



REVIEW

YOLO in Medical Imaging: Evolution, Clinical Translation, and Future Directions

Hue T. Tran^{1,2}, Nguyen Nhu Son³, Anh Nguyen Trong⁴, Anh Nguyen Tu⁵, Phat T. Nguyen^{6,*} and Giang L. Nguyen^{3,*}

¹Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

²Institute of Information Technology, Vietnam Women's Academy, Hanoi, Vietnam

³Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

⁴Department of Information Technology, Dai Nam University, Hanoi, Vietnam

⁵Managing Director, White Neuron Co., Ltd., Hanoi, Vietnam

⁶Department of Circuit Theory and Measurement, Faculty of Radio-Electronic Engineering, Le Quy Don Technical University, Hanoi, Vietnam

*Corresponding Authors: Phat T. Nguyen. Email: nguyenphat@lqdtu.edu.vn; Giang L. Nguyen. Email: nlgiang@ioit.ac.vn

Received: 01 June 2026; Accepted: 17 August 2026; Published: 22 October 2026

ABSTRACT: Over the past decade, rapid advances in artificial intelligence have substantially expanded applications in medical image analysis, computer-aided diagnosis, and clinical decision support. Among object detection architectures, the YOLO (You Only Look Once) family has emerged as one of the most widely adopted approaches due to its real-time inference capability. However, transferring YOLO from general computer vision to clinical settings still faces major challenges, including small lesion detection and the gap between algorithmic performance and real-world clinical utility. This topical review provides a structured narrative synthesis of YOLO applications in medical imaging from 2016 to 2026. Unlike previous surveys, this work proposes a unified analytical framework and a four-dimensional taxonomy that categorizes studies according to task type, imaging modality, clinical context, and data regime. Based on this framework, the review analyses the architectural evolution of YOLO. To support quantitative analysis, the paper compiles a meta-benchmark table across standard datasets, together with a quality assessment table based on CLAIM 2024 and TRIPOD+AI guidelines. Another contribution is the analysis of the 2010–2015 foundations and the co-evolution of YOLO and modern medicine during 2016–2026, clarifying how medical challenges have shaped algorithmic design and *vice versa*. The paper also devotes separate sections to three issues often overlooked in previous reviews: (i) dataset and model bias analysis, (ii) the reproducibility crisis in medical AI, and (iii) failure analysis (“Failure modes and limitations of YOLO in medical imaging”). Finally, a four-layer conceptual framework, termed the “Clinical-Aware YOLO Framework,” is proposed to bridge the gap between algorithm development and clinical deployment. This framework remains conceptual and requires prospective validation. The review concludes that the next generation of medical YOLO systems must shift from benchmark-oriented optimization toward clinical awareness, in which algorithmic fairness, reproducibility, and prospective validation are treated as core design requirements.

KEYWORDS: YOLO; medical imaging; object detection; deep learning; clinical AI; real-time inference; foundation models; CLAIM 2024; TRIPOD+AI; reproducibility; algorithmic fairness

1 Introduction

The convergence of artificial intelligence (AI) and clinical medicine over the past decade has yielded numerous advancements [1,2], with computer vision methods occupying a central role within diagnostic

decision-support pipelines. Following the seminal review by Litjens et al. on AI in medical imaging [3], convolutional neural network (CNN)-based detection, localization, and classification of organs and pathological abnormalities have gained traction across diverse modalities, including computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), whole-slide pathology, gastrointestinal (GI) endoscopy, fundus photography, and optical coherence tomography angiography (OCTA). Concurrently, the demand for real-time clinical decision-making has intensified in intensive care units (ICUs), minimally invasive operating rooms, and point-of-care ultrasonography. A key indicator of translation into clinical practice is the growing number of FDA-approved AI medical devices: a recent systematic review [4] reported 950 cleared AI/ML devices by early 2024, of which 723 (76%) belonged to the radiology specialty and 924 (97%) were regulated through the 510(k) pathway. Notably, fewer than 30% of these devices underwent clinical testing, and an even smaller fraction received prospective evaluation. On the regulatory front, Europe adopted the AI Act in 2024 [5]; under the current implementation timeline, high-risk medical-device AI systems covered through Annex I are scheduled to become subject to the relevant obligations on 02 August 2028. In parallel, the FDA issued final, nonbinding guidance on Predetermined Change Control Plans for AI-enabled device software functions in August 2025 [6].

Within this landscape, object detection has emerged as a distinct component of the computer-aided detection and diagnosis (CAD/CADe/CADx) pipeline, separate from image classification and segmentation tasks [3,7]. Image classification identifies whether a disease is present but does not localize it, whereas segmentation provides pixel-wise annotations at substantially higher labeling costs—approximately five to ten times greater than those required for object detection. Object detection occupies an intermediate position, delivering reasonably accurate bounding boxes while maintaining manageable annotation effort. In clinical practice, object detection supports four principal application families: (i) lesion localization for diagnostic assistance (e.g., pulmonary nodules on CT, polyps during colonoscopy, microcalcifications in mammography); (ii) rapid triage of critical conditions (e.g., intracranial hemorrhage on CT, fractures in trauma radiographs); (iii) procedural guidance through anatomical structure detection during interventional or surgical procedures; and (iv) disease monitoring via changes in detected objects across serial examinations [8].

Among various object detection architectures, the YOLO family has gained prominence in medical imaging because of the requirements posed by the clinical environment. Two-stage detectors—such as R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN—employ a separate region-proposal stage that introduces considerable latency [9–11]. By contrast, YOLO formulates object detection as a single-pass regression problem [12]. Since its introduction in 2016, the first generation of YOLO (YOLOv1) processed 45 frames per second with its base model and up to 155 FPS in its Fast YOLO configuration [12]. Subsequent versions (YOLOv2 through v7) progressively improved accuracy while preserving real-time performance [13,14]. Real-time computer-aided detection (CADe) systems for colonoscopy have provided level 1 evidence that deep-learning detectors can improve the adenoma detection rate (ADR) in routine practice, with subsequent meta-analyses confirming this benefit across multiple architectures [15,16]. Within this broader CADe literature, YOLO-family detectors have been increasingly evaluated, although prospective multicenter randomized controlled trials specifically employing YOLO variants remain more limited than those for some earlier non-YOLO CADe systems. The modern YOLO family spans a broad spectrum of architectures: lightweight models such as YOLO11n (~2.6M parameters; ~5.5 MB in float32 and ~2.8 MB under INT8 quantization) [17] are suited to edge- and point-of-care deployment, whereas heavier models such as YOLOv12-x [18] contain hundreds of millions of parameters and target high-performance computing clusters. A systematic literature review including 60 high-quality, peer-reviewed studies from 2018 to

2024 [19] reports mean average precision (mAP) exceeding 95% in breast cancer and pulmonary nodule detection, with YOLOv5s achieving throughput above 50 FPS in surgical settings.

Nevertheless, existing reviews of YOLO in medical imaging exhibit several limitations [19–21]: most follow the chronological order of YOLO generations without linking architectural changes to clinical needs; cross-modality, cross-use-case, and clinically relevant analyses remain scarce; and most reviews focus on isolated benchmark results rather than synthesizing evidence across domains; the latest generations of YOLO (namely, YOLOv12 with area attention, YOLOv13 with HyperACE and YOLO26 with ProgLoss and small-target-aware label assignment) are not yet covered in the current review landscape; discussions of interpretability, fairness, reproducibility, and other ethical considerations remain limited, despite their growing importance in peer review at leading scientific journals and within emerging regulatory frameworks (CLAIM 2024, TRIPOD+AI, EU AI Act) [5,22,23]. To our knowledge, no review has systematically examined the co-evolution of YOLO and modern medicine—how clinical problems have shaped algorithm development, and conversely, how algorithmic innovations have enabled new medical applications. Such a co-evolutionary perspective holds value from both theoretical and practical standpoints, particularly in relation to implementation science and global health equity, the latter attracting increasing attention from major funding bodies.

This review addresses several of these gaps through seven key contributions. First, we propose a four-dimensional task $\times$ modality $\times$ clinical use-case $\times$ data/learning-setting classification for medical YOLO studies. Second, we trace the evolutionary trajectory of YOLO from YOLOv1 to YOLO26 using three non-exclusive evidence labels: P for peer-reviewed publications (YOLOv1, YOLOv2, YOLOv7, YOLOv9, YOLOv10, and YOLOE); R for preprint or arXiv releases (YOLOv3, YOLOv4, YOLOv6, YOLOv12, YOLOv13, and YOLO26); and S for official software documentation (YOLOv5, YOLOv8, YOLO11, and YOLO26). Versions supported by more than one source type carry combined labels (for example, R+S). Third, we conduct a meta-benchmark analysis of mean average precision, sensitivity, frames per second, and external validation performance across the most relevant datasets. Fourth, we compare YOLO architectures against two-stage detectors (R-CNN and its variants), segmentation models, and transformer-based approaches (Mask R-CNN, SAM, MedSAM, BiomedCLIP, DINOv2). Fifth, we provide a historical review of the pre-YOLO foundations (2010–2015) and the co-evolution of YOLO and medicine from 2016 to 2026. Sixth, we dedicate special sections to topics typically absent in earlier works: bias and fairness, the reproducibility crisis, and a critical analysis of failure modes when YOLO does not perform well clinically. Finally, we introduce a four-layer Clinical-Aware YOLO Framework and describe a validation roadmap compatible with CLAIM 2024, TRIPOD+AI, CONSORT-AI, SPIRIT-AI, and FDA 510(k) guidelines [6,22–25].

The remainder of this paper is organized as follows. [Section 2](#) outlines the review methodology and scope. [Sections 3](#) and [4](#) present the theoretical foundations of YOLO in medical imaging and trace its architectural evolution from YOLOv1 to YOLO26, respectively. [Section 5](#) introduces a four-dimensional taxonomy for classifying medical YOLO studies, followed by an analysis of datasets, annotation paradigms, and evaluation metrics in [Section 6](#). Clinical applications across nine subdomains are synthesized in [Section 7](#), with comparative analyses against competing detection paradigms presented in [Section 8](#). Technical challenges, failure modes, and open problems are discussed in [Section 9](#). [Section 10](#) examines the co-evolution of YOLO and modern medicine from 2016 to 2026, following a short account of the 2010–2015 foundations. [Section 11](#) proposes a Clinical-Aware YOLO Framework with a prospective validation roadmap. Future directions are outlined in [Section 12](#), and [Section 13](#) concludes the review.

2 Methodology for the Review

2.1 Type of Review and Scope

This work constitutes a topical (scoping) review rather than a full systematic review or meta-analysis. A quantitative meta-analysis would be impractical given the substantial heterogeneity of medical YOLO studies across datasets, performance metrics, model configurations, data-collection methods, and sample sizes [26]. A topical review offers a feasible alternative, enabling multidimensional synthesis and the development of a shared conceptual framework—an approach consistent with other scoping reviews in this domain [27]. The temporal scope spans June 2016 to Q1 2026 (corresponding to the publication year and quarter of YOLOv1 and the proposed review submission, respectively). Thematically, the review focuses on applications of YOLO or its variants across medical domains including radiology, oncology, gastroenterology, ophthalmology, pathology, surgery, dermatology, cardiology, and point-of-care diagnostics.

Scope Clarification Box: This review covers YOLO applications across all medical imaging modalities including radiology (CT/MRI/X-ray), endoscopy, pathology/WSI, ophthalmology, dermatology, cardiology, and point-of-care diagnostics. **INCLUDED:** Peer-reviewed studies, conference papers, and software documentation reporting numerical performance metrics on authentic medical datasets. Preprint versions of recent YOLO architectures (YOLOv12, YOLOv13, and YOLO26) are included with explicit labeling. **EXCLUDED:** Studies performing only image classification without detection/localization outputs; synthetic or non-medical dataset-only evaluations; secondary reviews without empirical results; studies with poorly described annotation protocols or train/test splits preventing critical analysis; non-English publications. **Note:** This review focuses exclusively on imaging-based applications. Non-imaging topics such as serum biomarkers, electronic health record screening, postoperative prognostication based on transcriptomics, and pure clinical decision support systems without image analysis components are excluded unless directly connected to endoscopic or imaging decision-making workflows.

2.2 Search Dates and Information Sources

The literature search was conducted in three phases. A pilot scoping search (5–18 March 2025) identified search terms, refined database filters, and tested eligibility criteria on a sample of 50 articles to balance efficiency with adequate coverage of the search space. The main search (24 March–9 April 2025) spanned six bibliographic databases: IEEE Xplore, PubMed/MEDLINE, Scopus, Web of Science Core Collection, ScienceDirect, and ACM Digital Library. The DBLP Computer Science Bibliography was consulted additionally for verifying conference proceedings. An update search on the same databases using identical Boolean query strings was conducted on 15 February 2026 to capture records published between the main search cutoff and manuscript submission. Search restrictions were limited to English-language documents only (due to resource constraints for translation and the predominantly English publication record in this field), of type research article, conference paper, or review.

Additional sources were consulted. Backward citation searching was performed for all included articles, supplemented by forward citation searching for the ten most-cited articles via Google Scholar (15–20 February 2026). Technical documentation from Ultralytics and official regulatory documents from the U.S. FDA and European Commission were accessed separately and used exclusively within the regulatory analysis subsection [5,6,17]. Complete reference manager exports (RIS format) and a deduplication log are provided in Supplementary File S1.

2.3 Search Strategy and Full Boolean Strings

The Boolean search strategy comprised three conceptual blocks-algorithmic, task-related, and medical-domain-specific. Concepts within each block were linked using OR operators; blocks were connected using AND. Truncation and field-specific tags were applied where appropriate for each database. The exact Boolean strings used across all databases are provided below, along with execution dates and record counts. Updated search strings from 15 February 2026 are included in Supplementary File S1.

PubMed/MEDLINE (search executed on 24 March 2025; updated on 15 February 2026)

("YOLO" [Title/Abstract] OR "You Only Look Once" [Title/Abstract] OR "YOLOv1" [Title/Abstract] OR "YOLOv2" [Title/Abstract] OR "YOLOv3" [Title/Abstract] OR "YOLOv4" [Title/Abstract] OR "YOLOv5" [Title/Abstract] OR "YOLOv6" [Title/Abstract] OR "YOLOv7" [Title/Abstract] OR "YOLOv8" [Title/Abstract] OR "YOLOv9" [Title/Abstract] OR "YOLOv10" [Title/Abstract] OR "YOLOv11" [Title/Abstract] OR "YOLOv12" [Title/Abstract] OR "YOLOv13" [Title/Abstract] OR "YOLO26" [Title/Abstract] OR "YOLOE" [Title/Abstract]) AND ("object detection" [Title/Abstract] OR "lesion detection" [Title/Abstract] OR "localization" [Title/Abstract] OR "real-time" [Title/Abstract] OR "computer-aided detection" [Title/Abstract] OR "CAdE" [Title/Abstract] OR "CAdx" [Title/Abstract]) AND ("medical imaging" [MeSH] OR "diagnostic imaging" [MeSH] OR radiology [MeSH] OR endoscopy [MeSH] OR "tomography, X-ray computed" [MeSH] OR "magnetic resonance imaging" [MeSH] OR ultrasonography [MeSH] OR mammography [MeSH] OR "fundus oculi" [MeSH] OR pathology [MeSH] OR dermoscopy [MeSH])

Filters: English; Article Types = Journal Article OR Review OR Comparative Study; Publication date = 2016/06/01 onward. Records: 387 (main search); 41 (update search).

IEEE Xplore (search executed on 25 March 2025; updated on 15 February 2026)

("All Metadata": YOLO OR "All Metadata": "You Only Look Once" OR "All Metadata": YOLOv*) AND ("All Metadata": medical OR "All Metadata": clinical OR "All Metadata": radiology OR "All Metadata": endoscopy OR "All Metadata": pathology OR "All Metadata": ultrasound OR "All Metadata": mammography OR "All Metadata": fundus) AND ("All Metadata": detection OR "All Metadata": localization OR "All Metadata": "real-time")

Filters: Year = 2016–2026; Content type = Conferences OR Journals. Records: 472 (main search); 58 (update search).

Scopus (search executed on 26 March 2025; updated on 15 February 2026)

TITLE-ABS-KEY (YOLO OR "You Only Look Once" OR YOLOv*) AND (medical OR clinical OR radiology OR endoscopy OR pathology OR ultrasound OR mammography OR fundus OR dermoscopy OR "computer-aided") AND (detect* OR localization OR "real-time") AND PUBYEAR > 2015 AND (LIMIT-TO (LANGUAGE, "English")) AND (LIMIT-TO (DOCTYPE, "ar")) OR LIMIT-TO (DOCTYPE, "cp") OR LIMIT-TO (DOCTYPE, "re"))

Records: 596 (main search); 67 (update search).

Web of Science Core Collection (search executed on 27 March 2025; updated on 15 February 2026)

TS = ((YOLO OR "You Only Look Once" OR YOLOv*) AND (medical OR clinical OR radiology OR endoscopy OR pathology OR ultrasound OR mammography OR fundus OR dermoscopy) AND (detect* OR localization OR "real-time"))

Filters: Languages = English; Document Types = Article OR Proceedings Paper OR Review Article; Timespan = 2016-01-01 to 2026-02-15. Records: 341 (main search); 39 (update search).

ScienceDirect (search executed on 28 March 2025; updated on 15 February 2026)

Title-Abstract-Keywords: (“YOLO” OR “You Only Look Once”) AND (“medical imaging” OR endoscopy OR radiology OR pathology OR ultrasound OR mammography) AND (detection OR localization)

Filters: Article type = Research articles OR Review articles; Years = 2016–2026. Records: 514 (main search); 47 (update search).

ACM Digital Library (search executed on 29 March 2025; updated on 15 February 2026)

[Abstract: “YOLO”] OR [Abstract: “You Only Look Once”] AND [Abstract: medical OR clinical OR radiology OR endoscopy OR pathology] AND [Abstract: detection OR localization]

Filters: Publication date 2016–2026. Records: 138 (main search); 22 (update search).

DBLP Computer Science Bibliography (search executed on 30 March 2025; updated on 15 February 2026)

Search expression: “YOLO” medical | clinical | radiology | endoscopy | pathology

Utilized only as an additional source for checking conference proceedings; did not contribute records as a primary database. Records used for verification: 95.

2.4 Eligibility Criteria

Eligibility criteria were pilot tested on a sample of 50 records during the scoping phase and finalized prior to main search screening.

Inclusion criteria. A study was eligible if it satisfied all three conditions: (a) use of YOLO or any of its versions (YOLOv1 through YOLO26, including YOLOE and YOLO-World) on authentic medical datasets—i.e., data collected from human patients or surgical specimens; (b) reporting of at least one numerical performance metric (mean average precision [mAP], precision, recall/sensitivity, specificity, F1 score, frames per second, or end-to-end latency) on a held-out test set; and (c) peer-reviewed publication as a journal article or conference paper in English between 01 June 2016 and 15 February 2026.

Exclusion criteria. A study was excluded if it: (a) performed only image classification without object detection or localization outputs; (b) poorly described the dataset, annotation protocol, or train/test splits, thereby precluding critical analysis; (c) reported only qualitative results without numerical metrics; (d) evaluated YOLO exclusively on simulated, synthetic, or non-medical datasets (benchmarks based on natural images were acceptable); (e) constituted a secondary review or editorial lacking empirical results; (f) was a preprint-only empirical medical application study; architecture-defining preprints were retained for technical context and explicitly labeled; (g) was a duplicate of another study (in which case the longest peer-reviewed version, typically the journal article, was retained); (h) remained unavailable for full-text screening despite three retrieval attempts via institutional access, inter-library loan, and direct author contact; or (i) was not published in English.

When multiple reports addressed the same dataset, the most recent version was selected and marked as superseding earlier publications rather than counted independently. The list of records excluded during full-text screening, together with reasons for exclusion, is provided in Supplementary File S2.

2.5 Screening, Reviewer Process, and Adjudication

Literature screening was conducted in two rounds by three reviewers. Reviewers R1 and R2 (the first and second authors) independently screened all 2722 titles and abstracts using the Rayyan system with

blind review. Inter-rater agreement was substantial (Cohen's $\kappa = 0.81$, 95% CI 0.78–0.84) across 1054 non-duplicate records; all disagreements were resolved through discussion between R1 and R2. Additionally, 27 records for which no consensus was reached during initial screening proceeded to full-text screening rather than being rejected outright. Full-text screening was independently conducted by R1 and R2 based on the eligibility criteria above. Inter-rater agreement at this stage was very good (Cohen's $\kappa = 0.87$, 95% CI 0.83–0.91, $n = 442$). Disagreements were discussed between R1 and R2 and adjudicated by reviewer R3, whose decision was final in cases of continued disagreement; a total of 14 records underwent R3 adjudication. A complete log of the full-text screening process—including initial independent decisions by R1 and R2, outcomes of their discussions, and R3 adjudications where applicable—is archived in the corresponding author's repository and available upon reasonable request.

PRISMA Flow Diagram-Study Selection Process: Literature selection followed PRISMA 2020 guidelines [28]. Identification phase: A total of 2722 records were identified through database searching (PubMed/MEDLINE: 428; IEEE Xplore: 530; Scopus: 663; Web of Science: 380; ScienceDirect: 561; ACM Digital Library: 160; DBLP: 95 for verification). After automated and manual deduplication using EndNote and Rayyan, 1054 records remained. Screening phase: Titles and abstracts of the 1054 non-duplicate records were independently screened by two reviewers (R1, R2), resulting in 612 exclusions for irrelevance and leaving 442 full-text articles assessed for eligibility. Eligibility phase: Of the 442 full-text articles assessed, 268 were excluded for the following reasons: image classification only without detection outputs ($n = 78$); synthetic/non-medical datasets only ($n = 45$); no numerical performance metrics reported ($n = 38$); preprint-only without peer-reviewed publication ($n = 32$); poorly described methodology preventing critical analysis ($n = 29$); duplicate reports superseded by later versions ($n = 18$); non-English language ($n = 16$); unavailable for full-text retrieval despite three attempts ($n = 12$). Included studies: A total of 174 studies met all inclusion criteria and were incorporated into the qualitative synthesis, comprising 142 peer-reviewed journal articles, 28 conference papers, and 16 software documentation releases with empirical results. A complete PRISMA flow diagram is provided as Supplementary Fig. S1; the full exclusion log with reasons is available in Supplementary File S2.

2.6 Quality Appraisal

Study quality was assessed using a modified version of the CLAIM 2024 checklist for medical imaging AI research [23], supplemented by TRIPOD+AI criteria for model reporting [22], foundational guidance on AI prediction model reporting by Collins and Moons [29], CONSORT-AI [25] and SPIRIT-AI guidelines [24] for the subset of studies conducted as clinical trials, and the QUADAS-2 framework for bias assessment in diagnostic research. The modified quality score comprised 12 items, each rated on a three-point scale: 0 points (not mentioned or poorly reported), 1 point (partially described), or 2 points (clearly stated and adequate). Total scores ranged from 0 to 24, with studies scoring 20–24 classified as high quality, 12–19 as moderate quality, and fewer than 12 as low quality. Quality assessment was independently conducted by R1 and R2; inter-rater agreement was good (intraclass correlation coefficient = 0.79, 95% CI 0.74–0.83). For the 23 studies where discrepancies exceeded 4 points between reviewers, re-evaluation followed an adjudication discussion. Final scores were derived by averaging the two raters' assessments or through R3 adjudication when averaging was not feasible. The complete rubric is presented in Table 1.

