Journal of Prevention, Diagnosis and Management of Human Diseases
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1202, DOI: 10.55529/jpdmhd.62.63.72

Research Paper

From cells to patients: Validation, image quality and predictive values of dep learning for malaria detection in thin blood smears

Dr. Sourav Kundu[1*], Tashmatova Nazgul[2], Turgunova Muattarkhon[3]
[1*]Department of Tropical Medicine, School of Tropical Medicine, Kolkata, West Bengal, India.
[2]Department of Morphological Disciplines, Jalal-Abad International University, Osh State University, Osh International Medical University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.
[3]Department of Morphological Disciplines, Jalal-Abad International University, 720906, 24 Kurmanbek Str., Manas, Kyrgyzstan.

Article Info

Article History:
Received: 26 May 2026
Revised: 03 July 2026
Accepted: 11 July 2026
Published: 27 August 2026

Keywords:
Malaria
Thin Blood Smear
Deep Learning
Diagnostic Accuracy
Patient-Level Validation
Image Quality

ABSTRACT

Background: Deep learning models for malaria detection in thin blood smears report accuracies of 95-99%, but are usually validated on randomly split, high-quality cell images and evaluated per cell rather than per patient.
Objective: To re-examine these assumptions using real thin-smear data.
Methods: The NIH thin-smear dataset was used (27,556 segmented red blood cell images from 150 Plasmodium falciparum-infected and 50 healthy patients). A compact convolutional neural network was evaluated with image-level and patient-level five-fold cross-validation, under 12 simulated image degradations, and after aggregation of cell predictions into a patient-level diagnosis. Predictive values were computed for prevalences of 1-50%.
Results: On unseen patients, cell-level accuracy was 95.7%, sensitivity 93.8%, specificity 97.5% and AUROC 0.988. Image-level validation overestimated accuracy by only 0.4 percentage points but understated variation between test folds four-fold. Under patient-level validation, seven patients had an accuracy below 90% (minimum 72.8%). Blur, low resolution and stain shift reduced sensitivity to 74-83%, whereas noise and underexposure reduced specificity to 52-69%. Severe JPEG compression reduced sensitivity to 72%. At the patient level, a healthy individual with about 68 examined cells had a median of one false-positive cell. Calling any positive cell a positive diagnosis gave a sensitivity of 99.3% but a specificity of only 34%. Requiring five positive cells balanced both at 94.0%. Requiring ten positive cells removed all false positives but missed 15% of infections, mostly at low parasitaemia. At 1% prevalence, positive predictive values were 14% or less.

Conclusion: Automated malaria microscopy should be validated per patient, under the image quality of the intended device and at the prevalence of the intended population.

Corresponding Author:
Dr. Sourav Kundu
Department of Tropical Medicine, School of Tropical Medicine, Kolkata, West Bengal, India.
Email: drsouravkundu@gmail.com

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Malaria remains one of the most important preventable and treatable infectious diseases, with an estimated 263 million cases and 597,000 deaths in 2023, most of them in sub-Saharan Africa and among young children [1]. Prompt parasitological confirmation before treatment is a cornerstone of malaria control, and light microscopy of Giemsa-stained blood smears is still the reference method in most endemic regions [2], [3]. Its accuracy, however, depends on the skill and workload of the microscopist, and trained staff are scarce exactly where the disease burden is highest. Automated analysis of smear images, increasingly captured with smartphones attached to microscopes, has therefore attracted considerable interest [4], [5].

A widely used resource for this work is the National Institutes of Health (NIH) thin-smear dataset, which contains about 27,500 segmented red blood cell images labelled as parasitised or uninfected [6]. Convolutional neural networks (CNNs) trained on it routinely report accuracies of 95–99% [6], [7], [8]. Three aspects of these evaluations limit what they say about clinical performance. First, many studies split the images randomly, so that cells from the same patient and slide appear in both the training and the test set; this "record-wise" validation is known to overestimate performance in clinical machine learning [9], [10], [11]. Second, accuracy is measured on curated, high-quality images, whereas field images are often blurred, compressed or poorly exposed [12], [13]. Third, results are reported per cell, while clinical decisions are made per patient, and the predictive value of a positive result depends strongly on the prevalence of infection in the tested population [14], [15].

