Medical Devices: Evidence and Research

ISSN: 1179-1470 (Online) Journal homepage: www.tandfonline.com/journals/dmde20

Validation of an AI-Driven Sputum Analysis Platform for Accurate and Reliable Tuberculosis Detection in South-Western Uganda

Habert Tumwesigye, Daniel Atwine, Raymond Atwine, Abraham Birungi, Caroline Birungi, Esther Munguciada, Phionah Nabimanya, Joseph Akandwanaho, Benard Ochama & Yekosani Mitala

To cite this article: Habert Tumwesigye, Daniel Atwine, Raymond Atwine, Abraham Birungi, Caroline Birungi, Esther Munguciada, Phionah Nabimanya, Joseph Akandwanaho, Benard Ochama & Yekosani Mitala (2026) Validation of an AI-Driven Sputum Analysis Platform for Accurate and Reliable Tuberculosis Detection in South-Western Uganda, Medical Devices: Evidence and Research, , 639007, DOI: 10.2147/MDER.S639007

To link to this article: https://doi.org/10.2147/MDER.S639007

© 2026 Tumwesigye et al.

Published online: 15 Oct 2026.

Submit your article to this journal

Article views: 125

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=dmde20

Medical Devices: Evidence and Research

Dovepress
Taylor & Francis Group

Open Access Full Text Article

ORIGINAL RESEARCH

Validation of an AI-Driven Sputum Analysis Platform for Accurate and Reliable Tuberculosis Detection in South-Western Uganda

Habert Tumwesigye[1], Daniel Atwine[2], Raymond Atwine[1], Abraham Birungi[1], Caroline Birungi[2], Esther Munguciada[2], Phionah Nabimanya[1], Joseph Akandwanaho[1], Benard Ochama[3], Yekosani Mitala[1]

[1]Department of Pathology, Mbarara University of Science and Technology (MUST), Mbarara, Uganda; [2]SRF Research and Training Centers, Mbarara, Uganda; [3]Department of Laboratory Services, Mbarara Regional Referral Hospital (MRRH), Mbarara, Uganda

Correspondence: Habert Tumwesigye, Department of Pathology, Mbarara University of Science and Technology, P.O. Box 1410, Mbarara, Uganda, Email htumwesigye@must.ac.ug

Background: Ziehl–Neelsen (ZN) sputum smear microscopy remains an accessible method for detecting acid-fast bacilli (AFB), but its interpretation is dependent on specimen quality and trained microscopy personnel. Artificial intelligence (AI) may provide a microscopy reading aid that supports consistent identification of AFB-positive and AFB-negative smear images.

Objective: To evaluate transfer-learning deep learning models for classification of AFB-positive and AFB-negative digitized ZN-stained sputum smear images generated in a Ugandan regional referral laboratory.

Methods: A retrospective laboratory-based cross-sectional study used 512 archived digitized ZN-stained sputum smear images from Mbarara Regional Referral Hospital. The dataset contained 256 AFB-positive and 256 AFB-negative images and was partitioned by stratified sampling into training (n=358), validation (n=77), and an internal hold-out test set (n=77). ResNet-50 V2, DenseNet-121, and Swin Transformer-Tiny models were trained using ImageNet-pretrained weights in PyTorch. Performance was assessed using sensitivity, specificity, accuracy, positive predictive value (PPV), negative predictive value (NPV), F1-score, and ROC-AUC.

Results: On the internal test set, Swin Transformer-Tiny achieved 97.40% accuracy, 97.37% sensitivity, 97.44% specificity, 97.37% PPV, 97.44% NPV, 97.37% F1-score, and ROC-AUC 0.9939. DenseNet-121 achieved 88.31% accuracy, 84.21% sensitivity, 92.31% specificity, and ROC-AUC 0.9730. ResNet-50 V2 achieved 85.71% accuracy, 81.58% sensitivity, 89.74% specificity, and ROC-AUC 0.9555. Wilson 95% confidence intervals were calculated for the threshold-based metrics.

Conclusion: The models demonstrated promising internal performance for classification of microscopy-defined AFB status. Swin Transformer-Tiny had the highest observed performance, but the small, balanced, single-centre internal test set and use of routine microscopy labels rather than an independent microbiological reference standard limit clinical interpretation. The platform should therefore be considered an AI-assisted microscopy reading aid rather than an autonomous tuberculosis diagnostic system. Prospective, multicentre evaluation using patient/specimen-level separation, independent microbiological reference standards, and multiple blinded readers is required before clinical deployment.

Keywords: tuberculosis, acid-fast bacilli, Ziehl–Neelsen microscopy, artificial intelligence, deep learning, swin transformer, DenseNet-121, ResNet-50 V2, sputum microscopy, Uganda