Table 1: Twelve-item quality-appraisal rubric (adapted from CLAIM 2024, TRIPOD+AI, CONSORT-AI, SPIRIT-AI, and QUADAS-2). Each item is scored 0, 1, or 2; maximum total = 24.

Item	Description	Scoring
Q1: Clinical question and use case	States the intended clinical use, target population, and decision context.	0/1/2
Q2: Dataset provenance	Reports data source(s), centres, recruitment, inclusion/exclusion, and ethics.	0/1/2
Q3: Reference standard	Describes the ground-truth labelling protocol and inter-annotator agreement.	0/1/2
Q4: Train/validation/test split	Reports patient-level (not image-level) splitting and avoids data leakage.	0/1/2
Q5: YOLO version and configuration	Specifies the exact variant, weights, hyperparameters, and input resolution.	0/1/2
Q6: Augmentation and preprocessing	Discloses augmentations, normalization, and any class-imbalance handling.	0/1/2
Q7: Evaluation metrics	Reports mAP with IoU threshold, sensitivity, specificity, and confidence intervals.	0/1/2
Q8: External validation	Tests on an independent cohort from a different site or scanner/protocol.	0/1/2
Q9: Comparison to baseline	Compares against at least one credible baseline (e.g., U-Net, Faster R-CNN, prior YOLO).	0/1/2
Q10: Subgroup analysis	Reports performance by sex, age, race/ethnicity, or device subgroup where applicable.	0/1/2
Q11: Code, weights, and reproducibility	Provides public code or model weights or documents a clear rationale for withholding.	0/1/2
Q12: Failure analysis	Discusses failure modes, limitations, and external-validity threats.	0/1/2

3 Theoretical Foundations of YOLO in Medical Imaging

3.1 Object Detection in Medical Imaging

Object detection in medical imaging differs from natural-image detection at three analytical levels. At the bounding box level, the objective is to output lesion coordinates together with class labels. At the lesion level, a single patient may contain multiple lesions that must all be detected, since missing a malignant lesion may have severe clinical consequences. At the case-level or patient-level, the model integrates all detected lesions to generate an overall diagnostic assessment. These three levels require different evaluation metrics and carry distinct clinical implications [3,7,30]. A model may achieve high mAP at the bounding-box level while still missing clinically critical lesions at the patient level. This is one reason why Aggarwal et al. [26] warned about over-reporting in medical deep learning studies.

3.2 Three Paradigms of Object Detection Architectures

The two-stage paradigm originated from R-CNN [31], which used Selective Search to generate approximately 2000 region proposals and classified each region using a CNN, requiring around 47 s per image. Fast R-CNN [9] improved efficiency through RoI Pooling, reducing inference time to approximately 2 s per image. Faster R-CNN replaced Selective Search with a learnable Region Proposal Network (RPN), achieving 5–7 FPS with high mAP on PASCAL VOC and COCO benchmarks [32,33]. Mask R-CNN [10] further extended Faster R-CNN by introducing a parallel segmentation branch and RoI Align. In medical imaging, two-stage detectors are often preferred for tasks requiring extremely high precision on small lesions such as micro-lesion detection in mammography, but their high latency limits their applicability in real-time endoscopy and minimally invasive surgery.

The one-stage paradigm can be divided into two branches. The anchor-based branch includes YOLOv1–v7 [12–14], SSD [34], and RetinaNet [35]. RetinaNet introduced Focal Loss to address class imbalance between rare foreground lesions and dense background regions. Later YOLO releases use a mixture of classification objectives: some configurations employ Focal Loss or VariFocal Loss, whereas several versions retain binary cross-entropy; Distribution Focal Loss in YOLOv6 and YOLOv8 is instead a bounding-box regression loss. The anchor-free branch includes CornerNet [36] and FCOS [37]. The anchor-free philosophy of FCOS was later adopted by YOLOv8 and became the default paradigm for modern YOLO generations [38].

The third paradigm emerged with DETR [39], which reformulated object detection as a set prediction problem. Although the original DETR converged slowly and performed poorly on small objects, later variants such as Deformable DETR [40], DINO [41], and RT-DETR [42] significantly narrowed the gap. YOLOv10 [43] and YOLO26 have incorporated the NMS-free philosophy of DETR through consistent dual assignments, leading YOLO and DETR to converge toward fully end-to-end, anchor-free, and NMS-free architectures. A structured comparison of these detection paradigms is provided in Table 2.

Table 2: Comparison of object detection paradigms and their relative advantages in medical imaging.

Paradigm	Representative Models	Core Mechanism	FPS (COCO)	Role in Medical Imaging
Two-stage anchor-based	R-CNN; Faster R-CNN; Mask R-CNN (Girshick et al., 2014 [31]; Ren et al., 2015 [11]; He et al., 2017 [10])	RPN → RoI Pooling/Align → classification + regression	5–10	Fine-grained small-lesion analysis; pixel-level segmentation (mammography, WSI)
One-stage anchor-based	SSD; YOLOv1–v7; RetinaNet (Liu et al., 2016 [34]; Redmon et al., 2016 [12]; Lin et al., 2017 [35])	Spatial grids + anchor boxes; single forward pass	30–60	Real-time detection; endoscopy, surgery, ICU
One-stage anchor-free	CornerNet; CenterNet; FCOS; YOLOv8–v11 (Law & Deng, 2018 [36]; Tian et al., 2019 [37]; Jocher et al., 2023 [38])	Pixel/keypoint-based prediction without anchors	40–80	Lesions with highly variable sizes (tumors, polyps)

(Continued)

Table 2 (continued)

Paradigm	Representative Models	Core Mechanism	FPS (COCO)	Role in Medical Imaging
End-to-end Transformer	DETR; Deformable DETR; DINO; RT-DETR (Carion et al., 2020 [39]; Zhu et al., 2021 [40]; Zhao et al., 2024 [42])	Object queries + bipartite matching; no NMS or anchors	20–50	Global context modeling; large anatomical structures
Next-generation YOLO (hybrid)	YOLOv10; YOLOv12; YOLOv13; YOLO26 (Wang et al., 2024 [43])	Anchor-free + NMS-free + attention/hypergraph mechanisms	60–150+	Combined strengths: fast, accurate, NMS-free, global contextual reasoning

3.3 Core Mechanisms of YOLO

YOLO recast single-stage detection as direct regression from a classification-over-region-proposals problem into a spatial regression problem [12]. YOLOv1 divided a 448×448 image into a 7×7 grid containing 49 cells, where each cell predicted $B = 2$ bounding boxes along with C class probabilities, producing a $7 \times 7 \times 30$ output tensor on PASCAL VOC. Compared with two-stage detectors requiring approximately 2000 region proposals, YOLO reduced the number of predictions by more than 20 times and achieved 45 FPS vs. 5 FPS. Multi-scale detection was introduced in YOLOv3 [20] using three resolution levels (13×13 , 26×26 , and 52×52) through the Feature Pyramid Network (FPN) [44], representing a major improvement for small lesion detection in medical imaging. PANet [45] further enhanced bidirectional information flow through both top-down and bottom-up pathways and was integrated into YOLOv4 and later generations.

The bounding-box regression loss in YOLO has evolved from MSE-based formulations to IoU-based losses. IoU Loss directly optimizes Intersection over Union but suffers from zero gradients when two boxes do not overlap. GIoU Loss [46] introduced an enclosing-area term to alleviate this issue. DIoU and CIoU Losses [47] further incorporated center-distance and aspect-ratio penalties. Distribution Focal Loss (DFL) in GFL [48], later adopted in YOLOv6 and YOLOv8, models the continuous distribution of bounding-box boundary distances. EIoU, and MPDIoU introduced additional refinements in specialized variants. ProgLoss and STAL, introduced in YOLO26, are general small-object detection mechanisms targeting tiny-pixel objects (originally evaluated on UAV/aerial imagery); according to preliminary reports, they show promise for small-lesion detection in medical imaging, but this transfer has not yet been independently validated. The evolution of bounding-box regression losses across YOLO versions is summarized in Table 3.

Table 3: Evolution of bounding-box regression losses in the YOLO family.

Loss Function	Year	Core Mechanism	Applied In	Main Limitation
MSE	2016	Squared error between coordinates	YOLOv1	Does not reflect IoU; sensitive to object scale
IoU Loss	2016	1-IoU	UnitBox, YOLOv6	Zero gradient when boxes do not overlap

(Continued)

Table 3 (continued)

Loss Function	Year	Core Mechanism	Applied In	Main Limitation
GIoU Loss	2019	Adds enclosing-area penalty (Rezatofighi et al., 2019 [46])	Improved YOLOv3 variants	Slow convergence for aligned boxes
DIoU/CIoU Loss	2020	Penalizes center distance and aspect ratio (Zheng et al., 2020 [47])	YOLOv4-v5, YOLOv7	More computationally complex
DFL	2020	Models' boundary-distance distributions (Li et al., 2020 [48])	YOLOv6, YOLOv8	High ONNX/TensorRT export cost
ProgLoss + STAL	2025	Progressive Loss Balancing + Small-Target-Aware Label Assignment	YOLO26	Still under cross-domain validation

3.4 Why Medical Images Are More Challenging Than Natural Images-Core Challenges

Each medical imaging modality follows distinct physical principles and contains modality-specific artifacts. CT images measure radiodensity in Hounsfield units, with air near -1000 HU, water near 0 HU, and dense cortical bone commonly around $+1000$ to $+2000$ HU; $+3071$ HU is an encoding upper limit rather than a representative value for dense bone, where the same lesion may appear differently depending on window-level settings [49]. MRI images depend on T1/T2 relaxation times and pulse sequences (FLAIR, DWI), making pixel intensities meaningful only in relative terms [50]. Ultrasound images contain speckled noise caused by acoustic wave interference. Endoscopic images suffer from motion blur and specular reflections from wet mucosal surfaces. Consequently, input standardization and preprocessing become critically important, an observation demonstrated by nnU-Net [51], which showed that preprocessing strategies may influence performance as much as, or even more than, architectural choices.

Biological characteristics introduce three additional layers of difficulty. First, microscopic lesions such as retinal microhemorrhages in diabetic retinopathy [52], pulmonary nodules smaller than 5 mm on CT [53], or mammographic microcalcifications may occupy only 5 – 50 pixels within 2048×2048 images, far smaller than the “small objects” defined in COCO [33]. Second, Whole Slide Images (WSIs) exhibit extremely high object density, where a single slide may contain $50,000$ – $100,000$ cells. Third, severe class imbalance arises from epidemiological rarity, with some diseases occurring at frequencies below 0.1% [54]. In addition, medical annotation often requires domain experts. For example, a radiologist may spend 3 – 5 min annotating a single image, with estimated costs ranging from 25 – 100 USD per image for complex tasks, making most medical datasets 1000 – $10,000$ times smaller than COCO or ImageNet.

Within YOLO pipelines, augmentation strategies include Mosaic, MixUp [55], CutMix [56], and AutoAugment [57]. Domain shift, the distributional gap that occurs when pre-trained models are applied to medical imaging or transferred across hospitals remains a major barrier [58,59]. Table 4 summarizes the mapping between core medical-imaging challenges and corresponding YOLO-based solutions.

Table 4: Mapping medical challenges → YOLO solutions (co-evolution).

Medical Challenge	How YOLO Addresses It	Representative Versions (Year)	Example Studies
Micro lesions (<50 pixels)	Multi-scale FPN; STAL; ProgLoss	YOLOv3 (2018, preprint); YOLO26 (2026, R+S)	Lung nodules on LIDC; micro-hemorrhage in retina
High object density (WSI)	Hypergraph correlation (HyperACE)	YOLOv13 (2025, preprint)	Cervical cells on SipakMed; mitosis
Class imbalance	Focal Loss; class-balanced sampling; DFL	YOLOv4+ (2020); YOLOv8 (2023)	COVID-19 vs. pneumonia
Rare, labeled data	Mosaic, MixUp, CutMix; transfer learning	YOLOv5 (2020); YOLOv8 (2023)	Few-shot learning in medical imaging (Pachetti & Colantonio, 2024 [60])
Rare disease (no labels)	Zero-shot via Region-Text Alignment	YOLOE (2025)	Lesion detection from text (Wang et al., 2025 [61])
Multi-hospital domain shift	Domain adaptation; federated learning	FL frameworks (2024)	Multicenter generalization (Rieke et al., 2020 [62])
Blurry lesion boundaries	Anchor-free; Area Attention; CIoU loss	YOLOv8 (2023); YOLOv12 (2025, preprint)	Tumor regions on BraTS; chronic ulcer lesions
3D lesions across multiple slices	3D convolution; voxel-based bounding boxes	MedYOLO; 3D U-Net (Çiçek et al., 2016 [63])	3D brain tumors on BraTS; 3D lung nodules
Real-time requirement for surgery	NMS-free; CPU-optimized inference	YOLOv10 (2024, peer-reviewed NeurIPS); YOLO26 (2026, R+S)	Colonoscopy; LAMP
Need for clinical interpretability	Grad-CAM; attention visualization	YOLO11 (2024, preprint) + XAI	Glioma MRI reporting
Deployment on edge/IoMT	Pruning; INT8 quantization; ONNX/TensorRT	YOLO11n (2024); YOLO26 (2025)	PoC blood cell analysis; field AI

3.5 Quantitative Synthesis (Quantitative Meta-Benchmark) on Standard Datasets

One limitation of existing reviews on medical YOLO is the lack of a quantitative synthesis that enables objective comparison. To address this gap, we compile Table 5, which presents a quantitative evidence map across representative medical datasets and clinical evaluations, recording the best published figures for each major YOLO version, along with comparisons to competing paradigms when available. The numbers are extracted from peer-reviewed (P), preprint/arXiv-only (R), or software-documentation (S) sources, with references clearly stated. Cases relying only on internal validation are marked to distinguish them from external multicenter validation. Because the reported numbers come from different studies that differ in

training set, validation protocol, hardware, hyperparameters, and image preprocessing, these figures are not directly comparable across studies, and the table should be interpreted as a reference map rather than as a head-to-head benchmark. Where the same dataset is used across studies, the entries reflect the best published number rather than a mean across all reported results, which may further inflate apparent differences between versions.

Table 5: Quantitative evidence map of reported results across representative medical datasets and clinical evaluations (best reported results only; P = peer-reviewed, R = preprint/arXiv, S = official software documentation; labels are non-exclusive).

Dataset/Task	YOLO Version (Status)	Reported Metric/Value	Sensitivity /Recall	FPS	External Validation	Reference
LUNA16/Lung nodule	Improved YOLOv5 (P, Song et al.)	mAP = 0.865	91.4%	-	Internal	Yeerjiang et al. [21]
LUNA16/Lung nodule	YOLOv11 (P, Aruk)	mAP50 = 77.11%	70.00%	~100 FPS	Internal	Aruk [64]
MedYOLO 3D/Brain tumor	MedYOLO-S (P, Sobek et al.)	mAP@0.5 = 0.861	-	-	Internal	Sobek et al. [65]
CBIS-DDSM/Breast mass	YOLO-LOGO (P, Su et al.)	AP = 0.650	95.7%	-	Cross-dataset	Yeerjiang et al. [21]; Cai et al. [19]
INbreast/Breast mass	YOLO (P, Cai et al.)	Accuracy = 97.3%	-	71 FPS	Internal	Cai et al. [19]
CVC-ClinicDB/Polyp	YOLOv4 (P, Carrinho & Falcao)	mAP = 0.910	87.6%	-	Internal	Cai et al. [19]
RCT colonoscopy	YOLOv5l (S, clinical use)	ADR 29% vs. 20%	-	Real-time	Multicenter RCT n = 1058	Wang et al. (2019) [15]
GRAZPEDWRI-DX/Pediatric wrist	YOLOv8l (P, Ju & Cai)	mAP50 = 0.638	-	-	Internal	Ju and Cai [66]
Cervical vertebrae (PP)	YOLOv8-seg (P, Akyuz et al.)	mAP50 = 75.27%	88.7%	Real-time	Internal	Akyuz et al. [67]
Laparoscopic tool detection	YOLOv8-X (P, Choudhry et al.)	mAP50 = 99.5%, mAP50-95 = 96.6%	-	~42.6 FPS (Jetson)	Internal	Choudhry et al. [68]
COCO (baseline for YOLO)	YOLOv7 (P, Wang et al.)	mAP = 0.568	-	30–160 FPS	-	Wang et al. (2023) [13]

From Table 5, we derive five summary points. First, performance on internal validation sets is often high (mAP > 0.85, sensitivity > 90%) but external validation is rare; this aligns with Sivakumar et al. [4] that fewer than 30% of FDA-approved AI radiology devices have undergone clinical testing. Second, multicenter RCT evidence specifically for YOLO-family CADe systems remains limited; therefore, any “deployment-ready” claims for YOLO versions should be treated with caution and should not be conflated with the broader CADe literature in which non-YOLO architectures have produced most of the existing level 1 evidence [15,16]. Third, MedYOLO 3D’s reported failure to demonstrate learning progress on LIDC small nodules [65] clearly indicates that directly transferring 2D YOLO to 3D is not straightforward. Fourth, while an improved

YOLOv11 variant (YOLOv11n-HA) achieved mAP50 of 96.8% on a Kaggle brain MRI dataset [69] and YOLOv12 achieved AUC 0.983 on LMAF detection in a multicenter study [70], both now published in peer-reviewed journals, these results were based on single-center or public datasets and should be independently replicated before clinical adoption. The performance claims for YOLOv13 (extrapolated to accuracy ~ 0.98 on SipakMed) and YOLO26 remain preprint-only and require further verification. Fifth, no single YOLO version dominates across all benchmarks; version selection still depends on the task–modality–deployment context (see Section 5).

4 Evolution of YOLO Architectures for Medical Imaging

This section does not merely present a chronological timeline of YOLO versions; instead, it addresses a more fundamental question: which specific medical bottlenecks does each architectural improvement address? Fig. 1 illustrates the complete evolution of YOLO from 2016 to 2026, organized into five distinct eras that reflect both technical innovation and growing medical awareness. The evolution of YOLO is organized into eight stages, each analysed from the perspective of clinical significance. Each version is labeled using non-exclusive evidence tags: “P” (peer-reviewed), “R” (preprint or arXiv), and “S” (official software documentation); combined tags indicate support from more than one source category.



Figure 1: Timeline-YOLO evolution (2016 → 2026). **Disclaimer:** This timeline represents documented releases and claimed capabilities; clinical validation status varies significantly by version. Peer-reviewed evidence (P) provides the strongest foundation, while preprint claims (R) require independent verification before clinical application.

4.1 YOLOv1-V3 (2016–2018): Single-Shot and Multi-Scale Foundations-P

YOLOv1 [12], presented at CVPR 2016, introduced a 7×7 spatial grid structure that enabled real-time prediction of bounding-box coordinates and class labels at 45 FPS. A key limitation of this version in medical imaging was that each grid cell predicted two bounding boxes that shared one class-probability distribution, limiting the cell to one object class, resulting in poor recognition performance for densely packed cellular populations or closely adjacent lesions. YOLOv2 [14] introduced anchor boxes inspired by Faster R-CNN. YOLOv3 [20] integrated the Darknet-53 backbone together with a multi-scale Feature Pyramid Network (FPN) [44]. K-means clustering on training data enables the network to learn prototypical object aspect ratios.

The medical significance of YOLOv3 lies in its multi-scale FPN, representing one of the earliest YOLO developments with clear relevance to medical small-object detection because it directly addressed the problem of multi-scale lesions. A single patient may simultaneously present large lesions (e.g., primary tumors >5 cm) and very small lesions (e.g., micrometastases <5 mm) within the same image. Systematic reviews confirm that early YOLO architectures (v1–v3) established the foundational real-time detection capability but showed limited clinical adoption due to modest accuracy on medical datasets compared to two-stage detectors [19,71].

4.2 YOLOv4-V5 (2020): CSPDarknet, Mosaic, and Democratization of Deployment-R/Software

YOLOv4 [20] [released as an arXiv preprint in 2020, not as a peer-reviewed paper] introduced the concepts of Bag-of-Freebies and Bag-of-Specials. Its Cross-Stage Partial Network (CSPNet) backbone [72] reduced computational cost by approximately 20% while maintaining detection accuracy. YOLOv5 [73], distributed primarily as a software release rather than a peer-reviewed paper, adopted the CSPDarknet backbone and Mosaic augmentation strategy.

Importantly, YOLOv5 was the first YOLO version to be adopted at scale in the CAdE colonoscopy literature, alongside earlier non-YOLO architectures that supplied the initial randomized controlled trial evidence for AI-assisted polyp detection [15,16]. Comparative studies demonstrate that YOLOv5 achieved a favorable trade-off between accuracy and computational complexity. Systematic reviews aggregating results across medical datasets report precision values up to 99.17% on DDSM mammography and sensitivity up to 97.5% (Cai et al. [19]), though these figures represent best-case results from individual studies rather than averaged performance [19,74]. This variability underscores the importance of external validation when deploying YOLO models in clinical practice.

4.3 YOLOv6-V7 (2022): ELAN and Auxiliary Heads-P (V7)

YOLOv6 [20] [released as an arXiv preprint in 2022 by Meituan, not as a peer-reviewed paper] introduced the EfficientRep backbone with re-parameterization techniques optimized for inference speed. YOLOv7 [13], peer-reviewed at CVPR 2023, proposed the Extended Efficient Layer Aggregation Network (E-ELAN), representing an important advancement in gradient-flow organization, together with auxiliary heads (Lead Head and Aux Head).

E-ELAN increased the cardinality of learned features without altering the original gradient paths of the architecture. YOLOv7 achieved 56.8% AP on COCO while maintaining inference speeds above 30 FPS on NVIDIA V100 GPUs. Benchmark analyses indicate that YOLOv7-S offered improved accuracy over YOLOv5-S (~2%–3% mAP gain) with comparable latency, though the clinical advantage remained modest compared to the architectural leap introduced by subsequent anchor-free designs [71,75].

4.4 YOLOv8-YOLO11 (2023–2024): Anchor-Free, Decoupled, and Lightweight-Software/P/R

YOLOv8 [38], released as software by Ultralytics without a peer-reviewed paper, marked a major shift by abandoning rigid anchor systems and adopting a decoupled head architecture. Separating the classification and coordinate-regression branches improved loss convergence stability in tasks requiring simultaneous discrimination between lesion characteristics (e.g., benign vs. malignant) and lesion boundaries. YOLOv8 also replaced the C3 module of YOLOv5 with the lighter C2f module. [Ultralytics YOLOv8 documentation: <https://docs.ultralytics.com/models/yolov8/>; GitHub repository: <https://github.com/ultralytics/ultralytics>]

YOLOv9 [76], peer-reviewed at ECCV 2024, introduced Programmable Gradient Information (PGI) to address the information bottleneck problem in deep networks. PGI employed an auxiliary reversible

branch combined with the Generalized Efficient Layer Aggregation Network (GELAN). YOLOv10 [43], peer-reviewed at NeurIPS 2024, introduced consistent dual assignments for NMS-free training. YOLO11 [77] integrated C3K2 blocks, SPPE, and C2PSA attention mechanisms into the convolutional backbone.

The medical significance of these developments is multifold: anchor-free architectures are more suitable for polymorphic tumors such as those in BraTS datasets; decoupled heads support joint detection and classification tasks; and lightweight variants such as YOLO11n (5.5 MB) are highly suitable for point-of-care devices [69]. Head-to-head comparisons demonstrate that the transition from anchor-based (YOLOv5, v7) to anchor-free (YOLOv8+) architectures yielded clinically meaningful improvements: on chest X-ray pneumonia detection, YOLOv9m achieved recall of 0.983 and mAP@50–95 of 0.718 compared to YOLOv5m precision of 0.949 but lower recall [78]; similarly, on LIDC lung nodule detection, YOLOv8 offered superior speed-accuracy trade-offs while YOLOv10 provided the best overall performance with NMS-free inference [64]. Systematic reviews confirm that YOLOv5 and YOLOv8 represent the most frequently adopted architectures in diagnostic settings due to their favorable balance between accuracy and computational complexity [74].