This study addresses these three gaps using the real cell images of the NIH dataset. It (i) compares image-level and patient-level cross-validation of the same CNN; (ii) measures how cell-level performance changes under realistic image degradations; and (iii) aggregates cell predictions into a patient-level diagnosis and estimates positive and negative predictive values at prevalences typical of screening and clinical settings.

2. METHODOLOGY

2.1 Dataset

The NIH malaria dataset was collected at Chittagong Medical College Hospital, Bangladesh, from Giemsa-stained thin blood smears of 150 patients infected with Plasmodium falciparum and 50 healthy individuals, photographed through a smartphone attached to a light microscope [6]. Red blood cells were segmented automatically and labelled by an expert slide reader as parasitised or uninfected. The images were obtained from a public mirror of the dataset [16], which contains 13,778 parasitised and 13,778 uninfected cell images (27,556 in total; the original release lists 27,558). The patient (smear) identifier was extracted from each file name, which yielded exactly 200 patients: 150 with at least one parasitised cell and

50 with only uninfected cells, in agreement with the original description. The number of cells per patient ranged from 65 to 702 (median 86). Cell images are of variable size (median 130 × 130 pixels) and were resized to 64 × 64 pixels for analysis. Figure 1 shows real examples from 16 different patients. The images are anonymised and publicly released; no patient information was used.

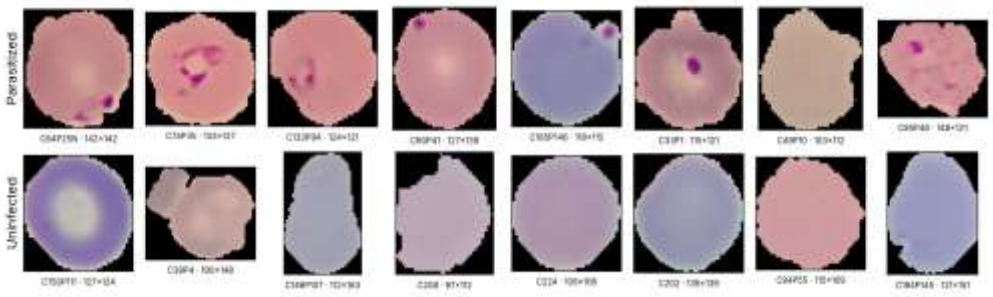


Figure 1. Real Segmented Red Blood Cell Images from the NIH Malaria Dataset, Each from a Different Patient (Patient Identifier and Original Size in Pixels Shown Below Each Image). Top: Cells Labelled as Parasitised, Most with a Stained Ring-Form or Trophozoite. Bottom: Uninfected Cells. Note the Variation in Stain Colour between Patients and That Some Cells Labelled as Parasitised Show No Clearly Visible Parasite (e.g. C49P10)

2.2 Classifier

A compact CNN was trained from scratch for each experiment: four convolutional blocks (16, 32, 64 and 64 channels; 3 × 3 kernels, batch normalisation, ReLU and 2 × 2 max pooling), global average pooling, dropout (0.2) and a single sigmoid output giving the probability that a cell is parasitised. Training used the Adam optimiser [17] with a one-cycle learning-rate schedule (peak 4 × 10−3), weight decay 10−4, batch size 64, binary cross-entropy loss, random flips and 90° rotations, for six epochs. The model was implemented in PyTorch [18] and trained on a two-core CPU.

2.3 Validation Schemes

Two five-fold cross-validation schemes were compared. In the image-level scheme, cells were assigned to folds at random (stratified by label), as in many published studies, so that cells from the same patient appeared in both training and test folds. In the patient-level scheme, all cells of a patient were kept in the same fold (stratified group K-fold), so that every test patient was unseen during training [9]. Every image was tested exactly once in each scheme. Cell-level performance was summarised by accuracy, sensitivity, specificity and the area under the ROC curve (AUROC) at a threshold of 0.5. Variability was assessed across folds and, for patient-level results, by a cluster bootstrap that resamples patients (1,000 replicates). Metrics were computed with scikit-learn [19].