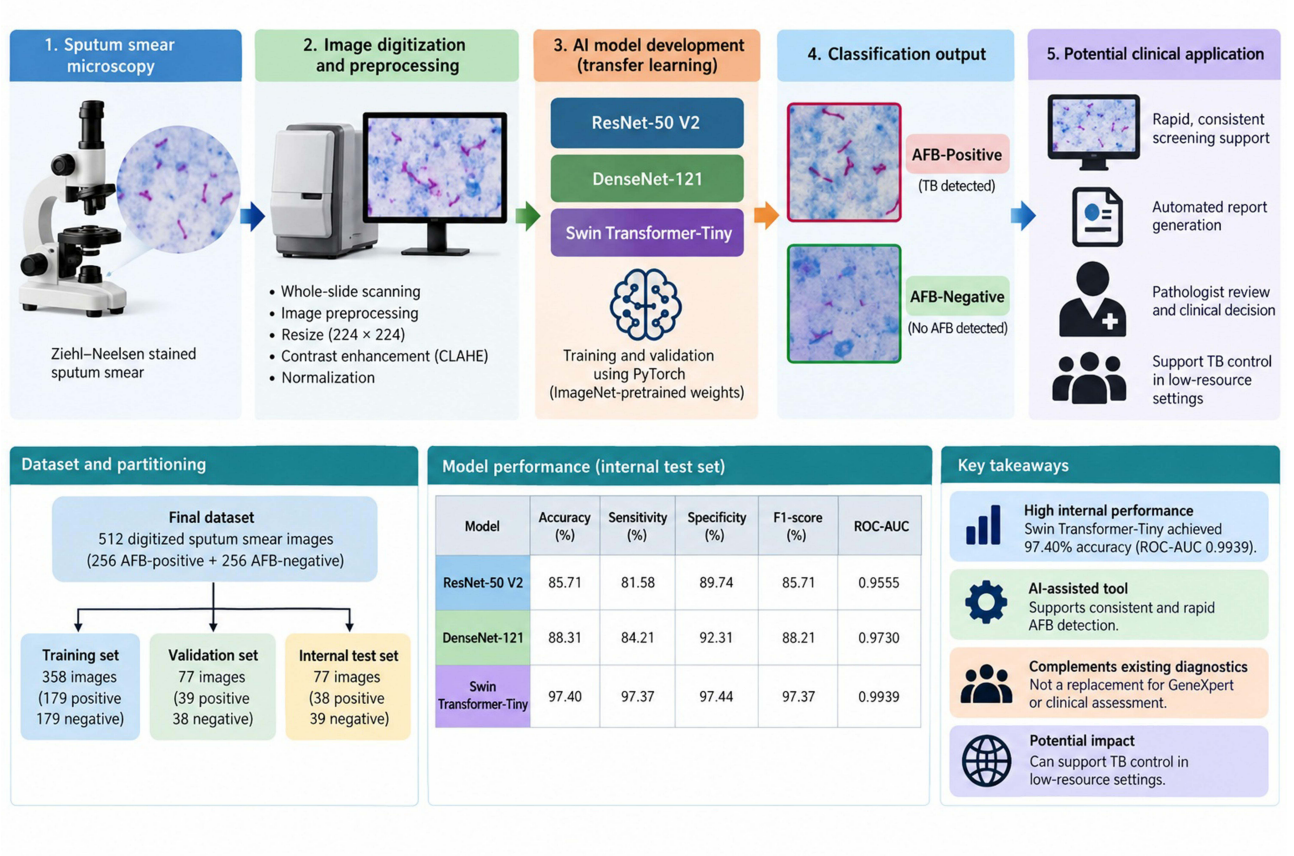
Introduction

Tuberculosis (TB) remains a major global infectious disease burden. In Uganda and other low- and middle-income settings, accessible laboratory diagnosis remains important for case detection and treatment initiation. Ziehl–Neelsen (ZN) sputum smear microscopy is inexpensive and requires comparatively limited infrastructure, but its sensitivity is reduced in paucibacillary disease and its performance depends on specimen quality, staining, microscopy conditions, and reader expertise.[1,2]

https://doi.org/10.2147/MDER.S639007
Received: 8 July 2026
Accepted: 23 September 2026
Published: 15 October 2026

© 2026 Tumwesigye et al. This work is published and licensed by Dove Medical Press Limited. The full terms of this license are available at https://www.dovepress.com/terms.php and incorporate the Creative Commons Attribution – Non Commercial (unported, v4.0) License (http://creativecommons.org/licenses/by-nc/4.0/). By accessing the work you hereby accept the Terms. Non-commercial uses of the work are permitted without any further permission from Dove Medical Press Limited, provided the work is properly attributed. For permission for commercial use of this work, please see paragraphs 4.2 and 5 of our Terms (https://www.dovepress.com/terms.php).

Graphical Abstract



Molecular tests such as Xpert MTB/RIF and related nucleic-acid amplification platforms provide more specific microbiological information and can detect TB in situations in which smear microscopy is negative. However, access, equipment requirements, maintenance, supply chains, and laboratory infrastructure can constrain their availability in some settings. The role of AI-assisted microscopy should therefore be understood as complementary to, rather than a replacement for, molecular testing and clinical assessment.

Deep learning has increasingly been applied to medical image analysis. Convolutional neural networks (CNNs), including ResNet and DenseNet, learn hierarchical image features directly from image data, while transformer-based architectures use attention mechanisms to model spatial relationships.[3–6] AI-assisted microscopy has also been investigated for AFB detection in ZN-stained sputum smears, with encouraging results in several settings.[7–16]

An important distinction is that ZN microscopy identifies acid-fast bacilli rather than definitively establishing Mycobacterium tuberculosis infection. Accordingly, an AI model trained against ZN microscopy labels principally learns to classify microscopy-defined AFB status. Without an independent microbiological reference standard such as Xpert MTB/RIF or culture, high agreement with microscopy should not be interpreted as definitive TB diagnostic accuracy.

The relevance of local evaluation is also important. Differences in staining procedures, microscopes, image acquisition, laboratory workflows, and patient populations can affect AI performance.[12,13,17–19] External validation across laboratories and patient populations is therefore necessary before clinical deployment.

This study evaluated three transfer-learning architectures ResNet-50 V2, DenseNet-121, and Swin Transformer-Tiny for classification of AFB-positive and AFB-negative digitized ZN-stained sputum smear images from Mbarara Regional Referral Hospital (MRRH), South-Western Uganda. The primary purpose was to assess the feasibility and internal

performance of AI-assisted microscopy classification in a locally generated dataset, rather than to establish definitive microbiological diagnosis of TB.

Materials and Methods

Study Design and Reporting

This retrospective laboratory-based cross-sectional study evaluated AI classification of AFB-positive and AFB-negative digitized ZN-stained sputum smear images. The study was reported with reference to the Standards for Reporting Diagnostic Accuracy Studies (STARD) 2015 statement, and the completed STARD 2015 checklist is provided in Supplementary File 1 and The overall workflow of image acquisition, preprocessing, model development, and evaluation is illustrated in Figure 1.

Study Setting

The study used archived sputum smear images obtained from the Tuberculosis Laboratory of Mbarara Regional Referral Hospital in South-Western Uganda. MRRH is a regional referral and teaching hospital serving South-Western Uganda and neighbouring regions. Routine TB laboratory services include ZN sputum smear microscopy.

Study Population and Sampling Unit

The study population comprised digitized archived ZN-stained sputum smear images generated during routine laboratory services. The analytical unit was the digitized smear image. The available manuscript records did not document a complete patient-level and specimen-level linkage sufficient to independently verify that every image represented a unique patient, specimen, and slide across all partitions. Consequently, patient/specimen-level independence cannot be assumed from the present retrospective dataset and is treated as a limitation. Future studies should enforce patient- and specimen-level separation before model development.

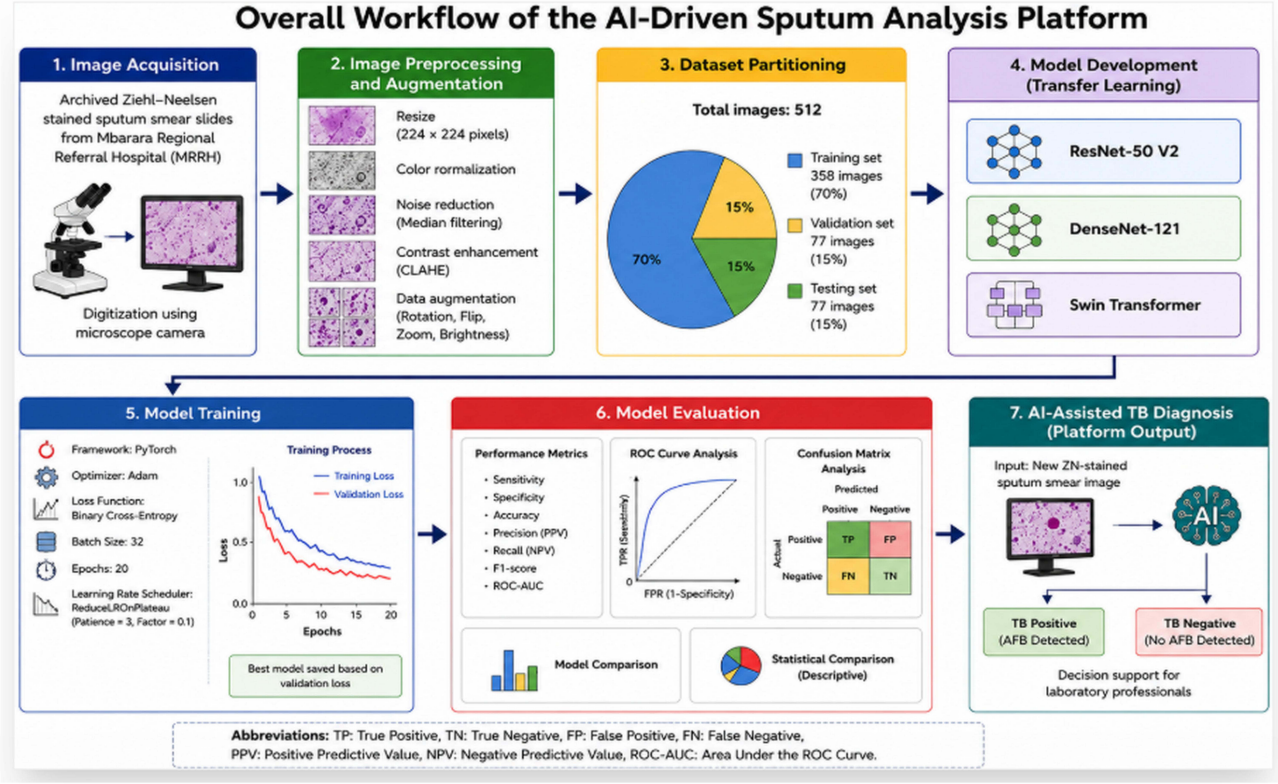


Figure 1 Overall workflow of the AI-assisted sputum smear microscopy platform.

