0.974 and its accuracy from 76.20% to 92.16% (Table 10, Fig. 6). Random Forest gained +0.065 AUC, so leakage affects shallow models on handcrafted features as well as deep networks. The inflated values are close to those reported for C-NMC in recent studies [1]– [3]. On malaria the gain is smaller (0.983 → 0.995) because performance is already near the ceiling. Removing CLAHE lowered the C-NMC AUC of EfficientNet-B0 from 0.8805 to 0.8190 but had no adverse effect on malaria (0.9867 vs 0.9830); removing augmentation lowered the C-NMC AUC to 0.8550. These variants were run once and indicate direction, not precise effect sizes.

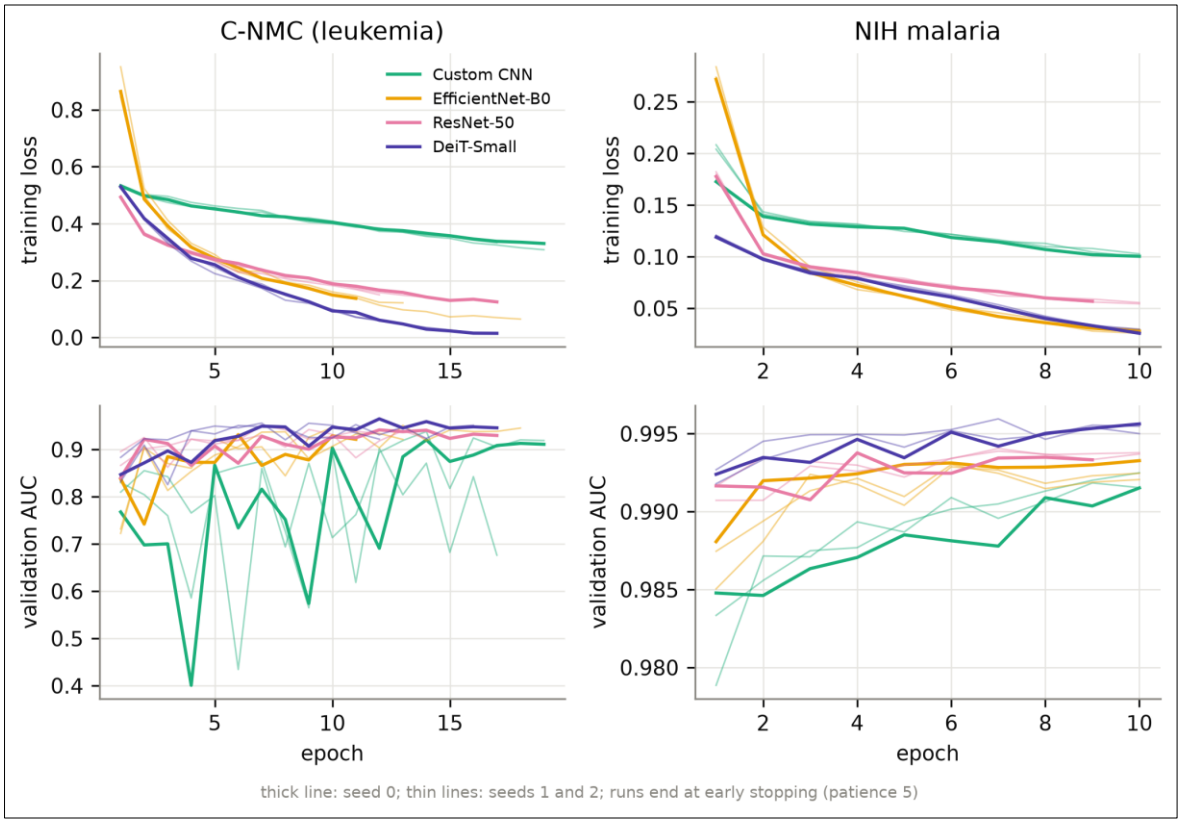


Fig 4: Training loss and validation AUC per epoch of the four networks (thick: seed 0; thin: seeds 1–2).