2.4 Image Degradations

To emulate image-quality problems that occur with smartphone microscopy and image transfer, each patient-level test fold was degraded in 12 ways and re-evaluated with the model trained on that fold: Gaussian blur (σ = 1 and 2 pixels, defocus), down-sampling by a factor of 2 and 4 followed by up-sampling (low resolution), JPEG compression at quality 30 and 10, additive Gaussian noise (σ = 10 and 25 grey levels), brightness reduced or increased by 30%, and a stain-colour shift that scales the red and blue channels in opposite directions by 15% or 30%. Degradations were applied to the 64 × 64-pixel images [13].

2.5 Patient-Level Diagnosis and Predictive Values

For each patient, the cell-level predictions from the patient-level scheme were aggregated into a patient score, defined as the number of cells with a predicted probability above 0.5. A patient was classified as infected if the score reached a threshold k (k = 1, 2, 3, 5 or 10 cells), and patient-level sensitivity and specificity were computed against the patient's true status (infected if any cell was labelled parasitised).

The patient-level AUROC was computed from the fraction of positive cells. Because the dataset contains 75% infected patients, far more than in most tested populations, positive and negative predictive values (PPV and NPV) were computed with Bayes' theorem for prevalences of 1%, 5%, 10%, 30% and 50%, representing community screening, outpatient testing and high-transmission settings [14], [15]. Reporting follows the relevant items of STARD 2015 [20].

3. RESULTS AND DISCUSSION

3.1 Cell-Level Performance and Validation Scheme

With patient-level cross-validation, the CNN classified the 27,556 cells with an accuracy of 95.7%, a sensitivity of 93.8% and a specificity of 97.5%, and an AUROC of 0.988 Table 1. Image-level cross-validation gave slightly higher values (accuracy 96.1%, AUROC 0.991). The optimistic bias of record-wise validation was therefore small in this dataset, about 0.4 percentage points of accuracy, and the patient-clustered 95% confidence intervals of the two schemes overlapped. The patient-level scheme, however, revealed much larger variation between test folds: the standard deviation of fold accuracy was 1.3 points, four times that of the image-level scheme (0.3 points), and fold accuracy ranged from 93.4% to 96.5%. Performance also varied between patients. Under patient-level validation, the median per-patient accuracy was 97.1%, but seven patients had an accuracy below 90%, with a minimum of 72.8%. Such patient-specific failures are the type of variation that patient-level evaluation is designed to expose. The count of poorly classified patients was not itself lower under image-level validation (nine patients below 90%); what image-level validation concealed was the between-fold variability, which it understated four-fold.

Table 1. Cell-Level Performance of the CNN under Image-Level and Patient-Level Five-Fold Cross-Validation (27,556 Cells, 200 Patients). Values Are Pooled over Folds; 95% Confidence Intervals from a Bootstrap That Resamples Patients

Validation Scheme	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUROC	SD of Fold Accuracy	Patients with Accuracy < 90%
Image-level (random cells)	96.1 (95.4–96.6)	94.4 (93.2–95.4)	97.7 (97.4–98.1)	0.991 (0.990–0.993)	0.3	9
Patient-level (unseen patients)	95.7 (94.8–96.4)	93.8 (92.2–95.1)	97.5 (97.0–97.9)	0.988 (0.985–0.991)	1.3	7

3.2 Effect of Image Degradation

Figure 2 shows how the degradations alter real cell images, and Table 2 and Figure 3 summarise their effect. Mild blur ($\sigma = 1$) and two-fold down-sampling had almost no effect. Stronger degradations that remove fine detail, such as blur at $\sigma = 2$, four-fold down-sampling and a 30% stain shift, mainly reduced sensitivity, to 83%, 75% and 74% respectively, while specificity stayed at 98–99%: small ring-stage parasites became invisible and infected cells were called uninfected. Degradations that add structure behaved in the opposite way. Strong noise ($\sigma = 25$) and a 30% reduction in brightness kept sensitivity at about 95% but reduced specificity to 52% and 69%, with very large variation between folds (SD 27 and 38 points), meaning that some models labelled most uninfected cells as parasitised. Severe JPEG compression (quality 10) reduced both, to 72% sensitivity and 82% specificity (AUROC 0.857). In contrast, the AUROC remained above 0.96 for most degradations, showing that much of the loss reflects a shift of the decision threshold rather than a complete loss of discriminative information.