Eligibility and Image Quality

Eligible images had acceptable ZN staining, adequate focus and illumination, and documented AFB-positive or AFB-negative microscopy status. Images with poor staining, inadequate illumination, excessive artefacts, poor focus, duplication, or incomplete documentation were excluded. Because images with poor quality were excluded before construction of the balanced dataset, the final dataset may under-represent difficult low-bacillary-load or otherwise challenging smears. This may lead to performance estimates that are more favourable than those achievable under routine conditions.

Reference Labels

AFB-positive and AFB-negative labels were based on routine ZN microscopy classification recorded in the archived laboratory data. ZN microscopy was not treated as an independent gold standard for definitive TB diagnosis. The present outcome is therefore classification of microscopy-defined AFB status. The available retrospective dataset did not provide a complete independent microbiological reference standard for all 512 images. Accordingly, the study does not establish sensitivity or specificity for culture- or molecularly confirmed M. tuberculosis infection.

A second independent blinded microscopist reading of the complete test set was not available in the retrospective dataset. Therefore, the study does not directly measure inter-observer agreement or demonstrate that the AI improves consistency relative to two independent human readers. Future prospective work should incorporate multiple blinded microscopists and predefined adjudication procedures.

Dataset Construction and Partitioning

A total of 512 digitized sputum smear images were included: 256 AFB-positive and 256 AFB-negative. The dataset was intentionally balanced to facilitate model development and comparative evaluation. Consequently, the 49.4–50.6% AFB-positive proportion in the test set does not represent the prevalence expected in routine clinical populations, and PPV and NPV should not be generalized directly to populations with different prevalence.

The final sample of 512 images reflected the availability of eligible archived digitized images meeting the predefined image-quality and documentation criteria; it was not prospectively powered using a prespecified confidence-interval precision calculation. The 77-image test set was selected as an internal hold-out subset from the same single-centre dataset.

Stratified random sampling produced 358 training images (179 positive, 179 negative), 77 validation images (39 positive, 38 negative), and 77 test images (38 positive, 39 negative). The complete specimen/image selection, exclusion, dataset construction, and allocation process is presented in Supplementary File 1. The test set was not used for training or model selection, but because it originated from the same single-centre dataset, it constitutes internal testing rather than external validation.

WHO Smear Grading

The present dataset and analysis were limited to binary AFB-positive/AFB-negative classification. Performance could not be stratified by WHO smear-grade categories (scanty, 1+, 2+, and 3+) because these categories were not incorporated into the available analysis dataset. This is a clinically relevant limitation, particularly because low-bacillary-load smears may be more difficult for both human readers and AI systems. Future prospective data collection should record smear grade and evaluate performance separately across grading categories.

Image Preprocessing and Augmentation

Images were resized to 224×224 pixels, converted to RGB tensors, and normalized using ImageNet preprocessing statistics to match the pretrained models. Online augmentation was applied only during training and included random horizontal flipping, rotation, zooming, brightness adjustment, and contrast adjustment. No augmentation was applied to validation or test images. The original training record did not preserve all augmentation parameter ranges and probabilities; these should be recorded prospectively in future experiments to improve reproducibility.

Deep Learning Models

Three transfer-learning architectures were evaluated: ResNet-50 V2, DenseNet-121, and Swin Transformer-Tiny. Each model used ImageNet-pretrained weights and was adapted for binary classification of AFB-positive versus AFB-negative images. Approximate parameter counts were 25.6 million for ResNet-50 V2, 8.0 million for DenseNet-121, and 28.3 million for Swin Transformer-Tiny. The model architectures and approximate parameter counts are presented in Table 1.

Model Training

Models were implemented in PyTorch 2.2 and trained using GPU acceleration in Google Colaboratory. The Adam optimizer was used with an initial learning rate of 0.0001, batch size of 32, Binary Cross-Entropy loss, and a maximum of 20 epochs. ReduceLROnPlateau was used to adjust the learning rate according to validation performance. The model with the lowest validation loss was retained for evaluation on the internal test set. The original analysis record did not retain a complete reproducibility log for random seed, exact scheduler parameters, or all layer-freezing decisions; these items should be prespecified and reported in future prospective model-development work. The model architectures and training parameters are summarized in Table 2

Intended Clinical Role and Computing Environment

The intended role of the system is as an AI-assisted microscopy reading aid that may support laboratory personnel in identifying AFB-positive fields on ZN-stained sputum smears. It is not intended to replace qualified laboratory personnel, molecular testing, clinical assessment, or treatment decisions.

Google Colaboratory GPU resources were used for model training and computational development, not as a requirement for routine clinical inference. The current study did not evaluate inference on an edge device, offline computer, or laboratory workstation and did not benchmark memory, energy consumption, or real-world turnaround time.

Table 1 Model Complexity

Model	Architecture	Approximate Parameters (Millions)
ResNet-50 V2	CNN	25.6
DenseNet-121	CNN	8.0
Swin Transformer-Tiny	Vision Transformer	28.3

Table 2 Model Architecture and Training Parameters

Parameter	ResNet-50 V2	DenseNet-121	Swin Transformer-Tiny
Architecture	CNN	CNN	Vision Transformer
Pre-trained weights	ImageNet	ImageNet	ImageNet
Input image size	224 × 224	224 × 224	224 × 224
Batch size	32	32	32
Initial learning rate	0.0001	0.0001	0.0001
Optimizer	Adam	Adam	Adam
Loss	Binary Cross-Entropy	Binary Cross-Entropy	Binary Cross-Entropy
Maximum epochs	20	20	20
Approximate parameters (million)	25.6	8.0	28.3

Claims concerning deployment in resource-limited settings are therefore prospective. Future work should benchmark the three models on affordable local or edge-computing hardware and assess offline operation where connectivity is limited.

Statistical Analysis

Performance was summarized using sensitivity, specificity, accuracy, PPV, NPV, precision, recall, F1-score, and ROC-AUC. Confusion matrices were used to summarize true-positive, true-negative, false-positive, and false-negative classifications. Wilson 95% confidence intervals were calculated for sensitivity, specificity, accuracy, PPV, and NPV because of the small test sample. F1-score confidence intervals were estimated by non-parametric bootstrap resampling of the observed test classifications. The available analysis record did not retain the original continuous prediction scores required to independently calculate a valid ROC-AUC confidence interval or conduct paired DeLong/AUC comparisons. Therefore, differences between models are described as observed point-estimate differences and are not presented as statistically significant superiority.

