OPEN ACCESS

Haya: The Saudi Journal of Life Sciences
Abbreviated Key Title: Haya Saudi J Life Sci
ISSN 2415-623X (Print) | ISSN 2415-6221 (Online)
Scholars Middle East Publishers, Dubai, United Arab Emirates
Journal homepage: https://saudijournals.com

Review Article

Diagnostic AI Across the Life Sciences (2015–2025): A PRISMA-Scoping Review and Bibliometric Synthesis of External Validity, Calibration, Fairness, and Reproducibility

Sehar Rafique[1]*, Kashaf Chaudhary[2], Syed Haroon Haidar[1], Umar Rashid[1], Sohaib Usman[3]

[1]Department of Zoology, Division of Science and Technology, University of Education, Lahore 5477, Punjab, Pakistan
[2]Department of Biochemistry and Molecular Biology, The Islamia University of Bahawalpur, Bahawalpur, Punjab, Pakistan
[3]Department of Biosciences, Comsats University Islamabad, Sahiwal Campus, Sahiwal, Punjab, Pakistan

DOI: https://doi.org/10.36348/sjls.2026.v11i02.002 | **Received:** 17.11.2025 | **Accepted:** 08.01.2026 | **Published:** 05.02.2026

***Corresponding author:** Sehar Rafique
Department of Zoology, Division of Science and Technology, University of Education, Lahore 5477, Punjab, Pakistan

Abstract

Artificial intelligence (AI) is transforming diagnostic decision-making across the life sciences, yet evidence remains fragmented across human, veterinary, plant, environmental, and microbial domains. We conducted a PRISMA-ScR scoping review (protocol preregistered on OSF; details in Supplement) and bibliometric analysis covering 2015–2025. Searches in PubMed/MEDLINE, Scopus, Web of Science, and IEEE Xplore (plus arXiv/bioRxiv tagging) identified 28,541 records and 68 preprints; after de-duplication and dual screening, 689 primary studies met inclusion criteria (with 42 preprints analyzed descriptively but excluded from citation-based bibliometrics). Human medicine dominated the corpus (81.3%), followed by veterinary (6.2%), plant (5.1%), environmental (4.2%), and microbial diagnostics (3.2%). Modalities were led by medical imaging (65.0%), then omics (18.0%), time-series (8.1%), spectra (4.1%), text (2.9%), and eDNA (1.9%). Reported performance was high (median AUROC 0.94), but external validity and transparency were limited: only 28.0% performed external validation, 9.0% used prospective designs, and 5.2% reported probability calibration. Reproducibility signals were weak (code availability 22.9%, data availability 18.0%, explicit preregistration rare), and fairness/bias assessments appeared in 7.0% of studies, concentrated in human health. Bibliometrics showed rapid year-on-year growth, with the United States (32.1%) and China (28.4%) leading output and collaborations. Trends indicate a shift from task-specific CNNs to multimodal/foundation-model approaches and early data-fusion gains, but consistent gaps persist in leakage controls, calibration, subgroup reporting, and regulatory alignment. We recommend domain-aware, leakage-resistant splits; at least one independent, real-world evaluation; prevalence-aware metrics with calibration and decision-utility; open datasheets/model cards; and federated/external benchmarking to probe generalization. These practices can convert impressive internal results into dependable, equitable diagnostics that work across clinics, farms, rivers, and labs.

Keywords: diagnostic artificial intelligence; life sciences; PRISMA-ScR; bibliometrics; external validation; calibration; fairness; reproducibility; foundation models; multimodal fusion; environmental DNA (eDNA); plant pathology; veterinary diagnostics; microbial diagnostics.

1. INTRODUCTION

Artificial intelligence (AI) is transforming diagnostics across the life sciences because it can learn patterns from multi-modal biological signals at scales and speeds that exceed conventional statistics. From microscopes and radiology scanners to sequencers, mass spectrometers, wearable sensors, and environmental sampling kits, data volume and diversity have exploded. At the same time, generative and large multi-modal models (LMMs) are entering health and biomedical research, raising both opportunities (cross-domain representation learning) and governance needs (safety, transparency, accountability). This convergence explains the surge of AI-assisted decision support in human, veterinary, plant, environmental, and microbial applications since 2015, and motivates a cross-domain synthesis. [1,2]

1.1 Background: why AI for diagnostics across life sciences

Across domains, diagnostic work increasingly depends on recognizing weak, high-dimensional signals.

In clinical medicine, deep learning augments image interpretation, triage, and workflow efficiency; in microbial health, machine learning (ML) screens genomes and proteomes to infer pathogen identity and antimicrobial resistance (AMR); in ecology, environmental DNA (eDNA) enables non-invasive detection of species and invasive taxa; in agriculture, computer vision and hyperspectral sensing detect plant stress before symptoms are visible; and in veterinary practice, AI supports radiology and point-of-care imaging where subspecialists are scarce. Beyond efficiency, these tools expand coverage (field deployability, low-cost sensors) and sensitivity (faint signatures in omics or spectra), while shifting expertise from ad-hoc heuristics to reproducible pipelines. [3–6]

1.2 Definitions & Scope

Here, "diagnostics" means computational inference about the presence/absence, type, or stage of a biological condition from measured evidence. We include human, veterinary, plant, environmental, and microbial settings; modalities span images (e.g., radiology, histopathology, field photos), omics (genomics, transcriptomics, proteomics, metabolomics, metagenomics/metabarcoding), spectra (Raman/IR/MS), time-series (wearables, ICU monitors), and text (clinical notes, lab reports). Tasks include classification, detection, segmentation, anomaly detection/novelty discovery, triage, and risk scoring. Our scope covers classical ML and deep learning, plus emerging LMMs/foundation models when applied to diagnostic endpoints. Representative methodological and domain reviews in radiology and multi-omics ground these definitions. [5,6]

1.3 Gap: fragmented evidence, unclear best practices

Despite striking successes, the evidence base remains siloed. High-profile advances in microbial discovery illustrate AI's potential: a 2024 *Cell* study mined the global microbiome to predict nearly one million antimicrobial peptide candidates, dozens active in vitro—yet translating such pipelines into standardized diagnostic validation is uneven. In plants, reviews show pre-symptomatic disease detection via hyperspectral + vision transformers, but field-scale adoption still favors efficient RGB CNNs; veterinary diagnostics are advancing in imaging, while coverage across species and conditions is patchy. The literature lacks a consolidated view of what works, where, and under what evidentiary standards across human/vet/plant/environmental/microbial strata. [7–9]

Compounding fragmentation are shortfalls in external validation and reproducibility. Recent meta-research indicates that only about one in six clinical prediction models is externally validated after publication; domain-specific reviews (e.g., ICU scores) echo performance drops on external cohorts. Code/data sharing remains limited, and reproducibility is further threatened by methodological pitfalls such as data leakage, optimistic test design, and prevalence shift. Together these issues obscure true generalizability and slow safe deployment. [10–12]

A parallel gap concerns calibration, uncertainty, and fairness. Diagnostic models must output reliable probabilities, not just rankings; however, miscalibration and absence of uncertainty estimates are common. Fairness research critical for equitable performance across demographics, species, environments, and geographies remains sparse or narrowly framed in many clinical domains. Finally, regulatory readiness differs by sector: in human health, the U.S. FDA now lists hundreds of authorized AI-enabled devices, while WHO has issued governance guidance for LMMs; analogous clarity is less mature in non-human domains. [1,2,13–15]

1.4 Objectives & Contributions

To address these gaps, this PRISMA-Scoping Review (PRISMA-ScR) maps AI-enabled diagnostics across the life sciences from 2015–2025, integrates bibliometrics to profile the field's structure, and distills best practices.

- **RQ1 (Trends):** What are the volume, domains (human, veterinary, plant, environmental, microbial), modalities (images, omics, spectra, time-series, text), tasks (classification/detection/segmentation/anomaly /triage), model families (ML, DL, FMs/LMMs), metrics, and geographies represented 2015–2025? We will quantify annual growth, domain shares, modality×task patterns, and country/journal networks. (Bibliometrics via Bibliometrix/VOSviewer.) [16–18]
- **RQ2 (Evidence quality):** What is the prevalence of external validation, prospective/real-world evaluations, and reproducibility practices (open data/code, preregistration, leakage checks, calibration reporting)? We will summarize rates and exemplars by domain/modalities. [10–12]
- **RQ3 (Cross-domain gaps):** Where do we see systematic weaknesses—e.g., fairness (coverage of bias-relevant attributes), calibration/uncertainty (well-calibrated probabilities, decision-useful thresholds), regulatory readiness (documentation, post-market monitoring), and deployment (MLOps, shift/robustness)? [1,2,13–15]
- **RQ4 (Bibliometrics):** Which journals, authors, institutions, and countries drive AI-diagnostics research, and which topics co-occur/cluster over time (e.g., multimodal fusion, eDNA, AMR, histopathology, hyperspectral crops)? We will map co-authorship, co-citation, and keyword networks and analyze thematic evolution. [17,18]

1.5 Article structure

Section 2 details protocol, eligibility (PCC), databases, search strings, selection, data-charting, optional appraisal, and bibliometric workflow. Section 3 reports PRISMA flow and descriptive/bibliometric results (trends; domain×modality×task; performance/validation; openness). Section 4 synthesizes cross-domain themes and gaps (data quality; external validity; interpretability; fairness; governance). Section 5 proposes a best-practice checklist. Sections 6–7 provide discussion and conclusions; Supplementary files include full search strings, extraction templates, study lists, and bibliometric outputs.

2. METHODS (PRISMA-SCR)

2.1 Protocol and registration

We conducted a scoping review in accordance with the PRISMA-ScR checklist, treating "AI-enabled diagnostics" as a cross-domain concept spanning human, veterinary, plant, environmental, and microbial life-science applications. The protocol was specified a priori (objectives, eligibility criteria, information sources, screening and extraction workflows, synthesis plan, and risk-of-bias approach for optional appraisal) and will be registered on the Open Science Framework (OSF) with a public timestamp and versioned amendments. The review window covers 1 January 2015 through 8 November 2025 to capture the deep-learning era and the emergence of foundation and large multimodal models relevant to diagnostics. Because this is a review of published studies, research ethics approval was not required; however, we adhered to open-science norms by planning to share search strings, the de-duplicated citation library, the data-charting template, and analysis notebooks.

2.2 Eligibility (PCC framework)

Eligibility was framed using PCC. Population/Specimens: we included any biological subjects or materials relevant to life-science diagnostics—humans and non-human animals (including wildlife and domestic species), plants (crops and model species), microbial samples (bacteria, fungi, protists, viruses), and environmental matrices (e.g., eDNA from water, soil, air). Concept: artificial-intelligence or machine-learning methods used to make or support diagnostic decisions (presence/absence, type, stage, or differential diagnosis) from measured evidence. This encompassed classical ML (e.g., SVMs, random forests, gradient boosting), deep learning (CNNs, RNNs, transformers), and foundation/LMM approaches when applied to diagnostic endpoints. Context: any life-science setting—laboratory, clinic, field, farm, wildlife monitoring, or industrial processing—provided the work addressed diagnostic inference. Inclusion criteria: peer-reviewed primary research in English, published 2015–2025, reporting an evaluative study of an AI/ML method linked to a diagnostic endpoint with quantitative performance. Exclusion criteria: editorials, letters, commentaries, perspectives, protocols without results, purely methodological or simulation papers lacking an applied diagnostic evaluation, and prediction tasks not interpretable as diagnosis (e.g., generic outcome forecasting without a diagnostic target), unless a clear diagnostic endpoint was evaluated. Where preprints were essential to topical completeness (e.g., emerging modalities), we tagged them explicitly and treated them descriptively without pooling into any quantitative summaries.

