Journal of Sustainability, Policy, and Practice

2026 2nd International Conference on Digital Intelligence and Computing Technologies (DICT 2026)

Article

Comparative Evaluation of Lightweight Machine Learning Classifiers for Rapid Infectious Disease Identification on Resource-Constrained Point-of-Care Devices: Accuracy, Latency, and Subgroup Performance

Yijie Wang [1,*]

[1] Epidemiology, University of Chicago, Chicago, IL, USA
* Correspondence: Yijie Wang, Epidemiology, University of Chicago, Chicago, IL, USA

Abstract: Portable point-of-care (POC) diagnostic devices hold considerable promise for improving the detection of infectious diseases in resource-limited healthcare settings, yet the performance of lightweight machine learning (ML) classifiers under realistic hardware constraints remains insufficiently characterized. This study presents a comparative evaluation of eight lightweight classifiers---five traditional ML algorithms (Random Forest, XGBoost, Support Vector Machine, k-Nearest Neighbors, Logistic Regression) and three compact deep learning architectures (quantized MobileNetV2, EfficientNet-B0, SqueezeNet) ---across two clinically relevant tasks: malaria parasite detection from thin blood smear images and tuberculosis identification from chest radiographs. Using the NIH Malaria Cell Images dataset (27,558 pre-segmented cell images from 200 patients) and the NLM tuberculosis chest X-ray datasets (800 radiographs), we assess classification accuracy, deployment-relevant latency, and age-stratified sensitivity patterns. EfficientNet-B0 achieved the highest accuracy on both tasks, while XGBoost provided the strongest baseline among classifiers operating on frozen ResNet-18 embeddings. On Raspberry Pi 4, XGBoost required 38 ms for the classification step alone, but its estimated end-to-end latency increased to approximately 143 ms once the shared ResNet-18 feature-extraction stage was included, compared with 210 ms for quantized MobileNetV2. Age-stratified analysis of the Shenzhen subset showed a consistent tendency toward lower sensitivity among patients over 60, with smaller observed declines for the lightweight deep learning models; however, these subgroup patterns should be interpreted with caution because the sample sizes were modest and no formal significance test was applied. Overall, the study provides a deployment-oriented comparison of lightweight diagnostic classifiers while highlighting the importance of fair latency accounting and cautious interpretation of subgroup differences.

Keywords: point-of-care diagnostics; lightweight classifiers; infectious disease detection; resource-constrained deployment

Received: 03 May 2026
Revised: 13 June 2026
Accepted: 25 June 2026
Published: 02 July 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

1.1. Background and Motivation

Artificial intelligence has the potential to transform healthcare delivery in resource-poor settings by enabling rapid, accurate diagnostic capabilities that do not depend on specialist expertise [1]. A systematic scoping review of AI applications in low- and middle-income countries identified clinical decision support and diagnostic assistance as the most promising deployment areas, yet noted that fewer than a dozen studies had

evaluated real-world implementations [2]. Broader reviews of clinical decision support systems have further highlighted the potential benefits and implementation challenges of deploying AI-driven tools in routine clinical workflows [3]. The emergence of portable POC diagnostic devices---ranging from smartphone-based lateral flow assay readers to handheld microscopy platforms---creates new opportunities to bring laboratory-grade analysis to community health centers, rural clinics, and mobile screening programs. Recent advances in ML-assisted POC testing have demonstrated that deep learning can interpret rapid diagnostic tests with accuracy comparable to or exceeding that of trained laboratory personnel [4].

The deployment of AI-assisted diagnostics on portable hardware introduces a fundamental tension between classification performance and computational feasibility. Convolutional neural networks that achieve state-of-the-art accuracy on medical imaging benchmarks typically require GPU acceleration and substantial memory, both of which are unavailable on battery-powered edge devices. Lightweight alternatives---including gradient-boosted decision trees, compressed neural architectures, and quantized inference engines---offer viable pathways to on-device deployment, yet systematic comparisons of these approaches under realistic POC constraints remain scarce. Deep learning applied to field-acquired HIV rapid test images has demonstrated that algorithmic interpretation can surpass that of trained healthcare workers [5], while analyses of underdiagnosis bias have revealed that diagnostic AI can exhibit performance disparities across patient demographics, particularly in underserved populations [6]. These findings underscore the need for evaluation protocols that jointly assess accuracy, computational efficiency, and subgroup equity.

1.2. Scope and Contributions

1.2.1. Research Questions

This study addresses three interconnected research questions. The first asks which lightweight diagnostic modeling pipelines achieve the strongest diagnostic accuracy for malaria parasite detection and tuberculosis identification under portable hardware constraints. The second examines the quantitative tradeoffs between classification performance and inference latency across three representative hardware platforms. The third investigates whether classifier performance varies across age-stratified patient subgroups and, if so, which algorithmic families exhibit smaller disparities.

