Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 15 April 2026
Revised 20 May 2026
Accepted 24 June 2026

Natural Resources for Human Health 2026, Vol. 6 (3s): 185-206

KEYWORDS:
tuberculosis diagnosis, sputum smear microscopy, object detection, federated learning, domain generalisation, bacillary load grading

Automated Detection, Quantification and Grading of Tuberculosis Bacilli in Ziehl-Neelsen Sputum Smear Microscopy: A Structured Review of Methods, Datasets, Multi-Centre Learning and Robustness

Vishakha Yadav[1,2], Thippeswamy G[3]

[1]Department of CSE, BMS Institute of Technology and Management, Bengaluru, India.
[2]Research Scholar, Nagarjuna College of Engineering and Technology (Visvesvaraya Technological University, Belagavi), Karnataka, India. Email: vishakhay@bmsit.in
ORCID ID: 0000-0002-3880-9361
[3]Principal and Professor, Department of CSE, Nagarjuna College of Engineering and Technology, Bengaluru, India. Email: swamy.gangappa@gmail.com
ORCID ID: 0000-0001-8424-2119

ABSTRACT: Sputum smear microscopy remains the frontline diagnostic test for pulmonary tuberculosis in high-burden settings, yet it is slow, demands a trained microscopist and reports limited sensitivity. Deep learning has been applied intensively to this task, but the literature has grown without a shared benchmark, a shared metric definition or a shared statement of the clinical target. This review organises the 2022-2026 literature into a six-category taxonomy spanning acquisition and image formation, detection and segmentation, quantification and bacillary load grading, learning under limited supervision, multi-centre privacy-preserving learning, and robustness. Primary studies, seven reviews and every reachable public dataset are compared in eight tables. Three findings emerge. Reported performance is inflated by task substitution, since patch classification studies report accuracies near ninety-nine percent whereas detection studies report mean average precision near fifty percent when averaged across localisation thresholds. Quantification is largely unsolved, because clumped bacilli are annotated away rather than resolved and only one study automates the five-point World Health Organization load scale. Multi-centre and robustness evidence is almost absent, with one of one hundred and fifty-two artificial intelligence tuberculosis studies reporting a domain-shift analysis and no published federated learning study on sputum smear microscopy. We formalise the evaluation gap, propose a foreground-density heterogeneity measure for federated smear analysis, and set out a research agenda centred on grading-aware, multi-centre and reliability-audited models.

1. INTRODUCTION

Pulmonary tuberculosis remains one of the leading causes of death from a single infectious agent, and in the settings where its burden is highest the diagnosis still rests on light microscopy of a Ziehl-Neelsen (ZN) stained sputum smear. The procedure is inexpensive and requires no cold chain, which is why it persists despite the availability of molecular assays. Its weaknesses are equally well documented: a microscopist must examine at least one hundred high-power fields per slide, throughput is low, and inter-reader agreement degrades with fatigue and workload [1], [2].

Critically, the clinical output of the procedure is not a binary decision. Under World Health Organization and International Union Against Tuberculosis and Lung Disease conventions the laboratory reports a five-point bacillary load grade, comprising negative, scanty, 1+, 2+ and 3+, derived from the number of acid-fast bacilli counted over a defined number of fields. This grade informs infectiousness assessment and treatment monitoring, so a system that reports only the presence or absence of bacilli has solved a proxy task rather than the clinical one [3], [4].

Automated analysis of ZN smears has consequently attracted sustained attention. A 2026 systematic review in the American Journal of Clinical Pathology screened the past decade and included forty-nine studies reporting accuracies between eighty and ninety-eight percent, while noting substantial heterogeneity in study design and limited external validation [5]. Earlier reviews reached compatible conclusions from different corpora: forty-five studies of which only five used a publicly available dataset [1], twenty-eight studies with an explicit repeatability component [2], and sixty-five studies spanning 2017 to 2023 [6]. Across all four, the recurring criticism is not that models perform badly but that their reported performance cannot be compared, reproduced or trusted to transfer.

This review differs from its predecessors in three respects. First, it is organised around a taxonomy that separates the clinical target (counting and grading) from the computational target (detection), because conflating the two is the main source of apparent progress. Second, it treats multi-centre learning and robustness as first-class categories rather than as future work, since deployment in a national tuberculosis programme necessarily spans many laboratories with different optics and staining practice. Third, it verifies dataset availability directly rather than by citation, which materially changes the picture, because the two most cited resources in this field are no longer served at their published addresses.

The contributions are as follows. We define a six-category taxonomy of automated ZN smear analysis and populate it with the 2022-2026 literature. We provide eight comparison tables covering each category and every reachable public dataset. We formalise the metric-substitution problem, the load-grading objective and a foreground-density heterogeneity measure for federated smear analysis. Finally, we identify five gaps that are verifiably open and propose concrete experimental designs to close them.

It is worth stating at the outset why a further review is justified when four already exist. The reviews cited above are bibliometrically thorough and methodologically conventional: they enumerate architectures, tabulate accuracies and conclude that heterogeneity limits comparison. What none of them does is ask whether the quantity being reported is the quantity that matters. A study that classifies cropped patches containing a single centred bacillus and a study that localises every organism in an uncurated field are solving problems that differ by orders of magnitude in difficulty, yet both are recorded in the same accuracy column and both contribute to the same pooled estimate. Once the reported quantity is separated from the clinical quantity, the apparent maturity of the field changes character, and the direction of useful work changes with it. That separation is the organising principle of this review.

A second motivation is temporal. The period covered here contains three developments that postdate the earlier surveys and that materially alter what is now feasible: the release of the first openly licensed corpora carrying focal stacks and load-grade labels, the first application of a pathology foundation model to smear screening under image-level supervision alone, and the first independent clinical validations of commercial smear-reading systems against culture. Each is discussed in the sections that follow, and together they shift the binding constraint of the field from data availability to evaluation discipline.

2. SCOPE AND REVIEW METHODOLOGY

Records were sought in PubMed, Scopus, IEEE Xplore, Web of Science and the arXiv application programming interface using combinations of tuberculosis, sputum smear, acid-fast bacilli, Ziehl-Neelsen, auramine, deep learning, object detection, federated learning and robustness. The primary window is January 2022 to August 2026, extended backwards where a method is foundational or where a dataset descriptor predates the window. Studies were included if they analysed smear or blood-film microscopy images computationally and reported quantitative results, or if they contributed a dataset, benchmark or methodological result that the smear literature depends upon. Chest radiograph tuberculosis studies were excluded, as they address a different modality and are covered elsewhere [7].

Every dataset cited in the included studies was resolved directly: the published uniform resource locator was requested, and the resource was recorded as reachable only if the files themselves were served. Every reference in this review was resolved against Crossref, and bibliographic fields were taken from the resolved record rather than transcribed.

Table 1 positions this review against the existing surveys. The distinguishing feature is coverage of categories C5 and C6, which no prior ZN-specific review addresses.

Table 1: Existing reviews compared with the present work

Review	Year	Studies	Window	Modality	Grading	Multi-centre	Robustness
Carvalho et al. [2]	2023	28	to 2022	ZN smear	No	No	Partial
Zachariou et al. [1]	2023	45	1998-2022	ZN and fluorescence	No	No	No
Tamura et al. [6]	2024	65	2017-2023	ZN smear and tissue	No	No	No
Lumamba et al. [8]	2024	10	2017-2022	ZN tissue only	No	No	No
Hansun et al. [7]	2025	152	2011-2023	All, mostly radiograph	No	No	Partial
Rastogi et al. [5]	2026	49	2015-2025	ZN and auramine	No	No	Partial
von Bahr et al. [9]	2026	22	2014-2024	Primary-care microscopy	No	No	No
This review	2026	68	2022-2026	ZN smear	Yes	Yes	Yes

3. A TAXONOMY OF AUTOMATED SMEAR ANALYSIS

Figure 1 presents the taxonomy. Six categories, denoted C1 to C6, partition the field, and Sections 4 to 9 address them in order. The partition is functional rather than chronological: a single study frequently contributes to two categories, and Tables 2 to 7 record such studies once per contribution.

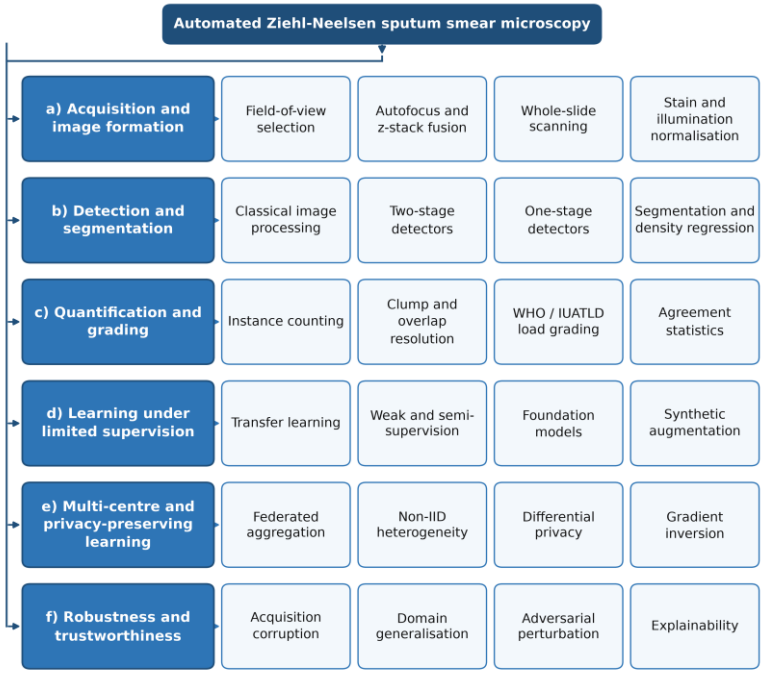

Fig. 1: Taxonomy of automated Ziehl-Neelsen smear microscopy research

Figure 2 shows the canonical processing pipeline that nearly every published system instantiates, together with the dominant failure mode reported at each stage. Two features of the pipeline deserve emphasis. Stages S1 to S3 determine the input distribution and are where cross-laboratory variation enters; stages S6 and S7 determine the clinical output and are where the literature is thinnest. The intermediate stages S4 and S5, which are where almost all methodological effort has been concentrated, are bounded above by S3 and are of no clinical value without S6 and S7.