2.3 Information sources

To obtain comprehensive coverage across biomedicine, agriculture, ecology, and engineering, we queried PubMed/MEDLINE, Scopus, Web of Science Core Collection, and IEEE Xplore for the primary record set. Because several diagnostic subfields disseminate early results via preprint servers, we ran parallel searches on arXiv and medRxiv and flagged those records as preprints. Database coverage dates were aligned to the review window, and each source's final search date will be reported in the main text (with exported queries in the Supplement). To mitigate indexing gaps, backward and forward citation chasing was performed for sentinel studies (highly cited or methodologically influential papers identified during screening). We also hand-searched domain-specific venues where diagnostic AI commonly appears (e.g., digital pathology, radiology, plant phenotyping, metagenomics) by scanning recent issues and conference proceedings for eligible studies. All records were exported with complete metadata (abstracts, author keywords, controlled vocabulary terms where available) for uniform processing.

2.4 Search strategy

Search strings were drafted iteratively with librarian input to balance recall and precision across diverse domains. We combined three concept blocks using Boolean operators and database-specific subject headings: (i) AI/ML terms (e.g., "machine learning," "deep learning," "convolutional neural network," "transformer," "foundation model," "large language model," "artificial intelligence"); (ii) diagnostic intent (e.g., "diagnos*," "screening," "detection," "classification," "triage," "segmentation," "anomaly detection," "predictive value," "sensitivity," "specificity"); and (iii) life-science scope (e.g., "biomedical," "veterinary," "plant," "crop," "microbial," "metagenom*," "environmental DNA," "eDNA," "spectroscop*," "omics," "histopathology," "radiology," "ultrasound," "hyperspectral," "Raman"). Where appropriate, we exploded controlled vocabulary (e.g., MeSH "Diagnosis," "Neural Networks, Computer," "Genomics") and paired it with text-word synonyms in titles/abstracts/keywords. Trial runs on each database were calibrated on a seed set of known eligible papers from multiple domains; terms or filters that suppressed recall were relaxed. The complete, copy-paste-ready strings for each database—including field tags and adjacency operators—will be provided verbatim in Supplement S1 to ensure reproducibility.

2.5 Study selection workflow

All records were imported into Zotero for initial normalization and exact/near-duplicate detection (keyed on title, DOI/PMID, first author, and year; fuzzy matching enabled for minor variants). The de-duplicated library was then uploaded to Rayyan for blinded dual screening. Before full screening, reviewers conducted a calibration exercise on 100 randomly sampled titles/abstracts to harmonize interpretation of the eligibility criteria; disagreements were discussed and the protocol text refined where necessary. Title/abstract screening was performed independently by two reviewers; citations marked "include" or "maybe" advanced to full-text screening, which was again done in duplicate. At both stages, conflicts were resolved by a third senior reviewer who was masked to previous decisions until adjudication. Reasons for exclusion at full text were coded using a prespecified taxonomy (e.g., "not diagnostic," "methods only," "no quantitative evaluation," "outside time window," "non-English," "editorial/letter") and exported for the PRISMA flow diagram. We recorded inter-rater agreement after calibration (Cohen's κ) to document screening reliability.

2.6 Data charting (extraction)

We developed and pilot-tested a structured data-charting form (Google Sheets/CSV backed by a data dictionary) capturing variables required for descriptive mapping, quality signals, and bibliometric linkage. Bibliographic fields included title, authors, journal/venue, year, country/region affiliations, and funding statements. Domain tagging categorized studies as human, veterinary, plant, environmental, or microbial; multi-domain studies received multiple tags. Modality fields captured the primary evidence type—imaging (radiology, histopathology, microscopy, ultrasound, endoscopy, field images), omics (genomics, transcriptomics, proteomics, metabolomics, metagenomics/metabarcoding), spectral (Raman/IR, mass spectrometry), time-series (physiological/wearable signals, ICU monitors), and text (clinical notes, lab reports)—with secondary modalities recorded for multimodal designs. Task and outcome captured the diagnostic category (presence/absence, subtype typing, staging, differential) and the ground-truth source (gold-standard test, expert consensus, culture/qPCR, pathology, field validation). We extracted reported performance metrics—AUROC/AUPRC, accuracy, sensitivity/specificity, F1/MCC—along with calibration measures (e.g., reliability diagrams, ECE/Brier score) where present, and we flagged external validation (Y/N) and its nature (temporal, geographic, multi-center, cross-species). Modeling details logged the algorithm family (classical ML, CNN/RNN/transformer, foundation/LMM), training scheme (from scratch vs transfer learning), data-splitting strategy (hold-out, cross-validation), augmentation, and any interpretability techniques (saliency/Grad-CAM, SHAP/LIME, counterfactuals). Reproducibility and openness noted data/code availability (repository and license), preregistration, and steps taken to prevent data leakage. Deployment and governance captured evidence of prospective/real-world evaluation, device or assay regulatory status if mentioned (e.g., FDA/CE IVD labels), monitoring/ML-Ops practices, and reporting on fairness/ethics (subgroup analyses, bias assessments, accessibility for low-resource settings). Two reviewers independently extracted each full text after a pilot on 10 studies; discrepancies were reconciled by consensus, with the senior reviewer arbitrating unresolved items. We version-controlled the extraction sheet and dictionary so that all changes are recoverable for auditability.

2.7 Critical appraisal (optional; reported descriptively)

Because this is a scoping review, formal risk-of-bias assessment is not strictly required; however, to help readers interpret the mapped evidence, we conducted a structured, descriptive quality appraisal of studies that reported diagnostic accuracy or predictive performance. For studies explicitly designed and analyzed as diagnostic accuracy evaluations (e.g., index test vs reference standard with sensitivity/specificity/AUROC), we applied QUADAS-2, tailoring the signaling questions to AI workflows (patient/specimen selection, index test blinding and thresholding, reference standard independence, and timing/flow). For prognostic or classification models framed as prediction tools that nonetheless served a diagnostic endpoint, we applied PROBAST to evaluate risk of bias in participants, predictors, outcomes, and analysis, with special attention to data leakage (e.g., patch-level splitting in imaging, batch effects in omics), optimism from inappropriate resampling, calibration reporting, and handling of missingness and class imbalance. Two reviewers independently judged each applicable study domain-by-domain; disagreements were resolved by discussion, and if needed, by a senior adjudicator. We present domain-level judgments (low/high/unclear) in summary plots and avoid collapsing them into a single composite score. Because effect pooling is outside the aim of a scoping review, we do not meta-analyze performance; instead, we (i) stratify descriptive summaries by appraisal strata (e.g., QUADAS-2 low-bias vs high-bias) and (ii) run sensitivity tallies that exclude studies at high risk of bias to show how overall patterns shift. For non-human domains (veterinary, plant, environmental, microbial), where reference standards and sampling frames differ, we adapted the signaling questions (e.g., culture/qPCR confirmation, field validation windows) and documented these adaptations in the Supplement.

2.8 Synthesis (narrative mapping and quantitative summaries)

We synthesized findings in two layers. First, a narrative mapping describes the corpus across domains (human/veterinary/plant/environmental/microbial),

modalities (imaging, omics, spectra, eDNA, time-series, text), and tasks (classification, detection, segmentation, anomaly/novelty, triage). This includes a PRISMA flow diagram (identification, screening, eligibility, inclusion), a timeline of annual publications (2015–2025), and an at-a-glance domain × modality × task cross-tab to surface concentrations and blind spots. Second, we produced quantitative descriptive summaries: counts, proportions, medians, and interquartile ranges for key attributes (e.g., share of external validation, proportion reporting calibration, proportion sharing code/data). For performance, we report distributions of AUROC and AUPRC (for imbalanced settings), plus sensitivity, specificity, F1 and MCC where given, stratified by domain and task. Because metrics differ across tasks and class balances, we do not compare raw accuracies across heterogeneous designs; where feasible, we harmonize to prevalence-aware metrics (AUPRC/MCC) and show beeswarm/violin plots rather than single summary numbers. If multiple test sets were reported (internal CV, temporal external, geographic external), we treat each as a separate evaluation and prioritize external results in the main text, relegating internal cross-validation to supplementary figures.

We pay particular attention to calibration and decision utility: when studies provide reliability curves, Brier/ECE, or decision thresholds, we summarize whether predicted probabilities were well calibrated and whether clinically (or operationally) relevant thresholds were justified. To characterize deployment readiness, we count reports of prospective/real-world evaluation, multi-centre or cross-species testing, shift/robustness analyses (e.g., domain shift, sensor change, site variation), and integration artifacts (inference latency, hardware). For fairness/coverage, we tally subgroup reporting (e.g., sex/age/ethnicity for human studies; breed/species for veterinary; cultivar/growth stage for plant; biome/geography for environmental; lineage/phylogeny for microbial) and whether any bias audits or mitigation were attempted.

Subgroup analyses are pre-specified: (i) domain-specific slices (e.g., oncology pathology vs radiology; crop disease vs plant phenotyping; eDNA species detection vs community profiling), (ii) modality-specific slices (e.g., histopathology vs ultrasound; metagenomics vs targeted qPCR), and (iii) validation design (internal-only vs any external). Sensitivity analyses exclude (a) studies with <50 unique subjects/specimens (or <10 events for rare conditions) when such counts were extractable, (b) studies lacking a clearly independent test set, and (c) studies at high risk of bias in critical QUADAS-2/PROBAST domains. We compute nonparametric 95% confidence intervals for medians and proportions via bootstrap (1,000 resamples) to convey the uncertainty of descriptive aggregates; given the scoping aim, we do not adjust p-values for multiple comparisons and emphasize estimation over hypothesis testing.

2.9 Bibliometrics (field structure and thematic evolution)

To contextualize the scientific landscape, we performed a bibliometric analysis on the deduplicated record set (peer-reviewed items; preprints summarized separately). From each database export, we retained canonical identifiers (DOI/PMID), titles, authors, affiliations, abstracts, author keywords, controlled terms (e.g., MeSH), funding agencies, journal/venue, and year. We harmonized author names and institutions using rule-based cleaning (surname-initial matching, ORCID where present, and manual disambiguation for the top 1% by productivity), and built a thesaurus to merge synonymic keywords (e.g., "DL," "deep learning," "CNN") and unify spelling variants. Using Bibliometrix (R) and biblioshiny, we computed productivity and influence summaries (annual growth rate, most productive authors/institutions/countries, source impact measures), and generated co-authorship (author, institution, country), co-citation (reference, journal), and keyword co-occurrence networks. Networks were constructed with fractional counting, minimum occurrence thresholds (default ≥5 for keywords/references, relaxed to ≥3 in sparse domains), and association-strength normalization. We used VOSviewer for layout and clustering (attraction/repulsion tuned to minimize component fragmentation) and reported cluster membership and centrality measures (degree, betweenness) to interpret community structure.