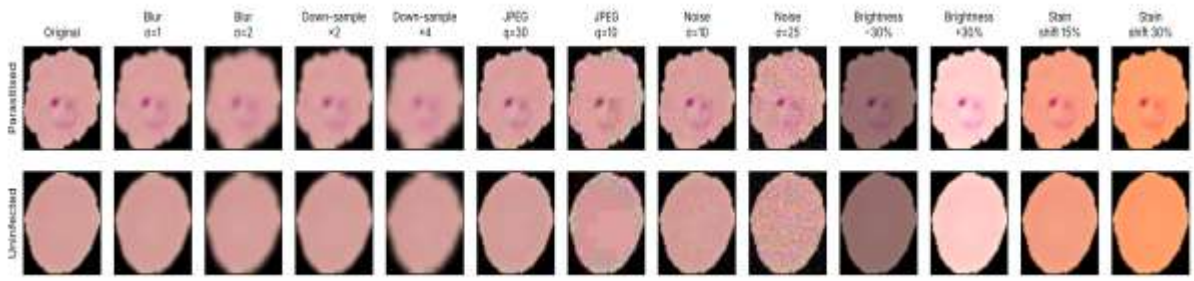


Figure 2. A Real Parasitised Cell (Top) and an Uninfected Cell (Bottom) from the NIH Dataset under Each of the 12 Image Degradations Used in the Robustness Analysis

Table 2. Cell-Level Performance on Unseen Patients under Image Degradation (Mean over Five Patient-Level Test Folds; SD in Parentheses)

Condition	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUROC
Original images	95.7 (1.3)	93.8 (2.9)	97.5 (0.8)	0.989
Blur σ = 1 / σ = 2	95.5 / 90.8	93.5 / 82.9	97.5 / 98.6	0.988 / 0.976
Down-sampling ×2 / ×4	95.2 / 87.1	92.5 / 75.4	97.8 / 98.8	0.987 / 0.965
JPEG quality 30 / 10	91.9 / 76.9	87.6 / 72.5	96.3 / 81.5	0.972 / 0.857
Gaussian noise σ = 10 / 25	94.3 / 73.0	94.9 / 94.6	93.6 / 51.5 (26.5)	0.982 / 0.914
Brightness −30% / +30%	82.4 / 92.1	95.4 / 86.0	69.2 (38.2) / 98.3	0.978 / 0.972
Stain shift 15% / 30%	93.6 / 86.7	88.4 / 74.1	98.7 / 99.4	0.986 / 0.981