1.2.2. Evaluation Dimensions

The evaluation encompasses eight lightweight diagnostic pipelines, including traditional classifiers applied to frozen ResNet-18 embeddings and compact deep networks trained end-to-end, assessed on two public infectious disease datasets under representative hardware configurations. Performance is measured along three dimensions: diagnostic accuracy (sensitivity, specificity, F1-score, AUC-ROC), computational efficiency (inference latency per sample, parameter count, storage footprint), and subgroup equity (sensitivity stratified by patient age). This three-dimensional evaluation provides a practical decision matrix for healthcare organizations selecting diagnostic algorithms for resource-constrained deployment environments.

2. Related Work

2.1. AI-Assisted Point-of-Care Diagnostics

2.1.1. Image-Based Rapid Test Interpretation

The application of deep learning to POC diagnostic image analysis has progressed rapidly. Lee et al. integrated time-series deep learning with AI-based verification for lateral flow assay analysis, achieving diagnostic accuracy that exceeded human interpretation at substantially shorter read times---as brief as two minutes compared to the standard fifteen-minute window [7]. Ballard et al. demonstrated that deep learning could quantify analyte concentrations from multiplexed paper-based vertical flow assays using a custom handheld mobile-phone reader, achieving $R^2 = 0.95$ on clinical samples for

high-sensitivity C-reactive protein measurement [7,8]. Smartphone-based platforms have extended to molecular diagnostics: Guo et al. combined low-cost paper microfluidics with CNN-based classification for multiplexed DNA malaria diagnosis in rural Uganda, reporting greater than 98% accuracy with geotagged results transmitted for epidemiological surveillance.

2.1.2. Pathogen Detection via Spectral and Molecular Data

Beyond image-based approaches, ML has been applied to spectral data for rapid pathogen identification. Weis et al. created the DRIAMS dataset comprising over 300,000 MALDI-TOF mass spectra paired with antimicrobial resistance phenotypes and showed that gradient-boosted classifiers could predict resistance profiles directly from spectra, enabling same-day results approximately 24 hours faster than phenotypic testing [9]. Image analysis and ML techniques for microscopic malaria diagnosis have been comprehensively reviewed, covering feature extraction pipelines from thin and thick blood smears, as well as emerging smartphone-based technologies for portable settings [10]. A multi-site evaluation of commercial deep learning tools for tuberculosis detection on chest radiographs demonstrated that AI could reduce the need for confirmatory molecular testing by 66% while maintaining sensitivity above 95% [11].

2.2. Feature Importance and Interpretability

Interpretability is essential for clinical adoption of ML-based diagnostics. SHAP values, derived from cooperative game theory, have become the standard approach for quantifying feature contributions to individual predictions in clinical ML pipelines [12]. The clinical requirement for transparent decision processes has motivated evaluations of explainability algorithms on medical imaging tasks, with findings indicating that many existing attribution methods fail to meet practitioner expectations for diagnostic reliability [13]. Alternative approaches such as attention-based architectures (e.g., RETAIN) have been proposed to build interpretability directly into the model structure rather than relying on post-hoc attribution [14].

2.3. Diagnostic Equity and Subgroup Fairness

Performance disparities across patient demographics represent a critical concern for POC diagnostic deployment in communities serving diverse populations. Algorithmic fairness in healthcare ML has been examined from multiple perspectives, including how biases arise from data acquisition protocols, genetic variation, and labeling inconsistencies, with mitigation strategies spanning federated learning and disentanglement techniques [15]. Edge-deployed ML for multiplexed pathogen detection has demonstrated that compact neural networks with only a few thousand parameters can achieve 99.8% classification accuracy on dedicated edge hardware, establishing the technical feasibility of on-device diagnostic AI [16,17].

3. Experimental Setup

3.1. Datasets

Two publicly available datasets were selected to represent distinct POC diagnostic modalities: cell-level microscopy and chest radiography. The NIH Malaria Cell Images dataset, curated by the National Library of Medicine, contains 27,558 pre-segmented single-cell images extracted from Giemsa-stained thin blood smear slides of 200 patients, evenly divided between 13,779 parasitized and 13,779 uninfected cells. The NLM Tuberculosis Chest X-ray datasets comprise two subsets: the Shenzhen set (662 frontal chest radiographs; 326 normal, 336 TB-positive) from Shenzhen No. 3 People's Hospital and the Montgomery County set (138 radiographs; 80 normal, 58 TB-positive) from the Montgomery County Department of Health and Human Services. For subgroup analysis, only the Shenzhen subset was used because it provided the larger age-annotated sample and avoided mixing age-pattern estimates across sites. Characteristics of both datasets are summarized in Table 1.