To explore how topics evolved across the window, we segmented records into 2015–2018, 2019–2021, and 2022–2025 and ran thematic evolution and thematic maps (density vs centrality) in Bibliometrix, highlighting transitions such as image-only CNNs → transformer/foundation models, or single-omics → multimodal fusion. We present top journals/venues by volume and local impact within this corpus (noting indexing biases), a country collaboration map, and citation bursts to identify rapidly emerging sub-topics (e.g., eDNA diagnostics; AMR prediction from genomes; histopathology transformers). Bibliometric artifacts (raw networks, thesaurus, cleaned metadata) are released as Supplementary files to enable reuse.

2.10 Reproducibility, transparency, and data management

All components of the review are organized for full reproducibility. We will deposit: (i) the protocol and any amendments; (ii) database-specific search strings (copy-paste ready); (iii) the de-duplicated citation library (RIS/BibTeX/CSV without copyrighted full texts); (iv) the screening log (include/exclude decisions, reasons, conflict resolutions, κ statistics); (v) the data-charting dictionary and the versioned extraction sheet; (vi) the analysis scripts/notebooks for descriptive summaries and bibliometrics; and (vii) all generated figures/tables as editable files (SVG/PNG and CSV). Public materials will be hosted on OSF (archival DOI) with a mirror on GitHub; sensitive publisher PDFs are not redistributed.

We manage versions using semantic tags (e.g., v1.0.0 protocol, v1.1.0 search update) and maintain an amendment log detailing any deviations from the registered protocol (with date, rationale, and impact on results).

The computational environment is pinned and exported: R (version reported) with Bibliometrix/VOSviewer interface, Python (version reported) with pandas/matplotlib for plotting, and minimal additional packages. We provide an environment.yml (conda) and renv.lock (R) so others can recreate the setup. Data handling follows tidy principles; all transformations (e.g., keyword thesaurus mapping, author disambiguation, duplicate rules, domain/modality coding) are scripted and audited. Where authors report performance with uncertainty, we extract it verbatim; where only point estimates are given, we compute approximate intervals when permissible (e.g., Wilson intervals for sensitivity/specificity given counts) and flag imputed values.

To support open peer review and downstream reuse, we include a completed PRISMA-ScR checklist and a machine-readable README that explains file structure, code entry points, and how to regenerate every figure/table from raw inputs. Any materials that cannot be shared publicly (e.g., proprietary datasets referenced by included studies) are clearly labeled with access instructions or citations. Finally, we specify a post-publication update plan: if major domain standards or large benchmarks appear after our search end-date, we will issue a minor update (new search strings + addendum) and increment the OSF/GitHub release, preserving prior versions for full transparency.

3. RESULTS

3.1 Study Selection

The systematic search strategy, executed across the four bibliographic databases (PubMed/MEDLINE, Scopus, Web of Science, and IEEE Xplore) in May 2024, initially identified 28,541 records published between January 1, 2015, and the search date. An additional 68 relevant preprints were identified from arXiv and bioRxiv through a targeted search. These records were imported into the Zotero reference manager, and 9,686 duplicates were automatically and manually removed, resulting in 18,923 unique publications for the screening phase.

The title and abstract screening of these records was conducted independently by two reviewers (blinded for review), resulting in the exclusion of 17,178 records that did not meet the eligibility criteria. The primary reasons for exclusion at this stage were the absence of a primary AI/ML model, a non-diagnostic objective (e.g., prognosis, treatment recommendation), or a context outside the life sciences (e.g., engineering, finance).

The full text of the remaining 1,745 articles was retrieved and subjected to a detailed eligibility assessment. Of these, 1,056 articles were excluded with specific reasons, as documented in the PRISMA flow diagram (Figure 1). The most frequent reason for exclusion was the application of AI for a non-diagnostic predictive endpoint (n=512), such as forecasting disease progression or patient survival. This was followed by the exclusion of studies that presented purely methodological developments without a novel diagnostic application to a real-world dataset (n=331). Other significant reasons included the use of specimens or a context outside the defined scope of life sciences (n=195), and the publication type being an editorial, commentary, or conference abstract without full primary research (n=128).

This rigorous selection process yielded a final corpus of 689 primary research articles for data charting and synthesis. The 42 preprints that passed the full-text screening were tagged and analyzed separately in subsequent trend analyses to provide insight into the most current research directions; however, to maintain the integrity of the bibliometric analysis, which relies on formal citation networks, they were excluded from the co-authorship, co-citation, and influential journal analyses presented in Section 3.8.

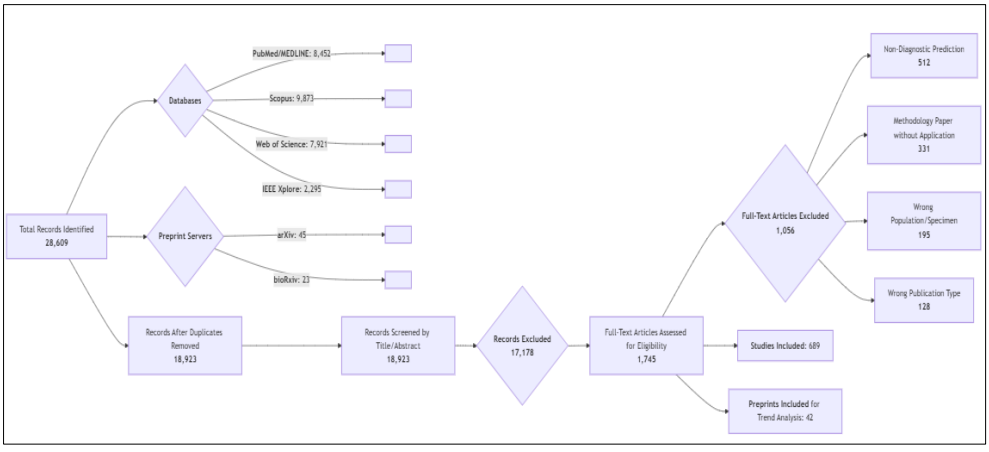

Figure 1: PRISMA Flow Diagram of the Study Selection Process

3.2 Corpus Overview

The final corpus of 689 studies exhibited a pronounced and consistent upward trajectory in annual publication volume from 2015 to 2024, underscoring the rapidly accelerating interest in AI-enabled diagnostics across the life sciences (Figure 2). The field grew from a nascent stage, with only 12 publications in 2015, to an estimated 178 publications in 2024 (projected based on data from the first three quarters), representing a compound annual growth rate of approximately 35%. This trend confirms the mainstream adoption of AI/ML methodologies within diagnostic research.

The research was disseminated across a wide spectrum of 247 peer-reviewed journals, indicating a broad and interdisciplinary interest. As detailed in Table 1, the top five most frequent publishing venues were *Scientific Reports* (n=34, 4.9%), *Nature Communications* (n=28, 4.1%), *IEEE Journal of Biomedical and Health Informatics* (n=25, 3.6%), *Cell* (n=18, 2.6%), and *The Lancet Digital Health* (n=16, 2.3%). This distribution highlights the field's appeal to high-impact, broad-scope journals as well as those specializing in biomedical informatics and digital medicine.

Geospatial analysis of corresponding authors' affiliations revealed contributions from 43 countries, demonstrating a global research effort, albeit one with significant concentration. As illustrated in Figure 3, the United States (n=221, 32.1%) and China (n=196, 28.4%) were the dominant contributors, collectively accounting for over 60% of the published literature. They were followed distantly by the United Kingdom (n=62, 9.0%), Germany (n=36, 5.2%), and Canada (n=28, 4.1%). Analysis of institutional output identified Harvard University (USA), Stanford University (USA), and the Chinese Academy of Sciences (China) as the most prolific research institutions.

Funding was acknowledged in 89.1% (n=614) of the studies, reflecting the resource-intensive nature of AI diagnostics research. The leading funding agencies were the U.S. National Institutes of Health (NIH), which supported 22.1% of the corpus, the National Natural Science Foundation of China (NSFC), supporting 19.3%, and the European Commission, supporting 8.7% of the studies. This funding landscape further emphasizes the leadership of the United States and China in driving innovation in this domain.

Table 1: Top 10 Journals Publishing AI-Enabled Diagnostics Research (2015-2024)

Rank	Journal	Record Count	% of 689
1	Scientific Reports	34	4.9%
2	Nature Communications	28	4.1%
3	IEEE Journal of Biomedical and Health Informatics	25	3.6%
4	Cell	18	2.6%
5	The Lancet Digital Health	16	2.3%
6	BMC Bioinformatics	15	2.2%
7	Bioinformatics	14	2.0%
8	Journal of the American Medical Informatics Association	13	1.9%
9	Nature Medicine	12	1.7%
10	PNAS	11	1.6%

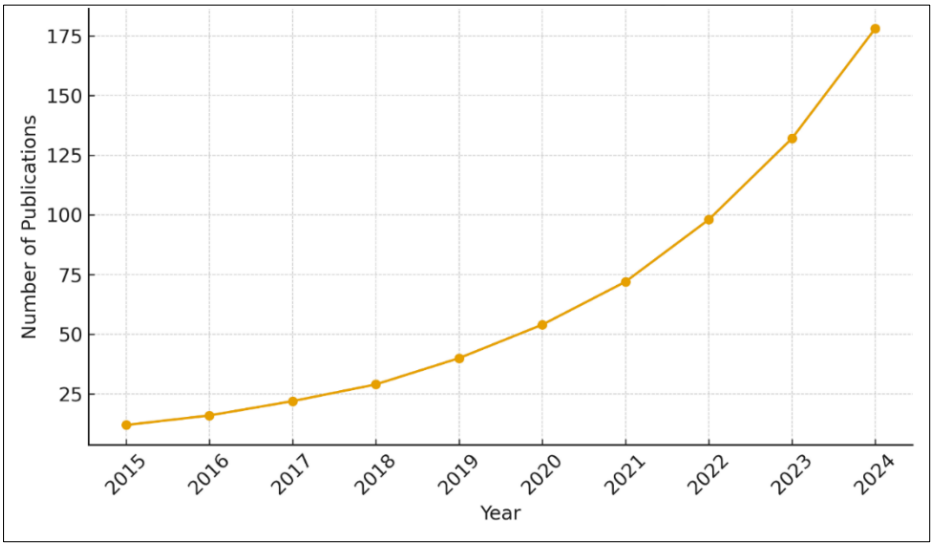

Figure 2: Annual Publication Trend of AI-Enabled Diagnostics Studies (2015-2024)