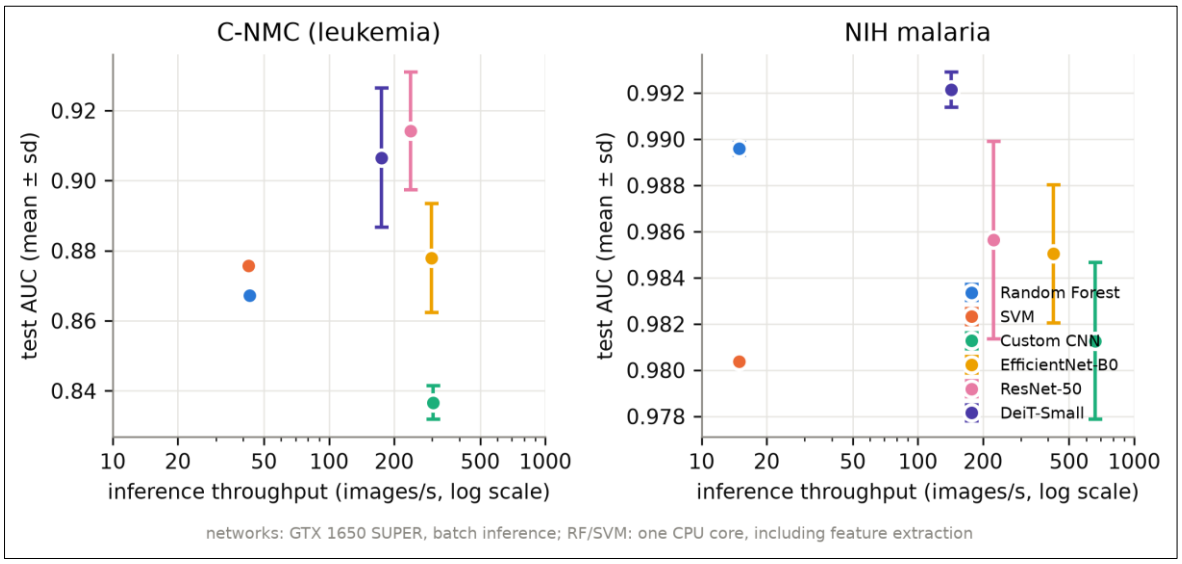


Fig 5: Test AUC (mean ± sd over seeds) against inference throughput (log scale).

Table 10: Ablation and leakage experiments (seed 0, single runs, threshold 0.5). The image-level split has its own test set, so it compares protocols, not models on one test set.

Task	Run	AUC	Acc. (%)	Sens. (%)	Spec. (%)
C-NMC	EfficientNet-B0, patient split	0.8805	76.20	68.05	89.12
	without CLAHE	0.8190	68.80	59.17	84.05
	without augmentation	0.8550	76.49	74.44	79.74
	image-level split	0.9739	92.16	93.09	90.23
	Random Forest, patient split	0.8661	81.20	90.89	65.85
	image-level split	0.9312	89.10	94.44	77.98
NIH malaria	EfficientNet-B0, patient split	0.9830	93.40	88.85	98.50
	without CLAHE	0.9867	93.24	88.28	98.79
	without augmentation	0.9829	92.85	87.38	98.98
	image-level split	0.9954	97.24	96.42	98.06
	Random Forest, patient split	0.9896	95.39	92.35	98.79
	image-level split	0.9914	95.98	94.73	97.24