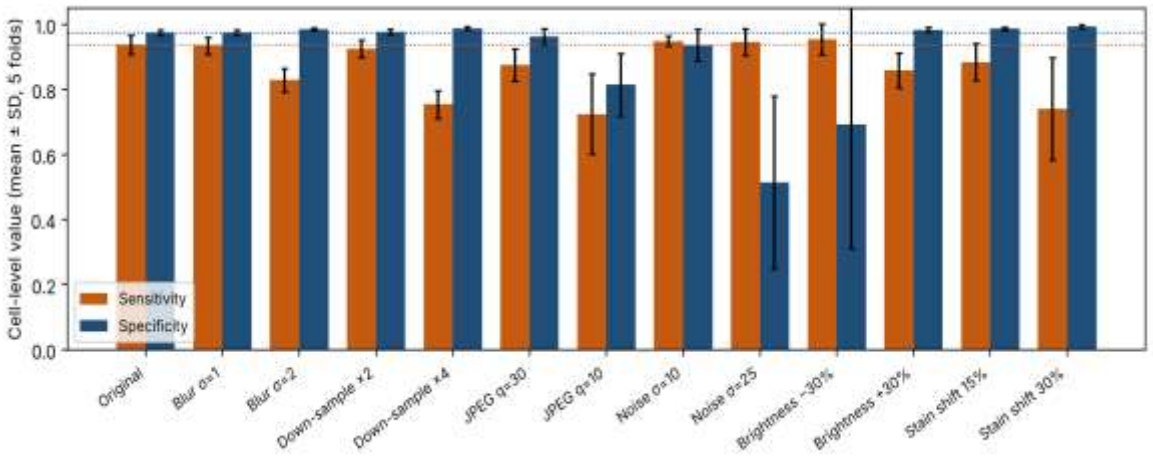


Figure 3. Cell-Level Sensitivity and Specificity on Unseen Patients for Each Image Degradation (Mean ± SD over Five Folds). Dotted Lines Mark the Values on Original Images

3.3 Patient-Level Diagnosis

When cell predictions were aggregated per patient, the number of cells predicted to be parasitised separated infected from healthy patients well (patient-level AUROC 0.983). However, healthy patients were not free of positive calls: with a false-positive rate of 2.1% among cells from healthy patients (2.5% across all uninfected cells, corresponding to the 97.5% specificity in Table 1, a healthy patient with about 68 analysed cells had a median of one and up to seven cells predicted as parasitised Figure 4, left. As a result, the choice of the decision rule strongly affected the diagnosis Table 3. Declaring a patient infected if any cell was positive (k = 1) detected 149 of 150 infected patients (sensitivity 99.3%) but only 17 of 50 healthy patients were correctly negative (specificity 34%). Requiring five positive cells balanced the two

(sensitivity 94.0%, specificity 94.0%), and requiring ten positive cells eliminated false positives (specificity 100%, 95% CI 92.9–100%) at the cost of missing 23 infected patients (sensitivity 84.7%). The missed patients were mainly those with low parasitaemia: among the 33 infected patients with ten or fewer labelled parasitised cells, the sensitivity of the k = 5 and k = 10 rules fell to 72.7% and 33.3%.

Table 3. Patient-Level Diagnosis by the Number of Cells Predicted to be Parasitised (k), with Exact 95% Confidence Intervals, and Predictive Values at Different Prevalences of Infection

Rule	Sensitivity (%)	Specificity (%)	PPV at 1%	PPV at 5%	PPV at 30%	NPV at 30%
k ≥ 1 cell	99.3 (96.3–100)	34.0 (21.2–48.8)	0.015	0.073	0.39	0.99
k ≥ 2 cells	99.3 (96.3–100)	66.0 (51.2–78.8)	0.029	0.13	0.56	1.00
k ≥ 3 cells	96.7 (92.4–98.9)	80.0 (66.3–90.0)	0.047	0.20	0.67	0.98
k ≥ 5 cells	94.0 (88.9–97.2)	94.0 (83.5–98.7)	0.14	0.45	0.87	0.97
k ≥ 10 cells	84.7 (77.9–90.0)	100 (92.9–100)	—*	—*	—*	0.94

*Not estimable because no false positives were observed among 50 healthy patients; with the lower confidence limit of specificity (92.9%), the PPV would be 0.11 at 1%, 0.39 at 5% and 0.84 at 30% prevalence.

Predictive values depended strongly on prevalence Figure 4, right. At the 75% prevalence of the dataset, all rules looked acceptable, but in a screening population with 1% prevalence, the PPV of the k = 1 rule was only 1.5% and that of the balanced k = 5 rule 14%: most positive results would be false. Even at the cell level, the 97.5% specificity translated into a PPV of only 27% if 1% of examined cells were parasitised. Negative predictive values remained above 0.93 for all rules and prevalences up to 30%.