Ethical Considerations

Ethical approval for this study was obtained from the Mbarara University of Science and Technology Research Ethics Committee (MUST-REC), Approval No. MUST-2025-618. The study was conducted in accordance with the ethical principles of the Declaration of Helsinki. Administrative permission to access archived Ziehl–Neelsen (ZN)-stained sputum smear images was granted by Mbarara Regional Referral Hospital. Only anonymized archived laboratory images were used, and no personally identifiable patient information was collected or processed. The requirement for individual informed consent was waived by MUST-REC for the retrospective analysis of de-identified archived images.

Study Registration

This retrospective study was not prospectively registered in a public clinical or diagnostic accuracy study registry before commencement.

Results

Dataset Characteristics

The final dataset comprised 512 digitized ZN-stained sputum smear images, including 256 AFB-positive and 256 AFB-negative images. Stratified random sampling produced 358 training images (179 positive, 179 negative), 77 validation images (39 positive, 38 negative), and 77 test images (38 positive, 39 negative). The distribution of images across the training, validation, and internal hold-out test sets is summarized in Table 3. The complete specimen/image selection, exclusion, dataset construction, and allocation process is presented in Supplementary Figure 1.

Diagnostic Performance

On the 77-image internal hold-out test set, Swin Transformer-Tiny achieved the highest observed performance across the reported threshold-based measures. It correctly classified 75 of 77 images, with one false-positive and one false-negative result. DenseNet-121 correctly classified 68 images, while ResNet-50 V2 correctly classified 66 images. The comparative

Table 3 Dataset Characteristics Used for Model Development and Evaluation

Subset	AFB-Positive	AFB-Negative	Total
Training	179	179	358
Validation	39	38	77
Internal hold-out test	38	39	77
Total	256	256	512

Table 4 Diagnostic Performance on the Internal Hold-Out Test Set

Metric	ResNet-50 V2	DenseNet-121	Swin Transformer-Tiny
Sensitivity (%)	81.58 (66.58–90.78)	84.21 (69.58–92.56)	97.37 (86.51–99.53)
Specificity (%)	89.74 (76.42–95.94)	92.31 (79.68–97.35)	97.44 (86.82–99.55)
Accuracy (%)	85.71 (76.20–91.83)	88.31 (79.25–93.73)	97.40 (91.02–99.28)
PPV (%)	88.57 (74.05–95.46)	91.43 (77.62–97.04)	97.37 (86.51–99.53)
NPV (%)	83.33 (69.40–91.68)	85.71 (72.16–93.28)	97.44 (86.82–99.55)
F1-score (%)	84.93 (bootstrap CI: 75.00–92.86)	87.67 (bootstrap CI: 78.57–94.74)	97.37 (bootstrap CI: 93.02–100.00)
ROC-AUC	0.9555	0.9730	0.9939

Notes: Values are percentages with 95% Wilson confidence intervals, except F1-score, for which bootstrap confidence intervals are reported; ROC-AUC values are point estimates.

diagnostic performance of the three models is presented in Table 4 and the confusion matrix and receiver operating characteristic (ROC) curve for ResNet-50 V2 are presented in Figures 2 and 3, respectively.

Values in parentheses are 95% confidence intervals. Wilson intervals were used for sensitivity, specificity, accuracy, PPV and NPV; F1-score intervals were obtained by bootstrap resampling. ROC-AUC confidence intervals and formal paired comparisons between AUCs were not estimated because the original continuous model scores were not retained in the available analysis record.

ResNet-50 V2

ResNet-50 V2 achieved 85.71% accuracy, with sensitivity of 81.58%, specificity of 89.74%, PPV of 88.57%, NPV of 83.33%, F1-score of 84.93%, and ROC-AUC of 0.9555. The confusion matrix comprised 31 true-positive, 35 true-negative, 7 false-negative, and 4 false-positive classifications. The confusion matrix and receiver operating characteristic (ROC) curve for ResNet-50 V2 are presented in Figures 2 and 3, respectively.

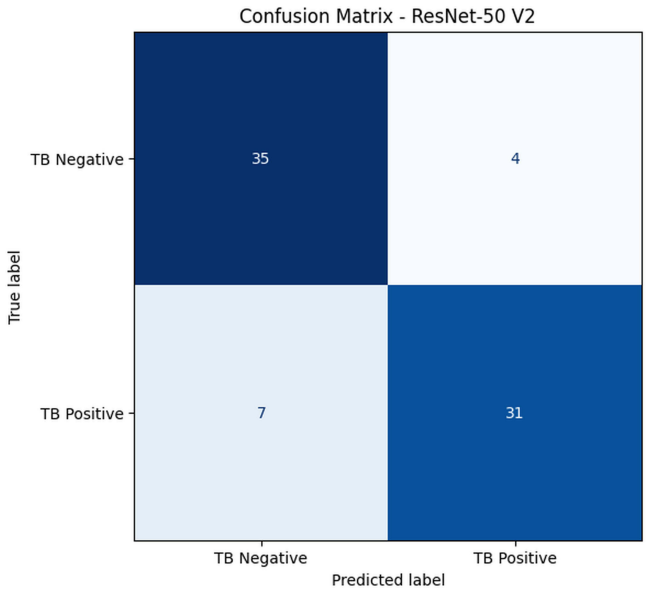


Figure 2 Confusion matrix of the ResNet-50 V2 model on the internal hold-out test set.

https://doi.org/10.2147/MDER.S639007

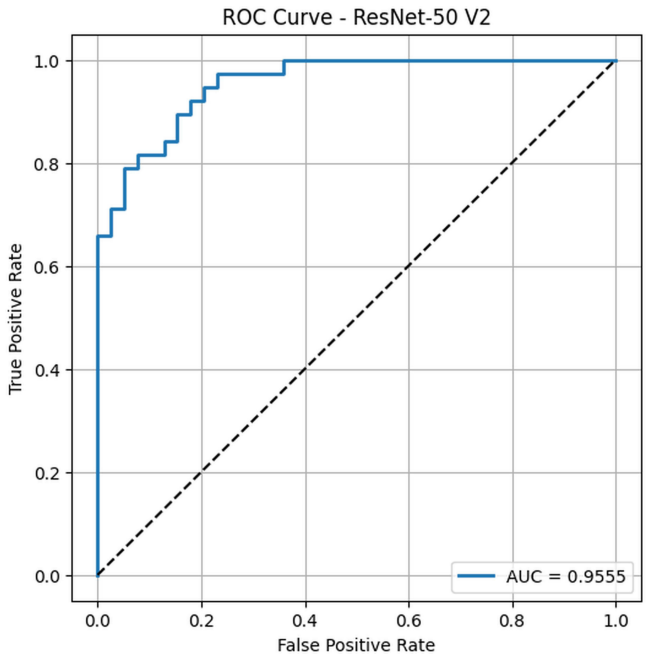


Figure 3 Receiver operating characteristic curve of the ResNet-50 V2 model (ROC-AUC = 0.9555).

DenseNet-121

DenseNet-121 achieved 88.31% accuracy, sensitivity of 84.21%, specificity of 92.31%, PPV of 91.43%, NPV of 85.71%, F1-score of 87.67%, and ROC-AUC of 0.9730. The model produced 32 true-positive, 36 true-negative, 6 false-negative, and 3 false-positive classifications. The confusion matrix and ROC curve for DenseNet-121 are presented in Figures 4 and 5, respectively.