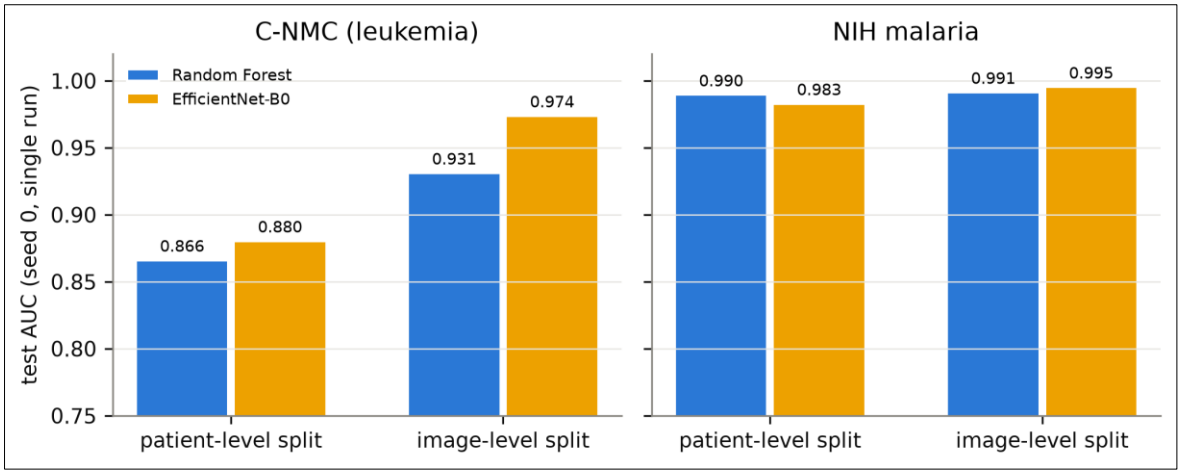


Fig 6: Test AUC with patient-level and image-level splits (seed 0, single runs).

4.6. Per-Patient Analysis of Malaria Errors

The 22 malaria test patients with parasitized cells contribute 2,313 positive cells. 95 of them are missed by all six models, and 75 of these come from one patient, C116 (Table 11, Fig. 7), whose 414 parasitized cells are detected with only 60–75% sensitivity, against a median of 97% for the other patients (DeiT-Small). The three worst patients hold 83% of these cells. When five model families with different inductive biases agree on the same errors, the cause is more likely the labels or the staining of that slide than the models; the sensitivity ceiling on this dataset is a data ceiling. We kept the patient in the test set.

Table 11: Malaria test patients with the most parasitized cells missed by all six models (seed 0, threshold 0.5); sensitivity in %.

Patient	Cells	RF	SVM	CNN	EffB0	R50	DeiT	Missed by all
C116	414	73.4	73.9	69.6	60.1	75.1	71.5	75
C95	24	79.2	79.2	83.3	83.3	87.5	87.5	2
C183	133	92.5	92.5	93.2	95.5	94.0	97.0	2
C159	75	88.0	90.7	88.0	93.3	89.3	97.3	2
C80	268	95.9	95.9	96.3	94.8	96.6	98.9	2
C70	229	99.1	99.1	99.1	99.1	99.1	99.1	2

Cells: parasitized cells of the patient in the test set; R50 = ResNet-50, EffB0 = EfficientNet-B0.

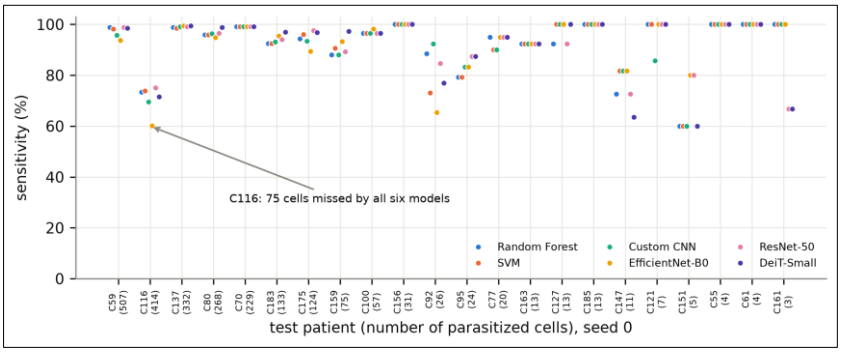


Fig 7: Sensitivity of each model for every malaria test patient with parasitized cells (seed 0).