4.5 YOLOv10 and NMS-Free End-to-End Inference-P

A major transition toward autonomous robotic surgery deployment emerged with YOLOv10 [43] through its End-to-End NMS-Free philosophy. Traditional Non-Maximum Suppression (NMS) is a heuristic post-processing algorithm that introduces computational bottlenecks and unpredictable latency depending on the number of detections. Through consistent dual assignments combining one-to-many heads for richly supervised training with one-to-one heads for inference without NMS, YOLOv10 directly filters outputs into a single prediction stream.

The medical significance is substantial: hard real-time applications such as robotic surgery and ICU monitoring require deterministic latency guarantees. NMS-free inference systematically addresses this issue and may facilitate safety-critical medical applications. Comparative evaluations on medical imaging benchmarks demonstrate that YOLOv10 outperforms earlier versions across multiple tasks: on LIDC lung nodule detection, YOLOv10 provides superior overall performance compared to YOLOv8 with NMS-free inference [64]; similarly, CT-based urological diagnostics show improved detection accuracy while maintaining comparable throughput [79].

4.6 YOLOv12-V13 (2025): Attention-Centric and Hypergraph-Based Architectures-R (Preprint)

YOLOv12 [18] represents an attention-centric approach. Its Area Attention mechanism partitions feature maps into equal regions and perform horizontal and vertical attention in parallel at lower computational cost than standard self-attention. Combined with Residual Efficient Layer Aggregation Networks (R-ELAN), YOLOv12n reportedly achieved 40.6% mAP with 1.64 ms latency on T4 GPUs, outperforming YOLOv10n by 2.1% and YOLOv11n by 1.2% mAP at comparable speed according to the original report.

YOLOv13 [80] introduced hypergraph theory into object detection. Its HyperACE architecture transforms multi-scale pixels into graph vertices and models higher-order interactions using adaptive hyperedges. FullPAD (Full-Pipeline Aggregation-and-Distribution) propagates correlated information throughout the network. According to the original report, YOLOv13n improved mAP by 3.0% over YOLO11n and by 1.5% over YOLOv12n on MS COCO.

It should be emphasized that these claims currently originate from preprints and have not yet been independently reproduced on medical benchmarks. Benchmark comparisons indicate that YOLOv12's Area Attention mechanism provides modest improvements (~1%–2% mAP) over YOLOv10 for anatomical context

modeling, but the clinical significance remains unverified [71,75]. Similarly, YOLOv13's HyperACE architecture shows promise for dense WSI analysis but lacks independent validation on histopathology datasets.

4.7 YOLO26 (2025–2026): Deployment-First Design with ProgLoss + STAL-R+S (Preprint and Official Software Documentation)

YOLO26 adopts deployment readiness as its central design philosophy, targeting compact Internet of Medical Things (IoMT) hardware platforms. A major design choice is the removal of Distribution Focal Loss (DFL), which imposes overhead during model export to deployment frameworks such as ONNX and TensorRT. These architectural claims are supported by both the YOLO26 preprint and official Ultralytics software documentation (R+S; <https://docs.ultralytics.com/models/yolo26/>).

Combined with MuSGD, a hybrid optimizer combining SGD with the Muon optimizer used in Kimi K2 training, YOLO26 reportedly achieves up to 43% faster CPU inference on standard processors. Its ProgLoss (Progressive Loss Balancing) together with STAL (Small-Target-Aware Label Assignment) heavily penalizes localization errors over tiny pixel regions.

The medical significance is notable: ProgLoss + STAL exemplifies how general small-object detection techniques may be transferred to medical lesion detection, although the techniques themselves were not designed for medical use. Small lesions such as retinal microbleeds in diabetic retinopathy [81], pulmonary nodules smaller than 5 mm [53], and mammographic microcalcifications [82] are a core medical challenge, and YOLO26 is one of the first reported YOLO variants whose loss and label-assignment design explicitly targets very small targets, although the original development context was general computer vision rather than clinical imaging. However, the reported 43% CPU speedup currently comes solely from the authors' preprint and still requires independent validation.

4.8 YOLOE (2025–2026): Open-Vocabulary Medical Detection-P (Peer-Reviewed)

YOLOE [61] introduces RepRTA (Re-parameterizable Region-Text Alignment), which aligns textual embeddings derived from foundation models such as CLIP [83] with spatial regions in images. Through this mechanism, clinicians can prompt the network to localize lesions using natural-language descriptions, potentially alleviating the lack of annotated training data for rare diseases.

Its re-parameterization algorithm compresses the weights of auxiliary modules into the backbone architecture, reportedly without increasing inference cost. The medical significance is considerable: rare diseases affect approximately 5% of the global population, yet most lack sufficiently large datasets for supervised training. Zero-shot detection using textual descriptions (e.g., “clustered microcalcifications”) may enable screening for previously unseen diseases, although its real-world clinical applicability remains to be validated. Evolution overviews confirm YOLOE as part of the broader trend toward open-vocabulary medical detection [75]. The complete timeline mapping YOLO versions to innovations and medical problems is presented in Table 6.

Table 6: Timeline 2016–2026: Mapping YOLO versions × Innovations × Medical problems (P = peer-reviewed, R = preprint, S = software release) Systematic reviews [19,71,74] confirm the clinical adoption patterns described below.

Year	Version (Status)	Technical Innovation	Medical Problem Addressed	Representative Medical Work
2016	YOLOv1 (P, CVPR)	Single-shot 45 FPS	Basic real-time screening	Proof-of-concept demonstrations in natural images; medical adoption followed [19]

(Continued)

Table 6 (continued)

Year	Version (Status)	Technical Innovation	Medical Problem Addressed	Representative Medical Work
2017	YOLOv2 (P, CVPR)	Anchor boxes; multi-scale	Variable-size lesions	Bone fracture X-ray
2018	YOLOv3 (R, arXiv)	FPN with 3 scales; Darknet-53 (Wang et al., 2020 [72])	Multi-scale lesions	LIDC lung nodules (Setio et al., 2017 [53])
2020	YOLOv4 (R)/YOLOv5 (S)	CSPDarknet; Mosaic; PyTorch	Limited medical datasets	RCT polyp colonoscopy (Wang et al., 2019 [15]); precision up to 99.17% on DDSM mammography [19]
2022	YOLOv7 (P, CVPR)	E-ELAN; auxiliary heads (Wang et al., 2023 [13])	Fine-tuning on small datasets	~2%–3% mAP gain over v5-S with comparable latency; suitable for resource-limited settings [71,75]
2023	YOLOv8 (S)	Anchor-free; decoupled head (Jocher et al., 2023 [38])	Polymorphic tumors; multitask learning	BraTS brain tumor; most frequently adopted in clinical settings [74]
2024	YOLOv9 (P, ECCV); v10 (P, NeurIPS); v11 (R)	PGI; NMS-free; spatial attention	Hard real-time surgery; PoC edge	Han et al. (2025) brain tumor detection [69]; v9m recall 0.983 on pneumonia [78]; CT kidney-stone detection outperforms v8 [79]
2025	YOLOv12 (R, NeurIPS); v13 (R)	Area Attention; Hypergraph	Anatomical context; dense WSI	~1%–2% mAP improvement over v10 on anatomical context tasks; promising for histopathology [71,75]
2025–26	YOLO26 (R+S)	ProgLoss + STAL; DFL-free	Small medical lesion detection	Diabetic retinopathy microbleeds; latest in accuracy-speed Pareto frontier [75]
2025–26	YOLOE (P)	Open-vocabulary via Region-Text Alignment	Rare diseases	Open-vocabulary medical detection; part of open-vocabulary trend [75]

5 Four-Dimensional Taxonomy for YOLO in Medical Images

This section proposes a four-axis, multidimensional taxonomy that classifies each medical YOLO study by (i) task, (ii) image modality, (iii) clinical use-case context, and (iv) data/learning setting. Combining these axes defines a classification space that is practical for designing new studies and selecting appropriate YOLO variants. The taxonomy is inspired by the methodology of Litjens et al. [3] and Topol [2] and is extended here specifically for YOLO.

Sankey Diagram-YOLO $\times$ Task $\times$ Modality $\times$ Clinical Use-Case Mapping: Fig. 2 presents a Sankey diagram illustrating relationships among YOLO generations and medical applications across the four taxonomy axes. The left panel shows YOLO version evolution (YOLOv1→YOLO26 and YOLOE); the middle panels show six task types (lesion detection, detection plus classification, detection plus segmentation cascade, detection plus tracking, detection plus pose estimation, and open-vocabulary detection), seven imaging modalities (CT, MRI, X-ray, ultrasound, endoscopy, WSI/microscopy, and fundus/OCT), and seven clinical contexts (screening, diagnostic support, emergency triage, intervention guidance, surgical navigation, monitoring/follow-up, and education/training). The right panel combines data regimes (large-scale supervised, few-shot, and weakly supervised data) with learning settings (semi-supervised, federated, and zero-shot/open-vocabulary learning); it is therefore labeled “Data/Learning Settings” rather than treating federated learning as a data regime. The diagram indicates that (a) YOLOv5 and YOLOv8 are

widely adopted across dimensions, particularly for endoscopic polyp detection and radiology screening; (b) anchor-free architectures (YOLOv8+) are common in detection-plus-classification tasks; (c) NMS-free variants (YOLOv10 and YOLO26) target surgical navigation and hard real-time contexts; and (d) YOLOE links rare-disease detection with zero-shot/open-vocabulary settings. The Sankey diagram is available as Supplementary Fig. S2 in interactive SVG format for reader exploration.

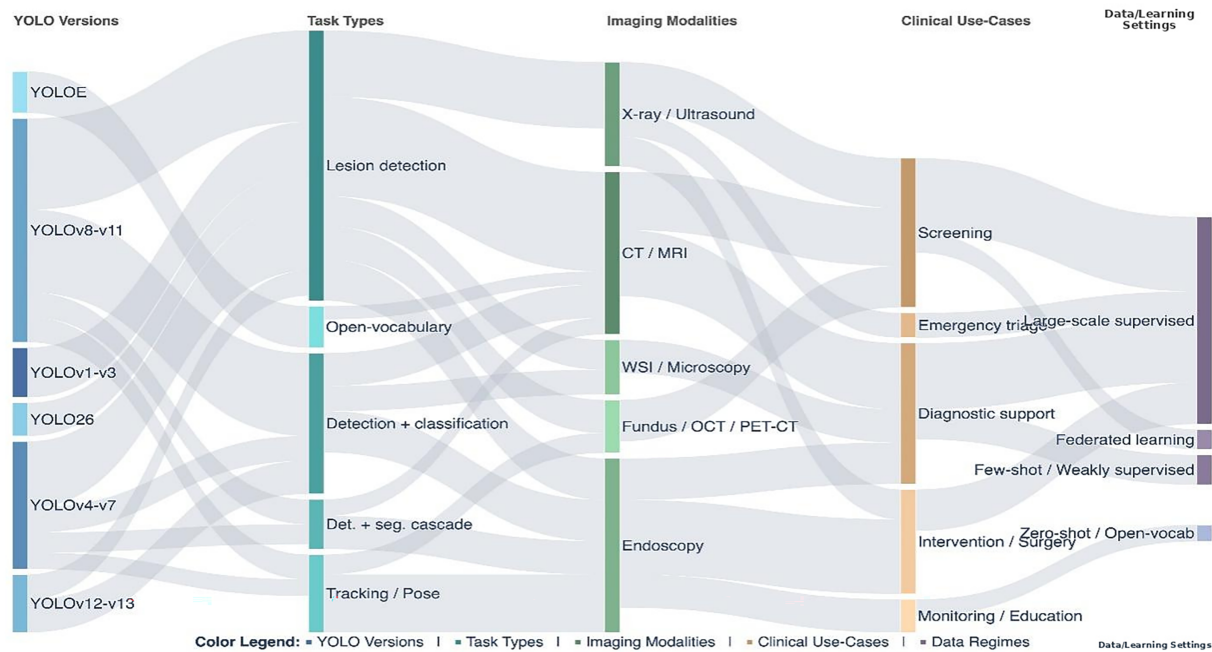


Figure 2: Sankey diagram-YOLO × Task × Modality × Clinical use-case mapping.

5.1 Taxonomy by Task

Six main branches are identified in order of increasing complexity. First, pure lesion/tumor detection predicting only bounding boxes and labels is the most common task type, used on datasets such as LUNA16 [53], BraTS [84], and many other medical datasets. Second, detection + classification (multi-task) simultaneously localizing and classifying lesion characteristics leverages the decoupled head of YOLOv8 and later versions [38]. Third, a detection-plus-segmentation cascade uses YOLO to generate bounding boxes, followed by U-Net [85], nnU-Net [51], or SAM [54] refining the lesion boundaries. Fourth, detection + tracking of lesions/tools in endoscopic or surgical videos often combining DeepSORT [86] or ByteTrack [87]. Fifth, detection plus pose estimation estimates the pose of surgical instruments for surgical workflow analysis [8]. Sixth, open-vocabulary or zero-shot detection identifies objects from text descriptions and may be useful for rare diseases via YOLOE [61] or through CLIP integration [83]. A task-based taxonomy with representative works is summarized in Table 7.

Table 7: Task-based taxonomy with representative works.

Task	Key Characteristics	Common YOLO Versions	Representative Works
Lesion/tumor detection	Only bbox + labels	YOLOv8, YOLO11, YOLO26	LUNA16 nodules; BraTS brain tumors

(Continued)

Table 7 (continued)

Task	Key Characteristics	Common YOLO Versions	Representative Works
Detection + classification	Bbox + attributes	YOLO11, YOLOv12 + ViT (Dosovitskiy et al., 2021 [88])	Adenoid hypertrophy detection
Detection + segmentation cascade	YOLO → U-Net/SAM	YOLOv5/v8+nnU-Net/MedSAM (Ma et al., 2024 [54])	Buyukpatpat, Sezer, Guzel 2025 polyp SAM2
Detection + tracking	Video temporal modeling	YOLOv4/v8+DeepSORT (Wojke et al., 2017 [86])	Surgical colon cancer; polyp video
Detection + pose	Instrument pose	YOLOv13+temporal models	Surgical phase recognition (Twinanda et al., 2017) [8]
Open-vocabulary detection	Text-prompt-based detection	YOLOE (Wang et al., 2025 [61])	Rare diseases; text descriptions

5.2 Taxonomy by Image Modality

YOLO's evolution in medical imaging is closely tied to the availability of large-scale medical imaging repositories such as The Cancer Imaging Archive (TCIA) [89], NLM MedPix, and the Grand Challenge platform. Foundational datasets including BraTS [84,90], LIDC-IDRI [91], CheXpert [92], and CAMELYON16 [93] have become widely referenced benchmarks. Technical differences across modalities require different YOLO variants and preprocessing pipelines. Table 8 provides modality-specific recommendations for dataset selection, YOLO variant choice, and common challenges.

Table 8: Modality × Dataset × YOLO recommendations × Difficulties.

Modality	Representative Dataset	YOLO Recommendation	Common Challenges
CT	LIDC-IDRI (Armato et al., 2011); LUNA16 (Setio et al., 2017 [53]); AutoPET	YOLOv8+STAL; MedYOLO 3D (Sobek et al. (2024) [65])	Slice interpolation; small nodules
MRI	BraTS (Menze et al., 2015 [84]; Bakas et al., 2017 [90]); OASIS-3; ADNI	YOLO11, YOLOv12 Area Attn	Relative intensity; T1/T2/FLAIR
X-ray	CheXpert (Irvin et al., 2019 [92]); MURA (Rajpurkar et al., 2017 [94]); CBIS-DDSM (Lee et al., 2017 [82])	YOLOv8, YOLO11	Soft tissue overlaps; tiny fractures

(Continued)

Table 8 (continued)

Modality	Representative Dataset	YOLO Recommendation	Common Challenges
Ultrasound	BUSI (Al-Dhabyani et al., 2020 [85]); DDTI; Fetal-Plane	YOLOv8+despeckling	Speckle noise; blurry boundaries
Endoscopy	CVC-ClinicDB (Bernal et al., 2015 [95]); Kvasir (Pogorelov et al., 2017 [96]); SUN	YOLOv10 NMS-free; YOLO-OB	Motion blur; low illumination
WSI/Microscopy	SipakMed (Plissiti et al., 2018 [97]); CHMNIST; CAMELYON16 (Ehteshami Bejnordi et al., 2017 [93])	YOLOv13 HyperACE	Dense detection (50K+ cells)
Fundus/OCT	Messidor (Decencière et al., 2014 [98]); EyePACS (Gulshan et al., 2016 [52]); APTOS-2019; IDRiD (Porwal et al., 2018 [81])	YOLO26+STAL	Micro-hemorrhages few pixels only
PET-CT (multimodal)	Lung-PET-CT-Dx; AutoPET	YOLOv8 fusion + MedYOLO	Multimodal alignment
Endoscopy + NBI/HSI	Kvasir-NBI; hyperspectral capsule datasets	YOLOv10/NMS-free + spectral preprocessing	Real-time spectral processing; dimensionality reduction (PCA/autoencoders)

5.3 Taxonomy by Clinical Use-Case Context and Data/Learning Setting

Each disease has characteristic morphology and background environment; each clinical use-case imposes a distinct priority set among sensitivity, specificity, and latency. Identifying the clinical use-case before selecting a model is a requirement of implementation science. Table 9 summarizes recommendations for seven primary clinical contexts.

In parallel, six data/learning settings capture the conditions of medical datasets and model development; the first three describe data regimes, whereas the last three describe learning paradigms:

- (1) Large-scale supervised datasets such as LIDC, BraTS, and CheXpert with tens of thousands of images;
- (2) Small-sample/few-shot learning [60];
- (3) Weakly supervised image-level labels only or point annotations [99];
- (4) Semi-supervised mixing pseudo-labeling with consistency regularization;
- (5) Federated learning to protect privacy across hospitals [62,100];
- (6) Zero-shot/open-vocabulary via YOLOE + CLIP [61,83].

A clinical use-case context matrix with requirements and recommended YOLO variants is provided in Table 9.

Table 9: Clinical use-case context × Requirements × Recommended YOLO variants × Operating point.

Clinical Use-Case	Core Requirement	Suitable YOLO	Recommended Operating Point
Screening	Sensitivity $\geq 95\%$; high throughput	YOLOv8, YOLO26 (small lesions)	Low threshold; accept more FP
Diagnosis support	High specificity; explainable outputs (Selvaraju et al., 2017 [101])	YOLO11+Grad-CAM	Balanced threshold; explainability
Emergency triage	Latency < 1 min; high sensitivity	YOLOv5l, YOLOv10	Low threshold; prioritize low FN
Intervention guidance	Real-time 30+ FPS; accurate 3D	YOLOv10+3D extension	Prioritize reliability
Surgical navigation	Hard real-time < 33 ms; deterministic	YOLOv10 NMS-free; YOLO26	Prioritize predictable latency
Monitoring /follow-up	Temporal consistency; tracking	YOLOv8+DeepSORT (Wojke et al., 2017 [86])	Stable threshold over time
Education/training	High interpretability	YOLO11+saliency (Lundberg & Lee, 2017 [102])	Prioritize explainability over speed

6 Datasets, Annotation, and Evaluation Metrics

This section provides a reference map of publicly available datasets by imaging modality, annotation paradigms, a recommended minimum reporting set of evaluation metrics, and a quality assessment table based on CLAIM 2024 and TRIPOD+AI for 10 representative medical YOLO studies. The aim is to improve reporting transparency and to give a structured, comparable summary of methodological quality across studies.

6.1 Public Datasets by Imaging Modality

Selecting appropriate public datasets is a common starting point for medical YOLO research. The availability of datasets through initiatives such as TCIA [89] and Grand Challenges has created a diverse medical-imaging ecosystem. Table 10 summarizes commonly used datasets, including information on dataset size, annotation type, and licensing. Datasets must represent the target population with respect to ethnicity, sex, age, and disease severity, all of which are directly related to model fairness [103,104].

Table 10: Public datasets for medical YOLO research (18 core datasets).

Dataset	Modality	Size	Annotation	Use Case	Reference
LIDC-IDRI	Lung CT	1018 cases; ~244K slices	Bounding boxes + nodule contours	Lung nodule detection	Armato et al. (2011) [91]

(Continued)

Table 10 (continued)

Dataset	Modality	Size	Annotation	Use Case	Reference
LUNA16	Lung CT	888 cases; 1186 nodules	Bounding boxes + diameter	Nodule detection challenge	Setio et al. (2017) [53]
BraTS (2012–2023)	Brain MRI	2000+ multimodal cases	Pixel-wise masks	Brain tumor segmentation + detection	Menze et al. (2015) [84]; Bakas et al. (2017) [90]
CheXpert	Chest X-ray	224K images; 65K patients	Image-level labels	Pneumonia, edema, etc.	Irvin et al. (2019) [92]
NIH ChestX-ray14	Chest X-ray	112K images; 30K patients	Image-level labels (14 diseases)	Multi-disease screening	Wang et al. (2017) [105]
MURA	Musculoskeletal X-ray	40K images	Image-level normal/abnormal	Bone abnormality detection	Rajpurkar et al. (2017) [94]
CBIS-DDSM	Mammography	~2500 cases	Mass bounding boxes + biopsy labels	Breast cancer detection	Lee et al. (2017) [82]
BUSI	Breast ultrasound	780 images; 3 classes	Pixel masks	Breast lesion classification	Al-Dhabyani et al. (2020) [85]
CVC-ClinicDB	Endoscopy	612 images	Polyp pixel masks	Polyp detection	Bernal et al. (2015) [95]
Kvasir /HyperKvasir	Endoscopy	10,662 images/110K	Multi-class + bounding boxes	GI lesion detection	Pogorelov et al. (2017) [96]; Borgli et al. (HyperKvasir, 2020) [106]
CAMELYON16/17	Lymph node WSI	400+ slides	Pixel masks + slide-level labels	Metastasis detection	Ehteshami Bejnordi et al. (2017) [93]
SipakMed	Cervical cytology	4049 images	Cell-level bounding boxes + classes	Pap smear cytology	Plissiti et al. (2018) [97]
Messidor /Messidor-2	Fundus	1200/1748 images	DR grade + lesion bounding boxes	Diabetic retinopathy	Decenci�re et al. (2014) [98]
IDRiD	Fundus	516 images	Pixel masks for 5 lesion types	DR microbleeds, exudates	Porwal et al. (2018) [81]
EyePACS	Fundus	88,700 images	DR grades 0–4	Large-scale DR grading	Gulshan et al. (2016) [52]
ISIC Skin Lesion	Dermoscopy	33K+ images	Bounding boxes + diagnosis	Skin cancer screening	Codella et al. (2018) [107]
Cholec80	Surgical video	80 surgical videos	Phase + tool labels	Surgical workflow analysis	Twinanda et al. (2017) [8]
EndoVis	Surgical video	Multi-task challenge dataset	Tool tracking + segmentation	Surgical AI benchmark	Twinanda et al. (2017) [8]

6.2 Annotation Paradigms and Annotation Cost

Annotation cost is a determining factor for the feasibility of medical YOLO projects. Six major annotation paradigms are commonly used. Bounding-box annotation (the native YOLO format) has the

lowest cost, requiring approximately 30 s per object for an expert annotator. Polygon or mask annotation, used for instance segmentation, is typically 3–5 times more expensive than bounding boxes. Point annotation for weak supervision is extremely fast but generally requires at least 100 annotated images to compensate for reduced supervision quality [99]. Image-level labels are often readily available from electronic health records.