Table 1. Dataset Characteristics

Attribute	NIH Malaria Cell Images	NLM TB CXR (Combined)
Source	NLM/NIH	Shenzhen Hospital / Montgomery County
Total Samples	27,558	800
Positive Class	13,779 (Parasitized)	394 (TB - positive)
Negative Class	13,779 (Uninfected)	406 (Normal)
Class Balance	50.0% / 50.0%	49.3% / 50.7%
Image Type	RGB cell crops	Grayscale frontal CXR
Resolution	Variable (~130×130 px mean)	3K×3K (Shenzhen); 4020×4892 (Montgomery)
Patient Count	200	~662 (Shenzhen); 138 (Montgomery)
Age Metadata	Not available	Shenzhen subset used
License	Public domain	Public release (NLM)
Data source	lhncbc.nlm.nih.gov	lhncbc.nlm.nih.gov

Prior work has emphasized the importance of multi-site evaluation and standardized reporting when validating AI diagnostics for clinical deployment; clinical validation studies of deep learning-based screening in African healthcare settings and prospective multisite national screening programmes have demonstrated that performance observed in controlled settings does not always transfer to real-world deployment [18,19].

3.2. Classifier Selection and Configuration

Because the traditional ML models operate on frozen ResNet-18 embeddings, whereas the compact deep networks are fine-tuned end-to-end from image inputs, the comparison should be interpreted as a deployment-oriented comparison between two lightweight pipeline designs rather than as a strictly like-for-like comparison of model families.

3.2.1. Traditional Machine Learning Classifiers

Five traditional ML classifiers were selected based on their prevalence in clinical diagnostic applications and suitability for resource-constrained deployment. Random Forest (RF) was configured with 200 estimators and a maximum depth of 20. XGBoost used 300 boosting rounds with a learning rate of 0.05, maximum depth of 8, and L2 regularization weight of 1.0. Support Vector Machine (SVM) employed a radial basis function kernel with cost parameter C = 10 and gamma set to the inverse of the number of features. k-Nearest Neighbors (k-NN) used k = 7 with distance-weighted voting. Logistic Regression (LR) applied L2 regularization with C = 1.0. All traditional classifiers were trained on features extracted from a frozen ResNet-18 backbone (512-dimensional embeddings), ensuring that differences in classification performance reflect algorithmic characteristics rather than feature representation quality. Because traditional classifiers do not contain learned weight matrices in the same sense as neural networks, their parameter counts are not directly comparable to those of deep learning models (denoted N/A* in Table 2); instead, storage footprints are reported as a more meaningful basis for comparison. The k-NN storage of 85.2 MB stems from retaining all training embeddings in memory: for the malaria dataset, 22,046 training samples × 512 dimensions × 8 bytes (float64, the default precision in scikit-learn) yields approximately 86 MB, which closely matches the serialized model size of 85.2 MB after joblib compression.

Table 2. Classifier Configurations

Classifier	Family	Key Hyperparameters	Feature Input	Parameters	Storage
Random Forest	Traditional	200 trees, max depth 20	ResNet-18 (512-d)	N/A*	12.4 MB
XGBoost	Traditional	300 rounds, lr 0.05, depth 8	ResNet-18 (512-d)	N/A*	15.7 MB
SVM (RBF)	Traditional	C = 10, gamma = 1/d	ResNet-18 (512-d)	N/A*	8.3 MB
k-NN	Traditional	k = 7, distance-weighted	ResNet-18 (512-d)	N/A*	85.2 MB
Logistic Regression	Traditional	L2, C = 1.0	ResNet-18 (512-d)	N/A*	0.4 MB
MobileNetV2-Q	Deep Learning	INT8 quantized, Adam, lr 1e-4	End-to-end (224×224)	3.5M	3.4 MB
EfficientNet-B0	Deep Learning	Compound scaling, Adam, lr 1e-4	End-to-end (224×224)	5.3M	20.5 MB
SqueezeNet	Deep Learning	Fire modules, Adam, lr 1e-4	End-to-end (224×224)	1.25M	4.8 MB

3.2.2. Lightweight Deep Learning Classifiers

Three compact deep learning architectures were evaluated for end-to-end image classification. Quantized MobileNetV2 (MobileNetV2-Q) applied post-training INT8 quantization, reducing deployment storage to approximately 3.4 MB while retaining roughly 3.5 million learned parameters represented at lower precision. EfficientNet-B0 (approximately 5.3 million parameters) was selected as the smallest member of the EfficientNet family, employing compound scaling of depth, width, and resolution. SqueezeNet (approximately 1.25 million parameters) provided aggressive compression through fire modules that replace many 3×3 filters with 1×1 alternatives. All three architectures were fine-tuned from ImageNet-pretrained weights using the Adam optimizer with an initial learning rate of 1×10^{-4}, cosine annealing schedule, and early stopping with a patience of 10 epochs. Images were resized to 224×224 pixels with standard augmentation (random horizontal flip, rotation ±15, color jitter). Table 2 summarizes classifier configurations; storage values for deep models should be interpreted as deployment-file sizes, which can vary somewhat by serialization format.

