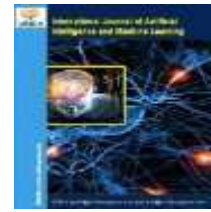




International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Machine Learning And Deep Learning For Glaucoma Detection: A Systematic Narrative Review Of Methods, Datasets, Evaluation Practice, And Clinical Translation

Sayema Asfahan Khusro¹, Shilpa Joshi²

¹JNEC MGMU Department of Information Technology MGM University. Email: sayemaasfahan@gmail.com (Corresponding author)

²CSE Department, JNEC MGMU. ssjd10@gmail.com sjoshi@mgm.ac.in

Abstract—Glaucoma is a leading cause of irreversible blindness and is often asymptomatic until substantial visual field loss occurs. Clinicians integrate intraocular pressure, optic nerve head assessment, optical coherence tomography (OCT), visual field testing, and risk factors—but screening and longitudinal care remain resource-intensive and variable across settings. Machine learning (ML), and especially deep learning (DL), promise automated assistance from fundus photographs, OCT volumes, and increasingly electronic health record text and multimodal combinations. This systematic narrative review synthesizes peer-reviewed work on ML/DL for glaucoma detection and related screening tasks, with emphasis on what is compared, on which data, and under which evaluation designs. We summarize trends across fundus-based classification and segmentation-plus-classification pipelines, OCT-based modeling, multimodal fusion, privacy-preserving federation, and explainability. We highlight persistent threats to validity, including small or homogeneous public benchmarks, weak external validation, spectrum bias, optimistic metrics, and limited reporting of calibration and fairness, and translate these into concrete recommendations for more clinically credible benchmarking.

Index Terms—glaucoma, machine learning, deep learning, fundus imaging, optical coherence tomography, systematic review, clinical validation, external validation, explainable AI.

I. INTRODUCTION

Glaucoma encompasses a group of optic neuropathies in which progressive retinal ganglion cell loss can lead to irre-versible visual disability [1]. Population modeling consistently projects a rising burden of disease, with the number of affected individuals projected to exceed one hundred million by 2040 [2]. Because early-stage disease may produce minimal symptoms, timely detection relies on structured examination and longitudinal monitoring combining structural and functional tests [3, 4]. The asymptomatic nature of early disease, combined with the irreversibility of optic-nerve damage once it occurs, makes glaucoma a prototypical screening problem: small gains in case-finding sensitivity at the population level can translate into substantial reductions in avoidable blindness over a lifetime, particularly in regions with limited access to subspecialty ophthalmology.

A. Clinical context for AI assistance

In contemporary practice, clinicians use intraocular pressure (IOP), optic nerve head (ONH) evaluation on color fundus imaging, perimetry, and optical coherence tomography (OCT)—often in combination—to stratify risk, diagnose, and monitor progression [5]. These modalities differ in availability, cost, test-retest behavior, and dependence on operator expertise, creating incentives for scalable decision support, particularly in screening and triage workflows where the supply of trained graders is the binding constraint. Tele-ophthalmology programs, optometry-led screening, and primary-care-based diabetic retinopathy services already routinely collect color fundus photographs that are not always reviewed by glaucoma specialists; this represents an actionable surface for opportunistic glaucoma case-finding if automated systems can be made robust and well calibrated [6].

B. Role of machine learning

Computational pipelines have evolved from classical machine learning on hand-crafted features (e.g., cup-to-

disc ratio (CDR), texture, vessel attributes) to end-to-end convolutional networks and hybrid systems combining learned representations with clinical metadata [7, 8]. Recent reviews emphasize that fundus-based AI can achieve strong within-dataset discrimination when benchmarks and training protocols align, but that translation to heterogeneous clinical populations remains a central open problem [9, 10]. The same pattern has been observed in adjacent ophthalmic AI tasks such as diabetic retinopathy grading, where impressive challenge-set performance has at times failed to replicate in deployment due to camera, illumination, and population shifts [11, 12].

C. Why another review?

There is no shortage of literature surveys on glaucoma AI. However, existing surveys vary widely in their selection criteria, level of methodological rigor, and the extent to which they engage with the evaluation layer rather than only the architecture layer. Many overview articles summarize architectures and datasets; fewer explicitly foreground evaluation validity (external testing, spectrum composition, leakage risks, reporting of confidence and calibration) as the primary determinant of whether reported performance is clinically meaningful. Others focus on a single sub-task (e.g., disc-cup segmentation) without situating it within an end-to-end screening pipeline, while still others mix high-tier peer-reviewed work with non-peer-reviewed sources without distinguishing between them. The result is that journal reviewers, clinical partners, and policymakers often lack a concise, integrative reference that makes the methodological landscape—and its open gaps—legible.