The asymmetry of effort across the pipeline is the structural problem this review documents. Of the studies in Tables 2 to 7, the overwhelming majority contribute to S4 and S5, a small number to S1 to S3, and a handful to S6 and S7. This distribution is understandable, since S4 and S5 are where standard computer vision transfers most directly and where public benchmarks in other domains have trained the community's intuitions. It is nonetheless the wrong allocation for the clinical task. An improvement in candidate classification cannot recover bacilli that were never imaged because the field was out of focus or the smear was too thick, and it cannot correct a load grade that is wrong because three overlapping organisms were counted as one.

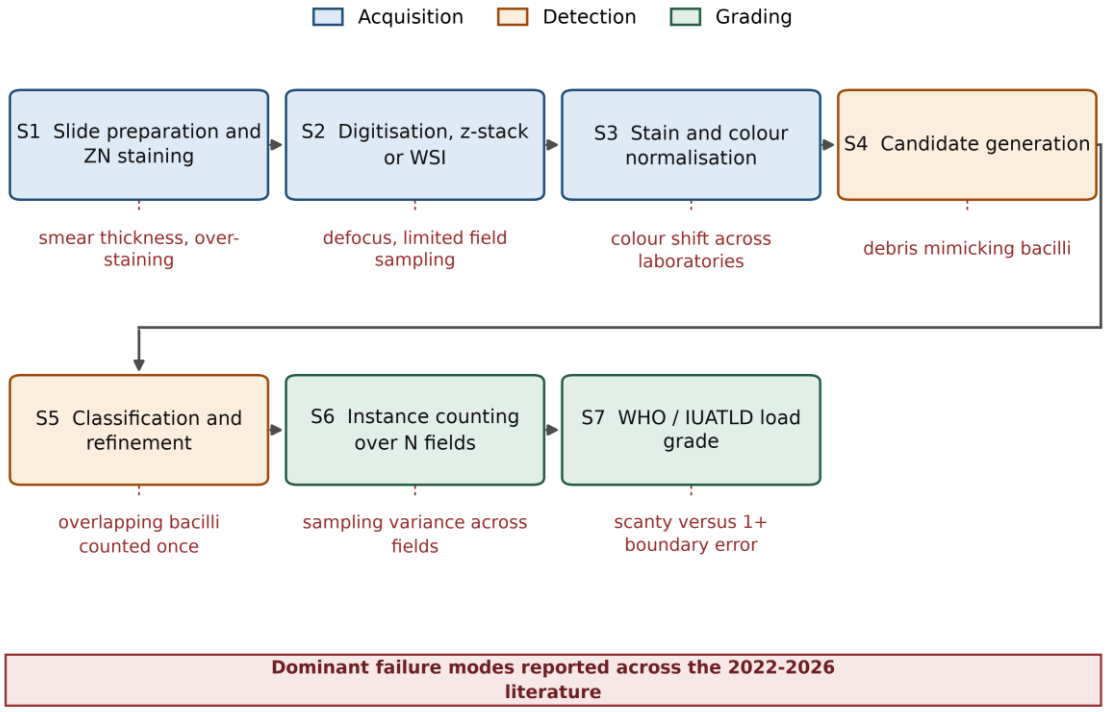

Fig. 2: Canonical pipeline with dominant failure mode per stage

4. ACQUISITION AND IMAGE FORMATION

4.1 Field-of-view selection and whole-slide scanning

A ZN slide contains far more material than can be imaged at oil-immersion magnification, so field selection is itself a decision with diagnostic consequences. Reported field budgets vary by more than a factor of three: three hundred fields for a commercial brightfield scanner [10], six hundred fields for a manual comparator read and one thousand fields for automated ZN in a prospective comparison [11]. Ravindran and Baskaran treat field recognition as a learned first stage on a motorised stage, reporting an area under the receiver operating characteristic curve of 0.9505 with precision 0.924 and recall 0.882 [12], and this remains one of very few studies to model field selection explicitly rather than assume it.

Field selection also interacts with the load grade in a way that is rarely acknowledged. Because the grade is defined per hundred fields, the number of fields examined is part of the measurement rather than an implementation detail: a system reading three hundred fields and one reading a thousand are estimating the same latent quantity with different variance, and

neither reports that variance. When the automated and manual arms of a prospective study read different field counts, the comparison confounds algorithmic sensitivity with sampling effort [11].

The consequence of leaving field selection unmodelled is quantified by two clinical evaluations. In a deployed scanner, one hundred and fifty of one thousand one hundred and fifty specimens, or thirteen percent, were rejected outright for staining, thickness or positioning faults, and the authors note explicitly that a smear containing bacilli may lie outside the scanned area [10]. In an independent retrospective validation of a commercial system, thirty-four of three hundred and twenty slides failed scanning altogether [13]. Neither failure is visible in the accuracy figures that dominate the methods literature.

4.2 Autofocus and multi-focal-plane fusion

Sputum is spread unevenly and uncovered, so a single focal plane cannot render all bacilli sharply. The literature handles this in three ways, none of them learned. Blurred fields are discarded before training [14]; extended depth-of-focus fusion is applied as a preprocessing step [15], [16]; or the problem is moved into scanner firmware, as in the curved-surface focus algorithm developed for oil-immersion whole-slide digitisation [17]. The 2026 systematic review confirms that multi-view and z-stack imaging is an emerging mitigation [5], yet no published detector consumes the focal stack directly, despite the public availability of eight thousand two hundred stacks of eleven depths [16].

The distinction matters because the three strategies are not equivalent. Discarding blurred fields biases the sample towards regions where the smear happens to be thin, which is precisely where bacilli are least likely to be concentrated. Extended depth-of-focus fusion is a fixed unsupervised operator applied before the network sees the data, so any information it discards is unrecoverable and its parameters cannot be tuned to the detection objective. Moving the problem into firmware solves it for one instrument and transfers to no other. A learned fusion, in which the stack is treated as an additional input dimension and the fusion weights are trained jointly with the detector, is the obvious remaining option and is unexplored in this domain.

4.3 Stain and illumination normalisation

Stain appearance is the dominant nuisance factor. One study is constructed entirely around thin, standard and thick ZN variants [18], and the evaluation of a deployed scanner states that colour is the load-bearing recognition feature [10]. The computational pathology literature is far more mature here and provides the baseline that smear work must respect: across four organs and nine laboratories, colour augmentation was found essential to reduce generalisation error, while colour normalisation could be omitted at negligible cost [19]. Recent work extends normalisation to self-supervised and privacy-preserving formulations [20], and a 2025 review catalogues the generative-adversarial and diffusion families now used for the task [21]. A recurring caution is that normalisation metrics do not predict downstream utility [21]: methods scoring well on image-similarity measures may perform no better, and occasionally worse, on the task the normalisation was intended to serve.

Two properties distinguish ZN smears from the haematoxylin and eosin images on which this literature was developed, and both should make the transfer easier rather than harder. ZN staining is effectively two-colour, with carbol-fuchsin bacilli against a methylene-blue counterstain, so the appearance manifold is lower-dimensional than in general histopathology. The foreground is also extremely sparse, which means colour statistics computed over a whole field are dominated by background and are therefore comparatively stable estimators of the staining condition. Neither property has been exploited: no smear study reports a stain-normalisation ablation of the kind now routine in pathology, and the field consequently has no evidence about whether the conclusion of Tellez et al. transfers to this modality.

Table 2: Acquisition and pre-processing approaches in tuberculosis smear microscopy

Reference	Year	Contribution	Field budget	Focus handling	Stain handling
Costa Filho et al. [15]	2019	Multi-focus fusion	Not stated	Fusion of stacks	None
Huang et al. [22]	2022	Automated scanner	Slide-level	Hardware autofocus	Fixed protocol

Reference	Year	Contribution	Field budget	Focus handling	Stain handling
Chen et al. [10]	2024	Deployed scanner	300 fields	Hardware	Colour-dependent
Desruisseaux et al. [13]	2024	Clinical validation	Whole slide	Scanner	Auramine
Ravindran and Baskaran [12]	2025	Learned field selection	Motorised stage	Not modelled	None
Yuniarti et al. [14]	2026	Raw dataset	Manual capture	Blurred fields excluded	None applied
Zhou et al. [17]	2026	Whole-slide AFB scanner	Full slide	Curved-surface focus	Fixed protocol
Rulaningtyas et al. [18]	2026	Stain-thickness robustness	Field-level	Not modelled	Thin, standard, thick
Zhang et al. [11]	2026	Prospective comparison	300 to 1000 fields	Scanner	ZN and auramine

5. DETECTION AND SEGMENTATION

5.1 From classical processing to learned detection

Early systems thresholded in a colour space chosen to separate the carbol-fuchsin signature, then filtered candidates on geometric descriptors such as eccentricity, area and axis ratio. These pipelines remain instructive because their failure modes are unchanged: debris that shares the chromatic signature of a bacillus is retained, and touching organisms merge into a single component. Panicker et al. showed that a shallow convolutional classifier applied to binarised candidate regions substantially outperformed handcrafted descriptors on the same candidates [23], which established the candidate-then-classify template that most later systems follow.