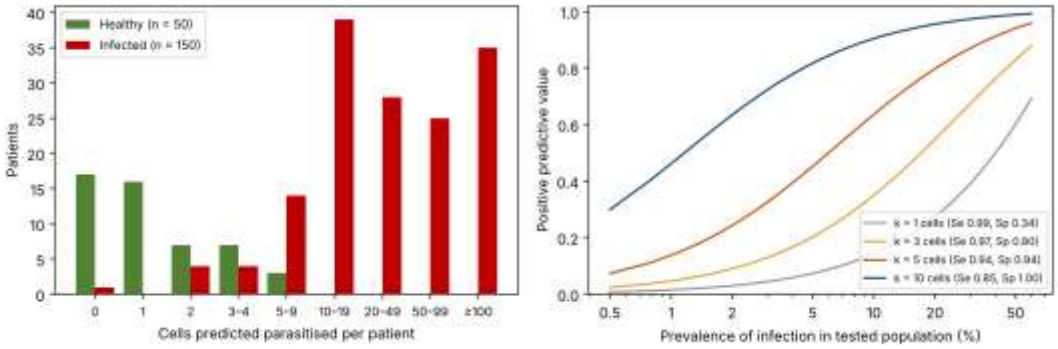


Figure 4. Left: Number of Cells Predicted to be Parasitised per Patient for 50 Healthy and 150 Infected Patients (Patient-Level Cross-Validation). Right: Positive Predictive Value of Patient-Level Decision Rules as a Function of the Prevalence of Infection. For k = 10, Where No False Positives Were Observed, Specificity Was Set to 0.99 (Half a False Positive among 50 Healthy Patients) to Allow the Curve to be Drawn

3.4 Discussion

This study re-examined CNN-based malaria detection on the real NIH thin-smear images from the perspective of clinical use. Three findings stand out. First, on these curated images, the CNN detected parasitised cells with high accuracy even in patients it had never seen (AUROC 0.988), comparable with

published results [6], [7], [8]. Contrary to expectation, image-level validation inflated accuracy only slightly in this dataset [9], [10]. Patient-level validation was nevertheless more informative: it exposed four times greater variability between test sets and a minority of patients in whom accuracy fell to 73–79%. For a diagnostic test, such patient-to-patient variation matters more than a fraction of a percentage point of average accuracy, and it can only be seen when patients are kept separate [11].

Second, performance depended strongly on image quality, and different degradations failed in different directions. Loss of detail from defocus, low resolution or stain changes mainly caused missed infections, which is the more dangerous error for a disease in which delayed treatment can be fatal [1]. Noise and underexposure mainly caused false alarms, which waste treatment and erode trust in the system. Because the AUROC was largely preserved under most degradations, quality control of the input images and recalibration of the decision threshold for the acquisition device may recover much of the loss [12], [21]. In practice, automated systems should reject or flag images whose quality falls outside the range seen during validation, in the same way that quality assurance is required for manual microscopy [3].

Third, excellent cell-level metrics do not translate directly into a reliable patient-level diagnosis. A specificity of 97.5% per cell sounds high, but a slide contains many cells: with about 68 cells per healthy patient, two thirds of healthy patients had at least one false-positive cell. In routine microscopy, far more cells are examined; at the same per-cell error rate, a smear in which 2,000 red cells are inspected would contain, on average, about 40 false-positive cells, so that a simple count threshold would no longer separate low-density infections from healthy individuals. Decision rules must therefore be defined and validated per patient, with the number of cells examined taken into account, and the positive predictive value must be judged at the prevalence of the intended setting [14], [15]. At the low prevalences typical of community screening and of elimination settings, even a well-performing classifier would produce mostly false-positive results unless its specificity is extremely high, so automated results should be confirmed by an expert or a second test before treatment, as recommended for other screening tools [22].

This study has limitations. The NIH dataset consists of pre-segmented cells and contains only P. falciparum infections from a single hospital; species identification, parasite staging and thick smears were not addressed. Some labels are known to be noisy [8], and a few cells labelled as parasitised showed no visible parasite Figure 1. Infected patients contribute more cells than healthy ones, and parasitised cells are enriched relative to a real smear, so the patient-level analysis is illustrative and cannot reproduce clinical parasite densities. The degradations were simulated on resized images and do not capture all real acquisition artefacts. Only one compact CNN trained from scratch was evaluated, and no external dataset was available for testing.