4.7. Explainability

Fig. 8 shows Grad-CAM maps of ResNet-50 and attention rollout of DeiT-Small for test cells chosen by a fixed rule (for each class, the first two cells in file order that both networks classify correctly and the first cell that ResNet-50 misclassifies). For correctly detected parasitized cells both models concentrate on the stained parasite, and for correctly detected lymphoblasts on the interior of the cell, where chromatin texture differs. The errors are informative: the missed parasitized cell shows no visible parasite, consistent with a doubtful label, and the uninfected cell called parasitized carries a stained fragment at its edge, on which both models focus. Grad-CAM shows evidence for the disease class and is normalised per image, so on negative cells, where that evidence is weak, it highlights background noise; rollout maps are coarser (14 × 14 patches) and spread over the whole cell or its border.

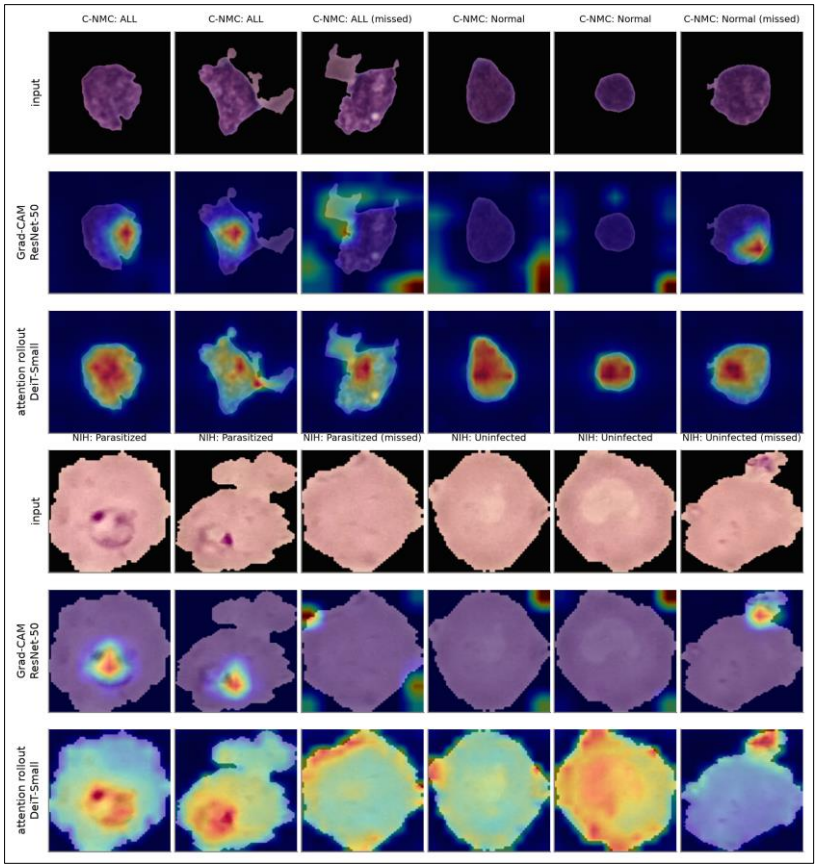


Fig 8: Grad-CAM (ResNet-50, evidence for the disease class) and attention rollout (DeiT-Small) on test cells selected by a fixed rule.

5. Discussion

Interpretation: Under a patient-disjoint protocol, the gap between model families depends on the task. Detecting a parasite is a colour-and-shape problem that 135 handcrafted features solve almost as well as a vision transformer; Random Forest is within 0.003 AUC of DeiT-Small and makes as few errors at a validated threshold. Separating lymphoblasts from normal lymphocytes depends on subtle chromatin texture, and there the pretrained networks gain about 0.04 AUC over handcrafted features, while a CNN trained from scratch does not.

Leakage: Reported C-NMC results of 92.5–97% accuracy and AUC up to 0.99 [1]_ [3] are approached in our experiments only when images are split instead of patients. We therefore read published C-NMC results without a stated patient-level split as optimistic by roughly 0.07–0.09 AUC. The duplicate cells filed under different patient IDs show that even a patient-level split needs a pixel-level audit.

Limitations: The C-NMC test part has only 14 subjects, so its confidence intervals are wide and the ResNet-50 versus DeiT-Small comparison cannot be settled. The Youden thresholds are estimated on validation parts of 14 and 30 subjects and transfer imperfectly. Statistical tests use seed 0 only. Ablation and leakage variants are single runs. The SVM scaler was changed after the first review of the results (Section 4.3), always on validation evidence. Healthy malaria cells come from the slides of infected patients, so the task is cell classification, not patient diagnosis. All timings are for one consumer GPU and a single desktop CPU, with feature extraction on one core.