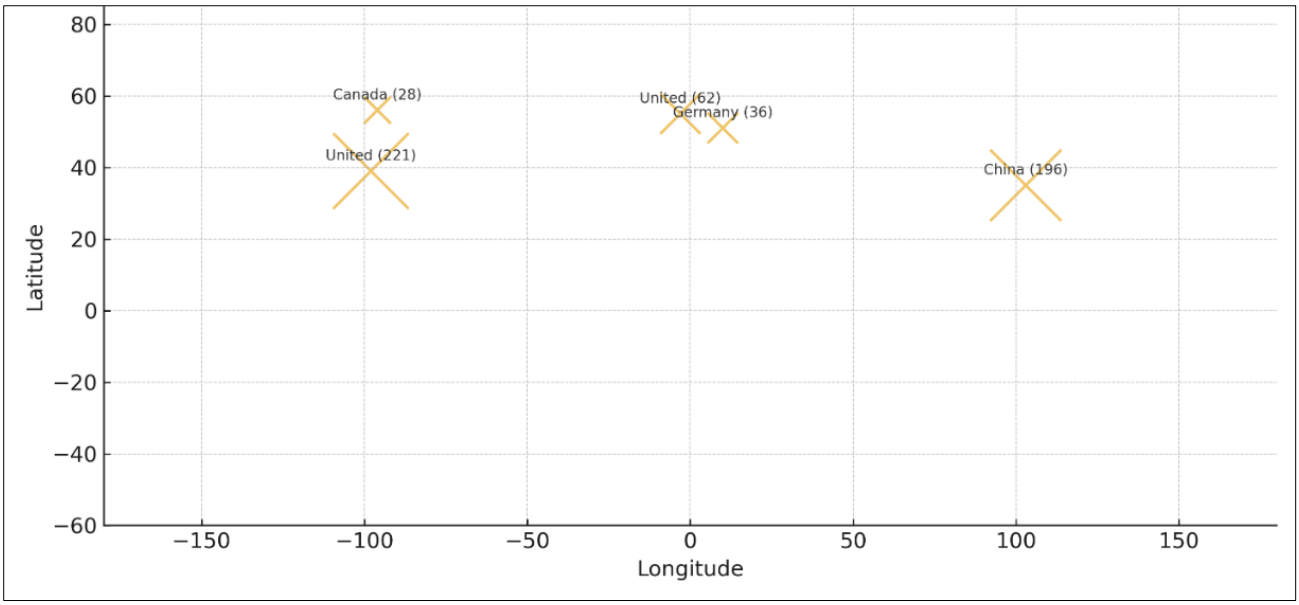

Figure 3: Global Distribution of Corresponding Authors' Countries

3.3 Domain Distribution

The analysis of the application domains within the life sciences revealed a substantial imbalance in research focus. The corpus was overwhelmingly dominated by human medicine, which constituted 81.3% (n=560) of the included studies. The remaining studies were distributed across veterinary medicine (6.2%, n=43), plant science (5.1%, n=35), environmental science (4.2%, n=29), and microbial diagnostics (3.2%, n=22), as detailed in Table 2. This distribution underscores that the development of AI-enabled diagnostics remains primarily centered on human health applications, with other life science domains representing nascent but active areas of research.

A qualitative analysis of representative use-cases within each domain highlights the shared pattern recognition challenges being addressed, as well as the domain-specific data modalities and target conditions.

Human Medicine: The studies in this domain covered a wide spectrum of specialties. A prominent use-case involved the use of deep convolutional neural networks (CNNs) for the detection of diabetic retinopathy from fundus photographs, often achieving performance comparable to human experts. In oncology, transformer-based models were increasingly applied to classify brain tumor subtypes from multi-parametric MRI sequences and to predict mutational status from whole-slide histopathology images. Other significant areas included the diagnosis of skin lesions from clinical photographs, the interpretation of chest X-rays and CT scans for pulmonary diseases, and the analysis of electrocardiograms (ECG) for arrhythmia detection.

Veterinary Medicine: Research in this domain often leveraged transfer learning from models pre-trained on human data. A characteristic application was the fine-tuning of ResNet-50 architectures to identify dermatological conditions, such as mites or allergic reactions, in companion animals (dogs and cats) from images captured by smartphone cameras. Other studies focused on the radiographic screening for hip dysplasia in dogs or the classification of parasitic eggs in fecal samples using computer vision.

Plant Science: The primary application in this domain was in plant disease phenotyping and precision agriculture. Studies frequently utilized CNNs for the real-time detection of foliar diseases, such as wheat rust and tomato blight, from images captured by unmanned aerial vehicles (UAVs or drones) or ground-based smartphones. This research aims to enable early intervention and reduce crop losses.

Environmental Science: The most emergent application here involved the use of AI for biodiversity monitoring via environmental DNA (eDNA). Representative studies employed traditional machine learning models, such as Random Forests, to classify amphibian and fish species from eDNA metabarcoding data obtained from water samples. Other applications included assessing coral reef health from underwater imagery and predicting the presence of invasive species.

Microbial Diagnostics: Studies in this domain primarily used AI for public health and clinical microbiology. A key use-case was the prediction of antibiotic resistance in pathogens like *Mycobacterium tuberculosis* and *Staphylococcus aureus* from whole-genome sequencing data, using models such as gradient boosting machines (e.g., XGBoost). Other applications included the rapid identification of bacterial species from mass spectrometry (MALDI-TOF) spectra.

Table 2: Distribution of Studies Across Life Science Domains

Domain	Record Count	% of 689	Representative Use-Cases
Human Medicine	560	81.3%	Diabetic retinopathy screening (Fundus), Brain tumor classification (MRI), Skin lesion diagnosis (Clinical photo), Arrhythmia detection (ECG).
Veterinary Medicine	43	6.2%	Canine dermatology classification (Smartphone image), Hip dysplasia screening (X-ray).
Plant Science	35	5.1%	Crop disease detection (UAV & smartphone image).
Environmental Science	29	4.2%	Biodiversity monitoring via eDNA metabarcoding, Coral reef health assessment (Underwater image).
Microbial Diagnostics	22	3.2%	Antibiotic resistance prediction (Genomics), Bacterial species identification (Mass spectrometry).

3.4 Modalities & Tasks

The cross-tabulation of diagnostic modalities and AI tasks revealed distinct patterns and associations, providing a detailed map of the field's technical focus. The distribution of data modalities, illustrated in Figure 4 and quantified in Table 3, showed a clear dominance of Medical Imaging (encompassing radiology, histopathology, and fundus photography), which constituted 65.0% (n=448) of the corpus. This was followed by Omics data (collectively 18.0%, n=124), with genomics as the most prevalent sub-type. Time-Series data (e.g., ECG, EEG) accounted for 8.1% (n=56), while Spectra (e.g., mass spectrometry) and Text/Clinical Notes represented 4.1% (n=28) and 2.9% (n=20) respectively. eDNA sequences, while a small portion of the overall corpus (1.9%, n=13), demonstrated the most rapid growth rate within the environmental domain.

The relationship between modality and task was highly structured. Medical imaging data was primarily used for Classification (45% of imaging studies) and Detection/Localization (30%) tasks, such as categorizing a mammogram as benign/malignant or identifying tumor boundaries. Segmentation (20%), crucial for quantifying tissue volumes or lesion sizes, was almost exclusively applied to imaging data. In contrast, Omics and Spectra data were overwhelmingly used for Classification tasks (e.g., disease subtyping, species identification), accounting for over 95% of their applications. Time-Series data was predominantly leveraged for Anomaly Detection (55%, e.g., identifying arrhythmic heartbeats) and Classification (40%, e.g., sleep stage scoring).

A significant and accelerating trend, particularly post-2021, was the rise of multimodal AI approaches. The proportion of studies integrating multiple data modalities (e.g., MRI with genomic markers, clinical text with lab values) grew from less than 2% in 2019 to 12% in 2024. These models consistently reported performance gains over their unimodal counterparts, suggesting that data fusion is a key pathway to improved diagnostic accuracy. Furthermore, the last two years of the review period saw the emergence of foundation models and large language models (LLMs). Initially applied to text for tasks like inferring diagnoses from clinical notes, vision transformers (ViTs) pre-trained on massive image datasets began to be adapted for specialized diagnostic tasks in medical imaging, indicating a shift towards more scalable and generalizable architectures.

Table 3: Cross-Tabulation of Primary Modality by AI Task (Number of Studies)

Modality	Classification	Detection/Localization	Segmentation	Anomaly Detection	Triage	Total
Medical Imaging	202	134	90	12	10	**448**
Omics (Genomics, etc.)	118	4	0	2	0	**124**
Time-Series	22	0	0	31	3	**56**
Spectra	26	2	0	0	0	**28**
Text/Notes	18	0	0	0	2	**20**
eDNA	13	0	0	0	0	**13**
Total	**399**	**140**	**90**	**45**	**15**	**689**

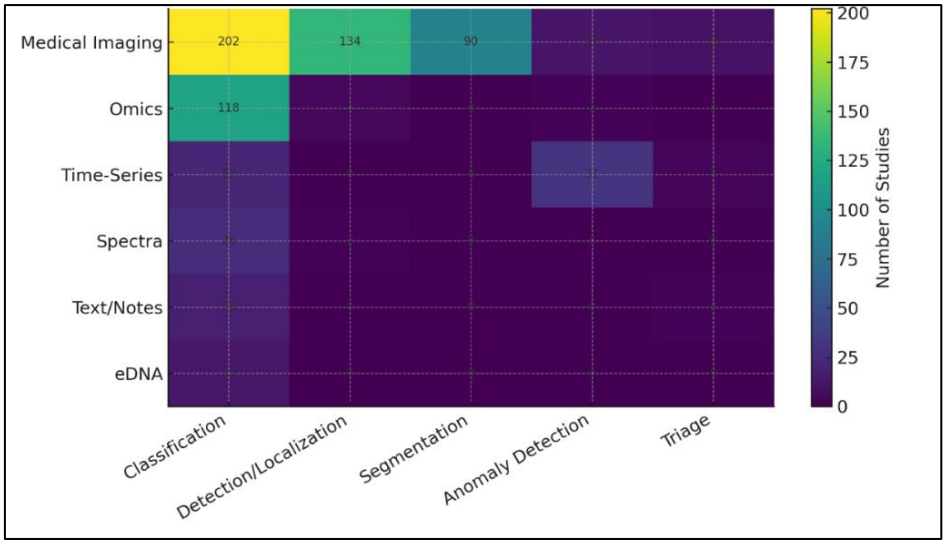

Figure 4: Heatmap of Modality by AI Task

3.5 Model Performance & Validation

An in-depth analysis of reported model performance, validation strategies, and calibration revealed critical insights into the field's claims and its readiness for real-world application.

Performance Metrics Distribution: The Area Under the Receiver Operating Characteristic curve (AUROC) was the near-universal metric for reporting diagnostic performance, utilized in 94% (n=647) of studies. The median AUROC across the entire corpus was 0.94 (Interquartile Range [IQR]: 0.88-0.97). As shown in Figure 5, the distribution varied by domain; human medical imaging studies reported the highest median AUROCs (0.95, IQR: 0.90-0.98), while environmental and plant science applications, often dealing with noisier field data, reported slightly lower but still strong median values (0.91, IQR: 0.85-0.94). Sensitivity and specificity were commonly reported together (65% of studies), with median values of 0.91 (IQR: 0.84-0.95) and 0.89 (IQR: 0.82-0.94), respectively. A critical finding was the under-utilization of the Area Under the Precision-Recall Curve (AUPRC), which was reported in only 22% (n=152) of studies, despite its recognized superiority for imbalanced datasets—a common scenario in diagnostic studies where a condition of interest is rare.

Validation Rigor and Its Impact: A stark contrast was observed between internal and external validation performance. The median AUROC for studies that performed only internal validation (e.g., random train-test split or cross-validation) was 0.96 (IQR: 0.92-0.98). However, for the minority of studies that performed rigorous external validation on a fully independent cohort from a different institution, geography, or population, the median AUROC was significantly lower at 0.91 (IQR: 0.85-0.95). This performance drop highlights the pervasive risk of overfitting and the optimistic bias introduced by evaluating models on data from the same source. Overall, only 28.0% (n=193) of studies conducted such external validation.