4. CONCLUSION

On real thin-smear images from 200 patients, a compact CNN detected malaria-parasitised red blood cells with an AUROC of 0.988 in patients unseen during training, and image-level validation overestimated accuracy only slightly. Patient-level validation, however, revealed substantial variation between patients, and realistic image degradations reduced sensitivity to 72–83% or specificity to 52–69%, depending on the degradation. Aggregated to patients, the same classifier offered a trade-off between missing low-density infections and false-positive diagnoses, with positive predictive values of 14% or less at 1% prevalence. Automated malaria microscopy should therefore be validated at the patient level, under the image quality of the intended device and at the prevalence of the intended population before it is used for screening or diagnosis. Future work will use whole-slide and smartphone images with clinical reference standards, include other Plasmodium species and evaluate patient-level decision rules prospectively.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Dr. Sourav Kundu	✓	✓	✓		✓				✓		✓			
Tashmatova Nazgul	✓	✓		✓		✓		✓		✓		✓		
Turgunova Muattarkhon			✓				✓				✓		✓	✓

C : **C**onceptualization	I : **I**nvestigation	Vi : **Vi**sualization
M : **M**ethodology	R : **R**esources	Su : **Su**pervision
So : **So**ftware	D : **D**ata Curation	P : **P**roject administration
Va : **Va**lidation	O : Writing - **O**riginal Draft	Fu : **Fu**nding acquisition
Fo : **Fo**rmal analysis	E : Writing - Review & **E**diting	

Acknowledgments

The authors acknowledge the Lister Hill National Center for Biomedical Communications of the U.S. National Library of Medicine and Chittagong Medical College Hospital for creating and releasing the malaria cell image dataset, and the maintainers of the public mirror used in this study, and the researchers whose works have been cited in this study.

Funding Information

This research has received no external funding.

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable. The study used only publicly available, anonymised images.

Ethical Approval

Not Applicable. The study is a secondary analysis of publicly available, anonymised images and involved no human participants.

Data Availability

This study uses the publicly available NIH malaria cell images dataset, obtained from a public mirror. No new datasets were created. The analysis code and patient-level fold assignments are available from the corresponding author on reasonable request.

Use of Generative AI

Generative AI tools were not used to generate the scientific content, data, analyses or results of this study. Any AI-assisted language editing was reviewed and revised by the authors, who take full responsibility for the content of this article.

REFERENCES

[1] World Health Organization, World Malaria Report 2024. Geneva, Switzerland: WHO, 2024. www.who.int/publications/i/item/9789240104440

[2] C. Wongsrichanalai, M. J. Barcus, S. Muth, A. Sutamihardja, and W. H. Wernsdorfer, "A review of malaria diagnostic tools: Microscopy and rapid diagnostic test (RDT)," American Journal of Tropical Medicine and Hygiene, vol. 77, no. 6 Suppl., pp. 119–127, 2007. doi.org/10.4269/ajtmh.2007.77.119

[3] World Health Organization, Malaria Microscopy Quality Assurance Manual, Version 2. Geneva, Switzerland: WHO, 2016. www.who.int/publications/i/item/9789241549394

[4] M. Poostchi, K. Silamut, R. J. Maude, S. Jaeger, and G. Thoma, "Image analysis and machine learning for detecting malaria," Translational Research, vol. 194, pp. 36–55, 2018. doi.org/10.1016/j.trsl.2017.12.004

[5] F. Yang et al., "Deep learning for smartphone-based malaria parasite detection in thick blood smears," IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 5, pp. 1427–1438, 2020. doi.org/10.1109/JBHI.2019.2939121

[6] S. Rajaraman et al., "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images," PeerJ, vol. 6, Art. no. e4568, 2018. doi.org/10.7717/peerj.4568

[7] S. Rajaraman, S. Jaeger, and S. K. Antani, "Performance evaluation of deep neural ensembles toward malaria parasite detection in thin-blood smear images," PeerJ, vol. 7, Art. no. e6977, 2019. doi.org/10.7717/peerj.6977

[8] K. M. F. Fuhad, J. F. Tuba, M. R. A. Sarker, S. Momen, N. Mohammed, and T. Rahman, "Deep learning based automatic malaria parasite detection from blood smear and its smartphone based application," Diagnostics, vol. 10, no. 5, Art. no. 329, 2020. doi.org/10.3390/diagnostics10050329