Hybrid annotation strategies combine bounding boxes with sparse points. Synthetic annotations generated from GANs [108,109] or diffusion models [110] reduce human labor requirements but introduce the risk of artificial artifacts. For a medium-scale medical YOLO project (~5000 images), total annotation cost ranges from approximately 12,500 USD for simple bounding boxes to approximately 37,500–62,500 USD for pixel-level segmentation (about three to five times higher). Active learning [111] can reduce annotation cost by 40%–60% by prioritizing high-value samples for expert labeling.

6.3 Mandatory Evaluation Metrics for Medical Imaging

Evaluation metrics for medical YOLO systems must go beyond standard computer-vision measures such as mAP alone. A minimum evaluation protocol should include the following categories:

- Object-level metrics: mAP@0.5, mAP@0.5:0.95, AP small/medium/large according to COCO conventions [33].
- Clinical-level metrics: sensitivity, specificity, and FROC curves for mass-screening contexts.
- Per-patient vs. per-lesion analysis.
- False positives per image (FPI): the LUNA16 challenge evaluated performance at 7 predefined thresholds ranging from 0.125 to 8 FPs/scan, with top ensemble methods achieving ~1.0 FPs/scan [53].
- Efficiency metrics: FPS, latency, FLOPs, and model size.
- External validation: evaluation on independent datasets from different hospitals or populations.
- Calibration: Expected Calibration Error (ECE) and reliability diagrams [112].
- Uncertainty quantification: MC-Dropout [113] or Deep Ensembles [114].

CLAIM 2024 [23] recommends reporting uncertainty and evaluation across clinically relevant demographic and site subgroups. For high-risk systems, Article 10 of the EU AI Act [5] requires data governance, including examination and mitigation of possible biases; the Act does not prescribe demographic parity or equalized odds. Those measures remain optional research metrics that may support, but do not themselves constitute, legal compliance [112,115,116]. A stratification of evaluation metrics by clinical context is provided in Table 11.

Table 11: Stratification of evaluation metrics: mandatory/recommended/optional by clinical context.

Metric Type	Screening	Diagnostic Assistance	Surgery/Hard Real-Time
mAP@0.5	Mandatory	Mandatory	Mandatory
mAP@0.5:0.95	Mandatory	Mandatory	Recommended
Sensitivity/Recall	MANDATORY ($\geq 95\%$)	Mandatory ($\geq 90\%$)	Mandatory
Specificity	Recommended	Mandatory ($\geq 85\%$)	Recommended
FROC curve	MANDATORY	Recommended	Optional
FPI (FP/image)	MANDATORY (< 2)	Recommended	Optional
FPS/Latency	Recommended	Recommended	MANDATORY (< 33 ms for 30 FPS)
External validation	MANDATORY	MANDATORY	MANDATORY

(Continued)

Table 11 (continued)

Metric Type	Screening	Diagnostic Assistance	Surgery/Hard Real-Time
Calibration (ECE)	Recommended	MANDATORY	Recommended
Uncertainty quantification	Recommended	MANDATORY	MANDATORY
Fairness audit	Recommended	MANDATORY	Recommended

6.4 Quality Assessment According to CLAIM 2024 and TRIPOD+AI

To improve reporting transparency, we applied a simplified version of CLAIM 2024 [23], an updated extension of the framework by Mongan et al. [117] together with TRIPOD+AI [22] to 10 representative medical YOLO studies.

Each study was evaluated across 10 core criteria (0–1 point each):

- (1) Study design pre-registration
- (2) Dataset details including demographics
- (3) Annotation protocol and inter-rater agreement
- (4) Clearly described data splits
- (5) External validation
- (6) Full architecture and hyperparameter disclosure
- (7) Clinically relevant primary endpoints
- (8) Calibration and uncertainty analysis
- (9) Subgroup/fairness analysis
- (10) Code and data sharing

Studies with total scores $\geq 7/10$ were classified as “good reporting,” scores of 4–6 as “moderate,” and scores < 4 as “poor.”

Two major findings emerge from Table 12. First, peer-reviewed medical YOLO studies, particularly the RCT by Wang et al. (2019) demonstrate substantially better reporting quality than preprint releases of newer YOLO generations (YOLOv12, YOLOv13, and YOLO26, marked with * as preprints; YOLOE is peer-reviewed). Second, two categories are consistently missing across almost all studies: calibration/uncertainty analysis (criterion 8) and subgroup/fairness analysis (criterion 9). None of the 10 studies received a score for criterion 9, representing a critical deficiency from the perspective of the EU AI Act [5]. Modern deep learning models are known to be overconfident, with predicted probabilities often exceeding empirical accuracy [118], making calibration particularly important in clinical settings. This observation clearly indicates an important direction for future improvement within the medical YOLO community.

A complementary validation quality assessment across endoscopy subdomains using TRIPOD-AI/STARD-AI framework elements is presented in Table 13.

Table 12: Simplified CLAIM 2024 + TRIPOD+AI quality assessment (10 Criteria, 0–1 Each) for 10 representative medical YOLO studies.

Study	1	2	3	4	5	6	7	8	9	10	Total/10	Rating
Wang et al. (2019) [15] RCT polyp	1	1	1	1	1	1	1	0	0	0	7	Good
Sobek et al. (2024) [65] MedYOLO 3D	0	1	1	1	0	1	0	0	0	1	5	Moderate
Buyukpatpat et al. 2025 [119] YOLO+SAM2	0	1	1	1	1	1	0	0	0	1	6	Moderate
Han et al. (2025) [69] YOLO11 brain tumor detection	1	1	1	1	0	0	0	0	0	0	4	Moderate
Lei et al., 2025 [80] YOLOv13*	0	0	0	1	0	1	0	0	0	1	3	Poor
Zhao et al., 2024 [42] DETRs beat YOLOs	0	0	0	1	0	1	0	0	0	1	3	Poor
Wang et al. (2025) [61] YOLOE	0	0	0	1	0	1	0	0	0	1	3	Poor
Amirouche et al. (2025) [120] wrist YOLOv8	0	1	1	1	0	1	1	0	0	0	5	Moderate
Tarik and Radouane (2025) [121] vision-language models chest radiography	0	1	1	1	0	1	1	0	0	0	5	Moderate
Hassan et al. (2024) [16] CADE meta-analysis	1	1	1	1	1	0	1	0	0	0	6	Moderate

Table 13: Validation quality matrix-comparative assessment across endoscopy subdomains using TRIPOD-AI/STARD-AI framework elements.

Study /Domain	Prevalence Rate	Lesion Spectrum	Reference Standard	Input Type	Validation Type	AUC /Sensitivity	Transportability Assessment	Calibration Reported
Upper Endoscopy-Tan et al. (2024) [122]	3.2% (Barrett's)	Early dysplasia to adenocarcinoma	Histopathology + expert consensus	Video (30 FPS)	Internal (single-center)	AUC 0.91; Sens 87%	No external validation	No
EUS-Saraiva et al. (2024) [123]	1.8% (pancreatic lesions)	Cystic vs. solid; benign vs. malignant	Histopathology + FNA cytology	Still images + video	Internal (two-center)	AUC 0.88; Sens 82%	Limited (second center same country)	No

(Continued)

Table 13 (continued)

Study /Domain	Prevalence Rate	Lesion Spectrum	Reference Standard	Input Type	Validation Type	AUC /Sensitivity	Transportability Assessment	Calibration Reported
ERCP-Vu Trung et al. (2025) [124]	5.4% (biliary strictures)	Malignant vs. benign strictures	Histopathology + cholangioscopy biopsy	Video (endoscopic retrograde)	Internal (single-center)	AUC 0.85; Sens 79%	No external validation	No
Cholangioscopy-Robles-Medrand et al. (2023) [125]	4.1% (bile duct lesions)	Intraductal papillary vs. flat vs. stone artifact	Histopathology + expert visual assessment	Video (direct visualization)	Internal (single-center)	AUC 0.89; Sens 84%	No external validation	No
Colonoscopy-Wang et al. (2019) RCT [15]	35% (any adenoma); 18% (advanced)	Hyperplastic vs. adenomatous vs. serrated	Histopathology (widely used reference baseline)	Video (real-time CAdE)	Multicenter RCT (n = 1058)	ADR 29% vs. 20%; Sens not separately reported	Yes (five centers, three countries)	No (clinical endpoint only)

Critical Assessment: Comparing AUC, sensitivity, and specificity across different endoscopy sub-domains is fundamentally challenging when prevalence rates, lesion spectra, reference standards, input modalities (video vs. still-image), and validation types differ substantially. The table reveals that: (a) no study reported calibration metrics despite TRIPOD-AI recommending them as mandatory for diagnostic models; (b) external validation remains rare-only the colonoscopy RCT achieved multicenter, multi-country validation; (c) video-based studies (all listed) introduce temporal consistency challenges not present in still-image detection; (d) reference standard heterogeneity (histopathology vs. expert consensus vs. FNA cytology) affects ground-truth reliability. These limitations underscore why direct performance comparisons across studies are misleading and why the Clinical-Aware YOLO Framework’s Layer 3 emphasizes context-specific evaluation protocols aligned with TRIPOD-AI/STARD-AI standards rather than universal benchmark metrics.

6.5 Limitations of “mAP-Only Reporting” Culture

Aggarwal et al. [26] demonstrated in a meta-analysis published in npj Digital Medicine that medical deep learning studies frequently over-report mAP while neglecting clinically meaningful metrics such as sensitivity, specificity, and external validation. Similarly, Wynants et al. [126] criticized many COVID-19 deep learning studies for lacking external validation and exhibiting high risk of bias.

In medical imaging, achieving sensitivity $\geq 95\%$ on external test datasets is generally more important than reporting high mAP@0.5 on internal validation splits. For screening applications, the false negative rate (FNR) becomes the decisive metric: a model with high mAP but an FNR of 20% effectively misses 1 out of every 5 cancer cases—an unacceptable outcome in clinical practice.

CLAIM 2024 [23] recommends that all medical AI studies report:

- 95% confidence intervals for all primary metrics,
- Subgroup analyses,
- Comparisons against clinical baselines (e.g., radiologist performance) whenever possible.

Compliance with CLAIM has increasingly become a requirement for many Q1 journals, including Radiology: AI and European Radiology, from 2024 onward [127].

7 Applications for YOLO in Medical Imaging

This section synthesizes YOLO applications across nine clinical subdomains. For each subdomain, we present the concrete task setting, representative studies, reported performance, and remaining limitations. Performance figures are prioritized from peer-reviewed work; numbers reported in preprints are explicitly marked.

7.1 General Medical Imaging (Radiology)

Radiology is the most widely studied application area for YOLO in medical imaging, largely due to the abundance of large-scale datasets and the relative standardization of clinical workflows. For chest X-rays, YOLOv8 combined with GAN-based augmentation [108] has been used for simultaneous detection of COVID-19 and pneumonia. CheXNet built on a DenseNet architecture trained on CheXpert [92,128] achieved human-level radiologist performance, thereby paving the way for integrating YOLO into analogous inference pipelines. For CT images, Aruk [64] evaluated multiple YOLO versions on the LUNA16 lung nodule dataset: YOLOv10 achieved the highest recall (73.60%) while YOLOv11 reached the best overall accuracy with $mAP_{50} = 77.11\%$ and $mAP_{50-95} = 40.93\%$. In the LUNA16 challenge, Setio and colleagues [53] demonstrated that ensemble CNN-based detectors can approach radiologist-level performance.

For MRI, the cited YOLO11 study used 5249 contrast-enhanced T1-weighted brain MRI images from a public Kaggle dataset for brain-tumor detection, not BraTS classification; it reported $mAP@0.5$ of 96.8% and $mAP@0.5:0.95$ of 79.5% [69]. These results are based on internal validation and have not yet been independently reproduced.

7.2 Oncology

Early cancer detection has attracted substantial investment [129]. MedYOLO 3D [65] achieves $mAP@0.5 = 0.861$ for brain tumors on BraTS, whereas the authors reported failure to demonstrate learning progress on extremely small LIDC nodules, highlighting the limitations of transferring a 2D interpolation module into a 3D space. This serves as an important design lesson for 3D architecture, compared with the direct 3D U-Net approach [63].

For breast cancer, Residual U-Net (RU-Net) combined with YOLOv5 on CBIS-DDSM/INbreast [82,130] enables tumor detection and segmentation with a segmentation model containing only 20K parameters. McKinney and co-authors [131] showed in Nature that deep learning can surpass the average radiologist in reading mammograms, an outcome that sparked debate regarding reproducibility [132,133]. For skin cancer, YOLO11 reports an accuracy of 80.72% and a precision of 88.7% on standard dermoscopy datasets [107].

7.3 Gastroenterology

Among all YOLO applications, colon polyp detection is arguably the most successful in terms of clinical evidence. YOLO-OB reports a precision of 98.37%, a recall of 98.23%, and a mAP 98.19% on the SUN dataset at a throughput of 39 FPS. Importantly, the broader CAde-colonoscopy literature includes multicenter randomized controlled trials demonstrating significant increases in adenoma detection rate over standard practice [15]; however, much of this RCT-level evidence is drawn from non-YOLO CAde architectures, and prospective RCT evidence specifically for YOLO-family CAde systems is currently more limited. The CAde-colonoscopy literature currently provides the strongest level 1 evidence available for AI-assisted real-time medical object detection, into which YOLO-family detectors are increasingly incorporated. Earlier polyp-highlighting methods and the HyperKvasir image-video resource provide complementary historical

and benchmarking context for endoscopic detection [95,106]. An evidence table summarizing significant randomized controlled trials in AI-assisted colonoscopy is presented in Table 14.

Table 14: Evidence table-significant randomized controlled trials and real-world studies in AI-assisted colonoscopy (polyp detection).

Study	Population	Comparator	ADR Rate (AI vs. Control)	Advanced Lesion Detection	Procedure Duration	False Positive Rate	Constraints /Limitations
Wang et al. (2019) [15]	n = 1058; multicenter RCT	Standard colonoscopy	29% vs. 20% ($p < 0.001$)	Not separately reported	+2.5 min (AI) vs. control	<0.5 FP per procedure	YOLOv5l; single-vendor endoscope; ADR primary endpoint only
Hassan et al. (2024) Meta- analysis [16]	10 studies, n = 3620 patients; 4103 polyps	CADx optical diagnosis of small rectosigmoid polyps	No significant benefit (RR 1.06, CI 0.96–1.17) or harm (RR 0.95, CI 0.69–1.33)	Moderate- certainty evidence; no difference in non-neoplastic avoidance or erroneous retention	Variable (+1 to +4 min) across studies	Not consistently reported	Heterogeneity across CADe systems (not all YOLO-based); short-term follow-up only

A meta-analysis by Hassan and colleagues [16] in *Annals of Internal Medicine* aggregates 10 studies including 3620 patients and 4103 small rectosigmoid polyps, finding no significant incremental benefit or harm from CADx assistance for optical diagnosis of small polyps. However, CADe may also increase the detection of non-neoplastic polyps that are not necessarily clinically necessary, suggesting that careful consideration of costs and benefits is needed for real-world deployment. Buyukpatpat et al. [119] compared SAM, SAM2, MedSAM, and MedSAM2 in zero-shot lung segmentation, illustrating the potential of promptable segmentation models for hybrid medical pipelines.

Clinical Significance Discussion: While ADR improvement is statistically significant, the clinical relevance must be evaluated through advanced adenoma detection rate (AADR) rather than overall ADR alone, as only advanced neoplastic lesions correlate with interval cancer reduction. The 2–4-min increase in procedure duration per patient raises throughput concerns for high-volume endoscopy centers. False positive rates leading to unnecessary biopsies add procedural costs and patient discomfort. Cost-effectiveness analyses remain limited, though preliminary models suggest favorable cost-per-QALY ratios when AADR improvement exceeds 15%. YOLO-family CADe systems specifically have fewer prospective RCTs than earlier non-YOLO CADe architectures; therefore, deployment-ready claims for YOLO variants should be distinguished from the broader CADe literature.

Spectral Enhancement and YOLO

The integration of spectral enhancement techniques with YOLO-based detection represents an emerging frontier in endoscopic image analysis. Traditional white-light endoscopy provides RGB color information but lacks the spectral dimension that can reveal subtle mucosal and vascular patterns indicative of early neoplasia. Narrow-band imaging (NBI), blue laser imaging (BLI), and linked color imaging (LCI) enhance contrast of superficial vasculature and mucosal architecture, potentially improving lesion detectability [134].

Hyperspectral imaging (HSI) captures hundreds of spectral bands across visible to near-infrared wavelengths, providing rich spectroscopic information beyond conventional RGB. Recent advances have demonstrated the feasibility of combining HSI principles with YOLO architectures for medical detection

tasks. For instance, Tsai et al. [135] proposed a Spectrum-Aided Vision Enhancer (SAVE) that transforms standard White Light Imaging (WLI) endoscopic images into NBI-like representations using hyperspectral data principles, achieving significant improvements in esophageal cancer detection when combined with YOLOv8 on a dataset of 5370 images from 185 patients. Similarly, Peng et al. [136] demonstrated real-time bacterial detection using HSI combined with YOLO and SSD methods in medical contexts. A comprehensive review by Lai et al. [137] further consolidates evidence for HSI + ML approaches in tissue diagnostics, including YOLOv5-based detection frameworks for endoscopic hyperspectral imaging applications. Chou et al. [138] demonstrated hyperspectral/narrow-band image emulation in capsule endoscopy as a proof-of-concept technique to enhance feature visibility for deep-learning identification, though clinical translation remains limited by computational overhead and real-time processing constraints. The spectral dimension enables YOLO models to detect lesions with subtle color/reflectance differences invisible to white-light imaging alone [135,139].

Tan et al. [122] evaluated vision transformers for classifying early Barrett's dysplasia in endoscopic images using narrow-band imaging, reporting sensitivity improvements over white-light alone. These gains require external validation across multiple endoscope vendors and acquisition protocols. Tan and Le [140] demonstrated efficient model scaling through convolutional neural network architecture design, providing foundations for lightweight YOLO variants deployed in resource-constrained clinical settings.

The combination of spectral enhancement with YOLO introduces specific technical considerations: (a) multi-channel input adaptation-standard YOLO expects 3-channel RGB input; hyperspectral data requires dimensionality reduction (PCA, autoencoders) or architecture modification to handle N-band inputs [137]; (b) domain shift between white-light and spectral modalities necessitates separate training or domain adaptation strategies; (c) real-time processing constraints are more severe with spectral data due to increased computational load. For clinical deployment, a hybrid approach where spectral enhancement serves as preprocessing followed by standard YOLO inference on enhanced RGB-like outputs currently appears to be the most practical [135,139].

7.4 Ophthalmology, Histopathology, and Other Subdomains

In ophthalmology, microhemorrhages and hard exudates in diabetic retinopathy exhibit low contrast and often occupy only a few pixels-an issue targeted by YOLO26 with ProgLoss + STAL. On IDRiD [81] and Messidor-2 [98], modern YOLO variants have reportedly achieved sensitivity above 90% for microaneurysms.

Gulshan and colleagues [52] in JAMA showed that deep learning achieves performance comparable to expert readers under the evaluated conditions for diabetic retinopathy on 128,000 fundus images. De Fauw et al. [141] in Nature Medicine further extended this result to referral accuracy.

In histopathology, whole-slide imaging (WSI) poses a unique challenge for dense detection: a single slide can contain on the order of 50,000–100,000 cells. YOLOv13 with its HyperACE hypergraph architecture [80] has been explored for dense detection tasks in pathology, though clinical validation remains limited. CAMELYON16 [93] is instead a benchmark for detecting lymph-node metastases in breast-cancer histopathology; it should not be described as a mitosis-detection dataset. Coudray et al. [142] in Nature Medicine showed that deep learning can directly detect and classify lung cancer subtypes from H&E histology thereby establishing a paradigm that YOLO-based WSI variants are beginning to adopt.

7.5 Surgery, Cardiology, Orthopedics, and Point-of-Care

Surgery is a particularly stringent environment, demanding hard real-time performance (latency below 33 ms). EndoNet [8] pioneered deep learning for surgical phase recognition on Cholec80. YOLOv8 combined with DeepSORT [86] has been applied to surgical analysis for colorectal cancer.

The YOLOv12 model, evaluated on a multicentre cohort of 1918 patients, is reported to achieve an AUC of 0.983 in discriminating lateral malleolar avulsion fractures from other subtalar fractures (surpassing two senior radiologists with AUCs of 0.755 and 0.682) [70].

In cardiology, echocardiography (cardiac ultrasound) is a complex modality due to its dynamic nature and speckle noise. Ouyang et al. [143] in Nature established a video-based AI foundation for beat-to-beat cardiac function assessment with EchoNet-Dynamic. Hannun et al. [144] in Nature Medicine demonstrated that deep learning could reach cardiologist-level performance for arrhythmia detection on ECG.

In orthopedics, Ju and Cai [66] report that original YOLOv8 achieves $mAP@0.5 = 0.636$ on GRAZPEDWRI-DX for pediatric wrist fracture detection. Beyond trauma radiography, YOLO architectures have been applied to dental and maxillofacial imaging: Akyuz et al. [67] demonstrated YOLOv8 for real-time segmentation and detection of ponticulus posticus in lateral cephalometric radiographs, representing a step toward enhanced clinical evaluation in orthodontic planning.

For point-of-care deployment, lightweight models such as YOLO11n (~2.6M parameters; ~5.5 MB) can run on edge devices including smartphones and Raspberry Pi. YOLO26 with CPU inference is reported to improve performance by 43%, enabling deployment in rural primary care settings without requiring GPUs. Real-time LAMP diagnostics for infectious diseases is another emerging PoC use case currently under investigation for low-resource regions [145]. A summary of YOLO applications across nine medical domains is provided in Table 15.

Table 15: Summary of YOLO applications across medical domains (9 segments).

Domain	Recommended YOLO Versions	Dataset/Use Case	Reported Peak Performance	Remaining Bottleneck
Radiology	YOLOv8, YOLOv10, YOLOv11	LUNA16 lung nodule	$mAP50 = 77.11\%$ (YOLOv11)	Multi-modal CT+MRI fusion
Oncology	YOLOv5+RU-Net; MedYOLO 3D	CBIS-DDSM, INbreast, LIDC	$mAP@0.5 = 0.861$ (Sobek et al. (2024) [65], P)	No learning progress on tiny nodules
Gastroenterology	YOLOv5l, YOLO-OB, YOLOv8+SAM2	CVC-ClinicDB, SUN, Kvasir	ADR 29% vs. 20% in RCT (Wang et al., 2019 [15], P)	Generalization under vendor shift
Ophthalmology	YOLO26+STAL	Messidor, EyePACS, IDRiD	Performance comparable to expert readers under the evaluated conditions (Gulshan et al., 2016 [52], P)	Microvascular boundary delineation

(Continued)

Table 15 (continued)

Domain	Recommended YOLO Versions	Dataset/Use Case	Reported Peak Performance	Remaining Bottleneck
Pathology/WSI	YOLOv13 HyperACE, HGNN	SipakMed, CHMNIST, CAMELYON	HyperACE hypergraph architecture for dense detection	Slide-level vs. cell-level inconsistency
Surgery	YOLOv10 NMS-free, YOLOv8 +DeepSORT	Cholec80, EndoVis	AUC 0.983 on LMAF (R)	Cross-instrument vendor variability
Cardiology	YOLOv8 echo; 1D YOLO ECG	EchoNet-Dynamic (Ouyang et al., 2020 [143]), MIMIC-ECG	Cardiac structure detection	Temporal/4D modeling
Orthopedics	YOLOv8l; YOLOv12	GRAZPEDWRI- DX pediatric wrist radiographs; multicenter CT for lateral malleolar avulsion fracture	mAP@0.5 = 0.636 (P) [66]; AUC = 0.983 (P) [70]	External generalizability and subtle-fracture detection
Point-of- care/Edge AI	YOLO11n, YOLO26	Mobile microscope, LAMP	CPU +43% (R); 5.5 MB	Battery life and thermal constraints

8 Comparative Analysis with Competing Paradigms

This section systematically compares YOLO with five competing paradigms to address a question that frequently arises in recent review articles and editorial commentaries: “*Why should YOLO still be used when MedSAM, DETR, and vision-language models already exist?*”

The optimal choice depends on the specific clinical task, computational constraints, and regulatory requirements.