6. Conclusion and Future Work

On two public blood-smear benchmarks evaluated with patient-disjoint splits, three seeds and significance tests, pretrained ResNet-50 and DeiT-Small lead on leukemia (AUC about 0.91) without a significant difference between them, while on malaria a Random Forest on handcrafted features is nearly as good as the best network at a fraction of the model size. Image-level splitting inflates leukemia AUC by up to 0.09, and the remaining malaria errors cluster in a few patients. Future work will add more patients and external test sites, multiclass leukemia subtypes and anemia morphology, calibration of decision thresholds, and whole-slide (patient-level) diagnosis.

Data and Code Availability

Both datasets are public [19, 20]. The preprocessing, split, training and evaluation code, the patient-level splits (fingerprint 2c8efd575bae) and all result files are available from the author [TODO repository link].

References

1. Mohan, S. B., Sathya, S., Rajalakshmi, S., *et al*. Efficient convolutional neural networks for acute lymphoblastic leukaemia prediction in computer vision. Scientific Reports. 2025;16(1):1135. doi:10.1038/s41598-025-30777-w
2. Jammal, F., Dahab, M., Bayahya, A. Y. Neuro-Bridge-X: A Neuro-Symbolic Vision Transformer with Meta-XAI for Interpretable Leukemia Diagnosis from Peripheral Blood Smears. Diagnostics. 2040;15(16):2025. doi:10.3390/diagnostics15162040
3. Jammal, F. An explainable meta-learned hybrid CNN-transformer model with dual attention for leukemia diagnosis from peripheral blood smears. Scientific Reports. 2026;16(1):27271. doi:10.1038/s41598-026-55606-6
4. Hendrawan, J., Halim, T. F., Gusman, J. B., *et al*. Hybrid deep learning model for acute lymphoblastic leukemia (ALL) detection using pre-trained ResNet-50 and Vision Transformer architecture. Procedia Computer Science. 2025;269:933-942. doi:10.1016/j.procs.2025.09.036
5. Ashikuzzaman, M., Julhash, M., Ali, M. F., *et al*. Automated leukemia detection from microscopic images using deep transfer learning with explainable AI-based analysis. Scientific Reports. 2026;16(1):24693. doi:10.1038/s41598-026-55415-x
6. Manalı, D., Akel, M., Demirel, H. Comparative analysis of CNN, vision transformer, and hybrid architectures for white blood cancer classification. Signal, Image and Video Processing. 2025;19(18):1452. doi:10.1007/s11760-025-05041-3
7. Kasim, S., Malek, S., Tang, J., *et al*. Multiclass leukemia cell classification using hybrid deep learning and machine learning with CNN-based feature extraction. Scientific Reports. 2025;15(1):23782. doi:10.1038/s41598-025-05585-x
8. Boit, S., Patil, R. An Efficient Deep Learning Approach for Malaria Parasite Detection in Microscopic Images. Diagnostics. 2024;14(23):2738. doi:10.3390/diagnostics14232738
9. Mujahid, M., Rustam, F., Shafique, R., *et al*. Efficient deep learning-based approach for malaria detection using red blood cell smears. Scientific Reports. 2024;14(1):13249. doi:10.1038/s41598-024-63831-0
10. Benachour, Y., Flitti, F., Khalid, H. M. Enhancing Malaria Detection Through Deep Learning: A Comparative Study of Convolutional Neural Networks. IEEE Access. 2025;13:35452-35477. doi:10.1109/ACCESS.2025.3542160
11. Shahin, O. R., Alshammari, H. H., Alabdali, R. N., *et al*. Automated multi-model framework for malaria detection using deep learning and feature fusion. Scientific Reports. 2025;15(1):25672. doi:10.1038/s41598-025-04784-w
12. Reddy, C. K. K., Anisha, P. R., Almushharaf, A., *et al*. An optimized transfer learning approach integrating deep convolutional feature extractors for malaria parasite classification in erythrocyte microscopy. Frontiers in Medicine. 2025;16:1684973. doi:10.3389/fmed.2025.1684973
13. Gasmi, K., Krichen, M., Alanazi, A., *et al*. CT-Malaria Detection via Adaptive-Weighted Deep Learning Models. Biomedicines. 2026;14(4):898. doi:10.3390/biomedicines14040898
14. Dehbi, F., Hamou, R. M., Brahami, M., *et al*. Deep learning-assisted malaria microscopy with sensitivity-aware threshold optimization. Frontiers in Medicine. 2026;13:1911275. doi:10.3389/fmed.2026.1911275
15. Özdemir, E. Y., Koç, C., Küçük, K., *et al*. Ensemble-Based Early Detection of Malaria via Explainable ViT-CNN Feature Fusion and SHAP. Journal of Imaging Informatics in Medicine. 2025;39(4):3300-3314. doi:10.1007/s10278-025-01719-9
16. Nogueira, M., Gomes, E. F. Data Leakage Concerns in Training and Evaluation Protocols for Oral Cancer