The persistence of this template into the deep learning era is itself informative. It survives because it encodes a real property of the task: recall at the candidate stage can be made very high at low computational cost, since bacilli are chromatically distinctive, whereas precision is the hard part and benefits from learned features. The template also has a less benign consequence, in that it fixes the definition of an object at the candidate stage. If a connected component containing three touching organisms is emitted as a single candidate, no downstream classifier can recover the correct count, and the error is invisible to any metric computed over candidates.

5.2 Detector families

Figure 3 contrasts the four architectural families now present in this literature. Two-stage detectors derived from Faster R-CNN [24] provide accurate localisation but impose one object per proposal, which is precisely the wrong inductive bias for agglomerated bacilli. One-stage detectors combining a feature pyramid [25] with focal loss [26] dominate the 2024-2026 work because bacilli are sparse, small and numerous, which is the regime focal loss was designed for. Transformer detectors dispense with hand-designed anchors and non-maximum suppression in favour of set prediction, which is attractive for dense small objects but data-hungry, and no application to ZN smears has been published. Density-regression models replace detection with estimation of a continuous field whose integral is the count, trading instance identity for robustness to overlap.

The choice among these families is usually framed as a speed-accuracy trade-off, but for this task the more consequential axis is how each handles agglomeration. Two-stage and one-stage detectors both assume a one-to-one correspondence between predicted boxes and objects, an assumption that non-maximum suppression actively enforces by removing overlapping predictions. For touching bacilli this is not a nuisance to be tuned away but a structural mismatch, because the correct output

genuinely contains heavily overlapping boxes. Set-prediction models relax the assumption in principle, and density regression sidesteps it entirely at the cost of the instance-level information needed for morphology and size checks.

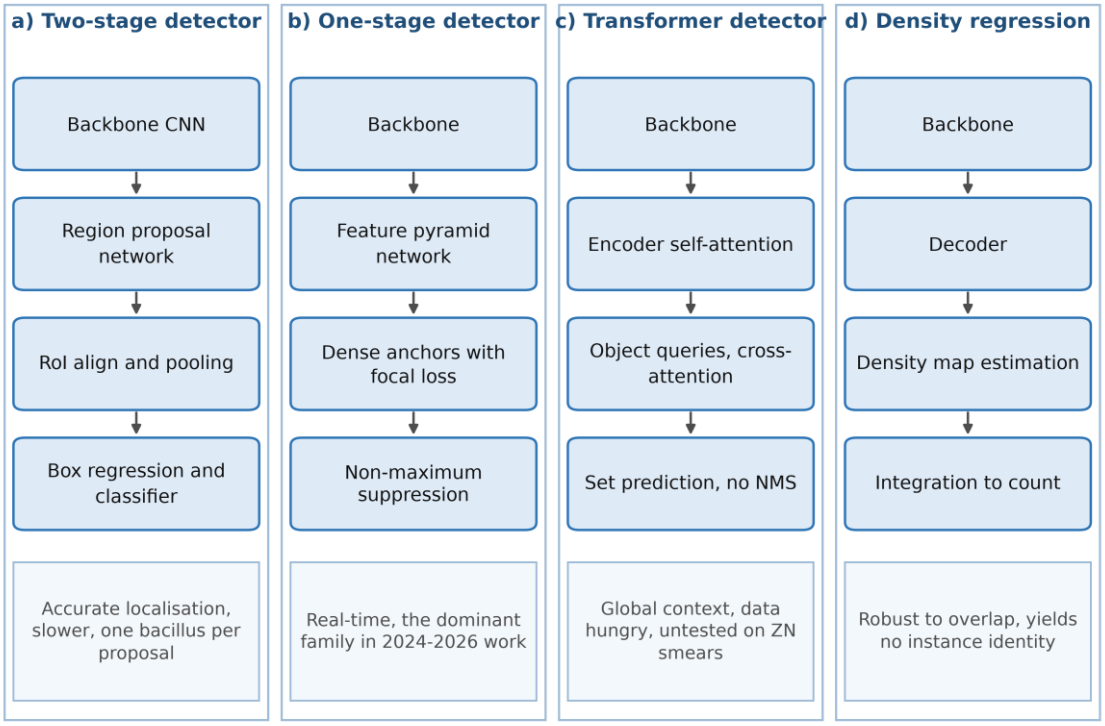

Fig. 3: Four detector architecture families for bacilli detection

Detection quality is scored by intersection over union between a predicted box B_p and a ground-truth box B_g,

$$\mathrm{IoU}(B_p, B_g) = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \tag{1}$$

from which precision and recall follow for a threshold τ,

$$P(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau)+\mathrm{FP}(\tau)}, \qquad R(\tau) = \frac{\mathrm{TP}(\tau)}{\mathrm{TP}(\tau)+\mathrm{FN}(\tau)} \tag{2}$$

and average precision is the area under the interpolated precision-recall curve,

$$\mathrm{AP}_\tau = \int_0^1 P_{\mathrm{interp}}\,(r)\,dr, \qquad P_{\mathrm{interp}}(r) = \max_{\tilde{r}\geq r} P(\tilde{r}) \tag{3}$$

The quantity that matters for small objects is the average over localisation strictness,

$$\mathrm{mAP}_{50:95} = \frac{1}{10}\sum_{k=0}^{9} \mathrm{AP}_{\tau_k}, \qquad \tau_k = 0.50 + 0.05k \tag{4}$$

Because bacilli occupy a very small fraction of each field, the class-imbalance term dominates training. Focal loss down-weights easy negatives through a modulating factor,

$$\mathcal{L}_{\mathrm{focal}}(p_t) = -\alpha_t(1-p_t)^\gamma \log p_t \tag{5}$$

with p_t the probability assigned to the true class and γ the focusing parameter. A smear study reports an explicit positive-to-negative ratio of one to four handled by class-weighted loss and minority oversampling [27].

5.3 Reported performance and the substitution problem

Table 3 collects the detection literature. The pattern that emerges is the central empirical finding of this review. Studies that report patch or image classification return figures near ninety-nine percent [28], whereas studies that report localisation on the same underlying problem return far lower numbers, and the gap widens as localisation strictness increases. BacilliFinder is the clearest instance, reporting mean average precision of 87.2 percent at a threshold of 0.5 but 50.5 percent when averaged from 0.5 to 0.95 [29]. Density-regression work reports average precision of 0.53 with an F1 score of 0.72 and is candid about it [30].

Table 3: Detection and segmentation methods for acid-fast bacilli, 2022-2026

Reference	Year	Family	Backbone or model	Reported result	Task scored
Panicker et al. [23]	2022	Candidate and classify	Shallow convolutional network	Recall 97.13, F1 0.98	Region classification
Chen et al. [31]	2024	One-stage	Improved YOLOv8s	mAP at 0.5 of 85.7, 26.9 frames per second	Detection
Soares et al. [32]	2024	Classification	Depthwise separable convolution with channel attention	F1 99.38	Patch classification
Venkatesh et al. [29]	2025	One-stage	BacilliFinder	mAP 0.5 of 87.2, mAP 0.5:0.95 of 50.5	Detection
Ding et al. [33]	2025	One-stage	YOLOv5, cloud deployed	mAP at 0.5 of 94.80	Detection
Sigit et al. [34]	2025	One-stage	YOLOv8	Precision 0.88, recall 0.77, F1 0.82	Detection
Kokane et al. [27]	2025	Classification	EfficientNetB0	Accuracy 92, sensitivity 89.1, AUC 94.5	Patch classification
Arasappan et al. [35]	2025	Pipeline	High-throughput digital microscopy	Reported at slide level	Screening
Genc et al. [28]	2026	Classification	InceptionV3 and Xception	99.00 across metrics	Patch classification
Saraf et al. [36]	2026	One-stage	YOLOv9-E with language model	Detection and reporting	Detection
Brito et al. [30]	2026	Density regression	Multi-task network	AP 0.53, F1 0.72	Detection and counting
Rulaningtyas et al. [18]	2026	Segmentation	Convolutional segmentation	Stain-thickness stratified	Segmentation

The substitution is not a matter of authors overclaiming. It is that the community reports whichever quantity its architecture naturally produces, and readers, including reviewers, then compare across incompatible units. A patch classifier evaluated on balanced crops is measured on a task where the object has already been found, localised and centred; a detector is measured on the task of finding it. Pooling both into a single accuracy range, as the systematic reviews necessarily do, produces an estimate that describes no realisable system.

A second and independent inflation arises from evaluation on curated fields. Several datasets in Table 8 comprise images selected by an expert as containing clearly visible bacilli, which removes from the test distribution exactly the fields that make

screening difficult: the over-stained, the debris-laden and the marginally focused. Reported sensitivity on such data is an upper bound whose distance from deployment performance is unmeasured. The independent clinical validations available [11], [13] are the only estimates of that distance, and both indicate it is substantial. Section 11 formalises the reporting standard that would remove the ambiguity.

6. QUANTIFICATION AND BACILLARY LOAD GRADING

6.1 Counting and the clump problem

Once bacilli are detected, the clinical quantity is a count over a defined field set. The obstacle is agglomeration. The annotation conventions of the field encode the difficulty rather than resolving it: ZNSM-iDB marks single bacilli with circles and occluded bacilli with rectangles [37], the largest annotated corpus labels approximately seven thousand one hundred individual bacilli and a further two hundred clusters as a distinct entity class [4], and the UFAM taxonomy separates true, agglomerated and doubtful objects [16]. Density regression avoids instance resolution entirely by integrating a predicted density map $D(x, y)$ over the field,

$$\hat{N} = \iint_{\Omega} D\,(x, y)\, dx\, dy \tag{6}$$

which is robust to overlap but discards instance identity [30]. A hybrid formulation, in which separable organisms yield instances and each cluster region R_j contributes a regressed count, gives

$$\hat{N} = \sum_{i=1}^{M} \mathbb{1}\,[s_i > \delta] + \sum_{j=1}^{J} g_{\phi}\,(R_j) \tag{7}$$

where s_i is the confidence of instance i, δ a threshold and g_{ϕ} a learned count head. No published smear system implements this, and the omission has a measurable clinical cost, described next.