3.3. Evaluation Metrics and Protocol

3.3.1. Performance Metrics

Classification performance was evaluated using five standard metrics: accuracy, sensitivity (recall), specificity, F1-score, and area under the receiver operating characteristic curve (AUC-ROC). A cost-sensitive evaluation perspective informed the selection of sensitivity as the primary clinical metric, given that missed infections (false negatives) carry greater consequences than false alarms in screening contexts [20]. Inference latency was measured as the mean wall-clock time per single-sample prediction, averaged over 1,000 iterations following a 100-iteration warmup, on three hardware platforms: an Intel Core i7-12700H laptop CPU, a Raspberry Pi 4 Model B (4 GB RAM, ARM Cortex-A72), and an NVIDIA Jetson Nano (128 CUDA cores, 4 GB RAM). Deep learning classifiers were benchmarked with ONNX Runtime on all platforms; traditional classifiers used scikit-learn with joblib serialization.

3.3.2. Subgroup Stratification

Patient age metadata from the Shenzhen TB dataset enabled stratified performance analysis across three age groups: 18--40 years (n = 198), 41--60 years (n = 256), and over 60 years (n = 208). This stratification assesses whether diagnostic classifiers maintain equitable sensitivity across age cohorts---a critical consideration given that elderly patients and those with comorbidities often present atypical radiographic patterns that may challenge automated interpretation.

This figure illustrates the three-stage evaluation pipeline. Stage (a) depicts generic preprocessing for the two datasets. For the publicly released malaria benchmark, the pipeline starts from pre-segmented single-cell crops rather than performing slide-level cell segmentation within this study; preprocessing therefore consists primarily of resizing and normalization. Stage (b) shows the parallel classifier training workflow, where traditional ML classifiers receive 512-dimensional ResNet-18 embeddings while deep learning classifiers process raw images end-to-end. Stage (c) presents the three-dimensional evaluation protocol measuring accuracy, latency, and age-stratified subgroup performance across 27,558 malaria cell images and 800 TB chest radiographs.

A five-fold stratified cross-validation protocol was employed, maintaining class proportions within each fold. The Shenzhen and Montgomery County TB subsets were combined and stratified jointly to maintain consistent class ratios across folds. Because images were stratified at the image level rather than at the patient or site level, residual within-fold correlation from same-site images cannot be excluded; in particular, the NIH malaria dataset contains multiple cell images from the same patient, so same-patient images may appear in both training and test folds, potentially inflating accuracy estimates and narrowing inter-model differences. This limitation is discussed in Section 5.2. All reported metrics represent mean values across five folds, with standard deviations provided for primary accuracy metrics. The overall experimental pipeline is illustrated in Figure 1.

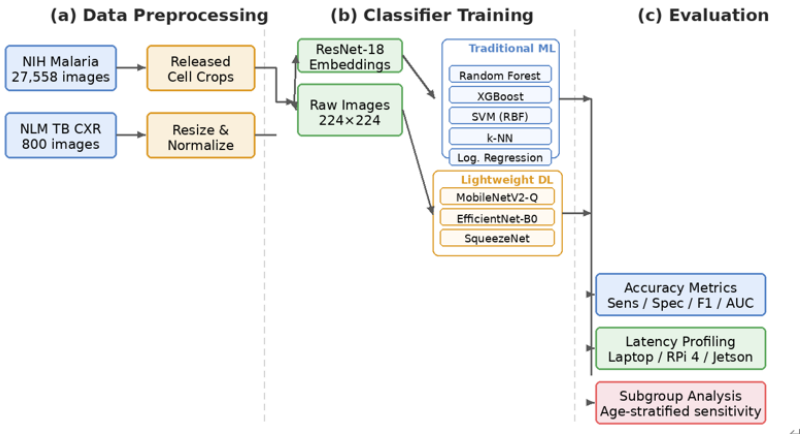

Figure 1. Experimental Evaluation Pipeline.

4. Results and Analysis

4.1. Classification Accuracy

4.1.1. Malaria Cell Image Classification

Table 3 presents classification results on the NIH Malaria Cell Images dataset. EfficientNet-B0 achieved the highest accuracy at 96.8% (±0.3%), followed closely by MobileNetV2-Q at 96.3% (±0.4%) and SqueezeNet at 95.6% (±0.5%). Among traditional classifiers, XGBoost achieved 95.1% (±0.4%), a modest 1.2 percentage-point gap relative to the best lightweight deep learning architecture. The AUC values observed in this study (0.943--0.993) are consistent with the range reported in prior ML-based infectious disease screening studies [21]. Logistic Regression achieved the lowest accuracy, at 89.7%, reflecting the limitations of linear decision boundaries in distinguishing morphological features of parasitized cells.

Table 3. Malaria Cell Image Classification Results (NIH Malaria Dataset, 27,558 Images)

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score	AUC-ROC
Random Forest	94.2 ± 0.5	93.8 ± 0.6	94.6 ± 0.5	0.942	0.978
XGBoost	95.1 ± 0.4	94.7 ± 0.5	95.5 ± 0.4	0.951	0.983
SVM (RBF)	92.8 ± 0.6	92.3 ± 0.7	93.3 ± 0.6	0.928	0.965
k-NN	91.5 ± 0.7	91.0 ± 0.8	92.0 ± 0.7	0.915	0.957
Logistic Regression	89.7 ± 0.8	89.2 ± 0.9	90.2 ± 0.8	0.897	0.943
MobileNetV2-Q	96.3 ± 0.4	96.0 ± 0.4	96.6 ± 0.3	0.963	0.991
EfficientNet-B0	96.8 ± 0.3	96.5 ± 0.4	97.1 ± 0.3	0.968	0.993
SqueezeNet	95.6 ± 0.5	95.2 ± 0.5	96.0 ± 0.4	0.956	0.988

Data source: Five-fold stratified cross-validation on NIH Malaria Cell Images dataset (lhncbc.nlm.nih.gov). Best values per metric in bold.