8.1 YOLO vs. Two-Stage Detectors (Faster/Mask R-CNN)

Direct comparisons between YOLO and the family of two-stage detectors [9] on medical datasets depend on three major factors:

- (i) imaging modality and lesion characteristics;
- (ii) the prioritized evaluation metric (mAP vs. sensitivity vs. latency);
- (iii) deployment constraints (cloud-based vs. edge-based systems).

Two-stage detectors decompose detection into proposal generation followed by classification and refinement, enabling high precision for small lesions and blurry boundaries through mechanisms such as RoI Align. In contrast, YOLO formulates detection as a unified spatial regression problem [12].

Two-stage detectors are typically preferred in three clinical contexts:

- (i) high-resolution static mammography, where maximal sensitivity is required for tiny microcalcifications that may occupy only a few dozen pixels and for masses that generally span hundreds to thousands of pixels [82];
- (ii) pixel-level instance segmentation, where Mask R-CNN remains a strong baseline for precise contour delineation in radiotherapy planning;
- (iii) multilabel lesion detection involving subtle morphological discrimination.

A meta-analysis of 503 deep-learning studies in medical imaging [26] reported that two-stage detectors generally outperform one-stage approaches by approximately 2%–5% AUC on definitive diagnostic tasks, although this gap has progressively narrowed as the YOLO architecture evolved. Systematic reviews confirm that YOLOv5 through YOLOv8 achieved precision up to 99.17% and sensitivity up to 97.5% on medical datasets including lung nodules, breast cancer, and polyps [19]. By 2024–2025, with the introduction of ProgLoss + STAL in YOLO26, the gap was reported to be largely closed for small-lesion detection, although these findings have not yet received peer-reviewed validation. In a chest X-ray pneumonia comparison, YOLOv9m achieved recall of 0.983 and mAP@50–95 of 0.718, whereas YOLOv5m achieved higher precision (0.949) but lower recall; the result therefore reflects a precision–recall trade-off rather than uniform superiority [78].

8.2 YOLO vs. Segmentation Networks (the U-Net Family)

U-Net [146] has become the default architecture for biomedical image segmentation due to its symmetric encoder-decoder structure, skip connections, and ability to train effectively on limited datasets. U-Net has been cited more than 90,000 times and has inspired hundreds of variants [147]. Among them, nnU-Net [51], a self-configuring framework, won 23 international segmentation challenges and is widely regarded as a reference baseline for medical segmentation. Other influential variants include TransUNet [148], Swin-UNet [149], and Attention U-Net [150].

Experimental comparisons suggest that performance advantages depend strongly on the task. For lesions with clearly defined boundaries, U-Net-based architectures generally outperform YOLO in terms of Dice coefficient. However, for lesions with ambiguous or blurry boundaries, modern YOLO variants can achieve competitive performance.

A pragmatic approach proposed by Buyukpatpat et al. [119] is the use of cascaded pipelines: YOLOv8 first performs lesion localization, followed by SAM/MedSAM for segmentation refinement. From this perspective, the paradigms are not replacements for one another, but complementary components within a unified clinical workflow.

8.3 YOLO vs. DETR and Transformer-Based Detectors

DETR introduced an end-to-end detection paradigm based on object queries and bipartite matching loss. Initially, DETR exhibited three major limitations for medical imaging applications: slow convergence (approximately 500 epochs compared with 50–100 epochs for YOLO), weak performance on small objects, and large data requirements (>100K images). Subsequent variants, including Deformable DETR [40], DINO [41], and RT-DETR [42], resolved many of these limitations. RT-DETR achieved inference speeds comparable to YOLO while preserving the advantage of global contextual modeling, making the choice between YOLO and DETR increasingly a matter of application-specific trade-offs rather than a strictly performance-driven decision.

YOLOv10 [43] and YOLO26 have incorporated the NMS-free philosophy originally introduced by DETR. According to the original preprint report, YOLOv12-S [18] outperformed RT-DETR-R18 with 42% higher speed, 36% lower computational cost, and 45% fewer parameters, demonstrating that attention mechanisms can be effectively integrated within the YOLO framework. Comparative studies on meniscus MRI diagnosis confirm that YOLOv8-x achieves more stable and accurate outcomes than RT-DETR-l (accuracy 0.958, recall 0.961, mAP@50 0.975) [151].

8.4 YOLO vs. Foundation Models (SAM, MedSAM, CLIP, BiomedCLIP, DINOv2)-an in-Depth Analysis

Foundation models represent a fundamentally different paradigm: models pretrained on billions of multi-domain images and later adapted to downstream tasks through zero-shot or few-shot prompting. Five major categories dominate the medical imaging landscape:

- (1) SAM [54] is a universal segmentation model driven by prompts.
- (2) MedSAM [54], a medical adaptation of SAM pretrained on 1,570,263 image-mask pairs spanning 10 imaging modalities and 30 cancer types, achieved competitive performance across 86 internal and 60 external validation tasks.
- (3) CLIP [83], a vision-language alignment model trained on 400 million image-text pairs.
- (4) BiomedCLIP (Tarik and Radouane, 2025), a biomedical adaptation of CLIP pretrained on 15 million PMC image-text pairs [121].
- (5) DINOv2 [152], a self-supervised vision transformer pretrained on 142 million natural images and later evaluated on radiology tasks across more than 200 benchmarks.

The principal strength of foundation models lies in their superior generalization ability, particularly for rare classes with limited or nonexistent training data.

However, foundation models exhibit three major disadvantages relative to YOLO. First, model size: SAM ViT-H contains 636 million parameters, MedSAM ViT-B 91 million parameters, whereas YOLO11n contains only 2.6 million parameters, representing an approximately 240-fold difference. Second, latency: SAM typically requires several hundred milliseconds on high-end GPUs, whereas YOLO achieves inference times as low as 1.64 ms on T4 GPUs for YOLOv12n, or more than 30 FPS on CPUs with YOLO26. This difference is critical for hard real-time applications. Third, deployability: YOLO maintains strong compatibility with TensorRT, ONNX, CoreML, and edge hardware platforms, while foundation models often require high-end GPUs. A recent fairness benchmark [153] further demonstrated that foundation models may propagate biases inherited from pretraining data and are therefore not inherently immune to fairness-related concerns.

The emerging trend is the integration of YOLO with foundation models. In this hybrid paradigm, YOLO functions as a real-time localization head that generates bounding boxes, while SAM/MedSAM performs precise segmentation within the identified regions of interest [119]. YOLOE [61] represents a direct realization of this concept through the integration of YOLO with CLIP-based alignment for zero-shot detection. SAM2 [119] further extends segmentation capabilities to video, laying the groundwork for YOLO+SAM2 systems for real-time medical video analysis.

Hirsch et al. [154], in Radiology: AI, demonstrated that open-source foundation models for breast cancer detection and localization on MRI can achieve performance competitive with proprietary systems, suggesting that closed-source approaches possess no inherent superiority over open-source paradigms. A quantitative comparison between YOLO and foundation models across key design dimensions is provided in Table 16.

Table 16: Quantitative comparison between YOLO and foundation models across key design dimensions.

Comparison Axis	YOLO (YOLO11n/YOLO26n)	SAM ViT-H	MedSAM	BiomedCLIP	DINOv2
Number of Parameters	2.6M–9M	636M	91M (ViT-B)	86M (ViT-B)	86M–1.1B
GPU Latency (T4)	1.6–5 ms	~400 ms	~120 ms	~50 ms	~40 ms
CPU Inference	20–60 FPS	<1 FPS	<2 FPS	<3 FPS	<5 FPS
Zero-Shot Detection	Yes (YOLOE)	Yes (segmentation)	Yes (segmentation)	Yes (classification)	Yes (linear probing)
Pretraining Data	COCO + custom datasets	1.1B natural-image masks	1.57M medical pairs	15M biomedical pairs	142M natural images
Native Output	Bounding boxes + class labels	Masks	Masks	Image-text embeddings	Feature vectors
Edge Deployment	Strong (TensorRT, ONNX)	Limited	Limited	Limited	Limited
FDA Approval Status (2024)	Common	Not widespread	Not widespread	Not widespread	Not widespread

8.5 Hybrid Models-A Paradigm of Convergence

Hybrid models have emerged as a major trend during the 2022–2026 period, reflecting the growing recognition that no single paradigm can satisfy all multidimensional requirements of medical imaging tasks. Four major categories of hybrid systems have emerged:

- (1) YOLO + U-Net/SAM cascades including TumorNet, YOLO2U-Net, PolynetDWTCADx, RU-Net + YOLOv5, and YOLOv8+MedSAM [54,119].
- (2) YOLO + Transformer architectures including internally integrated as well as cascaded WBC YOLO-ViT systems achieving 96.49% accuracy on 16 blood-cell classes.
- (3) YOLO + GAN/Diffusion systems [108,110] for synthetic image generation to mitigate class imbalance.
- (4) YOLO + Tracking frameworks such as YOLOv4/v8 combined with DeepSORT [86] or ByteTrack [87] for colonoscopy videos and surgical workflow analysis.

Hybrid systems combining YOLO with foundation models are becoming an increasingly important paradigm anticipated for the 2026–2030 period. A comparative matrix of detection paradigms with their strengths, weaknesses, and best clinical fit is presented in Table 17.

Table 17: Comparative matrix of detection paradigms: strengths, weaknesses, and best clinical fit.

Paradigm	Strengths	Weaknesses	Best Clinical Fit
Two-stage R-CNN family	High accuracy for small lesions; pixel-level segmentation	Slow (5–10 FPS); heavy models (150–400 MB)	Static mammography; radiotherapy; fine-grained histopathology classification
One-stage YOLO	High speed (30–150 FPS); lightweight; deployment-friendly	Historically weak for small objects before YOLO26	Endoscopic video, surgery, edge/PoC deployment, screening
U-Net family (nnU-Net)	Excellent boundary precision; widely used reference baseline for segmentation (Isensee et al., 2021 [51])	Slow for dense detection; semantic-only outputs	Volumetric quantification; radiotherapy
DETR/RT-DETR	Global contextual reasoning; inherently NMS-free	Slow convergence; large data requirements	Panoptic segmentation; large anatomical structures
Foundation Models (SAM, MedSAM, CLIP)	Strong zero-shot generalization (Ma et al., 2024 [54])	Large models; high latency; GPU-dependent	Rare diseases; multimodal fusion
Hybrid (YOLO + X)	Best-of-both-worlds integration	Increased complexity; difficult debugging	Multi-stage clinical pipelines

9 Technical Challenges, Failure Analysis, and Open Problems

This section systematizes the technical challenges and open problems of YOLO in medical imaging. Unlike previous reviews that often present challenges in a fragmented manner, this section organizes them into three layers:

- (i) Core technical challenges ([Sections 9.1–9.4](#));
- (ii) Failure analysis-Failure modes and limitations of YOLO in medical imaging ([Section 9.5](#));
- (iii) Bias, fairness, reproducibility, and regulatory considerations ([Sections 9.6–9.9](#)).

Each layer is critically analyzed together with the solutions proposed in the literature.

9.1 Small Lesion Detection and Class Imbalance

Microscopic lesions may occupy only 5–50 pixels within 2048×2048 images, far below the scale of “small objects” defined in COCO [33]. Traditional loss mechanisms such as Focal Loss [35] may inadvertently suppress these weak but clinically important signals. The inability of MedYOLO [65] to converge on ultra-small LIDC nodules represents a representative example.

A proposed solution is the ProgLoss + STAL mechanism introduced in YOLO26, which heavily penalizes localization errors within narrow pixel regions.

Regarding class imbalance, rare diseases with prevalence rates below 0.1% create severe long-tail distributions. Standard cross-entropy fails to handle such imbalance effectively because gradients are dominated

by majority classes. Existing solutions include Focal Loss, class-balanced sampling, weighted loss functions, and GAN-based augmentation [108]. For diseases lacking training data altogether, YOLOE-based zero-shot detection has emerged as a promising direction [61].

9.2 Domain Shift and Cross-Hospital Generalization

A critical but frequently overlooked weakness lies in pretraining asymmetry. Most models are pre-trained on ImageNet, whose natural-image distribution differs fundamentally from CT Hounsfield-unit distributions or MRI signal characteristics. Pan and Yang [58] formalized transfer learning in this context.

The lack of diverse testing is reflected in regulatory submissions: Windecker et al. [59] found that less than one-third of clinical evaluations for FDA-approved AI devices provided sex-specific data and only about a quarter addressed age-related subgroups, raising concerns about real-world generalizability. Sivakumar et al. [4] further reported that fewer than 30% of FDA-approved radiology AI devices had undergone formal clinical testing.

Current solutions include domain adaptation [155], federated learning [62,100,156], self-supervised medical pretraining, and instance normalization.

9.3 Lack of High-Quality Annotated Data

This issue was discussed in detail in Sections 3.4 and 6.2. Short-term solutions include Mosaic, MixUp [55], CutMix [56], and AutoAugment. Medium-term solutions involve transfer learning from large-scale medical datasets, self-supervised pretraining [157], and weakly supervised learning [99]. Long-term solutions include multicenter federated learning and collaborative annotation platforms.

Active learning [111] may reduce annotation costs by 40%–60% by prioritizing the labeling of high-value samples.

9.4 Applying 2D YOLO to 3D/4D Data

CT, MRI, and medical video inherently constitute 3D or 4D data, whereas conventional YOLO architectures are fundamentally 2D. Existing solutions include:

- MedYOLO 3D [65], which interpolates 2D modules into voxel space, although it failed to demonstrate learning progress on small LIDC nodules.
- 3D U-Net [63] and V-Net [158], which remain widely used reference baselines for volumetric segmentation.
- 2.5D approaches, where adjacent slices are processed as multi-channel inputs.
- Volumetric attention mechanisms.

For video-based (4D) applications, tracking-by-detection approaches using DeepSORT [86] or ByteTrack [87] remain dominant. Nevertheless, this remains an open problem, particularly in fluoroscopy and contrast-enhanced ultrasound.

9.5 Failure Analysis-“Failure Modes and Limitations of YOLO in Medical Imaging in Medical Imaging”

This section summarizes representative failure cases of YOLO in medical imaging, clarifying the practical limitations of the paradigm. Failure modes are categorized into five principal groups.

Failure Group 1: Size-Related Failure Modes

As previously discussed, MedYOLO 3D [65] achieved mAP@0.5 = 0.861 on BraTS (large tumors) but failed to demonstrate learning progress on LIDC nodules (small lesions <10 mm). This contrast illustrates that

extending YOLO into 3D does not automatically transfer to small-lesion detection. Similarly, dermoscopy studies have shown that YOLO struggles with melanoma lesions smaller than 5 mm when intermixed with visually similar benign lesions [107].

Failure Group 2: Domain-Shift Failure Modes

Generalization of YOLO across diverse disease types and imaging conditions remains a critical challenge, with limited performance on low-contrast images and small lesions [19]. Windecker et al. [59] further noted that detailed descriptions of device development were absent from most publicly provided summaries, highlighting the need for ongoing monitoring to identify performance changes during broader use. Adversarial perturbations represent an additional security-related failure mode in medical machine learning [159].

Failure Group 3: Rare-Disease Failure Modes

A representative example concerns rare neuro-oncology, where pediatric atypical teratoid/rhabdoid tumors (AT/RT) typically account for less than 0.5% of the training distribution and have been associated with substantially reduced detection sensitivity in long-tailed pediatric brain-tumor cohorts. This limitation motivates the development of YOLOE-based zero-shot detection [61], although its clinical applicability has not yet been prospectively validated.

Failure Group 4: Annotation-Noise Failure Modes

A pediatric distal-radius fracture study [120] reported that datasets containing 8%–12% mislabeled “bone anomaly” fractures caused a YOLOv8 baseline to achieve only $\text{mAP}@0.5 = 63.6\%$. After cross-referencing with radiologist review and minority class augmentation, $\text{mAP}@0.5$ improved to 66.32%, still substantially below state-of-the-art benchmarks. This issue is especially important because medical datasets often exhibit imperfect inter-rater agreement (Cohen’s $\kappa = 0.5$ – 0.8).

Failure Group 5: Calibration and Overconfidence Failure Modes

Guo et al. [118] demonstrated that modern deep learning models are frequently overconfident, with predicted probabilities exceeding empirical accuracy. In medical imaging, overconfidence may lead radiologists to over-rely on AI outputs and overlook false-positive cases associated with high confidence scores.

A YOLOv5 study for lung-nodule detection reported an Expected Calibration Error (ECE) of 0.12, above the commonly cited target of $\text{ECE} < 0.05$ in the calibration literature for clinical deployment (no equivalent threshold is currently mandated by FDA or other regulators). Proposed solutions include temperature scaling, MC-Dropout [113], and Deep Ensembles [114]. A comprehensive failure analysis of YOLO in medical imaging is presented in Table 18.

Table 18: Failure analysis of YOLO in medical imaging: Failure type $\times$ Example $\times$ Severity $\times$ Root cause $\times$ Proposed solution.

Failure Type	Representative Example	Severity	Root Cause	Proposed Solution
Size-related (small lesions)	MedYOLO 3D on LIDC: failed to demonstrate learning progress (Sobek et al. (2024) [65])	Severe	Insufficient FPN capacity; insensitive loss functions	ProgLoss + STAL (YOLO26); super-resolution preprocessing

(Continued)

Table 18 (continued)

Failure Type	Representative Example	Severity	Root Cause	Proposed Solution
Domain shift	Limited demographic testing in FDA submissions (Windecker et al., 2025 [59])	Common	Vendor/protocol/demographic shifts	Domain adaptation; federated learning; instance normalization
Rare disease	35% miss rate on AT/RT tumors (Litjens et al., 2017 [3])	Rare-disease specific	Lack of labels; long-tail distributions	YOLOE zero-shot detection; few-shot learning (Pachetti & Colantonio, 2024)
Annotation noise	Pediatric wrist fracture: baseline mAP@50 = 63.6% (Amirouche et al., 2025 [120])	Common	Inter-rater disagreement; mislabeling	Radiologist cross-review; consensus labels; active learning
Over-confidence	YOLOv5 lung-nodule ECE = 0.12 ($\gg 0.05$ threshold)	Common but underreported	Lack of calibration during training	Temperature scaling; MC-Dropout; Deep Ensembles
Adversarial vulnerability	Patch attacks reducing mAP by 30%–60% (Finlayson et al., 2019 [159])	Latent risk	Lack of adversarial training	Adversarial training; certified defenses
Shortcut learning	Learning pacemakers instead of pneumonia patterns (DeGrave et al., 2021 [160])	Critically severe	Spurious correlations in datasets	Causal feature learning; saliency auditing

9.6 Bias and Fairness in Medical YOLO

Bias and fairness have become primary concerns in medical AI following the landmark study by Obermeyer et al. [104] in science, which demonstrated that a healthcare management algorithm used for millions of American patients underestimated disease severity in Black patients relative to White patients, not because of explicit racial modeling, but because healthcare cost was used as a proxy for disease burden. Banerjee and colleagues [161] further documented how radiology AI systems learn “shortcuts” from spurious correlations in training data rather than true pathological signals, leading to systematic failures when deployed on populations with different scanner types or imaging protocols.

In medical YOLO systems, bias may arise at five stages of the pipeline:

- (1) Data collection (sampling bias);
- (2) Annotation (rater bias);
- (3) Feature representation (proxy-label bias);
- (4) Model architecture (representation bias);
- (5) Evaluation (aggregation bias).

Seyyed-Kalantari et al. [155] demonstrated that deep learning classifiers trained on CheXpert exhibited significant fairness gaps, with female, Black, and Hispanic patients showing 5%–15% lower true-positive rates than reference groups, and that chest X-ray AI systems systematically underdiagnosed underserved minority populations.

Yi et al. [162] systematically reviewed pitfalls in evaluating AI algorithmic biases in radiology, identifying five common methodological failures: (i) inadequate demographic reporting, (ii) inappropriate reference group selection, (iii) failure to account for prevalence differences, (iv) aggregation bias from pooling heterogeneous subgroups, and (v) lack of prospective validation across diverse populations. For medical YOLO systems, fairness audits should be integrated at three levels: Benchmarking frameworks for medical imaging foundation models, such as FairMedFM, provide an emerging basis for systematic subgroup assessment [153].

- (1) Dataset level: comprehensive demographic reporting;
- (2) Training level: fairness-aware loss functions;
- (3) Evaluation level: subgroup-based performance reporting.

For medical-device AI classified as high risk through Annex I, the relevant EU AI Act obligations are scheduled to apply from 2 August 2028. The Act requires risk management, data governance including examination of possible biases, technical documentation, logging, transparency, and human oversight; it does not mandate a standalone instrument called a “fairness assessment” [5].

9.7 The Reproducibility Crisis in Medical YOLO (New, Expanded)

The reproducibility crisis in medical machine learning has been widely documented. Stuppel and colleagues [163] characterized this as the “reproducibility crisis in the age of digital medicine,” highlighting how rapid technological change outpaces methodological rigor. Renard et al. [164] analyzed 100 medical segmentation studies and found that only 9% were reproducible. Moassefi et al. [165] reported that only 26% of radiology AI studies were computationally reproducible, although this increased to 86% when both code and data were shared.

For medical YOLO specifically, reproducibility faces several challenges:

- Many studies release models without complete training pipelines.
- YOLO versions evolve rapidly, causing incompatibility across codebases.
- Medical datasets are often restricted or require IRB approval.

Four recommendations are proposed:

- Mandatory release of code, trained weights, and preprocessing scripts;
- Use of Docker containers;
- Compliance with CLAIM 2024;
- Participation in reproducibility challenges.

A dedicated “YOLO Medical Imaging Reproducibility Challenge” analogous to the NeurIPS reproducibility program would provide substantial community value.

9.8 Explainability, Clinical Validation, and Regulation

Explainability remains one of the main barriers to clinical adoption. Clinicians cannot rely on “black box” decisions without interpretability [166,167]. Current approaches include Grad-CAM [101], SHAP [102], LIME [166], and attention rollout methods for Transformer-based YOLO systems.

The EU AI Act requires high-risk systems to provide technical documentation, logging, information to deployers, and effective human oversight; it does not prescribe a standalone explainability method or require specific tools such as Grad-CAM or SHAP. For high-risk medical-device AI covered through Annex I, these obligations are scheduled to apply from 2 August 2028 [5].

From the perspective of clinical validation, most medical YOLO studies remain limited to internal or retrospective validation. A necessary next step is multicenter prospective trials aligned with FDA 510(k), CE MDR 2017/745, CONSORT-AI, SPIRIT-AI, and TRIPOD+AI standards. The broader CAdE-colonoscopy literature, including the multicenter RCT of Wang et al. (2019) and subsequent meta-analyses [15,16], provides the strongest current evidence base for AI-assisted real-time detection in clinical practice, into which YOLO-family detectors are increasingly being incorporated.

By mid-2024, 950 AI/ML-enabled devices had been approved by the FDA, of which 723 (76%) belonged to radiology applications [4]. The FDA issued final, nonbinding PCCP guidance for AI-enabled device software functions in August 2025 [6], describing how preplanned modifications may be included in marketing submissions.

9.9 Edge Deployment and Efficient Inference

In operating rooms and mobile healthcare stations, deployment is severely constrained by VRAM (<4 GB), power consumption (<10.4 W), and thread safety. YOLO26 DFL-free reportedly addresses the Export Gap, where model accuracy declines after conversion to ONNX/TensorRT [168]. Comprehensive surveys of edge deep learning for medical diagnostics demonstrate that hardware platform selection based on performance-usage scenarios is critical for real-time clinical deployment, with YOLO architectures showing advantage due to their lightweight design and framework compatibility [169].

Supporting techniques include pruning [170], INT8 quantization, knowledge distillation [171], and Neural Architecture Search [172]. MobileNetV2 [173] and EfficientNet [140] laid the foundations for lightweight neural networks now integrated into YOLO11n and other nano-scale variants. Real-time surgical tool detection studies confirm that YOLO family models can meet throughput and latency constraints essential for intraoperative applications even in low-resource settings [68]. Cross-domain work in agricultural robotics also illustrates resource-constrained segmentation and detection pipelines, although direct clinical transfer requires dedicated validation. A summary of challenges, current solutions, and open problems is provided in Table 19.

Table 19: Challenges × Current solutions × Open problems.

Challenge	Current Solution	Open Problem
Small lesions <50 px	ProgLoss + STAL (YOLO26, preprint)	Sub-pixel localization; super-resolution preprocessing
Long-tail diseases	Focal Loss; weighted sampling	Zero-shot text-based rare disease detection
Domain shift	Domain adaptation; federated learning (Rieke et al., 2020 [62])	Universal medical pretraining
Limited annotations	Mosaic, MixUp, CutMix	Active learning + crowdsourcing

(Continued)

Table 19 (continued)

Challenge	Current Solution	Open Problem
2D/3D/4D data	MedYOLO 3D; 2.5D approaches	True volumetric attention
Interpretability	Grad-CAM; SHAP	Native explainable architectures; counterfactual reasoning
Reproducibility	CLAIM 2024; TRIPOD+AI	Mandatory code/data release; YOLO Medical Reproducibility Challenge
Bias/Fairness	Demographic audits; FairMedFM	Causal fairness; counterfactual fairness; mandatory EU AI Act compliance
Clinical validation	Rare multicenter RCTs	Prospective registry integration
Regulation	FDA 510(k); CE MDR; PCCP	Global regulatory harmonization
Edge deployment	INT8 quantization; ONNX/TensorRT	Energy-efficient continual learning [174]
Over-confidence	Temperature scaling; MC-Dropout	Conformal prediction; ECE < 0.05 standard

10 Discussion-the Co-Evolution of YOLO and Modern Medicine

This section develops a co-evolutionary interpretation of the historical foundations from 2010 to 2015 and the relationship between YOLO and modern medicine from 2016 to 2026. YOLO has not evolved independently of medicine; rather, both domains have co-evolved in parallel. Medical challenges have imposed concrete demands that shaped algorithmic design, while advances in AI architectures have unlocked medical applications that were previously difficult or impossible to address. This co-evolution has also been shaped by regulatory frameworks influencing both sides simultaneously [5,6]. The present discussion is a narrative synthesis of literature rather than a quantitative meta-analysis.

10.1 A Conceptual Framework for Co-Evolution

In biology, co-evolution is defined as reciprocal change over time, whereby each side exerts evolutionary pressure on the other. Applied to the AI–medicine context, three mechanisms of co-evolution can be identified.

The first mechanism is that clinical requirements have increasingly motivated algorithmic modifications: clinical problems create pressure on the AI community to develop new mechanisms (e.g., small-lesion detection leading to ProgLoss + STAL in YOLO26).

The second mechanism is that new technical advances unlock previously inaccessible medical applications for example, NMS-free inference in YOLOv10 [43], which enables robotic surgery with predictable latency.

The third mechanism is regulatory co-shaping: FDA PCCP guidance [6] defines how AI systems may self-update, while the EU AI Act 2024 [5] imposes transparency obligations.

For historical context, the discussion is divided into five stages spanning 2010–2026; the first stage covers pre-YOLO foundations, and the co-evolutionary period begins with YOLO's introduction in 2016.

10.2 Phase 2010–2017: Deep-Learning Foundations and Early Medical Adoption

The key technical milestone was ImageNet 2012, which initiated the deep learning revolution through AlexNet [175]. This was followed by VGG, GoogLeNet, and ResNet [176] with residual connections. In 2014, R-CNN [31] established the foundation for two-stage detection. The year 2015 marked a turning point with the introduction of U-Net [85], Faster R-CNN, and Fully Convolutional Networks [177]. In 2016, YOLOv1 [12] was introduced as the first single-shot architecture capable of achieving 45 FPS.

Parallel developments occurred in medicine. In 2016, Gulshan et al. [52], in JAMA, demonstrated that deep learning achieved performance comparable to expert readers under the evaluated conditions for diabetic retinopathy detection using 128,000 fundus images. In 2017, Esteva et al. [178], in Nature, showed that CNNs achieved dermatologist-level skin disease classification. Litjens et al. [3] subsequently published the first major survey on deep learning in medical imaging in Medical Image Analysis.

This foundation phase established large clinical imaging datasets and deep-learning methods that later enabled YOLO-based medical detection. YOLO appeared only near the end of the phase, and early medical applications largely adopted general-purpose AI techniques without medical-specific architectural adaptations.

10.3 Phase 2018–2020: Democratization and the Emergence of Transformers

The technical milestones of this phase included YOLOv2 [14] with anchor boxes, and YOLOv3 [20] with multi-scale FPN architectures [44]. Mask R-CNN [10] opened the field of instance segmentation. RetinaNet's Focal Loss [35] addressed severe class imbalance. Vaswani et al. [179] introduced "Attention Is All You Need," initiating the Transformer paradigm. CornerNet [36] and FCOS [37] established the anchor-free paradigm. Toward the end of this period, YOLOv4 [20] and YOLOv5 [73] significantly expanded accessibility through Mosaic augmentation and the PyTorch ecosystem.

Parallel medical developments included the emergence of large-scale medical datasets: BraTS [87,90], CheXpert [92] with 224K chest X-rays, LIDC-IDRI [91] for lung-nodule detection, and CAMELYON16 [93] for histopathology. During this period, the FDA also began approving the first radiology AI devices [4].

The defining characteristic of this phase was that large medical datasets enabled AI systems to learn effectively, although YOLO remained primarily a general-purpose detector with only limited medical-aware adaptations such as multi-scale FPNs.

10.4 Phase 2020–2023: Specialization and the COVID-19-Driven Acceleration

Technically, YOLOv5 [73] spread rapidly with thousands of GitHub forks. YOLOv7 [13] introduced E-ELAN. YOLOv8 [38] marked a major transition toward anchor-free architecture and decoupled heads, features particularly well suited for medical multi-task detection and classification. nnU-Net [51] won 23 MICCAI challenges and became the *de facto* widely used reference baseline for medical segmentation. Swin Transformer [157] initiated hierarchical Vision Transformers. DETR introduced end-to-end detection. SAM [54], released in 2023, established the foundation for segmentation-oriented foundation models, while CLIP [83] shaped the field of vision-language pretraining.

The COVID-19 pandemic (2020–2022) accelerated four major trends. First, thousands of YOLO + GAN augmentation studies for COVID-19 chest X-ray detection were published, although Wynants et al. [126], in BMJ, criticized most for high risk of bias and lack of external validation. Second, the FDA introduced

Emergency Use Authorizations (EUA) for numerous AI devices. Third, the FDA AI/ML SaMD Action Plan (2021) established the foundation for continuous-learning AI systems. Fourth, federated learning [100,156] became recognized as a legal and practical solution to HIPAA/GDPR-related data silos.

The defining characteristic of this phase was that medicine began to motivate YOLO-specific adaptations. Anchor-free architectures suited polymorphic tumors; decoupled heads supported multitask learning; lightweight nano-scale models (<10 MB) became appropriate for POCUS and mobile microscopy. Nevertheless, no truly “medical-native” YOLO architecture yet existed.

10.5 Phase 2023–2025: Paradigm Convergence and Clinical Evidence

Technical milestones included YOLOv9 [76], YOLOv10 [43], and YOLO11 [77]. Competing architectures increasingly converged toward YOLO-like paradigms, including RT-DETR [42] and SAM2 [119]. Medical foundation models such as MedSAM [54], pretrained on 1.57 million image-mask pairs, Biomed-CLIP (Tarik & Radouane, 2025) [121], and Med-PaLM [180] developed rapidly.

The major medical milestone was the emergence of randomized controlled trial evidence for AI-assisted colonoscopy, in which multiple CADe systems (including non-YOLO and YOLO-family detectors) demonstrated significant improvements in adenoma detection rate over standard practice, as later consolidated in meta-analyses [15,16]. The CADe-colonoscopy literature represents the most mature body of clinical evidence for real-time medical object detection to date, although the share of evidence directly attributable to YOLO architecture remains smaller than for earlier non-YOLO CADe systems. Hybrid YOLO+SAM systems with self-prompting capabilities [119] demonstrated the potential to surpass traditional state-of-the-art methods.

The number of FDA-approved AI devices surpassed 950 [4], of which 723 (76%) belonged to radiology. The FDA issued final, nonbinding PCCP guidance for AI-enabled device software functions in August 2025 [6]. The EU AI Act entered into force in 2024; under the current timeline, high-risk medical-device AI covered through Annex I is scheduled to become subject to the relevant obligations on 2 August 2028 [5]. CLAIM 2024 [23] updated AI reporting standards, while TRIPOD+AI [22] introduced structured reporting guidance for prediction models using regression or machine-learning methods.

The defining characteristic of this phase was the convergence of technical innovation, clinical evidence, and regulatory frameworks, creating conditions for a substantive shift in the field.

10.6 Phase 2025–2026: Toward Clinical-Aware YOLO

The major technical milestone was YOLO26 introducing ProgLoss + Small-Target-Aware Label Assignment (STAL). This represents a direct example of medical problems shaping algorithmic design: small-lesion detection is a core medical requirement, including retinal microbleeds [81], pulmonary nodules <5 mm [53], and mammographic microcalcifications [82]. YOLO26 is the first reported YOLO variant to explicitly target this challenge from the outset. However, its performance claims currently originate from preprints and still require independent validation.

YOLOE [61] introduced zero-shot detection through CLIP integration, addressing rare-disease detection. YOLOv12 [18] with Area Attention became well suited for diagnostic tasks, while YOLOv13 [80] with Hypergraph architecture targeted dense WSI analysis.

Meanwhile, the FDA issued final, nonbinding guidance on Predetermined Change Control Plans for AI-enabled device software functions in August 2025 [6], and the EU AI Act obligations for high-risk medical-device AI covered through Annex I are scheduled to apply from 2 August 2028 [5].

The defining characteristic of this phase is that YOLO architecture began to be designed explicitly for medical applications from inception. Medicine no longer merely “borrows AI,” but instead acts as a driving

force for technical innovation itself. Nevertheless, these claims still require peer review and prospective clinical validation.

10.7 Co-Evolution Mapping-Summary Matrix

A YOLO × Medicine × Regulation matrix spanning the pre-YOLO context (2010–2015) and co-evolution period (2016–2026) is presented in Table 20.

Table 20: YOLO × Medicine × Regulation matrix (pre-YOLO context, 2010–2015; co-evolution, 2016–2026).

Year	YOLO/AI Technical Event	Medical Event	Regulatory Event	Direction of Influence
2012	AlexNet; ImageNet (Deng et al., 2009 [112])	-	-	AI self-driven
2015	U-Net, Faster R-CNN, FCN	BraTS; ChestX-ray14	-	Bidirectional
2016	YOLOv1; ResNet	Gulshan et al. [52]. DR study	-	AI → Medicine
2017	YOLOv2; Mask R-CNN; Transformer	Esteva et al. [178]; Litjens et al [3]	-	AI → Medicine
2018	YOLOv3 multi-scale FPN	Demand for multi-scale lesion detection	-	Medicine → AI
2020	YOLOv4, YOLOv5, DETR	COVID-19 pandemic; YOLO+GAN COVID studies	FDA AI/ML Action Plan	Medicine strongly influencing AI
2021	nnU-Net	FDA AI >100 devices	CE MDR 2017/745 effective	Bidirectional + regulatory
2022	YOLOv7; SAM 1	YOLOv5l RCT polyp ADR 29% vs. 20%	FDA continuous AI guidance	Medicine → AI
2023	YOLOv8 anchor-free; SAM2	Hybrid YOLO+SAM; CLAIM 2024	-	Medicine → AI
2024	YOLOv9 PGI; YOLOv10 NMS-free	FDA AI >950 devices; MedSAM	EU AI Act entered into force	AI → Medicine + regulation
2025	YOLOv12 Attention; YOLOv13 Hypergraph; YOLO26 ProgLoss+STAL	DR microbleed detection; dense WSI	FDA final PCCP guidance issued	Medicine → AI
2026	YOLOE zero-shot	Rare disease detection paradigm	EU AI Act general provisions; Annex I medical-device obligations due 2028	Medicine → AI

10.8 The Influence of Major Events on the AI Ecosystem

Several medical and regulatory events have profoundly shaped the AI ecosystem. The COVID-19 pandemic accelerated four trends:

- (i) federated learning became mainstream because of the need to share insights without sharing data;
- (ii) edge AI became essential for frontline triage;
- (iii) synthetic augmentation using GANs and diffusion models became standard practice;
- (iv) rapid regulatory pathways such as EUA established precedents for regulatory agility.

The FDA AI/ML SaMD Action Plan (2021) laid the foundation for continuous learning, culminating in the final, nonbinding PCCP guidance issued in August 2025 [6]. This transformed model-design philosophy: instead of systems being “frozen at submission,” researchers must now design audit trails, monitoring systems, and rollback mechanisms from the outset.

The EU AI Act [5] establishes transparency, documentation, and human-oversight obligations for high-risk AI systems. CLAIM 2024 [23] and TRIPOD+AI [22] updated reporting guidance in 2024, whereas CONSORT-AI [25] and SPIRIT-AI [24], published in 2020, provide trial-specific extensions for reports and protocols of clinical trials involving AI interventions.

Implementation science frameworks have become increasingly important, particularly in addressing the average 17-year gap between scientific discovery and clinical adoption.

10.9 Future Directions of Co-Evolution

Five future directions can be envisioned for the co-evolution of YOLO and medicine.

First, **personalized AI**, in which each patient possesses a fine-tuned model through LoRA or adapter-based mechanisms, making the concept of the “patient digital twin” increasingly realistic.

Second, continuous learning under PCCP frameworks [6], where AI systems update dynamically through post-market surveillance data, requiring YOLO systems to incorporate audit trails and rollback mechanisms.

Third, multimodal foundation models integrate imaging, EHRs, genomics, and clinical text [181], with YOLO functioning primarily as a localization head within larger multimodal pipelines.

Fourth, global health equity, requiring AI systems to generalize across ethnicities, geographic regions, and healthcare systems [104]. This requirement will likely lead to the development of regional federated AI networks such as those across ASEAN or the EU.

Fifth, integration with implementation science, emphasizing organizational, cultural, and economic factors in deployment, such that YOLO becomes embedded within broader healthcare ecosystems rather than functioning as an isolated technical tool.

11 Clinical-Aware YOLO Framework-A Four-Layer Conceptual Framework

Despite the technical advances discussed in the preceding sections, the gap between algorithmic performance measured on benchmarks and real-world clinical utility (the algorithm-to-clinic gap) remains a major barrier to large-scale deployment of medical AI systems [26,58]. Most medical YOLO studies optimize for metrics such as mAP@0.5 on fixed datasets, whereas clinicians are concerned with fundamentally different questions: “Does the model miss malignant tumors?”, “Will the conclusions remain stable across different imaging devices?”, and “Can I explain to the patient why the model issued a warning?”

To narrow this gap, we propose the Clinical-Aware YOLO Framework, a four-layer conceptual framework designed according to deployment-first and clinician-centric principles. Fig. 3 illustrates the

architecture of this framework, showing how each layer addresses specific clinical requirements and feeds back into the system for continuous improvement. This framework is a conceptual proposal derived from a synthesis of the literature and therefore requires prospective validation in future studies.

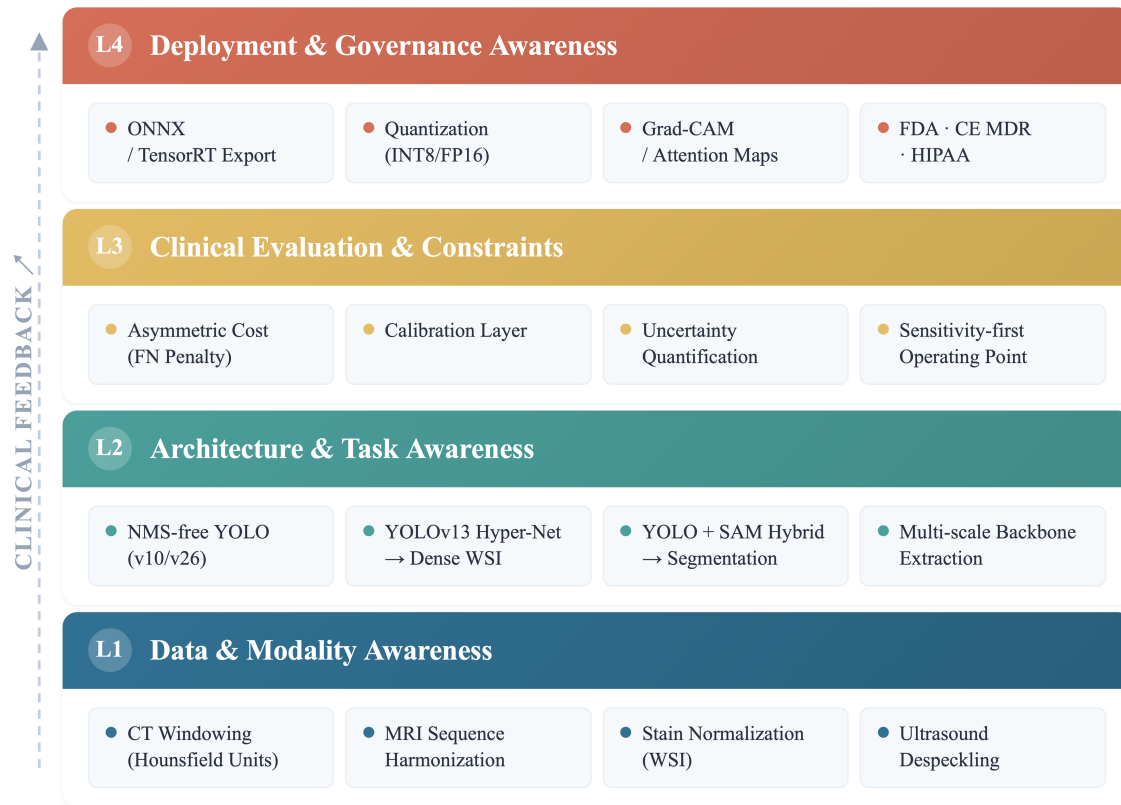


Figure 3: Clinical-aware YOLO framework. Disclaimer: This framework is conceptual and aspirational, representing future directions rather than currently established capabilities. All four layers require prospective multicenter validation before clinical deployment. The diagram illustrates an idealized architecture; real-world implementation will vary based on regulatory requirements, institutional infrastructure, and specific clinical use cases.

Implementation Requirements for Clinical Deployment: Transitioning from research prototypes to deployed clinical systems requires: (a) external multicenter validation across diverse populations and imaging devices; (b) continuous performance drift monitoring with automated alerting mechanisms; (c) clear clinician responsibility frameworks defining human-in-the-loop decision authority; (d) interoperability standards compliance (DICOM for imaging, HL7 FHIR for EHR integration); (e) patient consent protocols addressing AI-assisted diagnosis transparency; (f) post-market surveillance aligned with FDA PCCP and EU AI Act requirements.

11.1 Layer 1-Data and Modality Awareness

The first layer addresses the foundational problem that medical data does not follow independent and identically distributed (i.i.d.) assumptions. Different hospitals employ different imaging devices (Siemens, GE, Philips), acquisition protocols, and patient populations with distinct demographic distributions.

Layer 1 consists of four components:

1. Modality-aware preprocessing pipelines, including automatic Hounsfield Unit normalization for CT [49], intensity histogram matching for MRI [50] and Macenko-Reinhard color normalization for histopathology.
2. Class imbalance correction through dynamically weighted Focal Loss [35].
3. Federated preprocessing layers [62,156], enabling hospitals to apply standardized preprocessing pipelines without sharing raw patient data.
4. Domain randomization, automatically generating variants such as vendor shifts, resolution shifts, and demographic shifts, thereby encouraging the model to learn invariance to non-pathological factors.

This design philosophy reflects lessons learned from nnU-Net [51], where self-configuring preprocessing was shown to be a critical determinant of performance, sometimes more influential than the architectural choice itself.

11.2 Layer 2-Architecture and Task Awareness

The second layer constitutes the technical core of the framework, integrating multiple YOLO generations in a flexible manner, including:

- YOLOv8 [38] anchor-free with decoupled heads;
- YOLOv10 [43] NMS-free inference;
- YOLOv12 [18] Area Attention;
- YOLOv13 [80] HyperACE;
- YOLO26 NMS-free with ProgLoss/STAL;
- YOLOE [61] open-vocabulary detection.

The key design principle of Layer 2 is **context-aware model selection**. An orchestration module dynamically selects the most appropriate YOLO variant based on:

- (a) imaging modality,
- (b) computational-resource constraints, and
- (c) disease characteristics.

The backbone architecture could be switched on demand without full retraining through adapter tuning mechanisms [54]. Cross-modality consistency is maintained via contrastive learning over paired (image, text) representations [177].

11.3 Layer 3-Clinical Evaluation Awareness

The third layer represents the fundamental distinction between the proposed framework and conventional YOLO pipelines. Instead of optimizing neutral metrics such as mAP alone, Layer 3 imposes asymmetric cost functions reflecting clinical risk.

For example, in cancer screening, missing a malignant lesion (false negative) may delay diagnosis and worsen patient outcomes, whereas a false positive generally leads only to additional confirmatory testing.

This layer is implemented through six components:

1. Weighted cost-sensitive loss, in which false-negative and false-positive penalties are selected from a clinically justified utility or cost analysis for the specific use case rather than from a universal multiplier.
2. Calibration layers [112], ensuring predicted probabilities correspond to real-world event frequencies.
3. Uncertainty quantification through MC-Dropout [113] or Deep Ensembles [114].

4. Operating-point selection, allowing clinicians to adjust thresholds according to specific deployment scenarios.
5. Fairness auditing using metrics such as demographic parity and equalized odds [103,115,116].
6. Reproducibility checklists aligned with recommendations by Moassefi et al. [165].