The annotation record makes the scale of the problem explicit. In the largest annotated corpus, clusters account for roughly three percent of annotated entities but occur on the slides carrying the highest bacillary load [4], so their contribution to the count, and therefore to the grade, is far larger than their frequency suggests. Any system that treats a cluster as a single object systematically under-counts high-load slides, which compresses grades downward. That is precisely the direction of error which matters least for detecting disease and most for monitoring treatment response.

6.2 Load grading

The load grade is a monotone step function of the count per field set. Writing n for bacilli counted over F fields, the reporting rule has the form

$$G(n, F) = \begin{cases} 0 & n = 0 \\ 1 & 1 \le n \le 9 \text{ per } 100 \text{ fields} \\ 2 & 10 \le n \le 99 \text{ per } 100 \text{ fields} \\ 3 & 1 \le n \le 10 \text{ per field} \\ 4 & n > 10 \text{ per field} \end{cases} \tag{8}$$

with grades 0 to 4 denoting negative, scanty, 1+, 2+ and 3+. Two properties follow. The mapping is ordinal, so a scanty slide misreported as negative is a qualitatively different error from a 2+ slide misreported as 3+; and the decision boundaries are counts, so grading accuracy inherits counting error, which in turn inherits the clump problem.

Exactly one study automates this scale end to end, reporting mean precision, recall and F1 of 0.894, 0.896 and 0.895 across the five classes [3]. The clinical importance of the gap is shown by independent validation of a commercial system, which graded only 13.6 percent of 4+ smears and 35.7 percent of 3+ smears correctly while managing 81 percent of 2+ smears [13]: performance collapses exactly where bacilli are densest and most overlapped.

Because the target is ordinal, agreement should be reported with the quadratic weighted kappa,

$$\kappa_w = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}, \qquad w_{ij} = \frac{(i-j)^2}{(K-1)^2} \tag{9}$$

where O is the observed and E the expected agreement matrix over K grades. No study in Table 4 reports it.

Table 4: Counting and load-grading studies

Reference	Year	Output	Clump handling	Grading scale	Agreement statistic
Costa et al. [38]	2014	Annotated database	Separate class	None	None
Serrao et al. [3]	2024	Five-class grade	Not resolved	WHO five-point	Precision, recall, F1
Desruisseaux et al. [13]	2024	Grade, commercial	Not resolved	WHO five-point	Concordance 95.5 percent
Gomide et al. [4]	2025	Counts and boxes	Cluster as entity	Slide-level grade	None
Brito et al. [30]	2026	Count	Density integration	None	Mean absolute error
Zhou et al. [17]	2026	Slide decision	Not resolved	Positive or negative	Sensitivity

7. LEARNING UNDER LIMITED SUPERVISION

Exhaustive bacillus-level annotation is expensive and subjective, and unannotated regions may contain unlabelled positives, a hazard stated explicitly by the authors of the largest annotated corpus [4]. Four strategies appear in the literature.

a) Transfer learning: pretraining on natural images is near universal and is the implicit baseline of every study in Table 3.

b) Weak and semi-supervision: learning proceeds from image-level labels or sparse point annotations, and a graph-based semi-supervised approach has been applied to smears with limited expert annotation [39].

c) Foundation models: this is the most consequential recent development. A weakly supervised transformer built on a pathology foundation model achieves a precision-recall area under the curve of 0.943 to 0.974 using only image-level labels [40], which is the strongest published result on the largest smear corpus and sets the benchmark that supervised detectors must justify themselves against. Recent work adapts such models across centres continually [41].

d) Synthetic augmentation: generative modelling is established in adjacent bacterial microscopy, where a systematic comparison found downstream segmentation quality plateaued at approximately one hundred and ten real training images [42], and diffusion-based generation is now applied to pathology nuclei [43]. Conditional diffusion combined with graph regularisation has been applied to tumour segmentation in abdominal computed tomography [44], which shows that the same generative machinery transfers across modalities when a structural prior constrains the output. A necessary control is simple copy-paste augmentation, which suits small rod-shaped objects on near-uniform backgrounds almost ideally and costs nothing [45]; a generative pipeline that does not beat it has not demonstrated its value.

The foundation-model result deserves particular attention for what it implies about the supervision this field has been buying. Exhaustive bounding-box annotation of bacilli is the most expensive labelling task in the pipeline and the one whose subjectivity is least controlled, since expert readers disagree about doubtful objects and about where one organism ends and the next begins. That a model supervised only by image-level labels attains a precision-recall area under the curve above 0.94 [40] suggests the marginal value of exhaustive localisation labels for the screening decision is smaller than assumed. It does not follow that localisation is unnecessary, because counting and grading require it. It suggests instead a division of labour in which image-level labels are collected at scale for screening while scarce expert boxes are reserved for the counting stage, where they are indispensable.

Table 5: Supervision regimes in smear and adjacent microscopy analysis

Reference	Year	Regime	Supervision required	Domain	Reported result
Ghiasi et al. [45]	2021	Copy-paste augmentation	Instance masks	Natural images	Strong baseline gains
Pandit et al. [39]	2025	Semi-supervised graph	Sparse annotation	ZN smear	Reduced annotation need

Reference	Year	Regime	Supervision required	Domain	Reported result
Bedohazi et al. [40]	2025	Weakly supervised transformer	Image-level only	ZN smear	PR-AUC 0.943 to 0.974
Biofilm study [42]	2025	Generative augmentation	Annotated masks	Bacterial microscopy	Plateau near 110 real images
Zhu et al. [41]	2026	Foundation model adaptation	Continual, cross-centre	Pathology	Cross-centre retrieval
Zhang et al. [43]	2026	Diffusion augmentation	Text and mask conditioned	Pathology nuclei	Improved downstream task

8. MULTI-CENTRE AND PRIVACY-PRESERVING LEARNING

8.1 Motivation and topology

A national tuberculosis programme is organised as peripheral microscopy centres in districts reporting to a reference laboratory in a city. Slide images are patient data and cannot generally be pooled, which makes federated learning the natural training paradigm [46]. Figure 4 shows the resulting hub-and-spoke topology: clinics train locally and transmit parameter updates; the hospital aggregates and returns the global model; no image leaves its district.

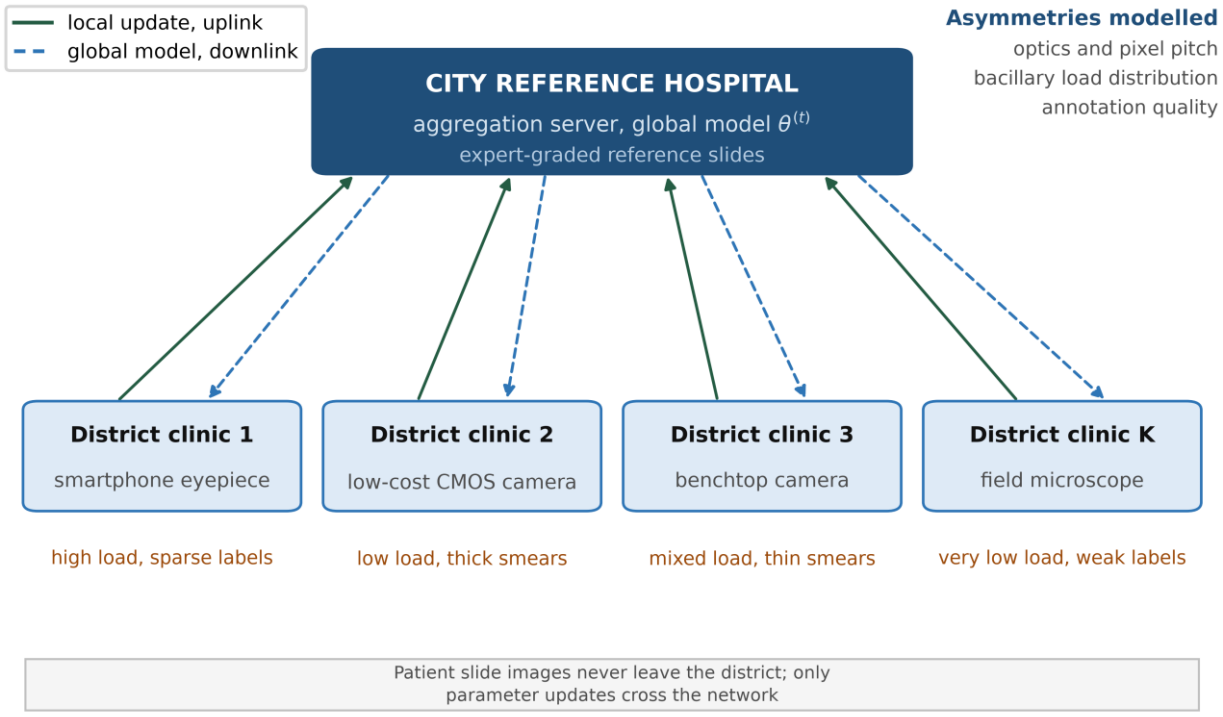

Fig. 4: Hub-and-spoke federated topology for smear microscopy

Federated optimisation minimises a weighted sum of local objectives,

$$\min_{\theta} \mathcal{F}(\theta) = \sum_{k=1}^{K} p_k \, \mathcal{F}_k(\theta), \qquad p_k = \frac{n_k}{\sum_{l=1}^{K} n_l} \tag{10}$$

where $\mathcal{F}_k$ is the empirical risk at clinic k holding n_k samples. The canonical aggregation step averages client parameters,

$$\theta^{(t+1)} = \sum_{k \in S_t} \frac{n_k}{\sum_{l \in S_t} n_l} \theta_k^{(t)} \tag{11}$$

over the participating subset S_t, and proximal variants add a penalty $\mu/2 \left\|\theta - \theta^{(t)}\right\|^2$ to each local objective to limit client drift. Feature-shift methods keep normalisation statistics local, motivated by the finding that batch normalisation is actively harmful under non-identically distributed data [47], while amplitude-harmonising methods align the frequency content of client images before local training [48]. Communication cost is a separate constraint on such deployments, and hybrid schemes that combine gradient sparsification with quantisation reduce the volume transmitted per round in medical imaging federations [49], which matters where peripheral centres connect over intermittent low-bandwidth links.