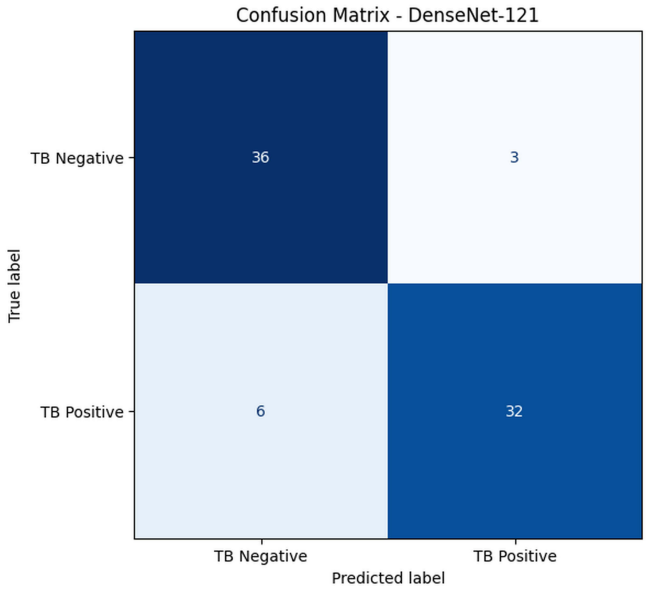


Figure 4 Confusion matrix of the DenseNet-121 model on the internal hold-out test set.

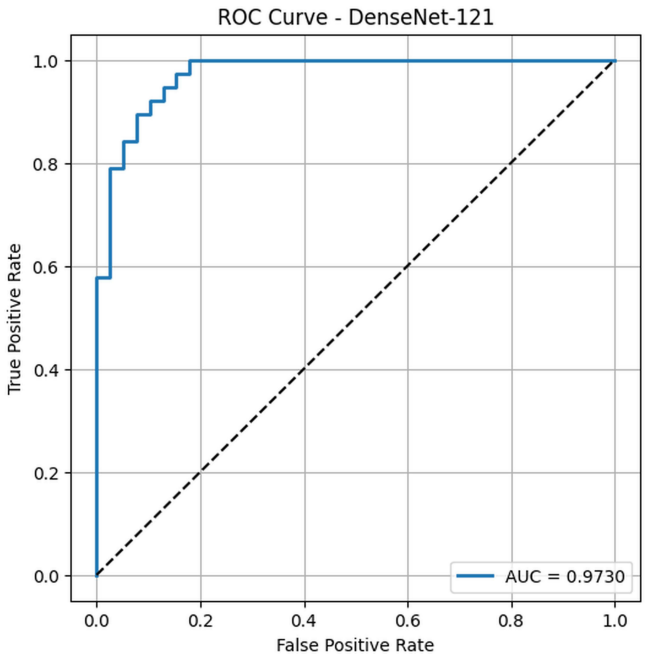


Figure 5 Receiver operating characteristic curve of the DenseNet-121 model (ROC-AUC = 0.9730).

Swin Transformer-Tiny

Swin Transformer-Tiny achieved 97.40% accuracy, sensitivity of 97.37%, specificity of 97.44%, PPV of 97.37%, NPV of 97.44%, F1-score of 97.37%, and ROC-AUC of 0.9939. The model produced 37 true-positive, 38 true-negative, one false-negative, and one false-positive classification. The confusion matrix and ROC curve for Swin Transformer-Tiny are presented in Figures 6 and 7, respectively.

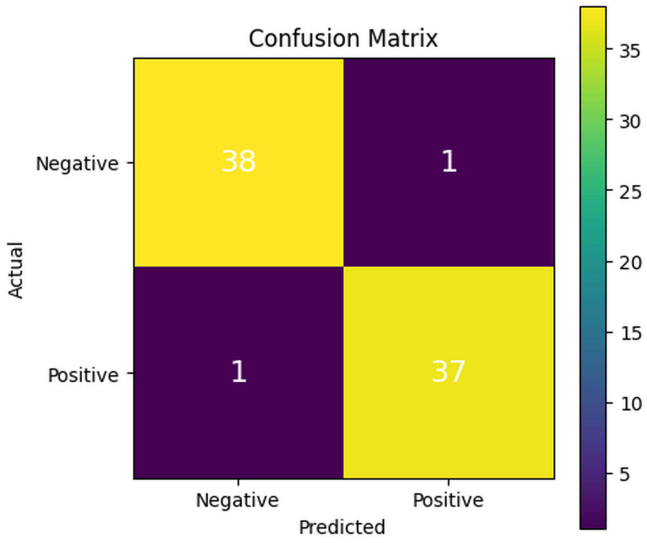


Figure 6 Confusion matrix of the Swin Transformer-Tiny model on the internal hold-out test set.

https://doi.org/10.2147/MDER.S639007

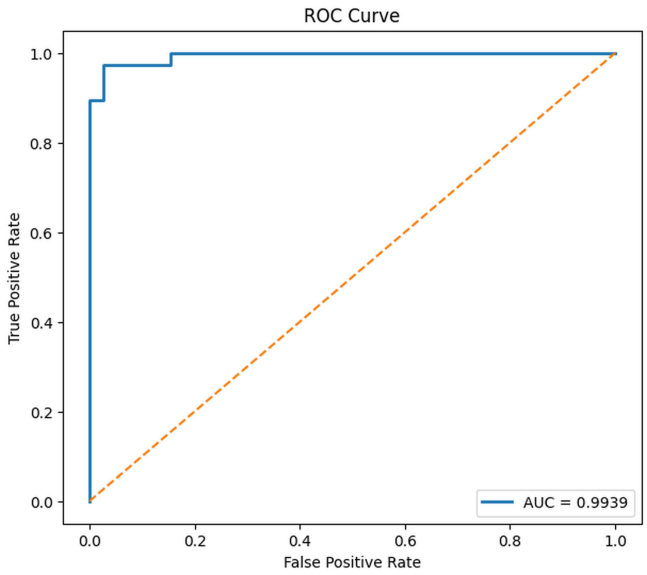


Figure 7 Receiver operating characteristic curve of the Swin Transformer-Tiny model (ROC-AUC = 0.9939).

Observed Model Comparison

Swin Transformer-Tiny had the highest point estimates for all reported performance measures, followed by DenseNet-121 and ResNet-50 V2. The classification errors observed for each model are summarized in Table 5. Because the test set contained only 77 images and formal paired statistical comparisons were not performed, the observed ranking should not be interpreted as evidence of statistically significant superiority. The results instead indicate that Swin Transformer-Tiny was the strongest-performing architecture in this internal evaluation. The comparative ROC curves and performance measures for the three evaluated models are shown in Figures 8 and 9, respectively.

Discussion

This study evaluated three transfer-learning models for classification of AFB-positive and AFB-negative digitized ZN-stained sputum smear images generated at a Ugandan regional referral laboratory. The principal finding was that Swin Transformer-Tiny achieved the highest observed performance on the 77-image internal hold-out test set, with 97.40% accuracy and ROC-AUC of 0.9939. DenseNet-121 and ResNet-50 V2 showed lower but still useful classification performance.