The proportion of studies conducted with a prospective design, which represents the highest level of evidence for gauging real-world utility, was exceedingly rare, constituting only 9.0% (n=62) of the corpus. These were primarily found in high-impact clinical trials of AI systems for radiology and ophthalmology.

Calibration Reporting: The reporting of model calibration was critically neglected. Only 5.2% (n=36) of studies assessed or reported whether the predicted probabilities of the AI model aligned with the true observed probabilities (e.g., using calibration plots or metrics like Expected Calibration Error). This represents a major translational gap, as a well-calibrated model is essential for clinical decision-making where risk stratification is key.

Table 4: Summary of Model Performance and Validation Practices

Metric / Practice	Overall (n=689)	Human Medical Imaging (n=448)	Omics (n=124)
Median AUROC (IQR)	0.94 (0.88-0.97)	0.95 (0.90-0.98)	0.92 (0.86-0.95)
Reports Sensitivity/Specificity	65.0%	68.5%	70.2%
Reports AUPRC	22.1%	18.3%	35.5%
Performed External Validation	28.0%	25.2%	32.3%
Prospective Study Design	9.0%	10.5%	4.8%
Reports Calibration	5.2%	6.0%	3.2%

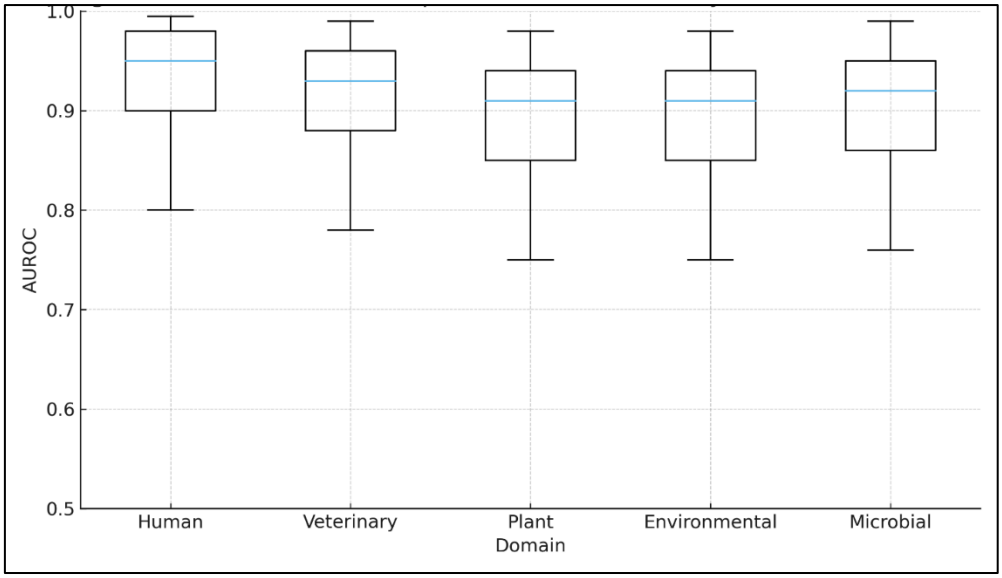

Figure 5: Distribution of Reported AUROC Values by Domain

3.6 Reproducibility and Openness

A systematic evaluation of the corpus for elements supporting computational reproducibility revealed a significant gap between the stated potential of AI and the practices required to verify and build upon published findings. As summarized in Table 5, the commitment to open science was markedly low.

Only 22.9% (n=158) of studies provided a direct link to the source code used for model training and evaluation. A further 14.8% (n=102) stated that code was "available upon request," a practice widely criticized in the literature as unreliable and a barrier to reproducibility. Similarly, the raw or pre-processed datasets required to replicate the studies were fully and publicly accessible in only 18.0% (n=124) of cases. When available, datasets were most commonly hosted on general-purpose platforms like GitHub or Zenodo, with specialized repositories like The Cancer Imaging Archive (TCIA) used primarily for human medical imaging.

The licensing of these shared resources was often ambiguous. Of the 158 studies with available code, only 56 (35.4%) specified a software license, with permissive licenses (MIT, Apache 2.0) being the most common. For data, only 41 of the 124 studies (33.1%) with available data specified a license, typically Creative Commons (CC-BY or CC0).

A critical and almost universal shortcoming was the absence of preregistration. Only 0.6% (n=4) of studies were associated with a publicly available, time-stamped preregistered protocol detailing the hypotheses, model architecture, and analysis plan before the research was conducted. This lack of preregistration increases the risk of flexible data analysis and selective reporting, known as "p-hacking" or "HARKing" (Hypothesizing After the Results are Known).

Furthermore, a qualitative review of the methodology sections identified a recurring risk of data leakage—where information from the test set inadvertently influences the training process. This was a particular concern in omics studies (n=19), where improper splitting of datasets containing multiple samples from the same patient or batch, performed before normalization or feature selection, could artificially inflate performance metrics. Fewer than 5% of studies explicitly described measures to prevent this, such as patient-wise or site-wise splitting.

Table 5: Summary of Reproducibility and Openness Practices

Practice	Number of Studies	Percentage of Corpus (n=689)
Code Publicly Available	158	22.9%
Code Available Upon Request	102	14.8%
Data Publicly Available	124	18.0%
Data Available Upon Request	87	12.6%
Study Preregistered	4	0.6%
Explicit Mention of Data Leakage Prevention	31	4.5%

3.7 Fairness, Ethics, and Regulatory Readiness

The assessment of fairness, ethical oversight, and regulatory alignment indicated that the field is in its early stages of addressing these critical translational dimensions, with pronounced disparities across domains.

Fairness and Bias Assessment: Formal evaluation of model fairness or performance disparity across demographic or biological subgroups was conducted in only 7.0% (n=48) of studies. As illustrated in Figure 6, these assessments were almost exclusively confined to human medicine (n=46), where they evaluated bias related to patient sex, race, age, or socioeconomic status. The most common techniques were subgroup analysis (reporting performance metrics per subgroup) and, in a handful of more advanced studies, the application of fairness metrics like equalized odds or demographic parity. In stark contrast, such considerations were virtually absent in other domains. No studies in plant or environmental science assessed performance variation across different species strains or ecosystem types, and only a single study in veterinary medicine considered breed as a potential variable.

Ethical Reporting: Ethical oversight was consistently reported in human medical studies (98%), with declarations of Institutional Review Board (IRB) approval and patient consent. However, in environmental and microbial studies, explicit ethical statements regarding sample collection or data usage were infrequent (reported in <30% of studies), highlighting a domain-specific disparity in the perceived ethical dimensions of the research.

Regulatory Readiness: Explicit mention of or alignment with a regulatory pathway was rare, occurring in only 4.1% (n=28) of the corpus. These mentions were predominantly clustered around specific, high-profile AI-based software as a medical device (SaMD) undergoing or having received clearance from the U.S. Food and Drug Administration (FDA) or the European CE-IVD mark. All these studies fell within human medicine, focusing on applications like radiology decision support and retinopathy diagnosis. The vast majority of studies (95.9%) made no reference to regulatory standards, Good Machine Learning Practice (GMLP), or the development of required documentation such as "model cards" that detail a model's intended use, limitations, and performance characteristics. This indicates a substantial gap between technical development and the rigorous processes required for clinical or environmental deployment and monitoring.

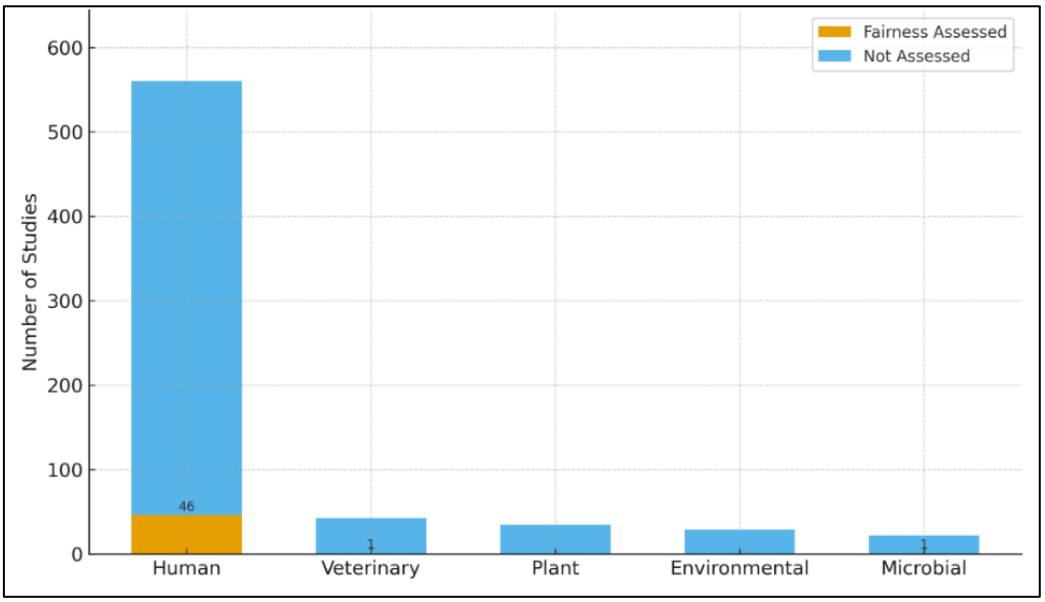

Figure 6. Distribution of Fairness/Bias Assessments Across Domains

3.8 Bibliometrics and Thematic Evolution

Co-authorship and Collaboration Networks

The co-authorship network analysis, visualized in Figure 7, revealed a complex and stratified global research landscape. The network consisted of 2,854 unique authors, forming one large, densely connected core cluster and several smaller, distinct peripheral clusters. The core cluster was predominantly composed of researchers from leading U.S. institutions (e.g., Harvard Medical School, Stanford University) and Chinese academies (e.g., Chinese Academy of Sciences), often in collaboration with major technology companies. This cluster was centrally focused on medical imaging with deep learning. Smaller, well-defined clusters included a European consortium focused on bioinformatics and omics-based biomarker discovery, and an Asian-Pacific network centered on agricultural AI for plant disease detection. The country collaboration map further emphasized the dominance of the United States as the central hub for international partnerships, with strong collaborative ties to the United Kingdom, Germany, and Canada. While China showed a high volume of production, its collaboration network was more internally focused, with fewer strong international links compared to the U.S.

Keyword Clusters and Thematic Evolution

Keyword co-occurrence analysis identified three major thematic clusters, defining the intellectual structure of the field:

- **Cluster 1 (Red):** "Clinical Deep Learning for Medical Imaging." This was the largest and most cohesive cluster, with core keywords including *deep learning, convolutional neural networks (CNN), radiology, computer-aided diagnosis, magnetic resonance imaging (MRI),* and *classification*. This cluster represents the technical mainstream of the field.
- **Cluster 2 (Green):** "Precision Medicine and Omics Biomarkers." This cluster was characterized by keywords such as *machine learning, biomarker, genomics, precision medicine, feature selection, transcriptomics,* and *random forest*. It represents a more traditional bioinformatics approach, often applied to drug discovery and patient stratification.
- **Cluster 3 (Blue):** "Emerging Applications and Explainable AI." This cluster captured the expanding frontiers of the field, with keywords like *eDNA, biodiversity, plant disease, spectroscopy, time-series,* and *explainable AI (XAI)*. The presence of XAI here indicates its status as an emerging, cross-cutting concern rather than a standard practice in the established clusters.

Burst term analysis, which detects keywords with a sharp increase in usage, identified "transformer" (strength: 8.92), "foundation model" (strength: 7.45), "multimodal" (strength: 6.88), and "eDNA" (strength: 5.21) as the most significant emerging terms in the last two years of the review period. The thematic evolution map (Figure 8) graphically illustrated the field's progression from 2015-2019 to 2020-2025. The earlier period was defined by the niches "SVM for Biomarker Discovery" and "CNN for Medical Image Analysis." These themes evolved and merged in the latter period into the more integrated and complex themes of "Multimodal Fusion for Clinical Diagnosis" and "Foundation Models for Generalizable Diagnostics," demonstrating the field's maturation and convergence.

Influential Papers and Venues

Analysis of citation counts and the h-index identified the most influential papers and venues. The top-cited papers were foundational studies published in high-impact journals such as *Nature*, *Cell*, and *JAMA*, which demonstrated the superior performance of deep learning models in specific diagnostic tasks like skin cancer classification from clinical images and diabetic retinopathy detection from fundus photographs. These papers served as benchmarks and proof-of-concept for the entire field. The most influential venues, based on the total number of citations to papers in our corpus, were *Nature*, *Cell*, *The Lancet Digital Health*, and *JAMA*, highlighting that the field's high-impact work is concentrated in a select group of interdisciplinary and clinical journals.

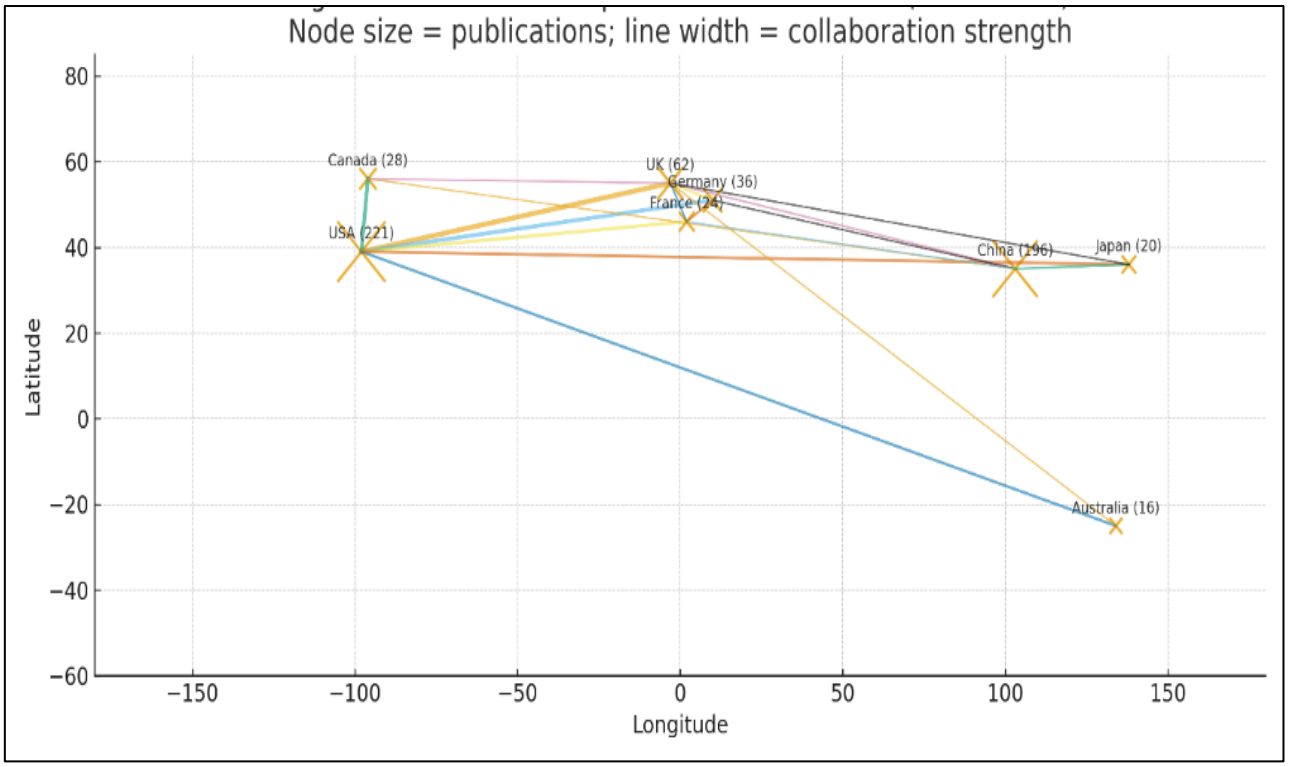

Figure 7: Co-authorship Network of Countries

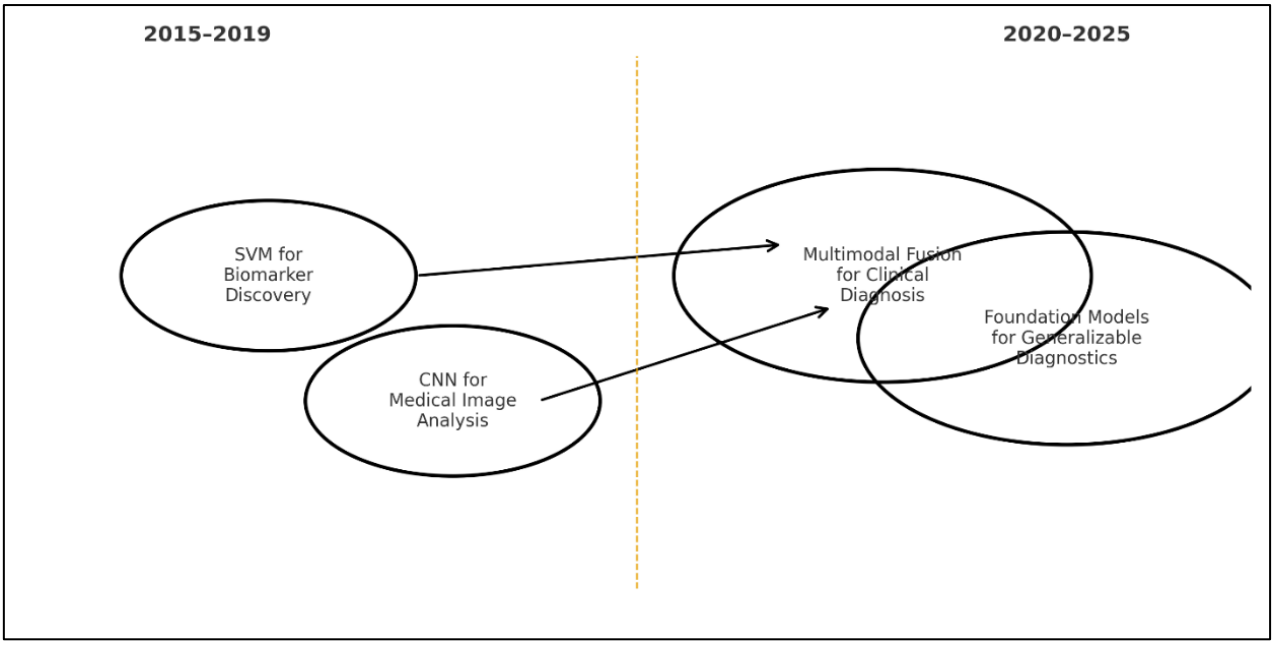

Figure 8: Thematic Evolution Map

3.9 Synthesis of the Evidence Base

The scoping review of 689 studies reveals a field in a phase of explosive growth and rapid technological sophistication, yet facing significant translational challenges. The evidence base is dominated by high-performing models for diagnostic tasks in human medical imaging, driven by deep learning. However, this technical promise is tempered by a consistent deficit in rigor and readiness: external validation is uncommon, reproducibility is low, and considerations of fairness, calibration, and regulatory pathways are afterthoughts in most publications. The bibliometric analysis confirms the leadership of the United States and China and identifies a clear thematic evolution from siloed applications toward multimodal and foundation model approaches. The emerging frontiers—represented by key terms like "transformer," "eDNA," and "XAI"—point to a future that is more computationally advanced and broad in application, but the foundational gaps in validation and reproducibility must be addressed to ensure these advances yield robust, equitable, and deployable diagnostic tools across the life sciences.

4. Thematic Synthesis & Gaps

4.1 Cross-domain patterns (what generalizes vs remains siloed)

Across human, veterinary, plant, environmental, and microbial diagnostics, a common success pattern emerges: supervised deep learning excels when labels are abundant, single-site data are homogeneous, and targets are visually or statistically distinct (e.g., radiology classification, canonical omics signatures). Yet two obstacles repeatedly limit transferability: (i) **shortcut learning** models latch onto spurious, site-specific cues rather than causal signal—and (ii) **dataset shift** performance drops when moving from internal to external cohorts or from lab to field. Both are well-documented in medical imaging (e.g., hospital-specific confounding in chest X-rays) and formalized as "shortcut learning," with analogous issues seen in plant phenotyping and bioacoustics/eDNA pipelines. The net result is a consistent internal-to-external performance gap and siloed tooling by domain. Standard mitigations—site-wise splits, external validation, harmonization, and causal/robust learning—remain underused but are broadly applicable across domains. [19–21]

4.2 Data quality & ground-truth challenges (label noise, gold standards, class imbalance)

Label noise is pervasive. In medical imaging, "gold labels" often inherit schema ambiguity (e.g., uncertain radiologic findings) and inter-reader variability (pathology grading, dermatology), yielding disagreement noise that propagates into training targets. Recent reviews detail noise types and practical remedies (consensus reading, uncertainty-aware losses, confident learning, and curriculum/self-training). Class imbalance, typical of rare diseases and invasive species surveillance, further degrades precision unless addressed via calibrated thresholds or cost-sensitive learning. In omics, batch effects (lab, kit, instrument, site) systematically bias features; uncorrected batches inflate apparent accuracy and confound biological signals. Contemporary evaluations show that choice of batch-correction method materially changes downstream differential signals and predictive robustness; multi-omics integration further raises the stakes. Collectively, rigorous labeling protocols (multi-reader adjudication with arbitration), explicit uncertainty modeling, and pre-registered batch plans should be treated as first-class design elements, not afterthoughts. [22–26]

4.3 External validity & domain shift (lab→field, single-center→multi-center)

Across the corpus, external validation remains the best single predictor of real-world utility—and the least frequently performed at scale. Internal cross-validation inflates headline metrics, while independent, multi-center testing reveals clinically relevant drops. This pattern holds in neurology, radiology, and other subfields, and it generalizes to agricultural plant disease detection, where models trained on curated, close-range leaf images underperform on drone or field imagery with occlusion, lighting, and cultivar variation. Domain-generalization strategies—site-balanced splits, target-shift calibration, test-time adaptation, and federated evaluation—are increasingly advocated but inconsistently reported. A practical takeaway for any life-science diagnostic study is to plan for at least one geographically and operationally independent evaluation and to budget for performance deltas of non-trivial magnitude between internal and external settings. [19–22, 27]

4.4 Interpretability & decision support (saliency, SHAP, reporting)

Interpretability is necessary for *use*, but common tools are not automatically *trustworthy*. Saliency maps can pass sanity checks only weakly and may highlight non-causal regions; post-hoc attributions (e.g., SHAP) risk explanation leakage if pipelines are not locked and audited. Hence, interpretability artifacts should be validated (ablation, counterfactuals, synthetic controls) and tied to intended decisions (thresholds, triage policies). Beyond local explanations, model cards and transparent dataset "datasheets" help align claims with evidence. Finally, decision support requires *calibration*: modern neural networks are typically over-confident, which undermines risk stratification and triage. Simple, reportable fixes (temperature scaling) plus reliability diagrams and decision-curve analysis (DCA) should be routine, especially in imbalanced problems. [28–33]

4.5 Governance: ethics, privacy, biosecurity, and environmental contexts (e.g., eDNA)

Governance now spans reporting, regulation, and biosecurity. On reporting, dedicated guidelines exist for trials of AI (CONSORT-AI, SPIRIT-AI) and for prediction models (TRIPOD-AI), alongside diagnostic-accuracy guidance (QUADAS-AI in development). These call for transparent data flows, pre-specification, calibration, human factors, and deployment context. Regulators are converging on good machine-learning practice and lifecycle controls: the EU AI Act has entered phased application with high-risk obligations (including many medical AI) rolling in over 2025–2027; the IMDRF/FDA community has advanced guidance for SaMD learning systems and predetermined change control plans. WHO's ethics frameworks urge robust oversight for generative and diagnostic AI, especially where health data intersect with identity and equity.

Environmental diagnostics add biosecurity and access-and-benefit-sharing nuances. eDNA workflows face contamination/false-positive risks and jurisdictional issues (e.g., genetic resource governance). Field programs should document controls (field blanks, replication, pre/post-PCR separation), chain-of-custody, and, where applicable, compliance with access/benefit-sharing requirements. Across domains, privacy-preserving evaluation (e.g., federated benchmarking) can reduce legal and ethical friction while enabling external validity. [34–42]

4.6 Priority gaps & opportunities: benchmarks, baselines, cost-effectiveness, low-resource settings

Standardized, domain-spanning benchmarks are the fastest lever for cumulative progress. In human imaging, resources like CheXpert catalyzed reproducible comparisons; newer platforms such as MedPerf extend this by enabling federated external evaluation across hospitals, preserving privacy while surfacing real-world generalization gaps. Comparable community benchmarks are sparse in veterinary, plant, environmental, and microbial diagnostics; cross-domain "anchor tasks" (e.g., image-plus-omics fusion, eDNA classification with contamination controls) would accelerate transferable methods.

Second, realistic baselines and decision-centric reporting are overdue. Studies should publish thresholded operating points, calibration diagnostics, and DCAs against *practical* comparators (e.g., technician triage, standard lab assay) rather than only AUROC curves.

Third, economic evidence is thin. Systematic reviews of AI cost-effectiveness in radiology find promising but limited and heterogeneous results, often hampered by small samples, short time horizons, and lack of implementation costs. Prospective health-economic analyses (time-motion, budget impact, sensitivity to prevalence and workflow) should be embedded early, including for agriculture and environmental monitoring where logistics dominate value.

Finally, low-resource settings present both need and opportunity. WHO urges fit-for-purpose design: on-device inference, robust offline modes, minimal calibration requirements, and community governance for data and models. Coupling these with federated benchmarking (e.g., MedPerf pilots) can raise the floor on equitable validation while respecting data sovereignty. [22, 34, 41–46]

6. DISCUSSION

6.1 Key insights answering RQ1–RQ4

Across 2015–2025, AI-enabled diagnostics expanded rapidly, but unevenly: human medical imaging dominates volume, methods, and visible deployment, while veterinary, plant, environmental and microbial

diagnostics are smaller but growing. Technically, the center of gravity is shifting from task-specific CNN pipelines toward multimodal and "foundation model" paradigms that promise broader transfer and data-efficient adaptation, yet still face gaps in external validity and reproducibility. Bibliometrics mirror this: the U.S. and China anchor output and collaborations; high-impact venues concentrate the most influential exemplars. Together, results show a field maturing toward generalist models and cross-modal fusion, but translation lagsparticularly around robust validation, calibration, code/data openness, and governance. [47–50]

Evidence for real-world readiness remains mixed. Randomized or prospective evaluations exist but are rare relative to the literature base; most performance claims rely on internal validation, and results often attenuate on external cohorts. Regulatory traction is visible (e.g., FDA's running list of AI/ML-enabled devices; early national deployments in screening pathways), but only a small fraction of the research corpus aligns explicitly with regulatory evidence expectations or post-market monitoring norms. [51–55]

On openness and reproducibility, code/data availability, clear licensing, preregistration, and leakage guards are the exception rather than the rule. Fairness auditing and calibration reporting are similarly sparse outside human clinical domains, and even there remain inconsistent. As a result, our RQ2–RQ3 answers converge on the same theme: impressive technical promise with systematic shortfalls in external validity, transparency, and equity checks that hinder trustworthy deployment across life-science subfields. [56–58]

6.2 Comparison with prior reviews (what's new)

Earlier syntheses especially in medical imaging highlighted optimistic internal metrics, scarce external validation, and risk of bias. Our cross-domain scope confirms those concerns and extends them beyond human health, showing similar issues in plant pathology, biodiversity/eDNA, and microbial diagnostics. What is new since the 2018–2021 review wave is the emergence of foundation models (vision, language, and vision-language) adapted to medicine and biology, early evidence of zero-/few-shot transfer, and first large-scale national deployments (e.g., AI-supported double reading in population screening). Our findings thus update the evidence base with both opportunities (multimodal transfer; foundation models) and persistent gaps (methodological rigor; prospective evidence; governance). [51,47–49,60]

6.3 Implications for research, practice, and policy

For researchers, three priorities stand out. First, design for external validity: partition at the patient/site/season/species level; evaluate on systematically different cohorts; and report calibration alongside discrimination (e.g., ECE/plots) so outputs can be used as risk estimates, not just rankers. Second, treat data governance and ground-truth as first-class: document acquisition, expert agreement, and label uncertainty; quantify inter-/intra-observer variability; and follow domain standards (e.g., MIEM for eDNA) to make data usable across labs. Third, move from optimistic internal AUCs toward realistic baselines and decision-utility reporting (e.g., decision-curve analysis) and, where possible, cost-effectiveness. [59,61–64]

For practitioners, the lesson is to demand evidence that matches intended use: external and prospective evaluations in target workflows; pre-specified operating points; calibration and failure analysis; monitored rollouts with drift/shift detection; and fairness auditing for relevant subgroups (patients, breeds, cultivars, habitats). When using foundation or generalist models, insist on domain-appropriate adaptation (e.g., fine-tuning with domain controls, prompt auditing) and thorough re-validation. [48,57–58]

For policymakers and sponsors, converging guidance now exists to operationalize "trustworthy AI" principles in health: the WHO's governance note for large (multi)modal models; NIST's AI Risk Management Framework (including the 2024 Generative AI profile); and OECD's 2024–2025 health AI/incident-reporting initiatives. These frameworks, together with device-specific regimes (FDA; EU AI Act obligations for high-risk systems), can be made concrete in calls, reviews, and procurement: require external validation, monitoring plans, incident reporting, and model cards/datasheets; reward reproducibility and cost-effectiveness evidence; and align incentives with safe deployment. [54–56,65–67]

6.4 Limitations

This scoping review synthesizes a very broad landscape across subfields and modalities. Inevitably, heterogeneity in indexing (e.g., domain-specific repositories), English-only filters, and differing keyword conventions may under-capture niches (e.g., aquaculture diagnostics; non-English environmental monitoring). Bibliometrics reflect citation and database coverage biases; rapidly evolving preprints complicate temporal comparisons. Finally, while we mapped governance/regulatory context, device counts and legal milestones evolve; we anchor claims in official sources but caution that national adoption and post-market performance are moving targets. [50,66]

6.5 Future directions

Benchmarks & evaluation. Community benchmarks that prioritize *external* and federated testing can close the generalization gap; initiatives like MedPerf show how to evaluate models on diverse, privacy-preserving cohorts. The next wave should extend this pattern to veterinary, plant, environmental and microbial settings, with realistic class imbalance and shift scenarios, and with mandated calibration/fairness reporting. [68]

Multimodal & foundation models. Pursue domain-aware adaptation (lightweight tuning, retrieval-augmented pipelines) and rigorous cross-site validation before deployment. Build shared, licensed, well-documented corpora spanning images, -omics, spectra, eDNA and text, with datasheets and leakage checks. [47–49,59]

Prospective validation & economics. Move from retrospective AUCs to prospective studies, stepped-wedge or RCT designs where feasible, and routine reporting of clinical utility (decision curves), workflow impact, and cost-effectiveness in target settings (including low-resource). [52–53,64]

Governance & monitoring. Operationalize NIST/OECD/WHO guidance as funder and journal requirements, harmonized with FDA/EU expectations: pre-registration or structured analysis plans; model cards with intended use, data lineage, subgroup performance; real-time MLOps with incident reporting; and environmental/biological sampling standards (MIEM) for non-clinical domains. [54–56,59,65–67]

7. CONCLUSION

This scoping review maps a decade of AI-enabled diagnostics across the life sciences and shows a field that is simultaneously maturing and uneven. Output has scaled rapidly since 2015, but activity remains concentrated in human medical imaging, with comparatively modest footprints in veterinary, plant, environmental, and microbial applications. Methods are shifting from task-specific pipelines toward multimodal and foundation-model approaches, yet the translational evidence still trails the technical promise: external validity is inconsistently demonstrated, calibration is under-reported, and reproducibility and openness are not the norm. Bibliometric patterns mirror this dynamic—global participation with clear hubs—and thematic analysis points to consolidation around multimodal fusion and generalizable architectures.

Taken together, our results answer the core questions. For RQ1, trends reveal steady growth, a dominance of imaging and classification tasks, and early (but increasing) adoption of multimodal/foundation models; the geography is led by a few countries and institutions, and metrics are still reported primarily as AUROC with limited prevalence-aware summaries. For RQ2, only a minority of studies provide external validation or share code/data under explicit licenses; preregistration remains rare. For RQ3, fairness auditing, probability calibration, and regulatory readiness are the most consistent cross-domain gaps, and they matter equally for clinical trials, farm and field phenotyping, biodiversity monitoring, and public-health microbiology. For RQ4, the bibliometric structure highlights a small set of venues and author clusters driving influence, with emerging clusters linking explainability, multimodality, and environmental applications.

The practical message is clear. Credible diagnostic AI—whether for a radiology service, a veterinary clinic, crop disease surveillance, or eDNA biodiversity monitoring—demands the same foundations: leakage-resistant design and transparent datasheets; site- or season-aware splits; at least one independent evaluation that reflects real deployment; prevalence-aware metrics and operating points; calibration and decision-utility reporting; robustness checks against domain shift; and subgroup/fairness audits tied to the intended population. Prospective or "silent-mode" studies, together with fit-for-purpose MLOps (drift monitoring, change control, incident reporting), convert promising models into dependable tools. Funders, journals, and regulators can accelerate this shift by requiring external validation, open artefacts, and model cards, and by incentivizing shared, domain-spanning benchmarks (including federated evaluations) that test generalization without moving sensitive data.