8.2 Heterogeneity specific to smear microscopy

The federated literature models label skew and quantity skew. Neither captures what varies across tuberculosis clinics. The clinically meaningful axis is bacillary load, which changes the ratio of foreground to background and therefore the loss geometry of a detector. We define the foreground-density skew of client k as

$$\rho_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{|\mathcal{B}_{k,i}|}{|\Omega_{k,i}|} \tag{12}$$

with $|\mathcal{B}_{k,i}|$ the annotated bacillus area and $|\Omega_{k,i}|$ the field area, and the federation-level heterogeneity as the dispersion of ρ_k across clients,

$$H_\rho = \sqrt{\frac{1}{K} \sum_{k=1}^{K} (\rho_k - \bar{\rho})^2}, \qquad \bar{\rho} = \frac{1}{K} \sum_{k=1}^{K} \rho_k \tag{13}$$

A Dirichlet partition of a single dataset cannot express H_ρ, which is why simulated federations are a poor proxy for this application.

Two further asymmetries follow from the topology rather than from the data. The first is optical: peripheral clinics operate lower-specification instruments, including smartphone-coupled eyepieces whose pixel pitch differs markedly from benchtop scanners, so the feature shift between hub and spoke is a device shift rather than the stain-colour shift the pathology literature usually models. The second is in supervision, since expert grading is concentrated at the reference centre, so the hub holds reliable labels while the periphery holds volume with weak or absent labels. The realistic problem is therefore semi-supervised federated learning with a labelled server, a setting that differs materially from the fully supervised federations of Table 6 and that has not been studied in any imaging modality with this structure.

8.3 Privacy guarantees and their limits

Federation alone is not a privacy guarantee. Differentially private training clips per-sample gradients and adds calibrated noise,

$$\tilde{g} = \frac{1}{B} \left(\sum_{i=1}^{B} g_i \,/ \max\left(1, \frac{\|g_i\|_2}{C}\right) + \mathcal{N}(0, \sigma^2 C^2 \boldsymbol{I}) \right) \tag{14}$$

yielding an (ε, δ) guarantee, and the utility cost is real and must be reported alongside ε, δ, the accountant and the clipping norm C [50], [51]. Gradient inversion attacks reconstruct training images from updates [52], though published attacks become far less effective once clients update batch-normalisation statistics [53], and defences continue to develop [54].

Table 6 summarises multi-centre learning evidence in medical imaging. The final row is the finding that motivates this review's agenda: no federated study exists on tuberculosis sputum smear microscopy.

Table 6: Federated and multi-centre learning evidence for smear microscopy

Reference	Year	Setting	Heterogeneity modelled	Privacy analysis	Real multi-site data
Sheller et al. [55]	2020	Brain tumour segmentation	Institutional	None	Yes
Rieke et al. [46]	2020	Position paper	Conceptual	Discussed	Not applicable
Dayan et al. [56]	2021	Clinical outcome prediction	Twenty sites	None	Yes

Reference	Year	Setting	Heterogeneity modelled	Privacy analysis	Real multi-site data
Liu et al. [48]	2022	Histopathology and radiology	Amplitude and feature shift	None	Yes
Adnan et al. [50]	2022	Histopathology	Institutional	Differential privacy	Yes
Hatamizadeh et al. [53]	2023	Medical imaging	Not applicable	Gradient inversion	Yes
Borazjani et al. [57]	2025	Cancer staging	Non-IID, unbalanced modality	None	Yes
Wang et al. [47]	2025	Theory and analysis	Non-IID normalisation	None	Simulated
Yu et al. [51]	2025	Medical imaging federation	Institutional	Privacy-utility trade-off	Yes
Nielsen et al. [52]	2026	Medical imaging	Not applicable	Gradient inversion framework	Yes
Tuberculosis smear microscopy	-	-	-	-	No study identified

9. ROBUSTNESS AND TRUSTWORTHINESS

Figure 5 organises reliability threats into three groups that differ in intent but share a clinical consequence: fewer detected bacilli means a lower reported grade, and a scanty-positive slide becomes a reported negative.

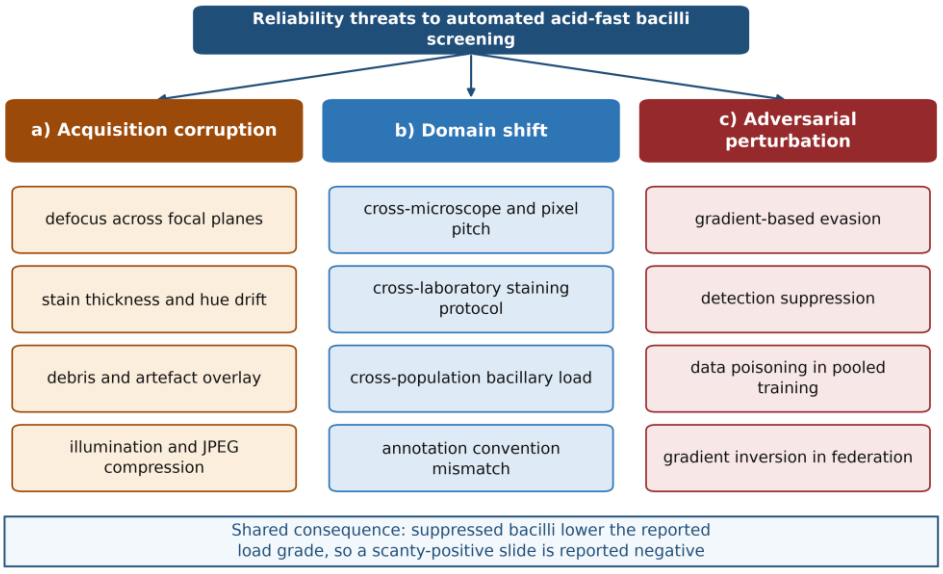

Fig. 5: Taxonomy of reliability threats to automated screening

a) Acquisition corruption: this is the credible and frequent threat, comprising defocus, stain-thickness drift, debris overlay, illumination change and compression. Benchmarks in the wider medical imaging community now evaluate models under systematic corruption suites [58], [59], and the smear literature has all the raw material to build a domain-specific equivalent, since real defocus can be sampled from published focal stacks rather than synthesised [16].

b) Domain shift: this is systematic rather than random and is the best-evidenced failure. Across one hundred and fifty-two artificial intelligence tuberculosis studies only one performed a domain-shift analysis [7]. In computational pathology the picture is quantified: adversarial and normalisation-based adaptation improve target-domain balanced accuracy substantially

over source-only training across four cancer types [60], survey guidance for domain generalisation is now available [61], test-time adaptation is emerging [62], and orthogonal-decomposition formulations refine adversarial adaptation [63]. Adversarial domain adaptation is often described as superseded, yet it remains demonstrably effective for optical rather than stain-colour shifts: a gradient-reversal network raised five-shot accuracy from 33.3 to 88.9 percent across magnification and modality changes in bacterial microscopy [64], which is precisely the shift separating a smartphone eyepiece rig from a benchtop scanner.

A practical difficulty in this category is that domain shift and acquisition corruption are not cleanly separable in smear microscopy. A laboratory that stains thickly does so consistently, so what appears as a corruption at the level of an individual image is a systematic domain property at the level of a site. This argues for evaluating the two together, with corruption severity sampled from the empirical variation observed across real sites rather than from a synthetic severity ladder, and it argues against reporting mean performance across sites without also reporting the worst site.

c) Adversarial perturbation: this is the worst-case bound. The standard threat model solves

$$\max_{\|\eta\|_\infty \le \epsilon} \mathcal{L}(f_\theta(x+\eta), y) \tag{15}$$

approximated by projected gradient descent,

$$\eta^{(t+1)} = \Pi_{\|\eta\|_\infty \le \epsilon}\left(\eta^{(t)} + \alpha\, \mathrm{sign}\big(\nabla_\eta \mathcal{L}(f_\theta(x+\eta^{(t)}), y)\big)\right) \tag{16}$$

Medical models are unusually fragile to this: at a perturbation budget of 0.002, accuracy fell by between 48.5 and 61.8 percent across computed tomography, mammography and magnetic resonance imaging, while natural-image benchmarks barely moved [65]. Two qualifications are essential for honest use in this domain. Medical adversarial examples are highly detectable, with detection area under the curve above ninety-eight percent [66]; and a systematic review of twenty-two studies concluded in a clinical journal that the threats remain more theoretical than practical [67], while an influential analysis found that no published medical robustness study had formulated an explicit threat model [68]. For sputum microscopy specifically the attack surface is narrow, since images are captured live from a physical slide that can be re-examined. The defensible use of adversarial perturbation here is therefore as a stress-testing instrument and worst-case bound rather than as a security narrative [69]-[71].

For a detector the clinically legible quantity is not a flipped label but a suppressed detection. Define the minimum perturbation that lowers the reported grade of slide s,

$$\epsilon^*(s) = \min\{\epsilon : G(\hat{N}(x_s+\eta)) < G(\hat{N}(x_s)),\ \|\eta\|_\infty \le \epsilon\} \tag{17}$$

Reporting ϵ^* stratified by true grade would express robustness in units a clinician can interpret. **d) Explainability:** this completes the category and remains a documented weakness of tuberculosis microscopy systems [1], with gradient-based attribution the usual instrument.