Image Classification. SN Computer Science. 2026;7(5):530. doi:10.1007/s42979-026-05141-y
17. Wang, L. Reliable Benchmarking of Breast Ultrasound Lesion Classification Requires Patient-Level Validation. J. Engineering and Science in Medical Diagnostics and Therapy. 2026;9(4):041204. doi:10.1115/1.4071712
18. Young, V. M., Gates, S., Garcia, L. Y., *et al*. Data Leakage in Deep Learning for Alzheimer's Disease Diagnosis: A Scoping Review of Methodological Rigor and Performance Inflation. Diagnostics. 2025;15(18):2348. doi:10.3390/diagnostics15182348
19. Gupta A, Gupta R. ALL Challenge dataset of ISBI 2019 (C-NMC 2019). The Cancer Imaging Archive. 2019. doi:10.7937/tcia.2019.dc64i46r
20. Rajaraman S, *et al*. Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images. PeerJ. 2018;6:e4568. doi:10.7717/peerj.4568
21. Zuiderveld K. Contrast limited adaptive histogram equalization. Graphics Gems IV, P. S. Heckbert, Ed. San Diego, CA, USA: Academic Press. 1994:474-485.
22. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans. Syst., Man, Cybern., vol. SMC-3. 1973:610-621. doi:10.1109/TSMC.1973.4309314
23. Ojala T, Pietikäinen M, Mäenpää T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 2002;24(7):971-987. doi:10.1109/TPAMI.2002.1017623
24. Breiman L. Random forests. Mach. Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324
25. Cortes C, Vapnik V. Support-vector networks. Mach. Learn. 1995;20(3):273-297. doi:10.1007/BF00994018
26. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. Proc. ICML. 2019:6105-6114.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proc. CVPR. 2016:770-778. doi:10.1109/CVPR.2016.90
28. Touvron H, *et al*. Training data-efficient image transformers & distillation through attention. Proc. ICML. 2021:10347-10357.
29. Wightman R. PyTorch image models (timm). GitHub repository. 2019. doi:10.5281/zenodo.4414861
30. Loshchilov I, Hutter F. Decoupled weight decay regularization. Proc. ICLR. 2019.
31. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32-35.
32. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988;44(3):837-845. doi:10.2307/2531595
33. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153-157. doi:10.1007/BF02295996
34. Holm S. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 1979;6(2):65-70.
35. Selvaraju RR, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proc. ICCV. 2017:618-626. doi:10.1109/ICCV.2017.74
36. Abnar S, Zuidema W. Quantifying attention flow in transformers. Proc. ACL. 2020:4190-4197. doi:10.18653/v1/2020.acl-main.385

How to Cite This Article
Mohammed ZR, Kadhim MH. Comparative performance evaluation of five machine learning models for diagnosing blood diseases from Microscopic images. Int J Future Eng Innov. 2026;3(6):54-64. doi:10.54660/IJFEI.2026.3.6.54-64.

Creative Commons (CC) License
This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms