

Artificial Intelligence in Clinical Microbiology: Applications, Challenges, and Future Directions

Authors:

Dr. Tanya Sachan¹, Dr. Riya Singh², Dr. Swati Saxena¹, Dr. Tanya Pandey³, Dr. Rajat Verma⁴

¹MD, Assistant Professor, Department of Microbiology, Autonomous State Medical College, Amethi, Uttar Pradesh, India.

²MD, Department of Microbiology, Maharani Laxmibai Medical College, Jhansi, Uttar Pradesh, India.

³Department of Microbiology, Autonomous State Medical College, Auraiya, Uttar Pradesh, India.

⁴MCh, Department of Neurosurgery, PIMS, Lucknow, Uttar Pradesh, India.

Corresponding Author:

Dr. Tanya Sachan, MD, Assistant Professor, Department of Microbiology, Autonomous State Medical College
Amethi, Uttar Pradesh, India. <https://orcid.org/0009-0008-0745-9275>

<https://doi.org/10.5281/zenodo.21868666>

Received: 16 July 2026, Accepted: 03 August 2026, Online First: 10 August 2026

ABSTRACT

Clinical microbiology sits at an unusual inflection point: the diagnostic targets have not changed, but the tools available to detect, interpret, and act on them have expanded faster than most laboratories can absorb. Artificial intelligence (AI) has moved from a speculative research interest to a technology with documented, if uneven, clinical utility across microscopy, mass spectrometry, molecular diagnostics, and antimicrobial resistance (AMR) prediction. This narrative review critically examines AI across the clinical microbiology workflow, highlighting AI-assisted Gram stain interpretation, MALDI-TOF-based resistance prediction, genotype-to-phenotype AMR modelling, tuberculosis screening, mycology-specific applications, and the emerging but largely unvalidated role of large language models (LLMs) in laboratory and clinical practice. Recurring weaknesses include small single-centre datasets, near-total absence of external validation, and a persistent gap between published accuracy and real-world deployment. We address bias, explainability, regulatory, and low-resource-setting barriers, and conclude that the field's next phase of progress depends less on model novelty than on rigorous external validation and workflow integration designed with practising microbiologists.

Keywords: *Artificial intelligence; machine learning; deep learning; clinical microbiology; antimicrobial resistance; MALDI-TOF mass spectrometry; digital microbiology; large language models*

1. INTRODUCTION

Clinical microbiology has always been a discipline built on pattern recognition: a technologist reading a Gram stain, or a microbiologist interpreting a MALDI-TOF spectrum, is performing classification against a mental library of morphotypes. This pattern-recognition character has made microbiology unusually receptive to machine learning, and the recent inflection point reflects data availability and computational infrastructure rather than any change in the underlying diagnostic logic.[1,2] The case for adoption is reinforced by real operational pressure: a 2025 survey of the Italian Association of Clinical Microbiologists found that while 25.4% of laboratories reported some current AI/ML use, only 13.6% of professionals had a good working

understanding of AI/ML concepts and just 2.5% had access to a trained data scientist.[5] A parallel American Society for Microbiology survey found only 17% of laboratories already using AI for image analysis and 8% for quality monitoring, with many not expecting meaningful implementation before 2027 a finding attributed to cost and workforce readiness rather than scepticism about the technology itself.[6] Three converging pressures make this review timely: rising clinical stakes from antimicrobial resistance, given the well-established relationship between delayed appropriate therapy and excess mortality in serious infection; maturation of deep learning into regulatory-cleared commercial products, meaning the relevant question for most practising microbiologists is no longer

whether the technology works in principle but whether a specific validated product improves accuracy, turnaround time, or cost in their own setting; and the emergence since 2022 of generative AI and LLMs, an entirely new application category whose limitations are considerably less well characterised than the image- and spectrum-based tools this review opens with. This review is organised around the practical architecture of a clinical microbiology laboratory rather than AI methodology in the abstract, on the view that clinicians and laboratorians are better served by a workflow-organised synthesis than a taxonomy of neural network architectures, prioritising critical appraisal over comprehensive cataloguing: where studies conflict, report implausibly high single-centre accuracy, or lack external validation, we say so explicitly.

• **Aim and Objective of the Review**

The aim of this narrative review is to critically evaluate the current and emerging applications of artificial intelligence across the clinical microbiology workflow, including microscopy, MALDI-TOF mass spectrometry, antimicrobial resistance prediction, molecular and genomic diagnostics, infectious disease surveillance, mycology, and large language models. Particular emphasis is placed on the quality of available evidence, external and prospective validation, clinical applicability, interpretability, implementation barriers, ethical and regulatory considerations, and the potential role of AI in resource-limited settings. The review also identifies key evidence gaps and proposes priorities for responsible clinical translation and future research.

• **Review Methodology**

A narrative literature review was conducted to synthesize current evidence on the applications, clinical utility, limitations, and future directions of artificial intelligence in clinical microbiology. Relevant literature was identified through searches of major biomedical and multidisciplinary databases, including PubMed/MEDLINE, Scopus, Web of Science, and Google Scholar, using combinations of terms related to artificial intelligence, machine learning, deep learning, clinical microbiology, antimicrobial resistance, MALDI-TOF mass spectrometry, microscopy, tuberculosis, mycology, molecular diagnostics, and large language models. Priority was given to peer-reviewed studies, systematic reviews, validation studies, clinical evaluations, and relevant benchmark or consensus literature. Recent publications were emphasized for rapidly evolving areas such as generative AI and large language models. Because this was a narrative review, studies were selected for clinical relevance and evidentiary value rather than according to a formal systematic-review screening protocol.

• **Search Strategy and Study Selection**

The literature search focused primarily on publications available up to July 2026. Searches were conducted using combinations of keywords related to artificial intelligence, machine learning, deep learning, clinical microbiology, antimicrobial resistance, MALDI-TOF mass spectrometry, microscopy, tuberculosis, fungal diagnostics, molecular diagnostics, whole-genome sequencing, metagenomics, natural language processing, and large language models. Boolean combinations of these terms were adapted according to the requirements of each database.

Priority was given to peer-reviewed original research articles, systematic and narrative reviews, clinical validation studies, multicentre studies, methodological benchmarking studies, consensus or guidance documents, and relevant high-quality reports addressing clinical implementation. Publications were considered particularly relevant when they evaluated diagnostic performance, external validation, prospective validation, clinical utility, workflow integration, antimicrobial resistance prediction, or patient-safety implications. Studies with limited clinical relevance, purely technical applications without a microbiological or diagnostic context, duplicate reports, and publications for which the available information was insufficient to assess the reported application were given lower priority or excluded from the narrative synthesis.

Because this was a narrative review, no formal systematic-review screening algorithm or meta-analysis was performed. Study selection and synthesis were based on clinical relevance, methodological quality, validation status, and applicability to contemporary clinical microbiology practice. Particular attention was given to studies reporting external or cross-site validation because generalisability remains a major limitation in clinical AI research.

2. Fundamentals of Artificial Intelligence for Clinical Microbiology

A working vocabulary matters because a paper's claim to use "AI" often obscures which of several quite different techniques was applied, and hence what validation ought to be expected. Machine learning (ML) random forest, gradient-boosted trees, support vector machines remains the dominant approach for structured data such as MALDI-TOF peak lists or genomic feature matrices, and frequently matches or outperforms deep neural networks on these tabular tasks, undercutting the assumption that "deeper is always better." [9,10] Deep learning (DL), specifically convolutional neural networks (CNNs) and, increasingly, vision transformers, is the workhorse for image-based tasks; transformers use self-attention to localise relevant image regions and have begun

outperforming CNNs on Gram stain classification while eliminating labour-intensive manual patch annotation.[11,12] Natural language processing (NLP) extracts structured information from unstructured clinical text and has a longer, more clinically embedded track record than generative AI proper. Generative AI and large language models (LLMs) produce novel text or content rather than classifying from fixed categories, and are addressed at length in Section 7 given both rapid clinical uptake since 2022 and a disproportionate number of open safety questions. (Table 1) Explainable

AI (XAI) SHAP values, saliency maps, attention visualisation is not a peripheral concern: a resistance-prediction model that cannot indicate which spectral peaks or genomic features drove a prediction is difficult for a microbiologist to trust or sanity-check against known resistance mechanisms. Computer vision, the applied subfield concerned with extracting clinically meaningful information from images, underlies essentially all microscopy, colony-counting, and digital mycology applications discussed in Section 3.

Table 1. Major AI Techniques Used in Clinical Microbiology

Technique	Description	Representative Microbiology Application	Illustrative Reference
Classical machine learning (random forest, SVM, gradient-boosted trees)	Learns statistical patterns from structured/tabular data; competitive accuracy with high interpretability	MALDI-TOF spectrum-based resistance prediction; genomic AMR genotype-to-phenotype prediction	Weis et al. 2022 (DRIAMS); Hu et al. 2024
Convolutional neural networks (CNNs)	Deep learning architecture specialised for image feature extraction via convolutional filters	Gram stain whole-slide classification; TB chest X-ray screening; fungal microscopy classification	Smith et al. 2024; TB CAD meta-analysis 2025
Vision transformers	Attention-based deep learning architecture; localises relevant image regions without manual patch annotation	Large-scale Gram stain WSI characterisation	McMahon et al. 2025
Natural language processing (NLP)	Extracts structured information from unstructured clinical/laboratory text	Electronic ordering support; healthcare-associated infection surveillance definitions	Institutional infection-prevention literature
Large language models (LLMs)	Transformer-based generative models producing human-like text from large training corpora	Clinical decision support; laboratory report drafting; medical education	GP-vs-LLM antibiotic study 2025 (JAC)
Explainable AI (SHAP, saliency maps, attention visualisation)	Post-hoc methods rendering model predictions interpretable to human experts	Interpreting resistance-prediction feature importance; fungal pathogen classification interpretability	Nature Scientific Reports 2026 XAI-fungal study
Federated learning	Trains a shared model across institutions without centralising patient-level data	Proposed solution to multi-site training data scarcity (Section 9)	Emerging direction, not yet microbiology-specific mature literature

3. AI Across the Clinical Microbiology Workflow (Table 2,3)

Pre-analytical applications NLP-based electronic ordering support and ML-based specimen rejection have a low-controversy but under-published track record relative to their evident operational value. Microscopy is the most extensively and rigorously studied analytical application. A widely cited CNN-based system classifying Gram stains from positive blood cultures into seven classes reported 92.5% overall accuracy, with sensitivity up to 97% for Gram-negative rods and clusters but weaker performance for Gram-positive chains, which were frequently misclassified as

clusters.[13] A 2025 vision transformer framework, trained without patch-level manual annotation, reported comparable accuracy (92.5%) on an independent 516-slide dataset.[11] The consistency of this ~92–95% figure across independently developed systems including separate CNN-based “big transfer” classifiers and image-processing pipelines reporting up to 99% accuracy on curated datasets is reassuring, but the foundational 2018 study in this space made an important caveat explicit: reported accuracy is frequently inflated where training and test data are not fully separated at the patient or slide level, a weakness not fully resolved in every subsequent study.[15] Gram stain-based

classification has been extended beyond blood cultures to urine microscopy, where a non-inferiority comparison of a smartphone-based computer-aided diagnosis app against ten microbiology specialists, trained on over 13,000 images from two hospitals, is notable for benchmarking directly against multiple human experts rather than a single reference standard, and to Nugent scoring for bacterial vaginosis, where one CNN achieved 82% sensitivity and 97% specificity for three-category classification but degraded meaningfully on four-category classification and altered vaginal flora.[16,17] MALDI-TOF organism identification is mature and near-universal; the more contested frontier is resistance prediction directly from spectra. The foundational DRIAMS dataset study demonstrated feasibility for

several species-antibiotic combinations,[9] but a 2025 independent validation explicitly designed to replicate this work found that cross-site AUROC dropped meaningfully relative to within-site performance, concluding that ML prediction from MALDI-TOF spectra is “promising but still not acceptable as an alternative for AST in routine diagnostics” without site-specific training and regular retraining a direct empirical demonstration of the external-validation problem this review returns to repeatedly.[19] Genotype-to-phenotype AMR prediction from whole-genome sequencing, automated colony/culture interpretation, molecular and metagenomic sequencing, and post-analytical stewardship decision support are addressed as cross-cutting themes in Sections 5–6 rather than repeated here.

Table 2. Applications of AI Across the Clinical Microbiology Workflow

Workflow Phase	Application	Predominant Technology	Current Evidence Level
Pre-analytical	Electronic order review, specimen rejection flagging	NLP, rule-based ML	Operationally embedded; limited standalone validation literature
Analytical Microscopy	Gram stain interpretation, AFB smear, malaria/blood smear, fungal microscopy	CNN, vision transformer	Multiple independently validated single/few-site studies; ~92–95% accuracy consistently reported
Analytical Culture	Automated colony detection/counting, mixed-culture recognition	Computer vision (CNN)	Embedded in commercial total lab automation; standalone peer-reviewed validation limited
Analytical MALDI-TOF	Organism ID (mature); resistance prediction from spectra (active research)	Classical ML (CatBoost, LightGBM, RF), MLP	Organism ID mature; resistance prediction promising but not yet routine-ready per cross-site validation data
Analytical Molecular	PCR/syndromic panel interpretation, metagenomic and nanopore sequencing	Rule-based ML, NLP, sequence classifiers	PCR interpretation mature; metagenomic surveillance emerging (single strong case example)
Analytical – AST	MIC prediction, genotype-to-phenotype resistance prediction, image-based rapid AST	ML/DL on genomic and imaging data	Active research area; multi-causal resistance mechanisms limit ceiling of marker-based prediction
Analytical Bioinformatics	Genome assembly, variant calling, outbreak phylogenetics	ML-assisted bioinformatic pipelines	Methodologically mature pipelines; ML components newer but well benchmarked
Post-analytical	Automated reporting, decision support, stewardship alerts, infection control	NLP, ML risk models, LLM-assisted reporting	Personalised AST-informed prescribing shows benefit; broader stewardship integration still emerging

Table 3. Commercial and Named AI Platforms Referenced in the Literature

Platform / System	Application Area	Notes
CAD4TB	Chest X-ray-based tuberculosis screening	Deployed in multinational migrant/community TB screening programmes; included in 2025 meta-analysis of 5 CAD products
qXR	Chest X-ray-based tuberculosis and abnormality screening	CE-marked; externally validated for normal/abnormal CXR stratification
Lunit INSIGHT	Chest X-ray-based	Evaluated alongside CAD4TB and qXR in multi-

Platform / System	Application Area	Notes
	tuberculosis/pulmonary abnormality detection	product TB screening comparisons
VITEK® (bioMérieux) + MS ML resistance-prediction add-ons	MALDI-TOF organism identification; research-stage resistance prediction	MS platform itself mature and widely deployed; ML resistance-prediction layer remains a research application, not yet a routine cleared feature in most markets
DRIAMS (Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra)	Public benchmark dataset (not a commercial product)	De facto research benchmark for MALDI-TOF AMR prediction studies; originated in Switzerland
BIOTIA-DX RESISTANCE	Genomic antimicrobial resistance phenotype prediction	Top-performing tool in the CAMDA AMR 2025 international benchmarking competition
Consumer/enterprise LLMs (ChatGPT, Claude, Gemini, Copilot, Med-PaLM, Mistral AI, Llama)	General and medical natural-language clinical decision support, reporting assistance, education	Not microbiology-specific products; performance and guideline-concordance vary by model and by national context (Section 7)

4. AI in Specific Infectious Diseases

Tuberculosis has, alongside Gram stain interpretation, the most mature AI evidence base of any syndrome, driven by global need: TB caused an estimated 1.25 million deaths in 2023, and radiologist shortages in high-burden countries motivate chest X-ray computer-aided detection (CAD) tools such as CAD4TB, qXR, and Lunit INSIGHT, now deployed in multinational migrant and community screening programmes.[25,26] A 2025 systematic review and meta-analysis pooling five such AI products, drawn from 21 studies identified across four databases, found generally strong diagnostic accuracy but clinically important variability across tools and screening contexts, concluding that threshold adjustment is likely necessary for different epidemiological settings rather than one universal operating point a nuance easily lost when marketing materials cite only best-case figures.[27] It is worth distinguishing this imaging-based screening literature from AI applied directly to sputum smear microscopy for acid-fast bacilli detection, sometimes described as microbiological rapid on-site evaluation, which remains a separate and comparatively less mature application for which purpose-built clinical datasets are scarce relative to the very large chest X-ray datasets available for the imaging application.[28] Bloodstream infection is the syndrome where AI-assisted Gram stain interpretation has the clearest stakes: each hour of delay to appropriate antimicrobial therapy is associated with measurably increased 30-day mortality in bacteraemia, providing the clinical rationale for rapid-diagnostics AI even where the AI papers themselves rarely measure mortality directly.[24] Respiratory infection applications beyond tuberculosis span chest imaging classification for pneumonia and COVID-19, a literature that expanded

enormously during 2020–2022 and has since matured into incremental, comparative benchmarking across CNN architectures, alongside automated interpretation of respiratory syndromic PCR panels. Sexually transmitted infection diagnostics remain comparatively underdeveloped, with most existing work concentrated on image-based genital lesion screening and considerably less published, externally validated evidence than exists for bloodstream infection or tuberculosis. Healthcare-associated infection surveillance, built on NLP extraction of structured and unstructured electronic health record data, is among the more operationally embedded infection-prevention AI applications currently in use, though rigorous head-to-head comparison against traditional manual surveillance definitions remains less common than the technology's maturity might suggest it should be.

5. AI in Mycology

Medical mycology has received disproportionately little dedicated AI investment relative to bacteriology, despite invasive fungal disease often carrying higher case fatality than bacterial bloodstream infection and despite emerging multidrug-resistant fungal pathogens.[29] Because *Candidozyma* (formerly *Candida*) *auris* cannot be reliably identified by conventional biochemical methods, it has attracted particular novel-diagnostic interest: a 2024 study combining Raman spectroscopy with machine learning achieved single-cell-level identification and clade differentiation, and separately built classifiers detecting antifungal resistance phenotypes and distinguishing aggregating from filamentous cell morphology directly from single-cell spectra a level of resolution well beyond what conventional phenotypic methods achieve.[30]

Electronic-nose technology, classifying volatile organic compound profiles by machine learning rather than mass spectrometry, has also been explored as a lower-cost alternative *Candida* identification route, though this remains an early-stage research direction rather than a deployed clinical tool.[31] A *Cryptococcus* proof-of-concept combining a switch-controllable magnetic nanoparticle capture system with CNN-based image analysis reported a detection limit of 1–4 cells/mL, 97% specificity, and six-minute total detection time, illustrating AI's frequent role interpreting the output of a novel capture technology rather than conventional culture material alone.[34] Deep learning-based *Aspergillus* detection directly from stereomicroscopy images was demonstrated as a proof-of-concept as early as 2021, and a 2024 systematic review specifically addressing AI utility in fungal detection and typing, with a stated eye towards future fungal cytology applications, together with a 2026 explainable-AI study applying interpretability methods to fungal pathogen computer vision classification, indicate a field still substantially at the proof-of-concept stage relative to bacteriology rather than one approaching routine clinical deployment.[33,35,37] Antifungal resistance prediction from MALDI-TOF spectra remains markedly less mature than its bacterial counterpart, with systematic multi-drug-class validation still largely absent one of the clearest actionable gaps identified in this review.[21]

6. AI and Antimicrobial Resistance (Table 4)

Genotype-to-phenotype resistance prediction from whole-genome sequencing has been demonstrated across many species-antibiotic combinations: representative examples include random forest-based prediction across twelve antimicrobial classes in *Pseudomonas aeruginosa*, gene-, SNP-, and k-mer-based feature representations predicting resistance to ten antibiotics in *Escherichia coli*, and comparative benchmarking using a real-world German dataset of 987 *E. coli* strains alongside external

public data.[38,39,40] A 2024 systematic benchmarking study makes an important point explicit: resistance is often genuinely multi-causal fluoroquinolone resistance in *E. coli*, for example, can arise from mutations in either *gyrA* or *parC*, among other mechanisms meaning the absence of a single known marker is frequently insufficient to rule out a resistant phenotype, limiting the ceiling of marker-only prediction regardless of data volume.[10] Similarly, MALDI-TOF-based prediction of methicillin resistance and multiple tier-1 antibiotics in *Staphylococcus aureus*, benchmarked across both the DRIAMS and an independent Taiwanese dataset, found that predictive performance varied by both database and algorithm choice, with LightGBM showing a specific advantage for oxacillin-resistant isolates again reinforcing that no single algorithm or dataset can be treated as a universal solution.[20] The CAMDA AMR competition has run annual international benchmarking challenges evaluating genomic resistance-phenotype prediction across thousands of genomes and multiple species-drug pairs, providing one of the few standardised, externally benchmarked comparisons in this literature.[42] One Health and wastewater surveillance represent a genuinely novel AI-enabled surveillance modality: a 2025 Quebec City study combined AI-driven social media “community analytics” with long-read nanopore metagenomic sequencing of communal wastewater to monitor non-typhoidal salmonellosis, with both surveillance streams detecting a signal peak in the same week, roughly five weeks before an official food recall was issued an approach with no clear pre-2020 precedent.[22] On stewardship, a 2025 systematic review of AI/ML tools guiding antibiotic prescribing, spanning 23 eligible studies across six databases, concluded that AI shows genuine promise, particularly for resistance prediction and optimising antibiotic selection, but that current limitations mean infectious diseases specialists retain an essential role that automated tools cannot yet replace.[7]

Table 4. AI Approaches for Antimicrobial Resistance

Data Type	Method	Example Pathogen/Drug	Reported Performance / Finding	Reference
MALDI-TOF spectra	CatBoost, Random Forest, SVM, Logistic Regression	<i>E. coli</i> (general resistance)	AUROC 0.91, F1 0.78 for <i>E. coli</i> (single-site)	Int J Mol Sci 2025
MALDI-TOF spectra (DRIAMS benchmark)	Multiple ML learners (LR, RF, XGB, LGBM, SVM, MLP)	<i>S. aureus</i> , <i>E. coli</i> , <i>K. pneumoniae</i>	AUROC ≥0.8 within-site; degraded cross-site performance	Weis et al. 2022; Münster validation 2025
MALDI-TOF spectra	LightGBM, SHAP attribution, MLP feature	<i>S. aureus</i> (<i>mecA</i> oxacillin resistance)	Accuracy/AUROC/AUPR up to 0.95/0.87/0.97 depending on antibiotic	Wang et al. 2024
MALDI-TOF spectra	PCA + multilayer perceptron	7 yeast species, 3 antifungal classes	Best and most stable performance among	bioRxiv 2025 (antifungal)

Data Type	Method	Example Pathogen/Drug	Reported Performance / Finding	Reference
			dimensionality-reduction approaches tested	
Whole-genome sequencing	Random Forest, BioWeka, Logistic Regression	<i>Pseudomonas aeruginosa</i> (12 antimicrobial agents)	44.66% isolates resistant across panel; classification accuracy varied by drug	Comput Intell Neurosci 2023
Whole-genome sequencing	Gene-, SNP-, k-mer-based representations, multiple ML algorithms	<i>E. coli</i> (10 antibiotics)	Gene-based and integrated multi-feature approaches most robust	Frontiers Microbiol 2026
Whole-genome sequencing (benchmarking)	Rule-based marker detection vs. ML approaches, systematically compared	Multiple species-drug pairs	Multi-causal resistance mechanisms limit marker-only prediction ceiling	Hu et al. 2024, Brief Bioinform
Whole-genome sequencing (competition benchmark)	Multiple ML approaches, international competition format	9 species-drug pairs, 5,345 genomes	Top-performing tool: BIOTIA-DX RESISTANCE (CAMDA AMR 2025)	CAMDA AMR 2025

7. Generative AI and Large Language Models

Generative AI is the newest and least settled category in this review. A rigorous 2025 comparative evaluation tested six LLMs (ChatGPT, Gemini, Copilot, Mistral AI, Claude, Llama 3.1) against practising general practitioners on antibiotic-prescribing vignettes across four countries, benchmarked against each country's own guidelines; performance varied meaningfully across both models and countries, underscoring that LLM accuracy is not a fixed property of a given model but depends on the specific question and national guideline context – a finding with direct implications for any institution assuming a model validated elsewhere will perform equivalently locally.[46] LLMs have been extensively benchmarked against medical licensing examinations, with newer models reaching up to 93% accuracy on USMLE Step 2 CK, though the same study noted persistent overconfidence and hallucination alongside strong raw accuracy, and cautioned that prior training exposure to public exam content limits generalisability to genuinely novel clinical reasoning.[47] Subspecialty-level benchmarking is beginning to extend this evaluation paradigm beyond general medical knowledge: a 2025 comparison of Claude and ChatGPT-4 on cardiothoracic anaesthesia board-style questions is one such example, a direction infectious diseases and clinical microbiology would benefit from replicating with purpose-built, non-public question sets given the risk of training-data contamination on any publicly available question bank.[48] The most consistently reported LLM limitation is hallucination, and infectious diseases consultation has been specifically flagged as high-stakes: a widely discussed commentary described “falsehood mimicry,” in which a model improvises a plausible

answer to a false premise rather than challenging it, citing an example where ChatGPT failed to challenge a request for oral antibiotic discharge planning for bacteraemia caused by a fictitious organism illustrating why a compliant, agreeable conversational style becomes a patient-safety liability rather than a convenience in this context.[49] A 2025 study went further, demonstrating that multiple LLMs are highly vulnerable to adversarial hallucination attacks specifically during simulated clinical decision support tasks, meaning hallucination is not merely a passive failure mode but one that can, in principle, be deliberately triggered.[50] A broader 2024 systematic review of NLP and LLM applications in infectious disease diagnosis and prediction reached a similar conclusion from a different angle: performance on narrow, well-defined tasks can be strong, but generalisability and rigorous external validation remain the field's principal unmet need language nearly identical to this review's own conclusions about the MALDI-TOF and genomic AMR prediction literature, suggesting the external-validation gap is a property of the field as a whole rather than any single technique.[51]

8. Challenges

Training data across nearly every application in this review are geographically skewed towards well-resourced centres in North America, Western Europe, and East Asia, risking degraded performance precisely where infectious disease and AMR burden is highest. Dataset quality and small sample sizes remain common, particularly in mycology, and ground-truth labels for tasks such as mixed-culture recognition are themselves often a matter of expert disagreement rather than an

unambiguous standard. Lack of external validation is, in this review's assessment, the single most consistent weakness: the Münster MALDI-TOF study's cross-site performance degradation is the clearest documented example of a group directly testing and honestly reporting this gap rather than asserting generalisability.[19] Regulatory pathways for AI-based diagnostic software remain unevenly developed across jurisdictions, with the evidentiary bar for clearance of a resistance-prediction algorithm still being worked out case by case rather than following one settled framework. Ethical questions include the degree of clinician oversight required before an AI-flagged result can directly inform empiric therapy without independent confirmation, and how liability is apportioned when an AI-assisted decision contributes to harm. Explainability limitations reduce a microbiologist's ability to sanity-check resistance predictions against known mechanisms, particularly for deep learning as opposed to classical ML components. Laboratory information systems increasingly integrated with cloud-based AI inference introduce a new cybersecurity attack surface that has received comparatively little dedicated attention, and commercial cloud-hosted LLMs used with patient-identifiable data raise data-retention questions most institutions have not yet resolved at a policy level. Both the Italian and ASM surveys identified capital cost and absent data-science expertise as bigger adoption barriers than model accuracy itself.[5,6] Implementation in low-resource settings is structurally mismatched: the

countries with the greatest infectious disease and AMR burden are also least likely to have the digital scanners, MALDI-TOF instruments, or sequencing infrastructure most published tools assume.(Table 5)

• **Strengths of the Review**

This review has several strengths. First, it adopts a clinically oriented, workflow-based framework that links AI technologies with specific stages of clinical microbiology rather than discussing algorithms in isolation. Second, it integrates evidence across microscopy, MALDI-TOF mass spectrometry, molecular and genomic diagnostics, antimicrobial resistance prediction, infectious disease surveillance, mycology, and generative AI, thereby providing a broad overview of both established and emerging applications. Third, the review places particular emphasis on external validation, cross-site performance, clinical implementation, and patient-safety considerations rather than relying solely on reported accuracy from individual datasets. Fourth, it considers practical barriers including explainability, regulatory requirements, cybersecurity, workforce preparedness, cost, and implementation in low-resource settings. Finally, the review distinguishes promising proof-of-concept applications from technologies that have demonstrated sufficient evidence for clinical deployment, helping to avoid equating technical performance with clinical readiness.

Table 5. Advantages and Limitations of AI in Clinical Microbiology

Advantage	Corresponding Limitation
Rapid turnaround relative to conventional phenotypic methods (e.g., resistance prediction in hours vs. 24–48h culture-based AST)	Predictive performance not yet consistently sufficient to replace phenotypic AST in routine diagnostics
High within-site classification accuracy for image-based tasks (Gram stain, TB chest X-ray)	Accuracy frequently measured on non-independent train/test splits; true generalisation often unverified
Scalable, potentially reduces workforce burden amid documented staffing shortages	Requires capital investment and data-science expertise most laboratories currently lack
Enables genuinely novel surveillance modalities (wastewater metagenomics + social media analytics)	Depends on sequencing/informatics infrastructure uncommon outside well-resourced public health systems
Classical ML methods offer good interpretability alongside strong tabular-data performance	Deep learning components of the same pipelines often remain comparatively opaque
LLMs offer flexible, natural-language interaction with clinical/laboratory data and education support	Documented hallucination, falsehood mimicry, and vulnerability to adversarial manipulation in clinical decision support tasks
Growing multi-institutional benchmark datasets and competitions (DRIAMS, CAMDA AMR) improve comparability	Most published studies still rely on single-centre data without cross-site validation

9. Future Directions (Table 6)

Several developments follow reasonably from current trends, with varying confidence. Foundation models,

broadly pre-trained across institutions and pathogens, could address the small-dataset and external-validation problems above, though this remains a hypothesis rather

than an established finding for microbiology specifically. Digital microbiology laboratories full digital microscopy and plate imaging with AI-first interpretation and human confirmatory review reflect the explicit “lab of the future” framing used at ASM's 2024 Clinical Microbiology Open.[6] Multimodal AI, jointly interpreting image, spectral, genomic, and clinical text data rather than largely single-modality models, represents a natural next step for complex presentations where no single data type is independently diagnostic. Federated learning, technically mature in other medical AI domains, offers a plausible route to multi-site training data without the data-sharing agreements that have historically slowed multi-institutional microbiology research. Real-time,

multimodal surveillance combining genomic and non-traditional data streams, as in the Quebec City wastewater example, is likely to expand to additional pathogens, contingent on public health informatics investment. AI-guided precision antimicrobial stewardship, integrating resistance prediction, personalised risk modelling, and LLM-based decision support into one workflow, is desirable but contingent on the hallucination and explainability concerns above being addressed first. Robotic microbiology and digital twin modelling remain the most speculative near-term directions, with essentially no mature clinical microbiology-specific implementations published to date.