The results are broadly consistent with previous work demonstrating that AI can assist identification of AFB in digitized sputum smear microscopy.[7–13] However, the present study should be interpreted as an evaluation of AI-assisted microscopy classification rather than definitive TB diagnosis. ZN microscopy detects acid-fast organisms and does not independently establish M. tuberculosis infection. The absence of an independent microbiological reference standard for the complete dataset limits the extent to which the observed performance can be generalized to microbiologically confirmed TB.

Table 5 Classification Errors on the Internal Hold-Out Test Set

Model	False Positives	False Negatives	Total Errors
ResNet-50 V2	4	7	11
DenseNet-121	3	6	9
Swin Transformer-Tiny	1	1	2

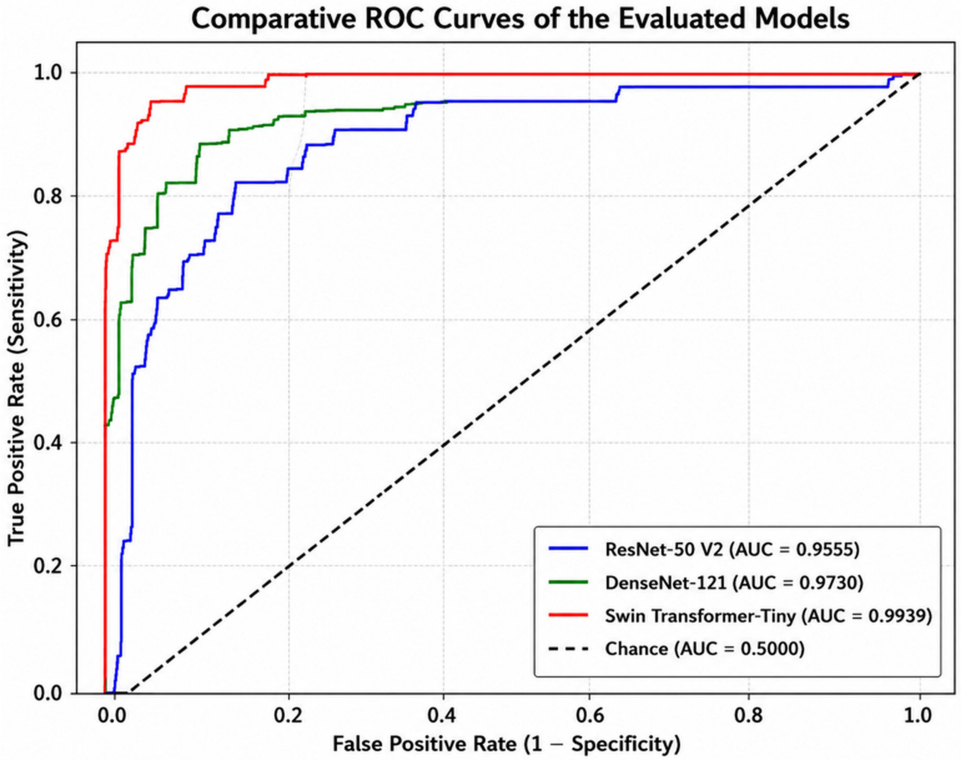


Figure 8 Comparative receiver operating characteristic curves for ResNet-50 V2, DenseNet-121, and Swin Transformer-Tiny.

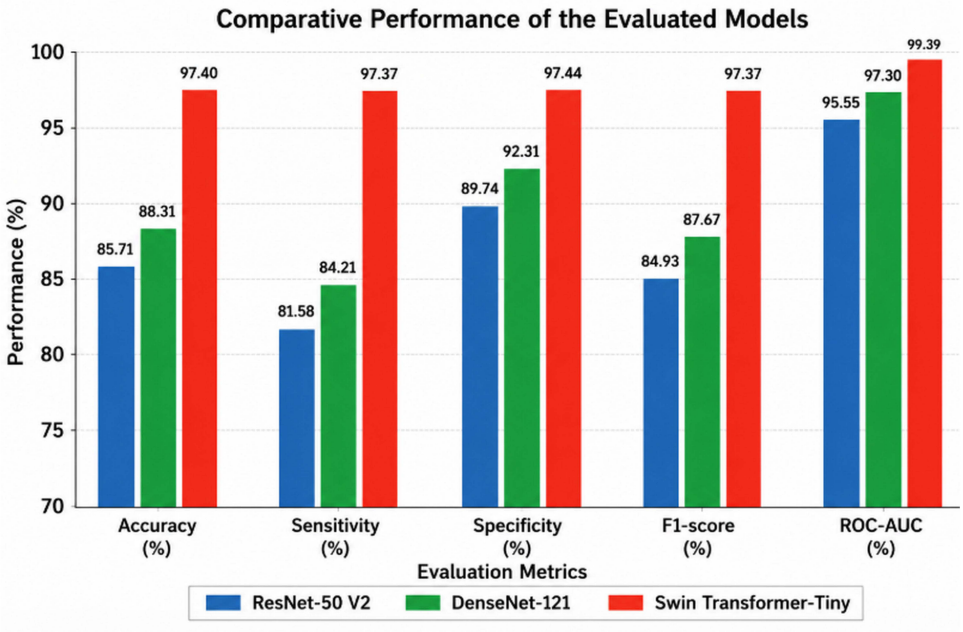


Figure 9 Comparative performance of the three evaluated models across accuracy, sensitivity, specificity, F1-score, and ROC-AUC.

The balanced dataset facilitated comparison among architectures but changes the interpretation of predictive values. In routine practice, the prevalence of AFB-positive smears will differ from the approximately 50% prevalence in this study sample. PPV and NPV are prevalence-dependent and therefore should not be transferred directly to clinical populations without evaluation in representative cohorts.

https://doi.org/10.2147/MDER.S639007

The exclusion of poorly focused, poorly illuminated, highly artefactual, or otherwise unacceptable images is another important consideration. Such quality-control exclusions are reasonable for technical model development, but they may remove some of the most difficult specimens encountered in routine microscopy, including paucibacillary or scanty smears. The resulting estimates may consequently be optimistic relative to real-world performance.

The absence of a second independent blinded microscopist also limits interpretation. The original rationale included potential reduction of observer variability, but this study did not directly compare AI performance with independent human readings of the same test images. Accordingly, it cannot establish that the AI improves inter-observer consistency. Future prospective evaluations should use multiple blinded microscopists, predefined adjudication, and reader-level analyses.

The 77-image test set provides useful internal evidence but remains small. The confidence intervals demonstrate that even apparently high point estimates have uncertainty. In addition, the test set was sampled from the same single-centre dataset and therefore does not represent external validation. Patient-, specimen-, laboratory-, microscope-, staining-, and image-acquisition-level differences may affect performance when the system is evaluated elsewhere.

Swin Transformer-Tiny had the strongest observed performance, which is compatible with the ability of hierarchical attention mechanisms to capture both local and broader spatial information.[6] Nevertheless, because formal paired comparisons were not performed, the results do not demonstrate statistically significant superiority. The finding should be treated as an architecture-selection result within this dataset.

From a resource-limited-setting perspective, the study demonstrates computational feasibility of model development but not clinical deployment. Training was performed with Google Colaboratory GPU resources, and the study did not evaluate offline inference, edge hardware, memory consumption, energy use, or operational turnaround time. Future work should determine whether a lightweight model can provide acceptable performance on affordable local hardware or an integrated digital microscope workstation.