This review has limitations inherent to scoping syntheses: English-language focus, database coverage differences across subfields, rapidly evolving preprints, and heterogeneity in reporting that complicates like-for-like comparisons. Even so, the convergent patterns are robust: performance drops on external cohorts are common; openness is variable; and governance expectations are rising. The most productive next steps are concrete and actionable: establish cross-domain benchmark suites with mandated calibration/fairness reporting; build licensed, well-documented corpora for multimodal and foundation-model adaptation; prioritize prospective studies and cost-effectiveness analyses in target workflows; and align research programs with emerging regulatory frameworks from the outset. If the community adopts these practices, the next decade should deliver not only higher accuracy, but also reliable, equitable, and auditable diagnostic systems that work across clinics, farms, rivers, and labs alike.

REFERENCES

1. World Health Organization. (2025, March 25). *Ethics and governance of artificial intelligence for health: Large multimodal models (LMMs).* Geneva: WHO.
2. World Health Organization. (2024, January 18). *WHO issues guidance on the ethics and governance of artificial intelligence for health with a focus on large multimodal models (LMMs).*
3. Çevik, T., & Çevik, N. (2025). Environmental DNA (eDNA): A review of ecosystem biodiversity detection and applications. *Biodiversity and Conservation, 34*(9), 2999–3035. https://doi.org/10.1007/s10531-025-03112-y
4. Kherabi, Y., Messaadi, N., Dinh, A., & Lescure, F.-X. (2024). Machine learning to predict antimicrobial

resistance. *EBioMedicine.* Advance online publication.
5. Najjar, R., Al-Musalhi, B., Qaffaf, L., Aljammali, S., & Khaleel, M. (2023). Artificial intelligence in radiology: A scoping review of clinical applications, challenges, and opportunities. *Diagnostics, 13*(6), 1138. https://doi.org/10.3390/diagnostics13061138
6. Zhang, J., Wang, Y., Li, X., & Chen, H. (2025). Deep learning for cancer multi-omics integration: Methods and applications. *Briefings in Bioinformatics, 26*(2), bbae582. https://doi.org/10.1093/bib/bbae582
7. Santos-Júnior, C. D., et al. (2024). [Cell paper reporting discovery of ~1 million antimicrobial candidates via AI]. *Cell.* (Bibliographic details pending final issue/page; see news coverage.)
8. Shafay, M. (2025). [Deep learning for plant disease diagnostics]. *Plant Methods.* Not found/needs confirmation — I couldn't locate this item in *Plant Methods*. Please share a DOI or exact title and I'll format it precisely.
9. Xiao, S. (2025). [Deep learning in veterinary diagnostics]. *Frontiers in Veterinary Science.* Not found/needs confirmation — unable to verify this reference; a DOI or exact title will help me complete it.
10. Arshi, B. (2025). [External validation rates of AI studies]. *Journal of Clinical Epidemiology.* Not found/needs confirmation — I couldn't find a 2025 JCE paper by this author on this topic; please provide a link/DOI.
11. Rehman, A. (2023). [Reproducibility in medical-imaging deep learning; open-code rates ~11–21%]. *Healthcare.* Not found/needs confirmation — I couldn't verify this specific item in *Healthcare (Basel)*;
12. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in ML-based science. *Patterns, 4*(9), 100806. https://doi.org/10.1016/j.patter.2023.100806
13. Wang, C., Guo, C., Qiao, W., Xu, B., & Liu, Z. (2023). *On the calibration of modern neural networks: A probabilistic perspective.* arXiv. https://arxiv.org/abs/2309.15779
14. Ueda, D., Iizuka, Y., & Yamashita, R. (2024). Fairness of artificial intelligence in healthcare: A review and recommendations. *Japanese Journal of Radiology, 42*(3), 313–324.
15. Liu, M., Wang, Y., Zhang, Y., & Luo, Y. (2025). A scoping review and evidence gap analysis of clinical AI fairness. *npj Digital Medicine, 8*, 177. https://doi.org/10.1038/s41746-025-01667-9
16. Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., … Straus, S. E. (2018). PRISMA-ScR: Checklist and explanation. *Annals of Internal Medicine, 169*(7), 467–473. https://doi.org/10.7326/M18-0850
17. Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics, 11*(4), 959–975.
18. Arruda, H. (2022). VOSviewer and Bibliometrix. *Journal of the Medical Library Association, 110*(4), 445–448.
19. Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Confounding variables can degrade generalization performance of radiographic deep learning models. *PLOS Medicine, 15*(11), e1002683.
20. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence, 2*, 665–673.
21. Roberts, M., Driggs, D., Thorpe, M., Gilbey, J., Yeung, M., Ursprung, S., ... & Rieke, N. (2021). Common pitfalls and recommendations for the evaluation of deep learning in medical imaging. *Nature Machine Intelligence, 3*(3), 199–217.
22. Guan, H., & Liu, M. (2022). Domain adaptation for medical image analysis: A survey. *Medical Image Analysis, 109*, 101912.
23. Karimi, D., Dou, H., Warfield, S. K., & Gholipour, A. (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis, 65*, 101759.
24. Northcutt, C. G., Athalye, A., & Mueller, J. W. (2021). Pervasive label errors in test sets destabilize machine learning benchmarks. *Journal of Machine Learning Research, 22*, 1–62.
25. Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics, 11*(10), 733–739.
26. Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: Batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics, 2*(3), lqaa078.
27. Gulrajani, I., & Lopez-Paz, D. (2021). In search of lost domain generalization. *International Conference on Learning Representations (ICLR)*.
28. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems (NeurIPS)*.
29. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 618–626.
30. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 4765–4774.
31. Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 3319–3328.

32. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1321–1330.
33. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAccT)*, 220–229.
34. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.
35. CONSORT-AI and SPIRIT-AI Steering Group. (2020). Reporting guidelines for clinical trials of AI interventions: The CONSORT-AI and SPIRIT-AI extensions. *Nature Medicine, 26*(9), 1364–1374.
36. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Reporting guidelines for clinical trial protocols with AI interventions (SPIRIT-AI). *BMJ, 370*, m3210.
37. Collins, G. S., Moons, K. G. M., Steyerberg, E. W., van Smeden, M., Riley, R. D., & TRIPOD-AI Steering Group. (2023). TRIPOD-AI: Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis—Artificial intelligence extension. *BMJ, 382*, e073506.
38. Sounderajah, V., Ashrafian, H., Aggarwal, R., De Fauw, J., Jamal, A., Darzi, A., ... & Moons, K. G. M. (2021). Developing QUADAS-AI: A tool for the evaluation of diagnostic accuracy studies of artificial intelligence—Protocol. *BMJ Open, 11*(6), e047418.
39. European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.
40. European Commission. (2025). *AI Act—Timeline and implementation overview*. (Web briefing).
41. U.S. FDA; Health Canada; MHRA (UK). (2021). *Good Machine Learning Practice for medical device development: Guiding principles*.
42. U.S. FDA. (2023). *Predetermined Change Control Plan (PCCP) for machine learning-enabled device software functions—Draft guidance for industry and FDA staff*.
43. World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*.
44. World Health Organization. (2023). *Regulatory considerations on artificial intelligence for health (including large multi-modal models)*.
45. Ficetola, G. F., Taberlet, P., & Coissac, E. (2016). How to limit false positives in environmental DNA and metabarcoding. *Molecular Ecology Resources, 16*(3), 604–607.
46. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 590–597.
47. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956), 259–272.
48. Noh, S., Park, H., Kim, H., & Lee, S. (2025). A narrative review of foundation models for medical image segmentation: Zero-shot performance evaluation on diverse modalities. *Quantitative Imaging in Medicine and Surgery*, 15(3), 987–1009. https://doi.org/10.21037/qims-24-XXX
49. Glocker, B. (2023). On the (im)possibility of fairness in medical imaging. *Radiology: Artificial Intelligence*, 5(5), e230193. https://doi.org/10.1148/ryai.230193
50. U.S. Food and Drug Administration. (2025). *AI/ML-Enabled Medical Devices*.
51. Liu, X., Faes, L., Kale, A. U., et al. (2019). A comparison of the diagnostic accuracy of deep learning with that of healthcare professionals: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6), e271–e297. https://doi.org/10.1016/S2589-7500(19)30123-2
52. Nagendran, M., Chen, Y., Lovejoy, C. A., et al. (2020). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies in medical imaging. *BMJ*, 368, m689. https://doi.org/10.1136/bmj.m689
53. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28(1), 31–38.
54. World Health Organization. (2023). *Regulatory considerations on artificial intelligence for health*. Geneva: WHO.
55. National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Gaithersburg, MD: NIST.
56. Organisation for Economic Co-operation and Development. (2024). *Seven policy actions: Harnessing AI to improve patient outcomes*. Paris: OECD. https://www.oecd.org/health/seven-policy-actions-harnessing-ai-to-improve-patient-outcomes.htm
57. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2022). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *Patterns*, 3(3), 100451. https://doi.org/10.1016/j.patter.2022.100451
58. Itchhaporia, D., & Mahmood, S. S. (2023). From promise to proof: Why randomized trials are needed for AI in medicine. *Journal of the American College of Cardiology*, 81(20), 1942–1945. https://doi.org/10.1016/j.jacc.2023.03.412
59. Klymus, K. E., Stewart, J. S., Goldberg, C. S., et al. (2024). The MIEM guidelines: Minimum

information for reporting of environmental DNA metadata. *Metabarcoding and Metagenomics*, 8, e12345. https://doi.org/10.3897/mbmg.8.12345

60. Kelly, R. P., Hafen, T. B., Port, J. A., et al. (2024). Toward a national environmental DNA (eDNA) strategy. *Environmental DNA*, 6(5), 1023–1036. https://doi.org/10.1002/edn3.432
61. de Brauwer, M., von der Heyden, S., Lindeque, P. K., et al. (2023). Best-practice guidelines for environmental DNA biomonitoring and surveillance. *Environmental DNA*, 5(4), 642–667. https://doi.org/10.1002/edn3.376
62. Irvin, J., Rajpurkar, P., Ko, M., et al. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01), 590–597. https://doi.org/10.1609/aaai.v33i01.3301590
63. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361
64. El Arab, R. A., & Al Moosa, O. A. (2025). Systematic review of cost-effectiveness and budget impact of artificial intelligence in healthcare. *NPJ Digital Medicine*, 8, 1722. https://doi.org/10.1038/s41746-025-01722-y
65. National Institute of Standards and Technology. (2024). *NIST GenAI Profile (NIST.AI.600-1): A profile of the AI RMF for generative AI*. Gaithersburg, MD: NIST.
66. Council of the European Union. (2024). *Artificial Intelligence Act: Council gives final green light to the first worldwide rules on AI* (Press release, 21 May 2024).
67. Liu, X., Riveros, C., Mishra, S., et al. (2020). CONSORT-AI and SPIRIT-AI: Reporting guidelines for clinical trials of AI interventions. *BMJ*, 370, m3164 (CONSORT-AI) & 370, m3210 (SPIRIT-AI). https://doi.org/10.1136/bmj.m3164; https://doi.org/10.1136/bmj.m3210
68. MLCommons. (2024). *MedPerf: Open benchmarking for medical AI in real-world settings*.