Table 6. Future Opportunities in AI for Clinical Microbiology

Opportunity	Description	Anticipated Timeframe
Foundation models for microbiology	Broadly pre-trained models adapted to genomic, spectral, or imaging tasks to improve cross-site generalisation	Medium-term; largely hypothesis-stage for microbiology specifically
Digital microbiology laboratories	Full digital microscopy/plate imaging with AI-first interpretation and human confirmatory review	Near- to medium-term; already the explicit stated goal of ASM's 2024 “Lab of the Future” discussions
Multimodal AI	Joint interpretation of image, spectral, genomic, and clinical text data	Medium-term; few microbiology-specific implementations published to date
Federated learning	Cross-institutional model training without centralising patient-level data	Medium-term; technically mature in other medical AI domains, not yet widely adopted in microbiology
Real-time, multimodal surveillance	Combined genomic/metagenomic and non-traditional data streams (e.g., social media) for early outbreak detection	Near-term; proof-of-concept already demonstrated (Quebec City salmonellosis example)
AI-integrated precision antimicrobial stewardship	Combining resistance prediction, personalised risk modelling, and LLM-based decision support into a single workflow	Medium- to long-term; contingent on resolving explainability and hallucination concerns first
Digital twins	Patient- or pathogen-specific simulation of infection or resistance trajectory under different treatment scenarios	Long-term / speculative; minimal mature clinical microbiology-specific work published to date

• **Limitations of the Review**

This review has several limitations. First, as a narrative review, it was not designed to provide an exhaustive systematic synthesis of all published evidence, and study selection may therefore be influenced by the authors'

judgement. Second, substantial heterogeneity exists across AI studies in terms of datasets, reference standards, algorithms, clinical settings, and performance metrics, making direct comparison difficult. Third, many published studies remain single-centre investigations

with limited external or prospective validation, particularly in mycology and emerging generative AI applications. Finally, the rapid pace of development in AI means that new models, regulatory decisions, and clinical validation studies may emerge after the literature considered in this review.

10. Practical Recommendations

Clinical microbiologists should treat published accuracy as a starting point for local evaluation, not a substitute for it, and maintain active oversight of AI-flagged resistance predictions rather than treating output as final. Infectious diseases physicians should approach LLM-based decision support with the same scepticism applied to any other tool, explicitly probing for hallucination and falsehood mimicry rather than trusting fluent, confident responses. Hospital administrators should budget for workforce data-science training and informatics infrastructure, and require evidence of external validation as a procurement condition rather than relying on vendor-supplied figures. Researchers should prioritise multi-site prospective validation over incremental single-site accuracy gains, report negative and cross-site degradation findings as transparently as the Münster MALDI-TOF study did, and treat the mycology evidence gap as a distinct priority. Industry should publish validation data disaggregated by site and population, invest in low-resource-compatible tool design, and engage transparently with explainability and regulatory questions rather than minimising them in marketing materials. Across all five stakeholder groups, the recurring practical implication of this review is the same: procurement, research, and clinical-use decisions should weight evidence of external, multi-site validation considerably more heavily than they currently appear to.

• Implementation Priorities

Before routine clinical deployment, AI-based microbiology tools should undergo independent external validation using representative local populations and laboratory workflows. Prospective evaluation should assess not only diagnostic accuracy but also turnaround time, clinical utility, workflow disruption, cost-effectiveness, and patient-safety outcomes. Laboratories should establish predefined criteria for human oversight, documentation of AI-assisted decisions, monitoring of model performance, and management of model drift after deployment. For resistance-prediction systems, AI-generated predictions should remain complementary to validated phenotypic or molecular antimicrobial susceptibility testing until sufficient evidence supports their replacement. Institutions implementing generative AI should additionally establish policies governing data privacy, model selection, hallucination monitoring, user

accountability, and the use of patient-identifiable information.

11. CONCLUSIONS

Artificial intelligence has moved, within roughly a decade, from research curiosity to a technology with genuine, if unevenly distributed, clinical utility across clinical microbiology. The strongest evidence exists for image-based applications Gram stain interpretation and tuberculosis chest radiograph screening where independently developed systems converge on similar, clinically meaningful accuracy. MALDI-TOF- and genome-based resistance prediction remain genuinely important but, on the explicit assessment of researchers working in the field, not yet ready to substitute for phenotypic susceptibility testing in routine practice. Mycology remains comparatively neglected across nearly every application category, a gap deserving dedicated investment given the severity of invasive fungal disease. Generative AI and LLMs offer real promise but also carry the most well-documented, clinically consequential failure modes, warranting the most cautious current deployment posture. Running through nearly every section is one recurring theme: external, multi-site, prospective validation remains the exception rather than the rule, and closing this gap will determine more than any single new algorithm whether the next decade of AI in clinical microbiology delivers on its promise.

Clinical Implications

The immediate clinical role of AI in microbiology should therefore be viewed as augmentation rather than replacement of microbiologists and infectious diseases specialists. Systems with demonstrated external validation may support prioritisation, screening, interpretation, and early decision-making, but final diagnostic and therapeutic decisions should remain under appropriately trained human oversight. The transition from promising algorithms to dependable clinical tools will require evidence generated in real-world, multicentre settings and continuous post-deployment monitoring.

Funding

No specific funding was received for this work.

Conflict of Interest

The authors declare that they have no conflict of interest.

Ethics Approval

Ethical approval was not required because this article is a narrative review of previously published literature and does not involve human participants, identifiable patient data, or animal subjects.

Author Contributions

All authors contributed to the conception and development of the review, literature identification and interpretation, and preparation and critical revision of the manuscript. All authors reviewed and approved the final version of the manuscript and agree to be accountable for the accuracy and integrity of the work.

Data Availability

No new datasets were generated or analyzed for this narrative review. All information discussed in this article was derived from previously published literature cited in the manuscript.

Abbreviations

AI, Artificial Intelligence; ML, Machine Learning; DL, Deep Learning; CNN, Convolutional Neural Network; NLP, Natural Language Processing; LLM, Large Language Model; AMR, Antimicrobial Resistance; MALDI-TOF, Matrix-Assisted Laser Desorption/Ionization–Time of Flight; AST, Antimicrobial Susceptibility Testing; TB, Tuberculosis; XAI, Explainable Artificial Intelligence; WSI, Whole-Slide Imaging; PCR, Polymerase Chain Reaction; MIC, Minimum Inhibitory Concentration; RF, Random Forest; SVM, Support Vector Machine; SHAP, SHapley Additive exPlanations; CAD, Computer-Aided Detection.