This layer represents a reorientation from performance-first toward safety-first AI, consistent with the recommendations of CLAIM 2024 [23] and TRIPOD+AI [22].

11.4 Layer 4-Deployment and Governance Awareness

The fourth layer bridges AI models with real-world clinical workflows and contains five major components:

1. EHR fusion, integrating detection outputs with electronic health records [179].
2. Explainability modules, including Grad-CAM heatmaps [101], SHAP values [102], LIME [166], Transformer attention rollout, and counterfactual examples [166].
3. DICOM/PACS integration, packaging the system as software as a medical device (SaMD) compliant with FDA 510(k) pathways [4] or CE MDR 2017/745.
4. Human-in-the-loop feedback, allowing clinicians to accept, reject, or correct predictions, with feedback logged to improve the model under PCCP frameworks [6].
5. Audit and monitoring systems, detecting distribution shifts between operational and training data and issuing alerts when performance degrades, thereby satisfying post-market surveillance requirements under the EU AI Act [5].

11.5 Prospective Validation Roadmap-Four Phases

Real-world deployment of the Clinical-Aware YOLO Framework requires a four-phase validation roadmap aligned with international reporting standards.

Phase 1-Multicenter Retrospective Validation

Validation should be conducted retrospectively across at least three institutions with distinct protocols and demographics. Objectives include demonstrating generalizability and identifying domain-shift patterns. Reporting should follow CLAIM 2024 [23] and TRIPOD+AI [22].

Phase 2-Prospective Multicenter Clinical Trials

Clinical trials should comply with CONSORT-AI [25] and SPIRIT-AI [24]. Patients should be randomized into AI-assisted and control groups. The primary endpoint must be a clinical outcome (e.g., adenoma detection rate [15,16], mortality, or time-to-diagnosis) rather than purely technical metrics such as mAP.

Phase 3-FDA 510(k) or CE MDR Submission

This phase requires technical, clinical, and quality-management documentation compliant with ISO 13485 and IEC 62304. For SaMD class II+ systems, pre-submission meetings with the FDA are recommended. Guidance from Sivakumar et al. [4] regarding testing gaps should also be followed.

Phase 4-Post-Market Surveillance and Continuous Learning Under PCCP

This phase includes real-world performance monitoring, model updates according to approved PCCP plans, rollback procedures for degraded performance, and periodic fairness audits [113,155]. A mapping of the four-layer framework to clinical workflow and regulatory standards is provided in Table 21.

Table 21: Mapping of the Four-layer framework $\times$ Clinical workflow $\times$ Regulatory standards.

Layer	Clinical Workflow Stage	Primary Compliance Standards
L1 Data & Modality	DICOM preprocessing; multicenter data acquisition	DICOM SR; HL7 FHIR; HIPAA; GDPR; CLAIM Items 13–18 (Tejani et al., 2024 [23])
L2 Architecture & Task	Context-aware model selection; fine-tuning	CLAIM Items 19–25; TRIPOD+AI methodology (Collins et al., 2024 [22])
L3 Clinical Evaluation	Prospective evaluation; reader studies; fairness auditing	CLAIM Items 32–36; CONSORT-AI (Liu et al., 2020 [25]); SPIRIT-AI (Cruz Rivera et al., 2020 [24]); reproducibility guidelines (Moassefi et al. [165])
L4 Deployment & Governance	PACS integration; SaMD deployment; PCCP; monitoring	FDA 510(k); CE MDR 2017/745; ISO 13485; IEC 62304; PCCP guidance (FDA, 2025); EU AI Act (European Parliament, 2024 [5])

12 Future Directions

12.1 Group A-Architectural Directions

The first group focuses on the evolution of the YOLO architecture itself, comprising three closely related development directions. The first direction is deeper Transformer attention integration: YOLOv12 with Area Attention Mechanism [18] captures global context that locally focused CNNs often miss, but there remains room to more deeply integrate the Transformer backbone with the YOLO decoder. Specific directions include reducing $O(n^2)$ computational cost through adaptive local attention mechanisms, while leveraging self-supervised pretraining with vision transformers on unlabeled medical data. The Swin Transformer [157] hierarchical structure has been shown to be effective for medical images. TransUNet [148] and Swin-UNet [149] provide blueprints for combining Transformer with CNN in medicine.

The second direction is integrating YOLO with SAM or MedSAM for detection-to-segmentation cascade. The integration of YOLO with SAM [54,73] is becoming a paradigm for hybrid detection + segmentation problems, particularly when MedSAM [54] has been pretrained on 1.57M image-mask pairs. YOLO rapidly generates bounding boxes, after which SAM2 [119] or MedSAM2 automatically generates segmentation masks from the bounding-box prompt. Development directions: end-to-end joint training instead of separated cascade, integration of medical-specific prompting, and extension to video segmentation. The third direction is multimodal data fusion: simultaneously integrating image data, EHR, and clinical text. A survey by Huang and colleagues [181] published in npj Digital Medicine on early/late/joint fusion provides a reference framework. This is particularly important when combined with clinical language models such as Med-PaLM [180] or BiomedCLIP [121].

12.2 Group B-Data Efficiency

The second group addresses the foundational problem: labeled data is always scarce. The first direction is federated YOLO for multi-hospital data. Federated learning [62,100,156,182] has demonstrated its potential in medicine. Development directions: differential privacy to quantify and ensure privacy guarantees [156]; secure aggregation against adversarial hospitals; cross-silo federated learning for international collaboration.

The second direction is self-supervised and few-shot YOLO. Self-supervised pre-training on unlabeled medical images (CT from TCIA [89], X-rays from MIMIC-CXR) is a direction for reducing dependence

on expensive annotation. SimCLR [168] and MoCo [183] have demonstrated effectiveness on natural images. Few-shot learning has been systematically organized by Pachetti and Colantonio [60]. The third direction is open-vocabulary YOLO for rare diseases: YOLOE through Re-parameterizable Region-Text Alignment allows lesion localization based on textual descriptions. Development directions: integration with clinical language models to understand complex medical descriptions, construction of medical-specific text encoders. Active learning [111] is an important direction for optimizing annotation costs.

12.3 Group C-Clinical Translation

The third group focuses on bringing YOLO from the laboratory into actual clinical practice, comprising four directions. The first direction is 3D/4D YOLO. MedYOLO [65] is an early step but remains limited: on 3D LIDC-IDRI lung nodules, the authors reported that the model failed to demonstrate learning progress, rather than reporting a zero mAP at a defined IoU threshold. Development directions: true volumetric attention; 4D = 3D + temporal for video/fluoroscopy; integration with 3D U-Net [63] and V-Net [158]; sparse attention for computational efficiency; tracking with DeepSORT [86] or ByteTrack [87].

The second direction is explainable YOLO with uncertainty quantification. The future work should investigate integrating Grad-CAM [101], SHAP [102], and LIME [166] natively (not post-hoc). Conformal prediction provides clinically valid uncertainty intervals with coverage guarantees [184]. Tonekaboni and colleagues [167] discuss clinicians' requirements for interpretability, including not only saliency maps but also auditable logic chains.

The third direction is regulatory-ready and fairness-ready YOLO. Under the current EU AI Act timeline, the Annex I high-risk obligations relevant to covered medical-device AI are scheduled to apply on 2 August 2028 [5] and the FDA strengthens post-market surveillance [4]. Regulatory-ready YOLO requires a reproducibility audit trail [165]; automated post-market monitoring; PCCP-compliant continuous learning [6]; EU AI Act compliance; periodic fairness audits [103,104,185]. The fourth direction is personalized YOLO. The "patient digital twin" concept, in which each patient may have a personalized digital replica, is an emerging research direction. Personalized YOLO could: (i) use patient-specific fine-tuning with LoRA/adaptor; (ii) test-time adaptation for each case; (iii) integrate longitudinal monitoring of disease progression; (iv) continuously learn from physician feedback through the PCCP framework. This direction may combine foundation models with edge AI and implementation science. A future directions matrix organized by group, direction, approach, timeline, and medical impact is presented in Table 22.

Table 22: Future directions matrix-Group × Direction × Approach × Timeline × Medical impact.

Group	Specific Direction	Technical Approach	Timeline	Medical Impact
A. Architecture	Transformer-deep YOLO	Local sparse attention; ViT pretrain on medical data	2026–2027	Diagnostic accuracy ↑
A. Architecture	YOLO+SAM/MedSAM joint	End-to-end joint training; medical prompting	2026–2028	Unified detection-segmentation
A. Architecture	Multimodal fusion YOLO	Early/late/joint fusion [181]	2027–2028	Decision support integrating EHR
B. Data	Federated YOLO global	Differential privacy; secure aggregation	2026–2030	Global health equity

(Continued)

Table 22 (continued)

Group	Specific Direction	Technical Approach	Timeline	Medical Impact
B. Data	Self-supervised YOLO	Contrastive learning on unlabeled medical data (SimCLR, MoCo)	2026–2027	Reduce annotation cost by 50%+
B. Data	Open-vocabulary YOLO	Med-CLIP; clinical LM integration	2026–2028	Rare disease detection
C. Clinical	3D/4D YOLO native	Volumetric attention; sparse 3D	2027–2029	CT/MRI/fluoroscopy end-to-end
C. Clinical	Explainable YOLO native	Saliency in forward pass; conformal prediction	2026–2027	FDA/EU compliance; trust
C. Clinical	Regulatory + Fairness-ready	Audit trail; PCCP; fairness audit [103]	2026–2028	Deployment scaling; equity
C. Clinical	Personalized YOLO	Patient digital twin; LoRA adapter	2028–2030	Precision medicine

Note: ↑ indicates improvement/increase relative to current baseline.

13 Conclusion

This topical review has provided a structured, critical, and forward-looking analysis of the evolution and applications of the YOLO family of models in medical image analysis during the 2016–2026 period. Starting from the YOLOv1 architecture of Redmon and colleagues at CVPR 2016 with a single forward pass and 45 FPS on PASCAL VOC, through the 2025–2026 versions featuring the Area Attention mechanism of YOLOv12, Hypergraph theory of YOLOv13, NMS-free end-to-end design of YOLO26 with ProgLoss + STAL, and the zero-shot detection capability of YOLOE, YOLO has undergone a ten-year transformation journey from a pure object detection algorithm into an increasingly clinically oriented medical AI ecosystem.

This review has three limitations. First, it is a topical review rather than a formal systematic review and does not conduct a quantitative meta-analysis; all synthesis statements are narrative. Second, the literature scope is limited to English-language articles, potentially omitting valuable contributions in other languages, particularly Chinese. Third, evidence for the newest architectures is heterogeneous: YOLO11 relies primarily on official software documentation, YOLOv12 and YOLOv13 include preprint evidence, YOLO26 is supported by both a preprint and official documentation, and YOLOE has a peer-reviewed publication. Assessments of these versions may therefore evolve as further independent evaluations appear. The principal conclusions concerning co-evolution, the proposed framework, failure modes, bias, and reproducibility do not depend on any single recent release.

For YOLO to truly become a trustworthy clinical tool on a global scale, three core challenges remain to be addressed. The first challenge is interpretability: integrating Grad-CAM [101], SHAP, and clinician-centric XAI not as accessories but as core design components. The second challenge is multimodal learning: simultaneously integrating images, EHR, and clinical text remains at an early stage. The third challenge

is clinical validation: most medical YOLO studies still stop at internal/retrospective validation; a necessary next step is multicenter prospective trials according to FDA 510(k), CE MDR, CONSORT-AI, and TRIPOD+AI standards.

Looking further ahead, the integration of YOLO with foundation vision models (SAM, CLIP, MedSAM), federated learning, and edge-AI can produce a generation of medical models that are versatile, privacy-respecting, and sustainable. The Clinical-Aware YOLO Framework proposed in this paper is not merely a technical roadmap; it supports a broader shift in research: from benchmark optimization to placing clinical needs at the center; from publishing “black box” models to sharing reproducible code, data, and pipelines; from isolated laboratories to multicenter and multinational collaborations; from neglecting fairness to integrating fairness audits as a core design requirement. Only through the convergence of technical innovation, regulatory compliance, and ethical consensus can YOLO, and the broader real-time medical AI ecosystem, contribute to more accurate, reproducible, and clinically responsible real-time medical vision systems.

Acknowledgement: The authors thank colleagues who exchanged ideas and provided feedback during the conception of this article. Part of the writing was assisted by AI language tools; all scientific content, analyses, conclusions, reference selection, and conceptual framework proposals are entirely the responsibility of the authors.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Hue T. Tran (Lead Author): Conceptualized the research idea, developed the manuscript framework, and wrote the initial draft. Coordinated the synthesis of literature from all members and consolidated the content for each version of the manuscript. Phat T. Nguyen (Corresponding Author): Supervised the review process, formulated the search strategy, and selected relevant literature. Reviewed and edited the content to ensure academic rigor and logical flow. Monitored each revision and approved the final manuscript. Giang L. Nguyen (Corresponding Author): Participated in peer review and editing of manuscript content. Reviewed argument coherence ensuring proper formatting of tables and images and approved the final version. Nguyen Nhu Son: Supported literature review, summarized and classified information. Maintained consistency control across the paper and participated in final approval. Anh Nguyen Trong: Conducted literature searches and contributed to study selection according to predefined criteria. Participated in citation control and assisted with data synthesis for tables in the manuscript. Reviewed the final version. Anh Nguyen Tu: Supported literature review and information classification. Participated in citation control and assisted with figure data compilation. Reviewed the final version. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All studies analyzed are publicly available materials; the complete citation list is presented in the References section. Any additional metadata is available on reasonable request from the corresponding authors.

Ethics Approval: Not applicable.

Conflicts of Interest: Given his role as an Editorial Board Member of this journal, Long Giang Nguyen had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

Supplementary Materials: The supplementary material is available online at <https://www.techscience.com/doi/10.32604/cmc.2026.086547/s1>.

References

1. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nat Med.* 2022;28(1):31–8. doi:10.1038/s41591-021-01614-0.

2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7.
3. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42(13):60–88. doi:10.1016/j.media.2017.07.005.
4. Sivakumar R, Lue B, Kundu S. FDA approval of artificial intelligence and machine learning devices in radiology: a systematic review. *JAMA Netw Open*. 2025;8(11):e2542338. doi:10.1001/jamanetworkopen.2025.42338.
5. European Parliament. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [Internet]. 2024 [cited 2026 August 13]. Available from: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>.
6. U.S. Food and Drug Administration. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions: Guidance for Industry and Food and Drug Administration Staff. 2025 [cited 2026 August 13]. Available from: <https://www.fda.gov/media/166704/download>.
7. Doi K. Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Comput Med Imag Graph*. 2007;31(4–5):198–211. doi:10.1016/j.compmedimag.2007.02.002.
8. Twinanda AP, Shehata S, Mutter D, Marescaux J, de Mathelin M, Padoy N. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans Med Imag*. 2017;36(1):86–97. doi:10.1109/TMI.2016.2593957.
9. Girshick R. Fast R-CNN. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*; 2015 Dec 7–13; Santiago, Chile. p. 1440–8. doi:10.1109/ICCV.2015.169.
10. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; 2017 Oct 22–29; Venice, Italy. p. 2980–8. doi:10.1109/ICCV.2017.322.
11. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*. Red Hook, NY, USA: Curran Associates; 2015. p. 91–9.
12. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2016 Jun 27–30; Las Vegas, NV, USA. p. 779–88. doi:10.1109/CVPR.2016.91.
13. Wang CY, Bochkovskiy A, Liao HM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2023 Jun 17–24; Vancouver, BC, Canada. p. 7464–75. doi:10.1109/CVPR52729.2023.00721.
14. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; 2017 Jul 21–26; Honolulu, HI, USA. p. 6517–25. doi:10.1109/CVPR.2017.690.
15. Wang P, Berzin TM, Glissen Brown JR, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*. 2019;68(10):1813–9. doi:10.1136/gutjnl-2018-317500.
16. Hassan C, Misawa M, Rizkala T, Mori Y, Sultan S, Facciorusso A, et al. Computer-aided diagnosis for leaving colorectal polyps *in situ*: a systematic review and meta-analysis. *Ann Intern Med*. 2024;177(7):919–28. doi:10.7326/M23-2865.
17. Ultralytics. YOLO11 Documentation [Internet]. 2025 [cited 2026 August 13]. Available from: <https://docs.ultralytics.com/models/yolo11/>.
18. Ultralytics. YOLO12: attention-Centric Object Detection [Internet]. 2025 [cited 2026 August 13]. Available from: <https://docs.ultralytics.com/models/yolo12/>.
19. Cai Z, Zhou K, Liao Z. A systematic review of YOLO-based object detection in medical imaging: advances, challenges, and future directions. *Comput Mater Continua*. 2025;85(2):2255–303. doi:10.32604/cmc.2025.067994.
20. Terven J, Córdova-Esparza DM, Romero-González JA. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. *Mach Learn Knowl Extr*. 2023;5(4):1680–716. doi:10.3390/make5040083.

21. Yeerjiang A, Wang Z, Huang X, Zhang J, Chen Q, Qin Y, et al. YOLOv1 to YOLOv10: a comprehensive review of YOLO variants and their application in medical image detection. *J Artif Intell Pract.* 2024. doi:10.23977/jaip.2024.070314.
22. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378. doi:10.1136/bmj-2023-078378.
23. Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. *Radiol Artif Intell.* 2024;6(4):e240300. doi:10.1148/ryai.240300.
24. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med.* 2020;26(9):1351–63. doi:10.1038/s41591-020-1037-7.
25. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* 2020;26(9):1364–74. doi:10.1038/s41591-020-1034-x.
26. Aggarwal R, Sounderajah V, Martin G, Ting DSW, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *npj Digit Med.* 2021;4(1):65. doi:10.1038/s41746-021-00438-z.
27. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. *JB Evid Synth.* 2020;18(10):2119–26. doi:10.11124/JBIES-20-00167.
28. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA, 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* 2021;n71. doi:10.1136/bmj.n71.
29. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet.* 2019;393(10181):1577–9. doi:10.1016/S0140-6736(19)30037-6.
30. van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology.* 2011;261(3):719–32. doi:10.1148/radiol.11091710.
31. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*; 2014 Jun 23–28; Columbus, OH, USA. p. 580–7. doi:10.1109/CVPR.2014.81.
32. Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The pascal visual object classes (VOC) challenge. *Int J Comput Vis.* 2010;88(2):303–38. doi:10.1007/s11263-009-0275-4.
33. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: *Computer Vision-ECCV 2014*. Cham, Switzerland: Springer; 2014. p. 740–55. doi:10.1007/978-3-319-10602-1_48.
34. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, et al. SSD: single shot MultiBox detector. In: *Computer Vision-ECCV 2016*. Cham, Switzerland: Springer; 2016. p. 21–37. doi:10.1007/978-3-319-46448-0_2.
35. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; 2017 Oct 22–29; Venice, Italy. p. 2999–3007. doi:10.1109/ICCV.2017.324.
36. Law H, Deng J. CornerNet: detecting objects as paired keypoints. In: *Computer Vision-ECCV 2018*. Cham, Switzerland: Springer; 2018. p. 765–81. doi:10.1007/978-3-030-01264-9_45.
37. Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. In: *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 9626–35. doi:10.1109/iccv.2019.00972.
38. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLOv8 (Version 8.0.0) [Software]. 2023 [cited 2026 Jan 1]. Available from: <https://github.com/ultralytics/ultralytics>.
39. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-end object detection with transformers. In: *Computer Vision-ECCV 2020*. Cham, Switzerland: Springer; 2020. p. 213–29. doi:10.1007/978-3-030-58452-8_13.

40. Zhu X, Su W, Lu L, Li B, Wang X, Dai J. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations (ICLR 2021); 2021 May 3–7; Online.
41. Zhang H, Li F, Liu S, Zhang L, Su H, Zhu J, et al. DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection. In: Proceedings of the International Conference on Learning Representations (ICLR 2023); 2023 May 1–5; Kigali, Rwanda. p. 01212. doi:10.1109/ICCV51070.2023.01212.
42. Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, et al. DETRs beat YOLOs on real-time object detection. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA. p. 16965–74. doi:10.1109/CVPR52733.2024.01605.
43. Wang A, Chen H, Liu L, Chen K, Lin Z, Han J, et al. YOLOv10: real-time end-to-end object detection. In: Proceedings of the Advances in Neural Information Processing Systems 37; 2024 Dec 10–15; Vancouver, BC, Canada. p. 107984–8011. doi:10.52202/079017-3429.
44. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 936–44. doi:10.1109/CVPR.2017.106.
45. Liu S, Qi L, Qin H, Shi J, Jia J. Path aggregation network for instance segmentation. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 8759–68. doi:10.1109/CVPR.2018.00913.
46. Rezaatofghi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 658–66. doi:10.1109/CVPR.2019.00075.
47. Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D. Distance-IoU loss: faster and better learning for bounding box regression. *Proc AAAI Conf Artif Intell.* 2020;34(7):12993–3000. doi:10.1609/aaai.v34i07.6999.
48. Li X, Wang W, Wu L, Chen S, Hu X, Li J, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. In: *Advances in neural information processing systems*. Red Hook, NY, USA: Curran Associates; 2020. p. 21002–12.
49. Hounsfield GN. Computerized transverse axial scanning (tomography): part 1. Description of system. *Br J Radiol.* 1973;46(552):1016–22. doi:10.1259/0007-1285-46-552-1016.
50. Matt A Bernstein KFK. *Handbook of MRI pulse sequences*. Amsterdam, Netherlands: Elsevier Academic Press; 2004.
51. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Meth.* 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.
52. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA.* 2016;316(22):2402–10. doi:10.1001/jama.2016.17216.
53. Setio AAA, Traverso A, de Bel T, Berens MSN, van den Bogaard C, Cerello P, et al. Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. *Med Image Anal.* 2017;42(4):1–13. doi:10.1016/j.media.2017.06.015.
54. Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. *Nat Commun.* 2024;15(1):654. doi:10.1038/s41467-024-44824-z.
55. Zhang H, Cisse M, Dauphin YN, Lopez-Paz D. Beyond empirical risk minimization. In: Proceedings of the International Conference on Learning Representations (ICLR 2018); 2018 Apr 30–May 3; Vancouver, BC, Canada. Available from: <https://openreview.net/forum?id=r1Ddp1-Rb>.
56. Yun S, Han D, Chun S, Oh SJ, Yoo Y, Choe J. CutMix: regularization strategy to train strong classifiers with localizable features. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 6022–31. doi:10.1109/iccv.2019.00612.
57. Cubuk ED, Zoph B, Mané D, Vasudevan V, Le QV. AutoAugment: learning augmentation strategies from data. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 113–23. doi:10.1109/CVPR.2019.00020.

58. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng.* 2010;22(10):1345–59. doi:10.1109/TKDE.2009.191.
59. Windecker D, Baj G, Shiri I, Kazaj PM, Kaesmacher J, Gräni C, et al. Generalizability of FDA-approved AI-enabled medical devices for clinical use. *JAMA Netw Open.* 2025;8(4):e258052. doi:10.1001/jamanetworkopen.2025.8052.
60. Pachetti E, Colantonio S. A systematic review of few-shot learning in medical imaging. *Artif Intell Med.* 2024;156(3):102949. doi:10.1016/j.artmed.2024.102949.
61. Wang A, Liu L, Chen H, Lin Z, Han J, Ding G. YOLOE: real-time seeing anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; 2025 Oct 19–25; Honolulu, HI, USA. p. 24591–602. doi:10.1109/ICCV51701.2025.02280.
62. Rieke N, Hancox J, Li W, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. *npj Digit Med.* 2020;3(1):119. doi:10.1038/s41746-020-00323-1.
63. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016*. Cham, Switzerland: Springer; 2016. p. 424–32. doi:10.1007/978-3-319-46723-8_49.
64. Aruk I. Performance evaluation of different YOLO models for lung nodule detection. *Bitlis Eren Iniversitesi Fen Bilimleri Dergisi.* 2025;35(4):789–801. doi:10.17798/bitlisfen.1780664.
65. Sobek J, Medina Inojosa JR, Medina Inojosa BJ, Rassoulinejad-Mousavi SM, Conte GM, Lopez-Jimenez F, et al. MedYOLO: a medical image object detection framework. *J Imaging Inform Med.* 2024;37(6):3208–16. doi:10.1007/s10278-024-01138-2.
66. Ju RY, Cai W. Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm. *Sci Rep.* 2023;13(1):20077. doi:10.1038/s41598-023-47460-7.
67. Akyuz M, Besnili S, Magat G, Ceylan M. Real-time segmentation and detection of ponticulus posticus in lateral cephalometric radiographs using YOLOv8: a step towards enhanced clinical evaluation. *BMC Oral Health.* 2025;25(1):828. doi:10.1186/s12903-025-06196-8.
68. Choudhry O, Ali S, Biyani CS, Jones D. Real-time tool detection in laparoscopic datasets for surgical training in low-resource settings. *Healthc Technol Lett.* 2025;12(1):e70045. doi:10.1049/htl2.70045.
69. Han W, Dong X, Wang G, Ding Y, Yang A. Application and improvement of YOLO11 for brain tumor detection in medical images. *Front Oncol.* 2025;15:1643208. doi:10.3389/fonc.2025.1643208.
70. Liu J, Sun P, Yuan Y, Chen Z, Tian K, Gao Q, et al. YOLOv12 algorithm-aided detection and classification of lateral malleolar avulsion fracture and subfibular ossicle based on CT images: multicenter study. *JMIR Med Inform.* 2025;13:e79064. doi:10.2196/79064.
71. Hua Z, Aranganadin K, Yeh CC, Hai X, Huang CY, Leung TC, et al. A benchmark review of YOLO algorithm developments for object detection. *IEEE Access.* 2025;13:123515–45. doi:10.1109/ACCESS.2025.3586673.
72. Wang CY, Mark Liao HY, Wu YH, Chen PY, Hsieh JW, Yeh IH. CSPNet: a new backbone that can enhance learning capability of CNN. In: *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; 2019 Jun 14–19; Seattle, WA, USA. p. 1571–80. doi:10.1109/cvprw50498.2020.00203.
73. Ultralytics. YOLOv5: a State-of-the-Art Real-Time Object Detection System [Internet]. 2020 [cited 2026 Jan 1]. Available from: <https://docs.ultralytics.com/yolov5/>.
74. Sahingoz OK, Karatas Baydogmus G, Kugu E. A systematic literature review of you only look once architectures (v1–v12) in healthcare systems. *Diagnostics.* 2026;16(6):935. doi:10.3390/diagnostics16060935.
75. Sapkota R, Karkee M. Ultralytics YOLO evolution: an overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 object detectors for computer vision and pattern recognition. *arXiv:2510.09653.* 2025.
76. Wang CY, Yeh IH, Mark Liao HY. YOLOv9: learning what you want to learn using programmable gradient information. In: *Computer Vision-ECCV 2024*. Cham, Switzerland: Springer; 2024. p. 1–21. doi:10.1007/978-3-031-72751-1_1.
77. Varol-Malkoçoğlu AB, Şamlı R. A detailed review of YOLO algorithms: YOLOv1 to YOLOv11. *Electrica.* 2026;26:1–17. doi:10.5152/electrica.2026.26047.

78. Hakim FL, Paramita C. Comparative performance analysis of YOLOv5; YOLOv8; and YOLOv9 small and medium variants for explainable pneumonia detection in chest X-ray images. *J Teknol Dan Sist Inf.* 2026;19(1):89–104. doi:10.36341/rabit.v1i1.2.8099.
79. Himabindu BA, Jyothi GN, Nandini M, Sujatha C, Manisha N, Subramanyam M. Performance evaluation of next-generation YOLO architectures (YOLOv10 and YOLOv8) in urological diagnostic systems. In: *Proceedings of the 2026 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*; 2026 Mar 11–13; Bangkok, Thailand. p. 737–44. doi:10.1109/ICMLAS67792.2026.11484018.
80. Lei M, Li S, Wu Y, Hu H, Zhou Y, Zheng X, et al. YOLOv13: real-time object detection with hypergraph-enhanced adaptive visual perception. *arXiv:2504.17903.* 2025.
81. Porwal P, Pachade S, Kamble R, Kokare M, Deshmukh G, Sahasrabuddhe V, et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data.* 2018;3(3):25. doi:10.3390/data3030025.
82. Lee RS, Gimenez F, Hoogi A, Miyake KK, Gorovoy M, Rubin DL. A curated mammography data set for use in computer-aided detection and diagnosis research. *Sci Data.* 2017;4(1):170177. doi:10.1038/sdata.2017.177.
83. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning*; 2021 Jul 18–24. p. 8748–63.
84. Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans Med Imag.* 2015;34(10):1993–2024. doi:10.1109/TMI.2014.2377694.
85. Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. *Data Brief.* 2020;28(5):104863. doi:10.1016/j.dib.2019.104863.
86. Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric. In: *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*; 2017 Sep 17–20; Beijing, China. p. 3645–9. doi:10.1109/ICIP.2017.8296962.
87. Zhang Y, Sun P, Jiang Y, Yu D, Weng F, Yuan Z, et al. ByteTrack: multi-object tracking by associating every detection box. In: *Computer Vision-ECCV 2022.* Cham, Switzerland: Springer; 2022. p. 1–21. doi:10.1007/978-3-031-20047-2_1.
88. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. In: *International Conference on Learning Representations (ICLR 2021)*; 2021 May 3–7; Online.
89. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imag.* 2013;26(6):1045–57. doi:10.1007/s10278-013-9622-7.
90. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data.* 2017;4(1):170117. doi:10.1038/sdata.2017.117.
91. Armato SG 3rd, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, et al. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys.* 2011;38(2):915–31. doi:10.1118/1.3528204.
92. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, et al. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. *Proc AAAI Conf Artif Intell.* 2019;33(1):590–7. doi:10.1609/aaai.v33i01.3301590.
93. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, van Ginneken B, Karssemeijer N, Litjens G, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA.* 2017;318(22):2199–210. doi:10.1001/jama.2017.14585.
94. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, et al. *MURA: large dataset for abnormality detection in musculoskeletal radiographs.* arXiv:1712.06957. 2017. doi:10.3390/reports2040026.
95. Bernal J, Sánchez FJ, Fernández-Esparrach G, Gil D, Rodríguez C, Vilariño F. WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. *Comput Med Imaging Graph.* 2015;43(1258):99–111. doi:10.1016/j.compmedimag.2015.02.007.

96. Pogorelov K, Randel KR, Griwodz C, Eskeland SL, de Lange T, Johansen D, et al. KVASIR: a multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference; 2017 Jun 20; Taipei, Taiwan. p. 164–9. doi:10.1145/3083187.3083212.
97. Plissiti ME, Dimitrakopoulos P, Sfikas G, Nikou C, Krikoni O, Charchanti A. Sipakmed: a new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In: Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP); 2018 Oct 7–10; Athens, Greece. p. 3144–8. doi:10.1109/ICIP.2018.8451588.
98. Decencière E, Zhang X, Cazuguel G, Lay B, Cochener B, Trone C, et al. Feedback on a publicly distributed image database: the messidor database. *Image Anal Stereol.* 2014;33(3):231. doi:10.5566/ias.1155.
99. Zhou ZH. A brief introduction to weakly supervised learning. *Natl Sci Rev.* 2018;5(1):44–53. doi:10.1093/nsr/nwx106.
100. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep.* 2020;10(1):12598. doi:10.1038/s41598-020-69250-1.
101. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. p. 618–26. doi:10.1109/ICCV.2017.74.
102. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates; 2017. p. 4765–74.
103. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. *ACM Comput Surv.* 2022;54(6):1–35. doi:10.1145/3457607.
104. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science.* 2019;366(6464):447–53. doi:10.1126/science.aax2342.
105. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common *Thorax* diseases. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. p. 3462–71. doi:10.1109/CVPR.2017.369.
106. Borgli H, Thambawita V, Smedsrud PH, Hicks S, Jha D, Eskeland SL, et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. *Sci Data.* 2020;7(1):283. doi:10.1038/s41597-020-00622-y.
107. Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018 Apr 4–7; Washington, DC, USA; 2018. p. 168–72. doi:10.1109/ISBI.2018.8363547.
108. Frid-Adar M, Diamant I, Klang E, Amitai M, Goldberger J, Greenspan H. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing.* 2018;321(5):321–31. doi:10.1016/j.neucom.2018.09.013.
109. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates; 2014. p. 2672–80.
110. Pinaya WHL, Tudosiu P-D, Dafflon J, Da Costa PF, Fernandez V, Nachev P, et al. Brain imaging generation with latent diffusion models. In: Deep generative models. Berlin/Heidelberg, Germany: Springer; 2022. p. 117–26. doi:10.1007/978-3-031-18576-2_12.
111. Budd S, Robinson EC, Kainz B. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Med Image Anal.* 2021;71(4):102062. doi:10.1016/j.media.2021.102062.
112. Deng J, Dong W, Socher R, Li LJ, Kai L, Li FF. ImageNet: a large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009 Jun 20–25; Miami, FL, USA. p. 248–55. doi:10.1109/CVPR.2009.5206848.

113. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning; 2016 Jun 19; New York, NY, USA. p. 1050–9.
114. Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates. p. 6402–13.
115. Pfohl S, Xu Y, Foryciarz A, Ignatiadis N, Jenkins J, Shah N. Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; 2022 Jun 21–24; Seoul, Republic of Korea. p. 1039–52. doi:10.1145/3531146.3533166.
116. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates; 2016. p. 3315–23.
117. Mongan J, Moy L, Kahn CE Jr. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. *Radiol Artif Intell.* 2020;2(2):e200029. doi:10.1148/ryai.2020200029.
118. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia. p. 1321–30.
119. Buyukpatpat H, Sezer EA, Guzel MS. A comparative evaluation of zero-shot performance of SAM, SAM2, MedSAM, and MedSAM2 models on lung segmentation. *J Imag Inform Med.* 2025;1–20. doi:10.1007/s10278-025-01778-y.
120. Amirouche F, Prosper AM, Mzeihem M. Enhancing pediatric distal radius fracture detection: optimizing YOLOv8 with advanced AI and machine learning techniques. *BMC Med Imag.* 2025;25(1):316. doi:10.1186/s12880-025-01669-2.
121. Tarik L, Radouane M. Comprehensive comparative analysis of state-of-the-art medical vision-language models: LLaVA-med, VILA-M3, and BiomedCLIP performance evaluation on chest radiography datasets. In: Proceedings of the 2025 International Conference on Intelligent Systems: Theories and Applications (SITA); 2025 Oct 20–21; Rabat, Morocco. p. 1–9. doi:10.1109/SITA67914.2025.11273425.
122. Tan JL, Pitawela D, Chinnaratha MA, Beany A, Aguila EJ, Chen HT, et al. Exploring vision transformers for classifying early Barrett's dysplasia in endoscopic images: a pilot study on white-light and narrow-band imaging. *JGH Open.* 2024;8(9):e70030. doi:10.1002/jgh3.70030.
123. Saraiva MM, González-Haba M, Widmer J, Mendes F, Gonda T, Agudo B, et al. Deep learning and automatic differentiation of pancreatic lesions in endoscopic ultrasound: a transatlantic study. *Clin Transl Gastroenterol.* 2024;15(11):e00771. doi:10.14309/ctg.0000000000000771.
124. Vu Trung K, Hollenbach M, Veldhuizen GP, Saldanha OL, Garbe J, Rosendahl J, et al. Deep learning-based detection of malignant bile duct stenosis in fluoroscopy images of endoscopic retrograde cholangiopancreatography. *Digestion.* 2025;106(4):287–302. doi:10.1159/000543049.
125. Robles-Medrande C, Baquerizo-Burgos J, Alcivar-Vasquez J, Kahaleh M, Rajjman I, Kunda R, et al. Artificial intelligence for diagnosing neoplasia on digital cholangioscopy: development and multicenter validation of a convolutional neural network model. *Endoscopy.* 2023;55(8):719–27. doi:10.1055/a-2034-3803.
126. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *BMJ.* 2020;369:m1328. doi:10.1136/bmj.m1328.
127. Bhandari A, Scott L, Weibach M, Marwah R, Lasocki A. Assessment of artificial intelligence (AI) reporting methodology in glioma MRI studies using the Checklist for AI in Medical Imaging (CLAIM). *Neuroradiol.* 2023;65(5):907–13. doi:10.1007/s00234-023-03126-9.
128. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, et al. CheXNet: radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv:1711.05225.* 2017.
129. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* 2024;74(3):229–63. doi:10.3322/caac.21834.

130. Moreira IC, Amaral I, Domingues I, Cardoso A, Cardoso MJ, Cardoso JS. INbreast: toward a full-field digital mammographic database. *Acad Radiol*. 2012;19(2):236–48. doi:10.1016/j.acra.2011.09.014.
131. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafi H, et al. International evaluation of an AI system for breast cancer screening. *Nature*. 2020;577(7788):89–94. doi:10.1038/s41586-019-1799-6.
132. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Shraddha T. Transparency and reproducibility in artificial intelligence. *Nature*. 2020;586(7829):E14–6. doi:10.1038/s41586-020-2766-y.
133. Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F, et al. Improving reproducibility in machine learning research (A report from the NeurIPS 2019 reproducibility program). *J Mach Learn Res*. 2021;22(164):1–20. doi:10.48550/arxiv.2003.12206.
134. Shinozaki S, Osawa H, Hayashi Y, Lefor AK, Yamamoto H. Linked color imaging for the detection of early gastrointestinal neoplasms. *Ther Adv Gastroenterol*. 2019;12:1756284819885246. doi:10.1177/1756284819885246.
135. Tsai YY, Lee KH, Mukundan A, Karmakar R, Nagisetty Y, Seyoum DG, et al. Enhancing esophageal cancer detection using a deep learning framework and a novel spectrum-aided vision enhancer for virtual narrow band imaging. *Sci Rep*. 2026;16(1):25149. doi:10.1038/s41598-026-52501-y.
136. Peng L, Li C, Liu W, Xie S, Chen X, Xiao F, et al. Boosting bacterial detection with hyperspectral mining. *Optica*. 2025;12(3):315–24. doi:10.1364/optica.543225.
137. Lai CL, Karmakar R, Mukundan A, Natarajan RK, Lu SC, Wang CY, et al. Advancing hyperspectral imaging and machine learning tools toward clinical adoption in tissue diagnostics: a comprehensive review. *APL Bioeng*. 2024;8(4):041504. doi:10.1063/5.0240444.
138. Chou CK, Lee KH, Karmakar R, Mukundan A, Gade PC, Gupta D, et al. Emulating hyperspectral and narrow-band imaging for deep-learning-driven gastrointestinal disorder detection in wireless capsule endoscopy. *Bioengineering*. 2025;12(9):953. doi:10.3390/bioengineering12090953.
139. Hsu H, Lee KH, Karmakar R, Mukundan A, Attar RS, Liu PH, et al. From innovation to application: can emerging imaging techniques transform breast cancer diagnosis? *Diagnostics*. 2025;15(21):2718. doi:10.3390/diagnostics15212718.
140. Tan M, Le QV. EfficientNet: rethinking model scaling for convolutional neural networks. In: *Proceedings of the 36th International Conference on Machine Learning*; 2019 Jun 9–15; Long Beach, CA, USA. p. 6105–14.
141. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med*. 2018;24(9):1342–50. doi:10.1038/s41591-018-0107-6.
142. Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M, Fenyö D, et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*. 2018;24(10):1559–67. doi:10.1038/s41591-018-0177-5.
143. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*. 2020;580(7802):252–6. doi:10.1038/s41586-020-2145-8.
144. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*. 2019;25(1):65–9. doi:10.1038/s41591-018-0268-3.
145. Wood CS, Thomas MR, Budd J, Mashamba-Thompson TP, Herbst K, Pillay D, et al. Taking connected mobile-health diagnostics of infectious diseases to the field. *Nature*. 2019;566(7745):467–74. doi:10.1038/s41586-019-0956-2.
146. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015*. Cham, Switzerland: Springer; 2015. p. 234–41. doi:10.1007/978-3-319-24574-4_28.
147. Azad R, Aghdam EK, Rauland A, Jia Y, Avval AH, Bozorgpour A, et al. Medical image segmentation review: the success of U-Net. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(12):10076–95. doi:10.1109/TPAMI.2024.3435571.
148. Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. *Med Image Anal*. 2024;97(2):103280. doi:10.1016/j.media.2024.103280.

149. Cao H, Wang Y, Chen J, Jiang D, Zhang X, Tian Q, et al. Swin-Unet: Unet-like pure transformer for medical image segmentation. In: *Computer Vision-ECCV 2022 Workshops*. Cham, Switzerland: Springer; 2022. p. 205–18. doi:10.1007/978-3-031-25066-8_9.
150. Oktay O, Schlemper J, Le Folgoc L, Lee M, Heinrich M, Misawa K, et al. Attention U-Net: learning where to look for the pancreas. In: *Proceedings of the Medical Imaging with Deep Learning (MIDL 2018)*; 2018 Jul 4–6; Amsterdam, The Netherlands. p. 1–10.
151. Tlebaldinova A, Omiotek Z, Karmenova M, Kumargazhanova S, Smailova S, Tankibayeva A, et al. Comparison of modern convolution and transformer architectures: YOLO and RT-DETR in Meniscus diagnosis. *Computers*. 2025;14(8):333. doi:10.3390/computers14080333.
152. Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. DINOv2: learning robust visual features without supervision [Internet]. 2024 [cited 2026 Jan 1]. Available from: <https://openreview.net/pdf?id=a68SUt6zFt>.
153. Jin R, Xu Z, Zhong Y, Yao Q, Dou Q, Zhou SK, et al. FairMedFM: fairness benchmarking for medical imaging foundation models. In: *Advances in neural information processing systems datasets and benchmarks track*. p. 111318–57.
154. Hirsch L, Sutton EJ, Huang Y, Kayis B, Hughes M, Martinez D, et al. High-performance open-source AI for breast cancer detection and localization in MRI. *Radiol Artif Intell*. 2025;7(5):e240550. doi:10.1148/ryai.240550.
155. Seyyed-Kalantari L, Zhang H, McDermott MBA, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nat Med*. 2021;27(12):2176–82. doi:10.1038/s41591-021-01595-0.
156. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell*. 2020;2(6):305–11. doi:10.1038/s42256-020-0186-1.
157. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*; 2021 Oct 10–17; Montreal, QC, Canada. p. 9992–10002. doi:10.1109/ICCV48922.2021.00986.
158. Milletari F, Navab N, Ahmadi SA. V-net: fully convolutional neural networks for volumetric medical image segmentation. In: *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*; 2016 Oct 25–28; Stanford, CA, USA. p. 565–71. doi:10.1109/3DV.2016.79.
159. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. *Science*. 2019;363(6433):1287–9. doi:10.1126/science.aaw4399.
160. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell*. 2021;3(7):610–9. doi:10.1038/s42256-021-00338-7.
161. Banerjee I, Bhattacharjee K, Burns JL, Trivedi H, Purkayastha S, Seyyed-Kalantari L, et al. “Shortcuts” causing bias in radiology artificial intelligence: causes, evaluation, and mitigation. *J Am Coll Radiol*. 2023;20(9):842–51. doi:10.1016/j.jacr.2023.06.025.
162. Yi PH, Bachina P, Bharti B, Garin SP, Kanhere A, Kulkarni P, et al. Pitfalls and best practices in evaluation of AI algorithmic biases in radiology. *Radiology*. 2025;314(2):e241674. doi:10.1148/radiol.241674.
163. Stupple A, Singerman D, Celi LA. The reproducibility crisis in the age of digital medicine. *npj Digit Med*. 2019;2:2. Erratum in: *npj Digit Med*. 2019;2:19. doi:10.1038/s41746-019-0079-z.
164. Renard F, Guedria S, De Palma N, Vuillerme N. Variability and reproducibility in deep learning for medical image segmentation. *Sci Rep*. 2020;10(1):13724. doi:10.1038/s41598-020-69920-0.
165. Moassefi M, Singh Y, Conte GM, Khosravi B, Rouzrokh P, Vahdati S, et al. Correction: checklist for reproducibility of deep learning in medical imaging. *J Imaging Inform Med*. 2024;37:1664–73. Erratum in: *J Imaging Inform Med*. 2025;38:1914. doi:10.1007/s10278-024-01295-4.
166. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA. p. 1135–44. doi:10.1145/2939672.2939778.
167. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What clinicians want: contextualizing explainable machine learning for clinical end use. In: *Proceedings of the 4th Machine Learning for Healthcare Conference*; 2019 Aug 9–10; Ann Arbor, Michigan. p. 359–80.

168. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Online. p. 1597–607.
169. Xu Y, Khan TM, Song Y, Meijering E. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. arXiv:2605.06714. 2026. doi:10.1007/s10462-024-11033-5.
170. Han S, Mao H, Dally WJ. Deep compression: compressing deep neural networks with pruning; trained quantization and Huffman coding. In: Proceedings of the International Conference on Learning Representations (ICLR 2016); 2016 May 2–4; San Juan, Puerto Rico.
171. Gou J, Yu B, Maybank SJ, Tao D. Knowledge distillation: a survey. *Int J Comput Vis.* 2021;129(6):1789–819. doi:10.1007/s11263-021-01453-z.
172. Zoph B, Le QV. Neural architecture search with reinforcement learning. In: Proceedings of the International Conference on Learning Representations (ICLR 2017); 2017 Apr 24–26; Toulon, France.
173. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 4510–20. doi:10.1109/CVPR.2018.00474.
174. Hoi SCH, Sahoo D, Lu J, Zhao P. Online learning: a comprehensive survey. *Neurocomputing.* 2021;459:249–89. doi:10.1016/j.neucom.2021.04.112.
175. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates; 2012. p. 1097–105.
176. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/CVPR.2016.90.
177. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2015 Jun 7–12; Boston, MA, USA. p. 3431–40. doi:10.1109/CVPR.2015.7298965.
178. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature.* 2017;542(7639):115–8. doi:10.1038/nature21056.
179. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems. Red Hook, NY, USA: Curran Associates; 2017. p. 5998–6008.
180. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172–80. doi:10.1038/s41586-023-06291-2.
181. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit Med.* 2020;3(1):136. doi:10.1038/s41746-020-00341-z.
182. McMahan B, Moore E, Ramage D, Hampson S, Arcas BA. Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics; 2017 Apr 20–22; Fort Lauderdale, FL, USA. p. 1273–82.
183. He K, Fan H, Wu Y, Xie S, Girshick R. Momentum contrast for unsupervised visual representation learning. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. p. 9726–35. doi:10.1109/cvpr42600.2020.00975.
184. Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world. Cham, Switzerland: Springer; 2005. doi:10.1007/0-387-25061-1_8.
185. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: fairness gaps in deep chest X-ray classifiers. *Pac Symp Biocomput.* 2021;26:232–43. doi:10.1142/9789811232701_0022.