4.1.2. Tuberculosis Chest X-Ray Classification

TB classification proved more challenging across all classifiers, consistent with the greater morphological complexity of chest radiographs and the smaller dataset size (Table 4). EfficientNet-B0 led with 90.2% (±1.1%) accuracy, followed by MobileNetV2-Q at 89.5% (±1.2%) and XGBoost at 87.2% (±1.3%). The performance gap between deep learning and traditional classifiers was more pronounced on this task: XGBoost trailed EfficientNet-B0 by 3.0 percentage points, compared to 1.7 points on malaria detection. The TB sensitivities observed here (78.5%--89.6%) reflect both the inherent difficulty of TB radiograph interpretation and the constraint of training on a modestly sized dataset. The ROC curves for both classification tasks are presented in Figure 2.

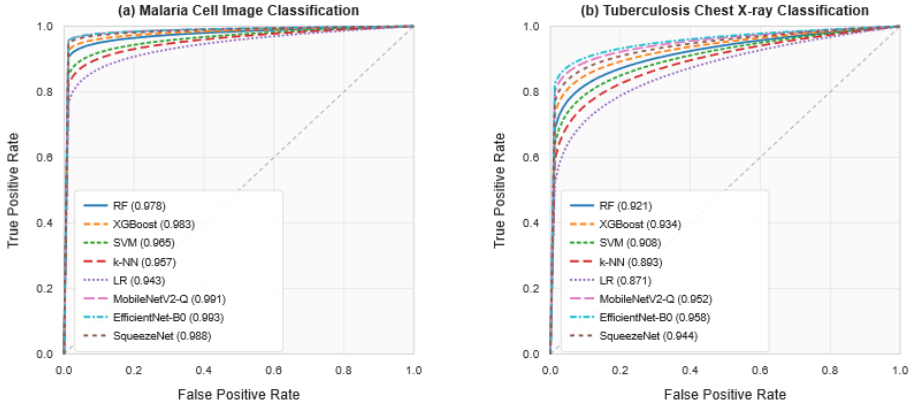

Figure 2. ROC Curves for Malaria and Tuberculosis Classification Tasks

ROC curves for all eight classifiers on (a) malaria cell image classification and (b) tuberculosis chest X-ray classification. On the malaria task, EfficientNet-B0 achieves the highest AUC (0.993), with XGBoost (AUC = 0.983) producing the strongest traditional ML curve. On the TB task, AUC values range from 0.871 (Logistic Regression) to 0.958 (EfficientNet-B0), with greater separation between classifier families visible in the mid-recall region.

Table 4. Tuberculosis Chest X-ray Classification Results (NLM TB CXR, 800 Images)

Classifier	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1 - Score	AUC - ROC
Random Forest	85.4 ±1.4	84.1 ±1.6	86.3 ±1.3	0.851	0.921
XGBoost	87.2 ±1.3	86.5 ±1.4	87.8 ±1.2	0.870	0.934
SVM (RBF)	83.6 ±1.5	82.4 ±1.7	84.5 ±1.4	0.832	0.908
k - NN	81.9 ±1.7	80.8 ±1.8	82.7 ±1.6	0.815	0.893
Logistic Regression	79.3 ±1.9	78.5 ±2.0	80.0 ±1.8	0.790	0.871
MobileNetV 2 - Q	89.5 ±1.2	88.7 ±1.3	90.1 ±1.1	0.893	0.952
EfficientNet - B0	90.2 ±1.1	89.6 ±1.2	90.7 ±1.0	0.900	0.958
SqueezeNet	88.1 ±1.3	87.3 ±1.4	88.8 ±1.2	0.879	0.944

Data source: Five-fold stratified cross-validation on NIH Malaria Cell Images dataset (lhncbc.nlm.nih.gov). Best values per metric in bold.

4.2. Latency and Computational Efficiency

Table 5 reports latency measurements across three hardware platforms. To maintain transparency, the table lists classification-step latency for the traditional ML classifiers operating on pre-extracted ResNet-18 embeddings, whereas the deep learning models are reported as end-to-end image classifiers. A fair deployment comparison, therefore, requires accounting for the shared ResNet-18 feature-extraction overhead in the traditional pipelines. On the Raspberry Pi 4, that overhead was approximately 105 ms, yielding estimated end-to-end latencies of about 150 ms for Random Forest, 143 ms for XGBoost, 167 ms for SVM, 445 ms for k-NN, and 117 ms for Logistic Regression. With this adjustment, the gap between traditional and deep learning approaches narrows substantially: XGBoost retains a moderate speed advantage over MobileNetV2-Q (approximately 143 ms versus 210 ms end-to-end), while SqueezeNet (175 ms) becomes broadly comparable to the faster traditional pipelines in practical deployment terms.