REFERENCES

- 1 Goodswen SJ, Barratt JL, Kennedy PJ, Kaufer A, Calarco L, Ellis JT. Machine learning and applications in microbiology. *FEMS microbiology reviews*. 2021 Sep;45(5):fuab015.
- 2 Alsulimani A, Akhter N, Jameela F, Ashgar RI, Jawed A, Hassani MA, et al. The impact of artificial intelligence on microbial diagnosis. *Microorganisms*. 2024;12(6):1051.
- 3 Mairi A, Hamza L, Touati A. Artificial intelligence and its application in clinical microbiology. *Expert Rev Anti Infect Ther*. 2025;23(7):469-490.
- 4 Mencacci A, De Socio GV, Pirelli E, et al. Laboratory automation, informatics, and artificial intelligence: current and future perspectives in clinical microbiology. *Front Cell Infect Microbiol*. 2023;13:1188684.
- 5 Rizzo A, Mensa E, Squarzon L, Clerici P, Lucis R. Artificial intelligence in clinical microbiology: results from the first National survey by the Italian association of clinical microbiologists. *European Journal of Clinical Microbiology & Infectious Diseases*. 2026 Jan;45(1):197-206.
- 6 Dien Bard J, Prinzi AM, Larkin PM, Peaper DR, Rhoads DD. Proceedings of the Clinical Microbiology Open 2024: artificial intelligence applications in clinical microbiology. *Journal of Clinical Microbiology*. 2025 Apr 9;63(4):e01804-24.
- 7 AlGain S, Marra AR, Kobayashi T, Marra PS, Celeghini PD, Hsieh MK, et al. Can we rely on artificial intelligence to guide antimicrobial therapy? A systematic literature review. *Antimicrob Steward Healthc Epidemiol*. 2025. doi:10.1017/ash.2025.47.
- 8 Miglietta L, Rawson TM, Galiwango R, Tasker A, Ming DK, Akogo D, et al. Artificial intelligence and infectious disease diagnostics: state of the art and future perspectives. *The Lancet Infectious Diseases*. 2025 Sep 16.
- 9 Weis C, Cuénod A, Rieck B, Dubuis O, Graf S, Lang C, et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nature medicine*. 2022 Jan;28(1):164-74.
- 10 Hu K, Meyer F, Deng ZL, Asgari E, Kuo TH, Münch PC, et al. Assessing computational predictions of antimicrobial resistance phenotypes from microbial genomes. *Brief Bioinform*. 2024;25(3):bbae206.
- 11 McMahon J, Tomita N, Tatishev ES, Workman AA, Costales CR, Banaei N, Martin IW, Hassanpour S. A novel framework for the automated characterization of Gram-stained blood culture slides using a large-scale vision transformer. *Journal of Clinical Microbiology*. 2025 Mar 12;63(3):e01514-24.
- 12 Visitsattapong S, Bunkum M, Pintavirooj C, Paing MP. A deep learning model for bacterial classification using big transfer (BiT). *IEEE Access*. 2024;12:15609-15621.
- 13 Walter C, Weissert C, Gizewski E, Burckhardt I, Mannsperger H, Hänselmann S, et al. Performance evaluation of machine-assisted interpretation of Gram stains from positive blood cultures. *Journal of Clinical Microbiology*. 2024 Apr 10;62(4):e00876-23.
- 14 Ebisawa KF, Ohji G, Yamamoto K, Miyatsuka I, Moriya A, Akamatsu K, et al. Comparison of Gram stain interpretation accuracy between a computer-aided diagnosis app and microbiology

specialists in blood culture samples. *Microbiology Spectrum*. 2026 Jun 23:e00081-26.

15. Smith KP, Kang AD, Kirby JE. Automated interpretation of blood culture Gram stains by use of a deep convolutional neural network. *J Clin Microbiol*. 2018;56(10):e01521-17.

16. Yamamoto K, Ohji G, Miyatsuka I, Furui-Ebisawa K, Moriya A, Maeta S, et al. Accuracy of classification of urinary Gram-stain findings by a computer-aided diagnosis app compared with microbiology specialists. *Journal of Medical Microbiology*. 2025 Apr 23;74(4):002008.

17. Watanabe N, Watari T, Akamatsu K, Miyatsuka I, Otsuka Y. Performance of deep learning models in predicting the nugent score to diagnose bacterial vaginosis. *Microbiology Spectrum*. 2025 Jan 7;13(1):e02344-24.

18. Nguyen HA, Peleg AY, Song J, Antony B, Webb GI, Wisniewski JA, et al. Predicting *Pseudomonas aeruginosa* drug resistance using artificial intelligence and clinical MALDI-TOF mass spectra. *MSystems*. 2024 Sep 17;9(9):e00789-24.

19. Wiesmann N, Enders D, Westendorf A, Koch R, Schaumburg F. Prediction of antimicrobial resistance from MALDI-TOF mass spectra using machine learning: a validation study. *Journal of Clinical Microbiology*. 2025 Dec 17;63(12):e01186-25.

20. Wang WY, Chiu CF, Tsao SM, Lee YL, Chen YH. Accurate prediction of antimicrobial resistance and genetic marker of *Staphylococcus aureus* clinical isolates using MALDI-TOF MS and machine learning—across DRIAMS and Taiwan database. *International Journal of Antimicrobial Agents*. 2024 Nov 1;64(5):107329.

21. Duroux D, Yang Y, Sattler J, Krauthammer M, Egli A. Early antifungal resistance prediction based on MALDI-TOF mass spectrometry and machine learning. *Scientific Reports*. 2026 Jun 2

22. Gauthier J, Mohammadi S, Kukavica-Ibrulj I, Boyle B, Landgraff C, Goodridge L, et al. Leveraging artificial intelligence community analytics and nanopore metagenomic surveillance to monitor early enteropathogen outbreaks. *Frontiers in Public Health*. 2025 Nov 26;13:1675080.

23. Howard A, Hughes DM, Green PL, Velluva A, Gerada A, Maskell S, et al. Personalised antimicrobial susceptibility testing with clinical prediction modelling informs appropriate antibiotic

use. *Nature communications*. 2024 Nov 21;15(1):9924.

24. Van Heuverswyn J, Valik JK, Desirée van der Werff S, Hedberg P, Giske C, Naclér P. Association between time to appropriate antimicrobial treatment and 30-day mortality in patients with bloodstream infections: a retrospective cohort study. *Clinical infectious diseases*. 2023 Feb 1;76(3):469-78.

25. World Health Organization. Consolidated guidance on tuberculosis data generation and use. Module 1. Tuberculosis surveillance. World Health Organization; 2024 Apr 29.

26. de Camargo TF, Ribeiro GA, da Silva MC, da Silva LO, Torres PP, Rodrigues DD, de Santos MO, Filho WS, Rosa ME, Novaes MD, Massarutto TA. Clinical validation of an artificial intelligence algorithm for classifying tuberculosis and pulmonary findings in chest radiographs. *Frontiers in Artificial Intelligence*. 2025 Feb 6;8:1512910.