The appropriate clinical position of the proposed platform is therefore as an AI-assisted microscopy reading aid. A future workflow could use the model to flag images requiring attention while retaining qualified human review and referral to molecular testing where clinically indicated. The current study did not evaluate downstream clinical decisions, treatment initiation, GeneXpert confirmation, or patient outcomes; these should not be inferred from the present results.

Strengths and Limitations

Strengths include the use of locally generated ZN-stained sputum smear images, comparison of CNN and transformer-based transfer-learning architectures using a common preprocessing and evaluation framework, and reporting of multiple threshold-based performance measures. The work also provides a locally relevant foundation for future prospective evaluation of AI-assisted microscopy.

Limitations include the small 77-image internal hold-out test set; intentional class balancing; retrospective image selection; exclusion of poor-quality images; absence of complete WHO smear-grade analysis; absence of a second blinded microscopist reading; lack of a complete independent microbiological reference standard; inability to perform formal AUC comparisons because continuous prediction scores were not retained; and the inability to definitively verify patient/specimen-level independence from the available retrospective records. The study was single-centre and was not prospectively registered. External multicentre validation and prospective evaluation under routine laboratory conditions remain necessary.

These limitations mean that the reported results should not be interpreted as evidence of autonomous TB diagnosis, routine clinical reliability, or generalizable performance across Ugandan laboratories.

Proposed Clinical Development Framework

Figure 10 should be interpreted only as a proposed future clinical-development framework. The stages from AI prediction to clinician review, molecular confirmation, and patient management were not evaluated in this study and therefore do not represent validated study outcomes.

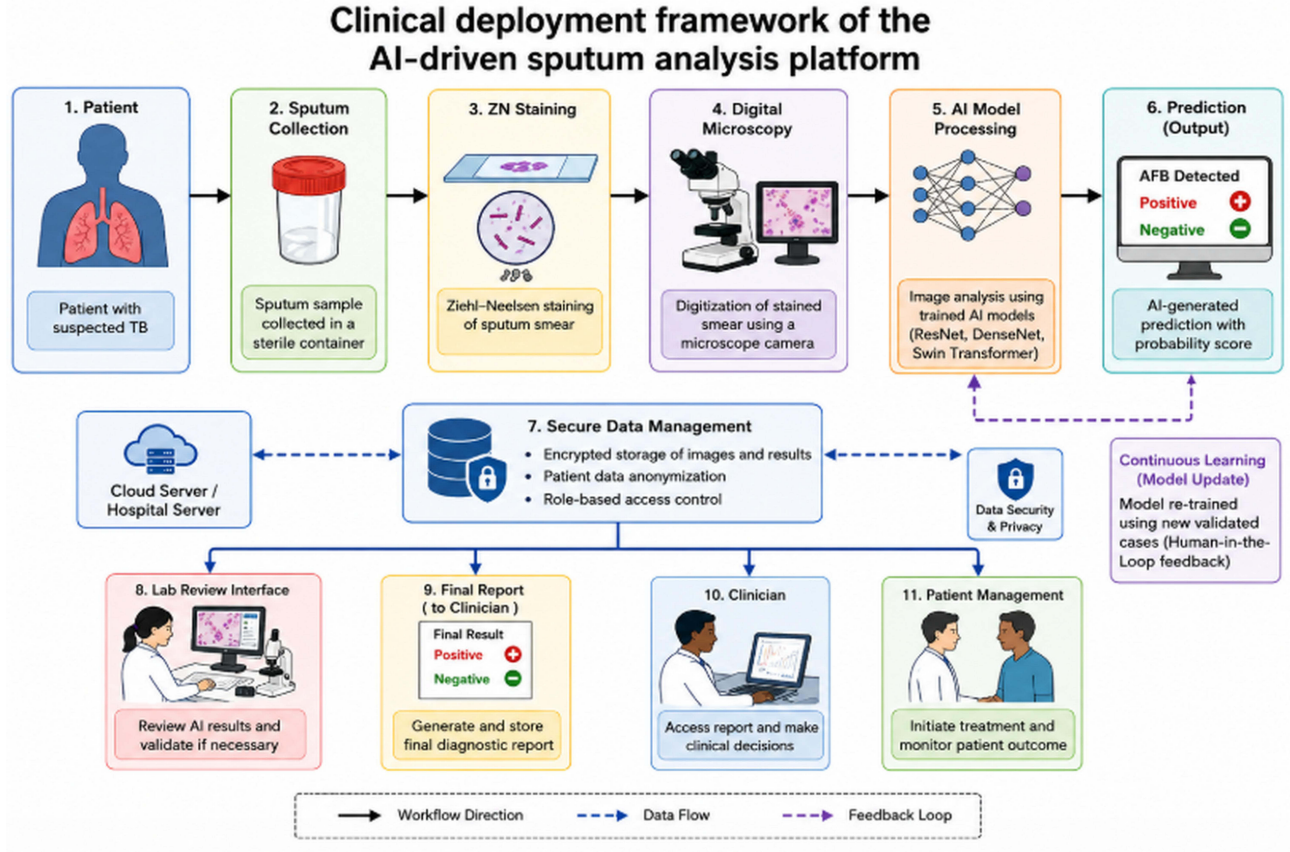


Figure 10 Proposed future clinical-development framework for AI-assisted sputum smear microscopy. The clinical workflow beyond image classification was not evaluated in the present study and is presented as a proposed framework rather than a validated outcome.

Conclusion

This study demonstrates that transfer-learning deep learning models can classify AFB-positive and AFB-negative digitized ZN-stained sputum smear images from a Ugandan regional referral laboratory. Swin Transformer-Tiny achieved the highest observed performance on the internal hold-out test set, with 97.40% accuracy, 97.37% sensitivity, 97.44% specificity, and ROC-AUC of 0.9939. However, the findings represent internal evaluation against routine microscopy-defined AFB status rather than definitive microbiological diagnosis of tuberculosis.

The platform should therefore be considered an AI-assisted microscopy reading aid that may support, but not replace, qualified laboratory personnel and established molecular or clinical diagnostic pathways. Larger prospective studies should use patient- and specimen-level separation, independent microbiological reference standards, multiple blinded microscopists, representative smear-grade distributions, diverse laboratories and imaging systems, and external multi-centre validation before clinical deployment.

Abbreviations

AFB, Acid-fast bacilli; AI, Artificial intelligence; AUC, Area under the curve; CNN, Convolutional neural network; F1-score, Harmonic mean of precision and recall; GPU, Graphics processing unit; ImageNet, Large-scale image database used for pretraining; MRRH, Mbarara Regional Referral Hospital; MUST, Mbarara University of Science and Technology; MUST-REC, Mbarara University of Science and Technology Research Ethics Committee; NPV, Negative predictive value; NTLP, National Tuberculosis and Leprosy Programme; PPV, Positive predictive value; ROC, Receiver operating characteristic; ROC-AUC, Area under the receiver operating characteristic curve; STARD, Standards for Reporting Diagnostic Accuracy Studies; TB, Tuberculosis; WHO, World Health Organization; XAI, Explainable artificial intelligence; ZN, Ziehl–Neelsen.

Data Sharing Statement

The dataset generated and analysed during the current study consists of de-identified archived sputum smear images. Subject to institutional policies, ethical approval requirements, and applicable data-protection regulations, the dataset may be made available by the corresponding author upon reasonable request. Model code and trained weights were not deposited in a public repository at the time of this study; future work will prioritize version-controlled code and model release where ethically and technically appropriate.