Table 7: Robustness and trustworthiness evidence

Reference	Year	Threat addressed	Domain	Key quantitative finding
Finlayson et al. [72]	2019	Adversarial, policy framing	Medical machine learning	Established threat framing
Tellez et al. [19]	2019	Stain variation	Pathology, nine laboratories	Colour augmentation essential
Ma et al. [66]	2021	Adversarial	Medical imaging	Detection AUC above 98 percent
Bortsova et al. [68]	2021	Adversarial, threat modelling	Multi-modality	No explicit threat models found
Joel et al. [65]	2022	Adversarial	Oncology imaging	48.5 to 61.8 percent accuracy loss

Reference	Year	Threat addressed	Domain	Key quantitative finding
Sun et al. [63]	2022	Domain adaptation	Medical imaging	Orthogonal decomposition
Yang et al. [62]	2022	Test-time adaptation	Cross-domain medical	Dynamic learning rate
Sorin et al. [67]	2023	Adversarial	Radiology, 22 studies	More theoretical than practical
Springenberg et al. [59]	2023	Robustness and generalisation	Medical imaging	CNN versus transformer comparison
Dong et al. [69]	2024	Attack and defence survey	Medical image analysis	Detection sparsely studied
AIDA study [60]	2024	Multi-centre generalisation	Histopathology	Adaptation beats source-only
Kanca et al. [70]	2025	Adversarial, transformers	Medical imaging	Adversarial training restores accuracy
Jahanifar et al. [61]	2025	Domain generalisation	Computational pathology	Survey and guidelines
Bhattacharya et al. [64]	2025	Optical domain shift	Bacterial microscopy	33.3 to 88.9 percent, five-shot
Shukla et al. [71]	2026	Adversarial defence	Medical imaging	Contrastive multitask defence

10. DATASETS AND BENCHMARKS

Table 8 compares every public Ziehl-Neelsen resource, with availability verified by direct request in August 2026 rather than by citation. Three observations follow.

First, availability is worse than the literature implies. The most cited resource, ZNSM-iDB, is no longer served at its published address, and studies published in 2025 still describe it as publicly available. Second, annotation conventions are mutually incompatible: one corpus marks circles for single organisms and rectangles for occluded ones [37], another separates true, agglomerated and doubtful objects [16], and a third draws tight boxes around each individually visible cell [14]. Cross-dataset numbers implicitly assume these define the same positive, which they do not. Third, only three resources ship a publisher-defined split, and no challenge, leaderboard or held-out test server exists for this task, in contrast to the tuberculosis radiograph community, where a fixed split and a public leaderboard have been available for several years. The absence of a held-out server matters more than the absence of a leaderboard, because it is what prevents the silent test-set reuse that inflates results over a literature's lifetime.

A further limitation cuts across all of these resources. Every one derives from a small number of collection sites, in several cases a single laboratory, so even the largest corpora provide little purchase on cross-site generalisation, and a model trained and tested within one of them measures interpolation rather than transfer. This is the structural reason the multi-centre evidence of Section 8 is thin, and it will not be resolved by any single institution releasing more images.

Table 8: Public Ziehl-Neelsen smear datasets and verified availability

Dataset	Year	Images	Annotation	Grading labels	Defined split	Availability
Costa et al. [38]	2014	1,320	Circles, rectangles, polygons	No	No	On request
ZNSM-iDB [37]	2017	About 1,800 in seven categories	Shape-coded overlays	No	No	Published host unreachable
UFAM ODS1 [16]	2026	90,200 in 8,200 focal stacks	None, focus modelling	No	No	Open, CC0

Dataset	Year	Images	Annotation	Grading labels	Defined split	Availability
UFAM DDS1 and DDS3 [16]	2026	31,484 patches, 8,000 mosaics	Masks	No	Yes	Open, CC0
UFAM DDS4 [16]	2026	250 field sets	Slide-level only	Five-class	No	Open, CC0
Gomide et al. [4]	2025	152,743 from 362 slides	7,100 bacilli, 200 clusters	Slide-level	Yes	On request
Yuniarti et al. [14]	2026	1,438	11,447 bounding boxes	No	Yes	Open, CC BY
Fluorescence or auramine	-	-	-	-	-	None identified

The absence of any public fluorescence dataset is consequential rather than incidental. A 2026 prospective comparison found automated ZN reading over one thousand fields detected 20.74 percent of cases against 18.75 percent for manual reading, whereas automated auramine reading over only three hundred fields reached 32.67 percent [11], which suggests brightfield ZN is contrast-limited rather than merely labour-limited. No public data exists to pursue that observation computationally.

11. EVALUATION AND REPORTING PRACTICE

The comparability problem identified in Sections 5 and 10 is a reporting problem with a precise remedy. We recommend four measures.

Detection should be reported as $\mathrm{mAP}_{50:95}$ rather than AP_{50}, since the difference between them is the small-object localisation error that clinical counting depends on. Screening operating points should be reported as sensitivity at a fixed false-positive rate per field,

$$\mathrm{Sens@FP}_f = R(\tau^*), \qquad \tau^* = \operatorname*{argmin}_{\tau} \left| \frac{\mathrm{FP}(\tau)}{F} - f \right| \tag{18}$$

because a threshold-free average precision hides the false-positive burden a microscopist would actually confirm. Counting should be reported as mean absolute error stratified by true load,

$$\mathrm{MAE}_g = \frac{1}{|\mathcal{S}_g|} \sum_{s \in \mathcal{S}_g} |\widehat{N}_s - N_s| \tag{19}$$

over the slides $\mathcal{S}_g$ of grade g, since aggregate error is dominated by high-load slides while clinical risk concentrates in low-load ones. Grading should be reported as quadratic weighted kappa with the full confusion matrix, and scanty-positive sensitivity should always be stated separately rather than absorbed into a macro average.

Three further practices would raise the evidential quality of this literature at negligible cost. Interval estimates should accompany every headline figure, since studies in this field frequently evaluate on fewer than two hundred images, where the sampling error on a sensitivity estimate is wider than most of the differences being claimed between methods. Results should be reported per acquisition site as well as in aggregate, because a mean over sites conceals precisely the variation that determines whether a model will work in the next clinic. And the unit of analysis should be stated explicitly: a result computed over patches, over fields and over slides are three different quantities, and the same model will score very differently on each because the number of opportunities to err differs by orders of magnitude between them.

None of these requires new data or new methods. They require only that the quantity being reported be named precisely enough for a reader to know what was measured, which is the minimum condition for the cumulative progress that a field of this size ought to be making.

12. OPEN CHALLENGES AND RESEARCH DIRECTIONS

a) Instance-level resolution of agglomerated bacilli: No published system resolves clumps; they are annotated as a separate class or dissolved into a density map. The hybrid objective of Section 6.1 is directly implementable on existing public data, and the commercial grading collapse at high load [13] establishes its clinical value.

b) Learned use of the focal stack: Focus is handled optically or by exclusion in every study reviewed. Eight thousand two hundred eleven-depth stacks are openly available [16], and no detector consumes them.

c) A shared benchmark: Harmonising the incompatible annotation conventions of Table 8 under one positive-object definition, one metric set and one frozen split is a prerequisite for meaningful comparison, and it is the single change that would most improve the field's evidential quality.

d) Federated learning with genuine heterogeneity: No federated study exists on this modality. The heterogeneity measure H_ρ of Section 8.2 is computable from published annotations, and datasets carrying per-image camera and stain-appearance metadata [14] permit physically real client partitions rather than Dirichlet simulation. The asymmetry worth exploiting is that expert grading exists principally at the reference centre, making the hub a labelled anchor and the periphery label-scarce, which is a semi-supervised federated problem rather than plain averaging.

e) Fluorescence microscopy and the modality gap: Every public resource in Table 8 is brightfield. Yet the prospective evidence indicates that the automated advantage is concentrated in fluorescence, where a third of cases were detected from three hundred fields against roughly a fifth from a thousand brightfield fields [11]. This asymmetry suggests that brightfield ZN is contrast-limited rather than merely labour-limited, and that the ceiling on brightfield automation may be lower than the field assumes. Releasing an annotated auramine corpus would be among the highest-value contributions available to any group with access to a fluorescence programme, and it is currently blocked by nothing more than the absence of a data-sharing decision.

f) Reliability auditing before deployment: Reporting ϵ^* stratified by grade, alongside performance under a domain-specific corruption suite and honest disclosure of the clean-accuracy cost of any robustness intervention, would give regulators and clinicians the evidence they currently lack. Any such intervention must be benchmarked against strong classical controls, specifically colour augmentation [19] and copy-paste augmentation [45], rather than against an unaugmented baseline.

13. CONCLUSION

Automated analysis of Ziehl-Neelsen sputum smears is a mature research area by volume and an immature one by evidence. The taxonomy developed here separates what has been solved from what has been substituted: detection of isolated bacilli in curated fields is largely solved, while counting under agglomeration, grading on the ordinal clinical scale, generalisation across laboratories and reliability under acquisition variation are not. The three quantitative anchors of this review, namely the collapse from 87.2 to 50.5 percent mean average precision under stricter localisation, the correct grading of only 13.6 percent of high-load smears by a commercial system, and the single domain-shift analysis among one hundred and fifty-two studies, describe the same underlying situation from three directions. Closing the gap requires no new imaging modality and no private data. It requires that the community measure the clinical quantity, on shared data, under the conditions of the clinics where the test is actually performed.

REFERENCES

[1] M. Zachariou, O. Arandjelović, and D. J. Sloan, "Automated Methods for Tuberculosis Detection/Diagnosis: A Literature Review," BioMedInformatics, vol. 3, no. 3, pp. 724-751, 2023, doi: 10.3390/biomedinformatics3030047.