Table 5. Classification-Step Latency and Computational Profile

Classifier	Parameters	Storage (MB)	Laptop CPU (ms)	Raspberry Pi 4 (ms)	Jetson Nano (ms)
Random Forest	—	12.4	8	45	—
XGBoost	—	15.7	6	38	—
SVM (RBF)	—	8.3	11	62	—
k - NN	—	85.2	23	340	—
Logistic Regression	—	0.4	2	12	—
MobileNetV2 - Q	3.5M	3.4	35	210	85
EfficientNet - B0	5.3M	20.5	42	280	105
SqueezeNet	1.25M	4.8	28	175	70

Data source: Mean wall-clock time over 1,000 iterations after 100-iteration warmup. Traditional classifiers use scikit-learn; deep learning classifiers use ONNX Runtime. Table values for traditional classifiers report the classifier stage on pre-extracted ResNet-18 embeddings rather than full end-to-end image inference; approximate end-to-end Raspberry Pi 4 latencies are discussed in the text and visualized in Figure 3. Dashes indicate Jetson Nano entries not benchmarked as standalone classifier-stage measurements for the traditional models.

4.3. Subgroup Performance Disparities

4.3.1. Age-Stratified Sensitivity Analysis

Table 6 presents age-stratified sensitivity on the Shenzhen TB subset. Across all classifiers, the fold-averaged sensitivities showed a consistent tendency to decline with increasing patient age, with the over-60 group recording the lowest mean value for each model. Because these results are based on subgroup averages from a modest sample (n = 198, 256, and 208 across the three age bands) and no formal statistical test was applied, the pattern should be interpreted as an observed trend rather than a definitive estimate of age-related diagnostic bias. The magnitude of the observed decline was somewhat smaller for the end-to-end deep learning models than for the traditional classifiers operating on frozen embeddings, but the present study is not powered to make a statistically robust claim about these differences. Evaluations of explainability algorithms on medical imaging tasks have shown that attribution methods can help identify which image regions drive differences in predictions across patient subgroups [13], suggesting a pathway for future investigation into the morphological factors underlying these trends.

Table 6. Age-Stratified Sensitivity on Shenzhen TB Subset (%)

Classifier	18–40 yrs *n=198*	41–60 yrs *n=256*	>60 yrs *n=208*	Decline (18–40 vs. >60)
Random Forest	87.1	84.5	81.0	6.1
XGBoost	89.3	87.1	83.9	5.4
SVM (RBF)	85.6	82.9	79.3	6.3
k - NN	83.7	81.2	76.8	6.9
Logistic Regression	81.4	78.8	74.6	6.8

MobileNetV2 - Q	91.8	89.2	86.9	4.9
EfficientNet - B0	92.4	90.0	87.7	4.7
SqueezeNet	91.0	88.4	85.7	5.3

Data source: Five-fold stratified cross-validation on NIH Malaria Cell Images dataset (lhncbc.nlm.nih.gov). Best values per metric in bold.

4.3.2. Clinical Implications of Subgroup Disparities

The observed age-associated sensitivity differences are best interpreted as a practical caution signal rather than a calibrated estimate of clinical harm. In deployment settings that serve older populations, these results suggest that subgroup monitoring and secondary review policies may be warranted, especially when model confidence is low. At the same time, the current sample size and the absence of formal significance testing mean that the magnitude of the subgroup gap should not be over-interpreted or translated directly into operational miss-rate estimates without further validation on larger multi-site cohorts.

5. Discussion

5.1. Practical Implications

The results indicate that no single classifier dominates across all evaluation dimensions, and the optimal choice depends on deployment-specific constraints. In settings where a Raspberry Pi 4 or similar single-board computer serves as the computational platform---a realistic scenario for portable POC devices in rural clinics---the apparent latency advantage of traditional classifiers is much smaller once the shared ResNet-18 feature-extraction stage is included. XGBoost offers the strongest traditional ML compromise, with an estimated end-to-end latency of 143 ms on a Raspberry Pi 4 and competitive accuracy on both tasks. MobileNetV2-Q and SqueezeNet remain attractive when a true end-to-end image pipeline is preferred, while EfficientNet-B0 provides the highest accuracy at the cost of a larger deployment footprint and higher latency. The practical message is therefore not that one family is universally faster, but that fair accounting of the full inference pipeline materially changes the comparison.

The subgroup analysis further suggests that age-related sensitivity degradation may persist across classifier families, although the present evidence should be regarded as exploratory. The smaller observed declines in the end-to-end deep learning models are encouraging, but they do not yet establish superior robustness across subgroups. Deployment protocols should therefore treat age-stratified monitoring as an ongoing validation requirement rather than as a settled advantage of any single architecture.