27. Han ZL, Zhang YY, Li J, Gao S, Liu W, Yang WJ, Xing ZH. A systematic review and meta-analysis of artificial intelligence software for tuberculosis diagnosis using chest X-ray imaging. *Journal of Thoracic Disease*. 2025 May 30;17(5):3223-37.

28. Wang X, Shi Y, Guo S, Qu X, Xie F, Duan Z, Hu Y, Fu H, Shi X, Quan T, Wang K. A clinical bacterial dataset for deep learning in microbiological rapid on-site evaluation. *Scientific Data*. 2024 Jun 8;11(1):608.

29. Terrero-Salcedo D, Powers-Fletcher MV. Updates in laboratory diagnostics for invasive fungal infections. *Journal of Clinical Microbiology*. 2020 May 26;58(6):10-128.

30. Xue J, Yue H, Lu W, Li Y, Huang G, Fu YV. Application of Raman spectroscopy and machine learning for *Candida auris* identification and characterization. *Applied and Environmental Microbiology*. 2024 Nov 20;90(11):e01025-24.

31. Bastos ML, Benevides CA, Zanchettin C, Menezes FD, Inácio CP, de Lima Neto RG, et al. Breaking barriers in *Candida* spp. detection with Electronic Noses and artificial intelligence. *Scientific Reports*. 2024 Jan 10;14(1):956.

32. Fursman F, Fitzsimmons N, Overton D, Blanton K, Hatton KW, Howard R, et al. *Candida auris* screening, positivity trends, and patient characteristics at the University of Kentucky between 2021 and 2024. *Antimicrobial Stewardship & Healthcare Epidemiology*. 2025 Jan;5(1):e258.

33. Ma H, Yang J, Chen X, Jiang X, Su Y, Qiao S, Zhong G. Deep convolutional neural network: a novel approach for the detection of *Aspergillus* fungi via stereomicroscopy. *Journal of Microbiology*. 2021 Jun;59(6):563-72.
34. Gu F, Hu S, Tian B, Ma T, Xu Y, Yang Y, Gu B. Intelligent diagnostic system for *Cryptococcus*: Switch-controllable nanocatcher and CNN-based artificial intelligence. *Chemical Engineering Journal*. 2023 Jun 1;465:142674.
35. Singla N, Kundu R, Dey P. Artificial Intelligence: Exploring utility in detection and typing of fungus with futuristic application in fungal cytology. *Cytopathology*. 2024 Mar;35(2):226-34.
36. Zieliński B, Sroka-Oleksiak A, Rymarczyk D, Piekarczyk A, Brzychczy-Włoch M. Deep learning approach to describe and classify fungi microscopic images. *PloS one*. 2020 Jun 30;15(6):e0234806.
37. Guthrie JD, Shankarnarayan SA, Charlebois DA. Peering inside the black box: Explainable AI to interpret advanced computer vision fungal pathogen prediction. *Scientific Reports*. 2026 Apr 18.
38. Noman SM, Zeeshan M, Arshad J, Deressa Amentie M, Shafiq M, Yuan Y, et al. Machine learning techniques for antimicrobial resistance prediction of *Pseudomonas aeruginosa* from whole genome sequence data. *Computational intelligence and neuroscience*. 2023;2023(1):5236168.
39. Wan F, Tong W, Wu W, Liu J, Huang H, Zhang P, et al. Utilizing whole genome sequencing data for machine learning driven prediction of antibiotic resistance in *Escherichia coli*. *Frontiers in Microbiology*. 2026 Apr 28;17:1842717.
40. Ren Y, Chakraborty T, Doijad S, Falgenhauer L, Falgenhauer J, Goesmann A, et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics*. 2022 Jan 15;38(2):325-34.
41. Anahtar MN, Yang JH, Kanjilal S. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *Journal of clinical microbiology*. 2021 Jun 18;59(7):10-128.
42. Fidler G, Wells H, Combs F, Papciak J, Couto-Rodriguez M, Rey S, et al. BIOTIA-DX RESISTANCE Achieved the Best Antimicrobial Resistance Phenotype Prediction Accuracy at CAMDA 2026. *bioRxiv*. 2026 May 13:2026-05.
43. Xu L, Zhao L, Tong Z, Peng K, Wang M, Li R, et al. Nanopore Sequencing in Veterinary Pathogen Detection: A Review of Technologies and Applications. *Veterinary Sciences*. 2026 Feb 25;13(3):216.
44. Li J, Silvester R, Williams RC, Chan-Herur V, Goldman M, Fidler D, Jones DL. Metagenomics-based source attribution of antimicrobial resistance in wastewater for improved epidemiological risk assessment. *Water research*. 2026 Mar 23:125810.
45. Vadakkan K, Ponnathodi KS, Raphael R, Jose B. A critical review on antibacterial mechanisms of carbon dots: a comprehensive approach to combating multidrug resistance. *Archives of Microbiology*. 2026 Aug;208(8):405.
46. Ngoc Nguyen O, Amin D, Bennett J, Hetlevik Ø, Malik S, Tout A, Vornhagen H, Vellinga A. GP or ChatGPT? Ability of large language models (LLMs) to support general practitioners when prescribing antibiotics. *Journal of Antimicrobial Chemotherapy*. 2025 May;80(5):1324-30.
47. Siam MK, Varela A, Faruk MJ, Cheng JQ, Gu H, Maruf AA, Aung Z. Benchmarking large language models on the United States medical licensing examination for clinical reasoning and medical licensing scenarios. *Scientific Reports*. 2025 Dec 3.
48. Beil J, Aggarwal I, Devarapalli M, Beil JA. Comparative Accuracy Assessment of Large Language Models in Cardiothoracic Anesthesia: A Performance Analysis of Claude and ChatGPT-4 on Subspecialty Board-Style Questions. *Cureus*. 2025 Jul 23;17(7).
49. Schwartz IS, Link KE, Daneshjou R, Cortés-Penfield N. Black box warning: large language models and the future of infectious diseases consultation. *Clinical infectious diseases*. 2024 Apr 15;78(4):860-6.
50. Omar M, Sorin V, Collins JD, Reich D, Freeman R, Gavin N et al. Multi-model assurance analysis showing large language models are highly vulnerable to adversarial hallucination attacks during clinical decision support. *Communications Medicine*. 2025 Aug 2;5(1):330.
51. Omar M, Brin D, Glicksberg B, Klang E. Utilizing natural language processing and large language models in the diagnosis and prediction of infectious diseases: A systematic review. *American Journal of Infection Control*. 2024 Sep 1;52(9):992-1001.