[9] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," GigaScience, vol. 6, no. 5, pp. 1–9, 2017. doi.org/10.1093/gigascience/gix019

[10] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, Art. no. 100804, 2023. doi.org/10.1016/j.patter.2023.100804

[11] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nature Machine Intelligence, vol. 3, no. 3, pp. 199–217, 2021. doi.org/10.1038/s42256-021-00307-0

[12] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in Proc. 8th Int. Conf. Quality of Multimedia Experience (QoMEX), 2016, pp. 1–6. doi.org/10.1109/QoMEX.2016.7498955

[13] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in Proc. Int. Conf. Learning Representations (ICLR), 2019. doi.org/10.48550/arXiv.1903.12261

[14] D. G. Altman and J. M. Bland, "Diagnostic tests 2: Predictive values," BMJ, vol. 309, no. 6947, p. 102, 1994. doi.org/10.1136/bmj.309.6947.102

[15] M. M. G. Leeflang, A. W. S. Rutjes, J. B. Reitsma, L. Hooft, and P. M. M. Bossuyt, "Variation of a test's sensitivity and specificity with disease prevalence," Canadian Medical Association Journal, vol. 185, no. 11, pp. E537–E544, 2013. doi.org/10.1503/cmaj.121286

[16] TheODDYSEY, "Malaria-Detection-Model" (mirror of the NIH malaria cell images dataset), GitHub repository. Accessed: Oct. 2026. github.com/TheODDYSEY/Malaria-Detection-Model

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. Int. Conf. Learning Representations (ICLR), 2015. doi.org/10.48550/arXiv.1412.6980

[18] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in Advances in Neural Information Processing Systems (NeurIPS), vol. 32, 2019, pp. 8024–8035. doi.org/10.48550/arXiv.1912.01703

[19] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. jmlr.org/papers/v12/pedregosa11a.html

[20] P. M. Bossuyt et al., "STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies," BMJ, vol. 351, Art. no. h5527, 2015. doi.org/10.1136/bmj.h5527

[21] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Machine Learning (ICML), PMLR vol. 70, 2017, pp. 1321–1330. doi.org/10.48550/arXiv.1706.04599

[22] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," BMC Medicine, vol. 17, Art. no. 195, 2019. doi.org/10.1186/s12916-019-1426-2

How to Cite: Dr. Sourav Kundu, Tashmatova Nazgul, Turgunova Muattarkhon. (2026). From cells to patients: Validation, image quality and predictive values of deep learning for malaria detection in thin blood smears. Journal of Prevention, Diagnosis and Management of Human Diseases (JPDMHD), 6(2), 63-72. https://doi.org/10.55529/jpdmhd.62.63.72

BIOGRAPHIES OF AUTHORS

	**Dr. Sourav Kundu**, is an Assistant Professor in the Department of Tropical Medicine at the School of Tropical Medicine (STM), Kolkata, since 2024. He completed his MBBS from Midnapore Medical College in 2010 and MD in Tropical Medicine from STM, Kolkata in 2020 under WBUHS. He also holds a Diploma in Diabetes from University of Boston, London (2014). He has over 15 years of experience in critical care and tropical medicine. He previously served as RMO-cum-Clinical Tutor at Burdwan Medical College (2012–2017) and at STM (2020–2024). His clinical expertise includes critical care, infectious diseases, diabetes, and respiratory tract infections. Email: drsouravkundu@gmail.com
	**Tashmatova Nazgul**, is an associate professor in the Department of Morphological Disciplines at Jalal-Abad International University and also holds affiliations with Osh State University and Osh International Medical University, Kyrgyzstan. Tashmatova Nazgul co-authored this evaluation of deep learning for malaria detection in thin blood smears, which carries the analysis from single cells to the diagnosis of individual patients.
	**Turgunova Muattarkhon**, is employed as a lecturer by the Department of Morphological Disciplines, Jalal-Abad International University, Manas. Turgunova Muattarkhon participated in this work, which tested how the validation scheme, simulated image degradation and the prevalence of infection change the apparent performance of an automated malaria classifier applied to stained red blood cells.