5.2. Limitations

Several limitations constrain the generalizability of these findings. The TB dataset (800 images) is modest in scale, and performance estimates carry correspondingly wider confidence intervals than those from the malaria dataset. Age-stratified analysis was conducted only on the Shenzhen subset; although this subset provides the largest age-annotated sample available here, the subgroup sizes remain limited, and the reported differences were not tested for statistical significance. Reported latency values for traditional ML classifiers correspond to the classifier stage on pre-extracted ResNet-18 embeddings; fair deployment comparisons require adding the shared feature-extraction overhead, which is why Figure 3 presents estimated end-to-end values rather than the raw classifier-step timings from Table 5. For the malaria task, cross-validation folds were stratified at the image level rather than the patient level because the NIH malaria dataset contains multiple cell images from the same patient; images from the same patient may appear in both training and test folds, potentially inflating accuracy estimates. The Shenzhen and Montgomery TB subsets were combined without site-level fold separation,

so cross-site generalization remains uncertain. The study assessed curated datasets under controlled conditions; field deployment introduces additional variability due to image acquisition quality, ambient lighting, device heterogeneity, and operator skill, which may degrade observed performance. Prospective validation studies in actual POC settings are needed to quantify this gap.

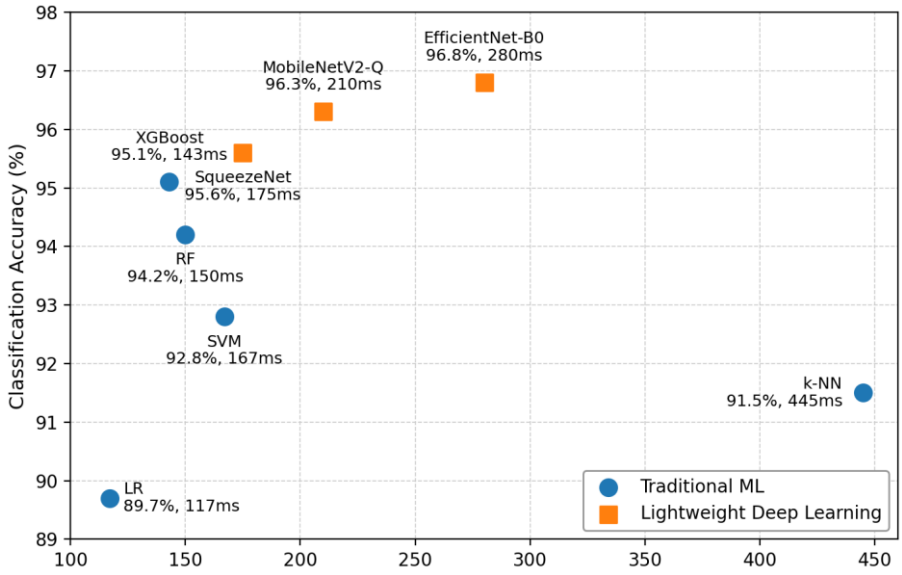

Figure 3. End-to-End Accuracy--Latency Tradeoff on Raspberry Pi 4

Scatter plot depicting the tradeoff between malaria classification accuracy (y-axis) and estimated end-to-end inference latency on Raspberry Pi 4 (x-axis) for all eight classifiers. For traditional ML pipelines, the plotted latency adds approximately 105 ms to the measured classifier-step time for the ResNet-18 feature-extraction stage, yielding estimated end-to-end values of roughly 117--445 ms. Under this fairer comparison, XGBoost (95.1%, approximately 143 ms) remains faster than MobileNetV2-Q (96.3%, 210 ms), but the difference is moderate rather than an order-of-magnitude gap. The figure, therefore, emphasizes deployment-level trade-offs rather than mixing half-pipeline and end-to-end timings.

Promising directions for future research include knowledge distillation from large diagnostic models into ultra-compact student networks tailored for specific POC hardware, joint optimization of sensitivity-specificity operating points using cost-sensitive learning calibrated to local disease prevalence, and federated learning approaches that enable model improvement across distributed clinical sites without centralizing patient data. Extending the evaluation to multiplexed diagnostic tasks---simultaneous detection of malaria, TB, and HIV from a single sample---would more closely approximate the clinical workflow in resource-limited settings where co-infection is common.