Ethics Approval and Consent to Participate

Ethical approval for this study was obtained from the Mbarara University of Science and Technology Research Ethics Committee (MUST-REC), Approval No. MUST-2025-618. Administrative permission to access archived ZN-stained sputum smear images was granted by Mbarara Regional Referral Hospital. The study used anonymized archived laboratory images and did not process personally identifiable patient information. MUST-REC waived the requirement for individual informed consent for this retrospective analysis.

Consent for Publication

Not applicable. The manuscript does not contain individually identifiable patient information or identifiable personal images requiring consent for publication.

Acknowledgments

The authors thank the management and staff of the Department of Laboratory Services, Mbarara Regional Referral Hospital, for access to archived sputum smear images and technical support. The authors also acknowledge the Department of Pathology, Mbarara University of Science and Technology, and SRF Research and Training Centers for academic and technical support.

Author Contributions

Habert Tumwesigye: Conceptualization, methodology, software, data curation, formal analysis, investigation, visualization, validation, writing, original draft, writing, review and editing, project administration.

Daniel Atwine: Conceptualization, methodology, supervision, writing, review and editing.

Raymond Atwine: Conceptualization, methodology, supervision, writing, review and editing

Abraham Birungi: Investigation, laboratory support, validation, writing, review and editing.

Caroline Birungi: Methodology, writing, review and editing.

Munguciada Esther: Methodology, writing, review and editing.

Nabimanya Phionah: Model development and training, validation, data interpretation, writing, review and editing.

Joseph Akandwanaho: Platform development, quality assurance, writing, review and editing.

Benard Ochama: Laboratory supervision, data acquisition, image labelling and validation.

Yekosani Mitala: Supervision, methodology, critical revision, writing, review and editing. All authors approved the final manuscript and accept responsibility for the work.

All authors made a significant contribution to the work reported, whether in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising, or critically reviewing the article; gave final approval of the version to be published; agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This work was supported by the Mbarara University of Science and Technology Faculty of Health Sciences Seed Grant. The funder had no role in study design, data collection, data analysis, interpretation of results, manuscript preparation, or the decision to submit the manuscript for publication.

Disclosure

The authors declare that they have no competing financial or non-financial interests related to this work.

This paper has been uploaded as a preprint to SOAR Research: https://www.soarresearch.org/project/on-going/7-validation-of-an-ai-driven-sputum-analysis-platform-for-accurate-and-reliable-laboratory-based-tuberculosis-detection-in-south-western-uganda-ai-sputum-tb-validation-study/validation-of-an-ai-driven-sputum-analysis-platform-for-accurate-and-reliable-laboratory-based-tuberculosis-detection-in-south-western-uganda-ai-sputum-tb-validation-study

References

1. World Health Organization. Global tuberculosis report 2024. Geneva: World Health Organization; 2024.
2. Smith KP, Kirby JE. Image analysis and artificial intelligence in infectious disease diagnostics. *Clin Microbiol Infect*. 2020;26(10):1318–1323. doi:10.1016/j.cmi.2020.03.020
3. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p.770–778. doi:10.1109/CVPR.2016.90.
4. Huang G, Liu Z, Van der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. p.4700–4708. doi:10.1109/CVPR.2017.243.
5. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale. International Conference on Learning Representations (ICLR); 2021.
6. Liu Z, Lin Y, Cao Y, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p.10012–10022. doi:10.1109/ICCV48922.2021.00986.
7. Gupta P, Khare V, Srivastava A, et al. A prospective observational multicentric clinical trial to evaluate microscopic examination of acid-fast bacilli in sputum by an artificial intelligence-based microscopy system. *J Investig Med*. 2023;71(7):716–721. doi:10.1177/10815589231171402
8. Chen WC, Chang CC, Lin YE. Pulmonary tuberculosis diagnosis using an intelligent microscopy scanner and image recognition model for improved acid-fast bacilli detection in smears. *Microorganisms*. 2024;12(8):1734. doi:10.3390/microorganisms12081734
9. Xiong Y, Ba X, Hou A, Zhang K, Chen L, Li Y. Automatic detection of Mycobacterium tuberculosis using artificial intelligence. *J Thorac Dis*. 2018;10(3):1936–1940. doi:10.21037/jtd.2018.03.182
10. Karlita T, Pastika FA, Ramadijanti N, et al. Mycobacterium tuberculosis detection on sputum smear microscopic images using attention-based multi-scale convolutional neural networks. In: 2024 International Electronics Symposium (IES). Piscataway (NJ): IEEE; 2024.
11. Shelomentseva IG, Chentsov SV. Classification of microscopy images stained by Ziehl–Neelsen method using different architectures of convolutional neural networks. In: Kryzhanovsky B, Dunin-Barkowski W, Redko V, Tiumentsev Y, editors. *Advances in Neural Computation, Machine Learning, and Cognitive Research IV*. Cham: Springer; 2021:423–434.
12. Zaizen Y, Kanahori Y, Ishijima S, et al. Deep learning-aided detection of mycobacteria in pathology specimens increases the sensitivity in early diagnosis of pulmonary tuberculosis compared with bacteriology tests. *Diagnostics*. 2022;12(3):709. doi:10.3390/diagnostics12030709
13. Witarto AB, Ceachi B, Popp C, et al. AI-based analysis of Ziehl–Neelsen-stained sputum smears for Mycobacterium tuberculosis as a screening method for active tuberculosis. *Life*. 2024;14(11):1418. doi:10.3390/life14111418
14. Bossuyt PM, Reitsma JB, Bruns DE, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. *BMJ*. 2015;351:h5527. doi:10.1136/bmj.h5527
15. Efficient and accurate tuberculosis diagnosis: attention residual U-Net and vision transformer-based detection framework. arXiv. 2025.
16. Kotei E, Thirunavukarasu R. Computational techniques for the automated detection of Mycobacterium tuberculosis from digitized sputum smear microscopic images: a systematic review. *Prog Biophys Mol Biol*. 2022;171:4–16. doi:10.1016/j.pbiomolbio.2022.04.005
17. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2009. p.248–255. doi:10.1109/CVPR.2009.5206848.
18. El-Melegy M, Mohamed D, ElMelegy T. Automatic detection of tuberculosis bacilli from microscopic sputum smear images using Faster R-CNN, transfer learning and augmentation. In: *Pattern Recognition and Image Analysis*. Cham: Springer; 2019:145–156.
19. Zurac S, Mogodici C, Poncu T, et al. A new artificial intelligence-based method for identifying Mycobacterium tuberculosis in Ziehl–Neelsen stain on tissue. *Diagnostics*. 2022;12(6):1484. doi:10.3390/diagnostics12061484

Medical Devices: Evidence and Research

Dovepress
Taylor & Francis Group

Publish your work in this journal

Medical Devices: Evidence and Research is an international, peer-reviewed, open access journal that focuses on the evidence, technology, research, and expert opinion supporting the use and application of medical devices in the diagnosis, monitoring, treatment and management of clinical conditions and physiological processes. The identification of novel devices and optimal use of existing devices which will lead to improved clinical outcomes and more effective patient management and safety is a key feature of the journal. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: https://www.dovepress.com/medical-devices-evidence-and-research-journal