[2] T. F. Mota Carvalho et al., "A systematic review and repeatability study on the use of deep learning for classifying and detecting tuberculosis bacilli in microscopic images," Progress in Biophysics and Molecular Biology, vol. 180-181, pp. 1-18, 2023, doi: 10.1016/j.pbiomolbio.2023.03.002.

[3] M. K. M. Serrão, M. G. F. Costa, L. B. M. Fujimoto, M. M. Ogusku, and C. F. F. Costa Filho, "Automatic bright-field smear microscopy for diagnosis of pulmonary tuberculosis," Computers in Biology and Medicine, vol. 172, p. 108167, 2024, doi: 10.1016/j.compbiomed.2024.108167.

[4] J. V. B. Gomide et al., "Image database with slides prepared by the Ziehl-Neelsen method for training automated detection and counting systems for tuberculosis bacilli," Journal of Medical Imaging, vol. 12, no. 03, 2025, doi: 10.1117/1.jmi.12.3.034505.

[5] K. Rastogi, L. Singh, A. Khan, P. P. Roy, and V. K. Iyer, "Deep learning-based detection of acid-fast bacilli in microscopy images: a systematic review," American Journal of Clinical Pathology, vol. 166, no. 1, art. no. aqag083, 2026, doi: 10.1093/ajcp/aqag083.

[6] G. Tamura, G. Llano, A. Aristizábal, J. Valencia, L. Sua, and L. Fernandez, "Machine-learning methods for detecting tuberculosis in Ziehl-Neelsen stained slides: A systematic literature review," Intelligent Systems with Applications, vol. 22, p. 200365, 2024, doi: 10.1016/j.iswa.2024.200365.

[7] S. Hansun et al., "Diagnostic Performance of Artificial Intelligence-Based Methods for Tuberculosis Detection: Systematic Review," Journal of Medical Internet Research, vol. 27, p. e69068, 2025, doi: 10.2196/69068.

[8] K. D. Lumamba, G. Wells, D. Naicker, T. Naidoo, A. J. C. Steyn, and M. Gwetu, "Computer vision applications for the detection or analysis of tuberculosis using digitised human lung tissue images - a systematic review," BMC Medical Imaging, vol. 24, no. 1, art. no. 298, 2024, doi: 10.1186/s12880-024-01443-w.

[9] J. von Bahr, A. Suutala, V. Diwan, A. Mårtensson, J. Lundin, and N. Linder, "AI-Supported Digital Microscopy Diagnostics in Primary Health Care Laboratories: Scoping Review," Journal of Medical Internet Research, vol. 28, pp. e78500-e78500, 2026, doi: 10.2196/78500.

[10] W. C. Chen, C. C. Chang, and Y. E. Lin, "Pulmonary Tuberculosis Diagnosis Using an Intelligent Microscopy Scanner and Image Recognition Model for Improved Acid-Fast Bacilli Detection in Smears," Microorganisms, vol. 12, no. 8, p. 1734, 2024, doi: 10.3390/microorganisms12081734.

[11] C. Zhang, H. Wang, Y. Chen, A. Shen, X. Yu, and H. Huang, "A smear test aiding with an automated slide reader acquires equivalent sensitivity to MGIT960 culture in pulmonary tuberculosis diagnosis," Microbiology Spectrum, vol. 14, no. 8, art. no. e03970-25, 2026, doi: 10.1128/spectrum.03970-25.

[12] D. J. S. RAVINDRAN and R. K. BASKARAN, "Enhancing tuberculosis diagnosis: A deep learning-based framework for accurate detection and quantification of TB bacilli in microscopic images," Tuberkuloz ve Toraks, vol. 73, no. 3, pp. 165-177, 2025, doi: 10.5578/tt.2025031113.

[13] C. Desruisseaux et al., "Retrospective validation of MetaSystems' deep-learning-based digital microscopy platform with assistance compared to manual fluorescence microscopy for detection of mycobacteria," Journal of Clinical Microbiology, vol. 62, no. 3, art. no. e01069-23, 2024, doi: 10.1128/jcm.01069-23.

[14] H. Yuniarti, C. Fatichah, R. Sigit, S. Arifin, and E. da Costa, "A comprehensive raw dataset of Ziehl-Neelsen-stained sputum smear microscopy images for Mycobacterium tuberculosis detection," Data in Brief, vol. 67, p. 112908, 2026, doi: 10.1016/j.dib.2026.112908.

[15] M. Costa, K. Pinto, L. Fujimoto, M. Ogusku, and C. Costa Filho, "Multi-focus image fusion for bacilli images in conventional sputum smear microscopy for tuberculosis," Biomedical Signal Processing and Control, vol. 49, pp. 289-297, 2019, doi: 10.1016/j.bspc.2018.12.018.

[16] M. G. F. Costa et al., "Collection: Datasets Applied to the Development of Automatic Tuberculosis Diagnostic Methods in Brightfield Sputum Smear Microscopy (Tb_UFAM)," IEEE Data Descriptions, vol. 3, pp. 463-469, 2026, doi: 10.1109/ieeedata.2026.3699270.

[17] Y. Zhou et al., "WSI AFB and AI deliver highest sensitivity for TB detection," Tuberculosis, vol. 158, p. 102747, 2026, doi: 10.1016/j.tube.2026.102747.

[18] R. Rulaningtyas et al., "Automated Mycobacterium tuberculosis Detection in Multivariant Digitized Ziehl-Neelsen Staining Using Faster R-CNN Method," International Journal of Biomedical Imaging, vol. 2026, no. 1, art. no. 6692222, 2026, doi: 10.1155/ijbi/6692222.

[19] D. Tellez et al., "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," Medical Image Analysis, vol. 58, p. 101544, 2019, doi: 10.1016/j.media.2019.101544.

[20] J. Wang et al., "Self-supervised stain normalization empowers privacy-preserving and model generalization in digital pathology," npj Digital Medicine, vol. 9, no. 1, art. no. 22, 2025, doi: 10.1038/s41746-025-02196-8.

[21] C. Xu et al., "Stain Normalization of Histopathological Images Based on Deep Learning: A Review," Diagnostics, vol. 15, no. 8, p. 1032, 2025, doi: 10.3390/diagnostics15081032.

[22] H. C. Huang, K. L. Kuo, M. H. Lo, H. Y. Chou, and Y. E. Lin, "Novel TB smear microscopy automation system in detecting acid-fast bacilli for tuberculosis - A multi-center double blind study," Tuberculosis, vol. 135, p. 102212, 2022, doi: 10.1016/j.tube.2022.102212.

[23] R. O. Panicker and M. K. Sabu, "Automatic Detection of Tuberculosis bacilli from Conventional Sputum Smear Microscopic Images Using Densely Connected Convolutional Networks," SN Computer Science, vol. 3, no. 4, art. no. 263, 2022, doi: 10.1007/s42979-022-01133-w.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 2017, doi: 10.1109/tpami.2016.2577031.

[25] T. Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 936-944, 2017, doi: 10.1109/cvpr.2017.106.

[26] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," in 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999-3007, 2017, doi: 10.1109/iccv.2017.324.

[27] C. Kokane, N. A. Deshpande, H. A. Bhute, H. J. Sarode, M. K. Kodmelwar, and S. Chavan, "Deep learning for automated sputum smear microscopy in tuberculosis diagnosis," Indian Journal of Tuberculosis, vol. 72, pp. S24-S30, 2025, doi: 10.1016/j.ijtb.2025.11.004.

[28] O. Genc, S. Genc, C. Ozdemir, and M. A. Gedik, "Detection of Mycobacterium tuberculosis in Ziehl-Neelsen Stained Sputum Smear Specimens Using Deep Learning Techniques," APMIS - Acta Pathologica, Microbiologica et Immunologica Scandinavica, vol. 134, no. 1, art. no. e70138, 2026, doi: 10.1111/apm.70138.

[29] N. Y, Venkatesh, R. G, and S. Basapur, "BacilliFinder: Revolutionizing Tuberculosis Detection with Computer Vision," Indian Journal of Tuberculosis, vol. 72, pp. S7-S17, 2025, doi: 10.1016/j.ijtb.2023.12.005.

[30] V. de Carvalho Brito, P. R. S. dos Santos, A. O. de Carvalho Filho, and J. O. B. Diniz, "Multi-task deep learning for bacilli detection and counting in sputum smear microscopy," Biomedical Signal Processing and Control, vol. 112, p. 108531, 2026, doi: 10.1016/j.bspc.2025.108531.

[31] H. Chen, W. Gu, H. Zhang, Y. Yang, and L. Qian, "Research on improved YOLOv8s model for detecting mycobacterium tuberculosis," Heliyon, vol. 10, no. 18, p. e38088, 2024, doi: 10.1016/j.heliyon.2024.e38088.

[32] R. R. Soares, F. P. Monteiro, G. M. Pinheiro, M. K. M. Serrão, C. F. F. Costa Filho, and M. G. F. Costa, "Evaluation of Depth-wise Separable Convolution and Channel Attention Mechanism to Bacilli Segmentation," in 2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1-5, 2024, doi: 10.1109/embc53108.2024.10781907.

[33] X. Ding et al., "Automated detection of mycobacterium tuberculosis based on cloud computing," BMC Infectious Diseases, vol. 25, no. 1, art. no. 1620, 2025, doi: 10.1186/s12879-025-11899-y.

[34] R. Sigit, H. Yuniarti, T. Karlita, R. Kusumawati, and F. H. Maulana, "Real-Time Tuberculosis Bacteria Detection Using YOLOv8," JOIV International Journal on Informatics Visualization, vol. 9, no. 5, p. 1988, 2025, doi: 10.62527/joiv.9.5.3147.