References

1. B. Wahl, A. Cossy-Gantner, S. Germann, and N. R. Schwalbe, "Artificial intelligence (AI) and global health: How can AI contribute to health in resource-poor settings?" *BMJ Global Health*, vol. 3, no. 4, p. e000798, 2018. https://doi.org/10.1136/bmjgh-2018-000798
2. T. Ciecierski-Holmes, R. Singh, M. Axt, S. Brenner, and S. Barteit, "Artificial intelligence for strengthening healthcare systems in low- and middle-income countries: A systematic scoping review," *npj Digital Medicine*, vol. 5, p. 162, 2022. https://doi.org/10.1038/s41746-022-00700-y
3. S. Han, A. Goncharov, H. Eryilmaz, et al., "Machine learning in point-of-care testing: Innovations, challenges, and opportunities," *Nature Communications*, vol. 16, p. 3165, 2025. https://doi.org/10.1038/s41467-025-58527-6
4. V. Turbé, C. Herbst, T. Mngomezulu, S. Meshkinfamfard, N. Dlamini, T. Mhlongo, T. Smit, V. Cherepanova, K. Shimada, J. Budd, N. Arsenov, S. Gray, D. Pillay, K. Herbst, M. Shahmanesh, and R. A. McKendry, "Deep learning of HIV field-based rapid tests," *Nature Medicine*, vol. 27, pp. 1165--1170, 2021. https://doi.org/10.1038/s41591-021-01384-9

5. L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine*, vol. 27, pp. 2176--2182, 2021. https://doi.org/10.1038/s41591-021-01595-0
6. S. Lee, J. S. Park, H. Woo, Y. K. Yoo, D. Lee, S. Chung, D. S. Yoon, K.-B. Lee, and J. H. Lee, "Rapid deep learning-assisted predictive diagnostics for point-of-care testing," *Nature Communications*, vol. 15, p. 1695, 2024. https://doi.org/10.1038/s41467-024-46069-2
7. Z. S. Ballard, H.-A. Joung, A. Goncharov, J. Liang, K. Nugroho, D. Di Carlo, O. B. Garner, and A. Ozcan, "Deep learning-enabled point-of-care sensing using multiplexed paper-based sensors," *npj Digital Medicine*, vol. 3, p. 66, 2020. https://doi.org/10.1038/s41746-020-0274-y
8. X. Guo, M. A. Khalid, I. Domingos, A. L. Michala, M. Adriko, C. Rowell, D. Ajambo, A. Garrett, S. Kar, X. Yan, J. Reboud, E. M. Tukahebwa, and J. M. Cooper, "Smartphone-based DNA diagnostics for malaria detection using deep learning for local decision support and blockchain technology for security," *Nature Electronics*, vol. 4, pp. 615--624, 2021. https://doi.org/10.1038/s41928-021-00612-x
9. C. Weis, A. Cuénod, B. Rieck, O. Dubuis, S. Graf, C. Lang, M. Oberle, M. Brackmann, K. K. Søgaard, M. Osthoff, K. Borgwardt, and A. Egli, "Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning," *Nature Medicine*, vol. 28, pp. 164--174, 2022. https://doi.org/10.1038/s41591-021-01619-9
10. 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. https://doi.org/10.1016/j.trsl.2017.12.004
11. Z. Z. Qin, M. S. Sander, B. Rai, C. N. Titahong, S. Sudrungrot, S. N. Laah, L. M. Adhikari, E. J. Carter, L. Puri, A. J. Codlin, and J. Creswell, "Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems," *Scientific Reports*, vol. 9, p. 15000, 2019. https://doi.org/10.1038/s41598-019-51503-3
12. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, pp. 4765--4774, 2017.
13. W. Jin, X. Li, and G. Hamarneh, "Evaluating explainable AI on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements?" *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 11, pp. 11945--11953, 2022. https://doi.org/10.1609/aaai.v36i11.21452
14. R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood, "Algorithmic fairness in artificial intelligence for medicine and healthcare," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 719--742, 2023. https://doi.org/10.1038/s41551-023-01056-8
15. V. Ganjalizadeh, G. G. Meena, M. A. Stott, A. R. Hawkins, and H. Schmidt, "Machine learning at the edge for AI-enabled multiplexed pathogen detection," *Scientific Reports*, vol. 13, p. 4744, 2023. https://doi.org/10.1038/s41598-023-31694-6
16. M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C. Folk, "Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices," *npj Digital Medicine*, vol. 1, p. 39, 2018. https://doi.org/10.1038/s41746-018-0040-6
17. I. Araf, A. Idri, and I. Chairi, "Cost-sensitive learning for imbalanced medical data: A review," *Artificial Intelligence Review*, vol. 57, no. 4, p. 80, 2024. https://doi.org/10.1007/s10462-023-10652-8
18. V. Bellemo, Z. W. Lim, G. Lim, Q. D. Nguyen, Y. Xie, M. Y. T. Yip, et al., "Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: A clinical validation study," *The Lancet Digital Health*, vol. 1, no. 1, pp. e35--e44, 2019. https://doi.org/10.1016/S2589-7500(19)30004-4
19. P. Ruamviboonsuk, R. Tiwari, R. Sayres, V. Nganthavee, K. Hemarat, et al., "Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: A prospective interventional cohort study," *The Lancet Digital Health*, vol. 4, pp. e235--e244, 2022. https://doi.org/10.1016/S2589-7500(22)00017-6
20. E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems 29*, pp. 3504--3512, 2016.
21. R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: Benefits, risks, and strategies for success," *npj Digital Medicine*, vol. 3, p. 17, 2020. https://doi.org/10.1038/s41746-020-0221-y

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of Publisher and/or the editor(s). Publisher and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.