[35] A. Arasappan Murugesan, S. Mohan, and B. Parthiban, "Development of Automated High-Throughput Digital Microscopy With Deep Learning for Enhanced Blood Smear Imaging," Microscopy Research and Technique, vol. 89, no. 4, pp. 663-676, 2025, doi: 10.1002/jemt.70101.

[36] V. Saraf and R. P. Srivastava, "Automated tuberculosis detection using YOLOv9-E and Bio-GPT: A deep learning framework for computer-aided diagnosis and medical report generation," Indian Journal of Tuberculosis, 2026, doi: 10.1016/j.ijtb.2026.07.009.

[37] M. I. Shah et al., "Ziehl-Neelsen sputum smear microscopy image database: a resource to facilitate automated bacilli detection for tuberculosis diagnosis," Journal of Medical Imaging, vol. 4, no. 2, p. 027503, 2017, doi: 10.1117/1.jmi.4.2.027503.

[38] M. G. F. Costa, C. F. F. C. Filho, A. Kimura, P. C. Levy, C. M. Xavier, and L. B. Fujimoto, "A sputum smear microscopy image database for automatic bacilli detection in conventional microscopy," in 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2841-2844, 2014, doi: 10.1109/embc.2014.6944215.

[39] P. Pandit et al., "Harnessing semi-supervised graph-based learning to advance automated bacilli detection in digital tuberculosis microscopy with limited expert annotations," Indian Journal of Tuberculosis, vol. 72, pp. S151-S158, 2025, doi: 10.1016/j.ijtb.2025.11.017.

[40] Z. Bedőházi, A. Biricz, N. Foster, Y. E. Lin, and I. Csabai, "Streamlining tuberculosis detection with foundation model-based weakly supervised transformer," Computers in Biology and Medicine, vol. 195, p. 110554, 2025, doi: 10.1016/j.compbiomed.2025.110554.

[41] X. Zhu, Z. Jiang, K. Wu, J. Shi, and Y. Zheng, "Adapting pathology foundation models for continual cross-center WSI retrieval," Medical Image Analysis, vol. 113, p. 104213, 2026, doi: 10.1016/j.media.2026.104213.

[42] A. A. Holicheva et al., "Deep generative modeling of annotated bacterial biofilm images," npj Biofilms and Microbiomes, vol. 11, no. 1, art. no. 16, 2025, doi: 10.1038/s41522-025-00647-4.

[43] Y. Zhang, Q. Liu, Q. Chen, and X. Bai, "CLIP-Guided Generative network for pathology nuclei image augmentation," Medical Image Analysis, vol. 109, p. 103908, 2026, doi: 10.1016/j.media.2025.103908.

[44] K. R. Radhika, S. Gupta, M. J. Oli, D. Sharma, H. Vinutha, and H. N. Shenoy, "Pancreatic Tumor Segmentation using Conditional Diffusion and Graph Regularization on Abdominal Computed Tomography Images," Journal of Intelligent Decision Making and Information Science, vol. 3, no. 1s, pp. 1123-1143, 2026, doi: 10.59543/jidmis.v3.489.

[45] G. Ghiasi et al., "Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2917-2927, 2021, doi: 10.1109/cvpr46437.2021.00294.

[46] N. Rieke et al., "The future of digital health with federated learning," npj Digital Medicine, vol. 3, no. 1, art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.

[47] Y. Wang, Q. Shi, and T. H. Chang, "Why Batch Normalization Damage Federated Learning on Non-IID Data?," IEEE Transactions on Neural Networks and Learning Systems, vol. 36, no. 1, pp. 1692-1706, 2025, doi: 10.1109/tnnls.2023.3323302.

[48] M. Jiang, Z. Wang, and Q. Dou, "HarmoFL: Harmonizing Local and Global Drifts in Federated Learning on Heterogeneous Medical Images," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 1, pp. 1087-1095, 2022, doi: 10.1609/aaai.v36i1.19993.

[49] S. Gupta, I. S. Rajesh, C. Singh, U. N. Ranjitha, K. Siddesha, and P. K. Sekharamantry, "A Hybrid CEFL Approach Using Gradient Sparsification and Quantization for Medical Imaging," International Journal of Computational Intelligence in Engineering, vol. 1, no. 1, pp. 1-15, 2026.

[50] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," Scientific Reports, vol. 12, no. 1, art. no. 1953, 2022, doi: 10.1038/s41598-022-05539-7.

[51] Z. Yu, "Optimizing the Trade-off between Privacy and Utility in Medical Imaging Federated Learning," Radiology: Artificial Intelligence, vol. 7, no. 4, art. no. e250434, 2025, doi: 10.1148/ryai.250434.

[52] C. Nielsen, M. Wilms, and N. D. Forkert, "A novel gradient inversion attack framework to investigate privacy vulnerabilities during retinal image-based federated learning," Medical Image Analysis, vol. 107, p. 103807, 2026, doi: 10.1016/j.media.2025.103807.

[53] A. Hatamizadeh et al., "Do Gradient Inversion Attacks Make Federated Learning Unsafe?," IEEE Transactions on Medical Imaging, vol. 42, no. 7, pp. 2044-2056, 2023, doi: 10.1109/tmi.2023.3239391.

[54] L. Jiang, L. Ma, and G. Yang, "Shadow defense against gradient inversion attack in federated learning," Medical Image Analysis, vol. 105, p. 103673, 2025, doi: 10.1016/j.media.2025.103673.

[55] M. J. Sheller et al., "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," Scientific Reports, vol. 10, no. 1, art. no. 12598, 2020, doi: 10.1038/s41598-020-69250-1.

[56] I. Dayan et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," Nature Medicine, vol. 27, no. 10, pp. 1735-1743, 2021, doi: 10.1038/s41591-021-01506-3.

[57] K. Borazjani, N. Khosravan, L. Ying, and S. Hosseinalipour, "Multi-Modal Federated Learning for Cancer Staging Over Non-IID Datasets With Unbalanced Modalities," IEEE Transactions on Medical Imaging, vol. 44, no. 1, pp. 556-573, 2025, doi: 10.1109/tmi.2024.3450855.

[58] J. Yang et al., "MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification," Scientific Data, vol. 10, no. 1, art. no. 41, 2023, doi: 10.1038/s41597-022-01721-8.

[59] M. Springenberg, A. Frommholz, M. Wenzel, E. Weicken, J. Ma, and N. Strodthoff, "From modern CNNs to vision transformers: Assessing the performance, robustness, and classification strategies of deep learning models in histopathology," Medical Image Analysis, vol. 87, p. 102809, 2023, doi: 10.1016/j.media.2023.102809.

[60] M. Asadi-Aghbolaghi et al., "Learning generalizable AI models for multi-center histopathology image classification," npj Precision Oncology, vol. 8, no. 1, art. no. 151, 2024, doi: 10.1038/s41698-024-00652-4.

[61] M. Jahanifar et al., "Domain Generalization in Computational Pathology: Survey and Guidelines," ACM Computing Surveys, vol. 57, no. 11, pp. 1-37, 2025, doi: 10.1145/3724391.

[62] H. Yang et al., "DLTTA: Dynamic Learning Rate for Test-Time Adaptation on Cross-Domain Medical Images," IEEE Transactions on Medical Imaging, vol. 41, no. 12, pp. 3575-3586, 2022, doi: 10.1109/tmi.2022.3191535.

[63] Y. Sun, D. Dai, and S. Xu, "Rethinking adversarial domain adaptation: Orthogonal decomposition for unsupervised domain adaptation in medical image segmentation," Medical Image Analysis, vol. 82, p. 102623, 2022, doi: 10.1016/j.media.2022.102623.

[64] S. Bhattacharya, A. Wasit, J. M. Earles, N. Nitin, and J. Yi, "Enhancing AI microscopy for foodborne bacterial classification using adversarial domain adaptation to address optical and biological variability," Frontiers in Artificial Intelligence, vol. 8, art. no. 1632344, 2025, doi: 10.3389/frai.2025.1632344.

[65] M. Z. Joel et al., "Using Adversarial Images to Assess the Robustness of Deep Learning Models Trained on Diagnostic Images in Oncology," JCO Clinical Cancer Informatics, no. 6, art. no. e2100170, 2022, doi: 10.1200/cci.21.00170.

[66] X. Ma et al., "Understanding adversarial attacks on deep learning based medical image analysis systems," Pattern Recognition, vol. 110, p. 107332, 2021, doi: 10.1016/j.patcog.2020.107332.

[67] V. Sorin, S. Soffer, B. S. Glicksberg, Y. Barash, E. Konen, and E. Klang, "Adversarial attacks in radiology - A systematic review," European Journal of Radiology, vol. 167, p. 111085, 2023, doi: 10.1016/j.ejrad.2023.111085.

[68] G. Bortsova et al., "Adversarial attack vulnerability of medical image analysis systems: Unexplored factors," Medical Image Analysis, vol. 73, p. 102141, 2021, doi: 10.1016/j.media.2021.102141.

[69] J. Dong, J. Chen, X. Xie, J. Lai, and H. Chen, "Survey on Adversarial Attack and Defense for Medical Image Analysis: Methods and Challenges," ACM Computing Surveys, vol. 57, no. 3, pp. 1-38, 2024, doi: 10.1145/3702638.

[70] E. Kanca, S. Ayas, E. Baykal Kablan, and M. Ekinci, "Evaluating and enhancing the robustness of vision transformers against adversarial attacks in medical imaging," Medical & Biological Engineering & Computing, vol. 63, no. 3, pp. 673-690, 2024, doi: 10.1007/s11517-024-03226-5.

[71] S. Shukla and P. Gupta, "Elevating adversarial robustness by contrastive multitasking defence in medical image segmentation," Neural Networks, vol. 194, p. 108182, 2026, doi: 10.1016/j.neunet.2025.108182.

[72] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," Science, vol. 363, no. 6433, pp. 1287-1289, 2019, doi: 10.1126/science.aaw4399.