D. Review question and contributions

This manuscript addresses the following research question: How do ML/DL approaches perform for glaucoma detection across fundus, OCT, and multimodal data under real-world benchmarking constraints, and which methodological gaps most impede safe deployment? We make the following contributions:

- 1) A transparent, PRISMA-oriented review protocol [13], with a reproducible extraction schema that can be re-applied as the literature evolves;
- 2) A modality-based synthesis (fundus classical ML, fundus deep learning, segmentation-first pipelines, OCT-centric models, multimodal and EHR text fusion, federated approaches, and explainability) that links method families to typical failure modes rather than to leaderboard rankings;
- 3) A heuristic but explicit appraisal rubric (Section II) that lets readers contextualize headline metrics against evaluation strength;
- 4) Concrete reporting recommendations aligned with modern medical-AI guidance (CONSORT-AI, SPIRIT-AI, TRIPOD/TRIPOD-AI, and FUTURE-AI) [14–16];
- 5) Curated reference integrity, deliberately restricted to peer-reviewed venues with stable bibliographic records and excluding off-topic citations that are common in informal survey drafts.

E. Scope and out-of-scope topics We focus on supervised ML/DL for glaucoma detection (binary or multi-class), referable disease screening, optic disc/cup segmentation as a clinically motivated intermediate task, OCT-derived structural-pattern modeling, and multimodal/text approaches that include glaucoma as a primary endpoint. We deliberately exclude: (i) purely intraoperative or device-engineering papers that do not report patient-relevant endpoints;

(ii) general fundus quality-control or vessel-segmentation papers that do not condition on glaucoma labels; (iii) cataract, AMD, and DR papers except where they form a directly comparable methodological precedent; and (iv) commercial systems for which only marketing material—not peer-reviewed performance data—is available. The temporal scope is 2015–2026, with foundational older references retained where they remain canonical.

F. Organization of the review

The remainder of the article is organized as follows. Section II describes the review methodology, including the search strategy, eligibility rules, extraction schema, and the appraisal rubric used throughout the synthesis. Section III summarizes the clinical background and translates it into computational task formulations. Section IV synthesizes the literature by modality and method family. Section V surveys datasets, public challenges, and benchmarking norms, including a critical view of leakage risks. Section VI discusses external validity, label/spectrum bias, fairness, and translational barriers. Section VII consolidates take-aways and contrasts them with prior surveys. Section VIII closes with reporting recommendations for primary studies. Appendices provide the search strings and the per-study

extraction schema.

II. REVIEW METHODOLOGY

A. Design and reporting standard

We conducted a systematic narrative review: systematic in prespecified sources, search logic, eligibility rules, and tabulated extraction; narrative in interpretive synthesis and emphasis on cross-study patterns rather than pooled meta-analytic effect sizes for heterogeneous metrics. Reporting follows PRISMA 2020 principles where applicable [13]; reg-istration and risk-of-bias tooling (e.g., PROBAST/PROBAST-AI [17]) may be added if the target journal requires a fully registered systematic review.

We chose a narrative format rather than a quantitative meta-analysis for two principal reasons. First, primary studies report fundamentally heterogeneous endpoints (binary glaucoma vs. referable disease vs. progression), heterogeneous label semantics (expert consensus vs. visual-field surrogate vs. chart review), and heterogeneous denominators (image-level vs. eye-level vs. patient-level), so pooled AUC summaries would conflate biologically and statistically distinct estimands. Second, threats to validity in this literature are dominated by qualitative factors—leakage, spectrum bias, choice of operating point, calibration to local prevalence—that a quantitative effect estimate cannot adequately surface.

B. Review question and eligibility criteria

Population and concept. Human glaucoma detection, screen-ing, or glaucoma-vs-healthy classification; progression or risk prediction when framed as ML/DL from ophthalmic imaging or multimodal clinical data.

Interventions and comparators. Supervised learning mod-els (classical ML, CNNs, hybrid models, transformers for text or multimodal fusion, foundation-model fine-tuning) compared to other models, clinician grading, or historical baselines as reported.

Outcomes. Discrimination metrics (AUC, accuracy, sensi-tivity/specificity at a defined operating point), segmentation

overlap (Dice, IoU) where applicable, and—when reported—calibration (Brier score, reliability diagrams), fairness stratifi-cation, or external validation performance.

Eligibility. Peer-reviewed full articles and high-quality peer-

TABLE I HEURISTIC APPRAISAL TIERS USED TO CONTEXTUALIZE STUDY CLAIMS IN THIS NARRATIVE REVIEW

(ILLUSTRATIVE; NOT A SUBSTITUTE FOR A

DOMAIN-SPECIFIC RISK-OF-BIAS INSTRUMENT SUCH AS PROBAST-AI

[17] OR QUADAS-AI FOR A REGISTERED SYSTEMATIC REVIEW).

reviewed survey papers in English, 2015–2026, presenting

primary algorithmic results or rigorous meta-analyses. Inclusion required: a clearly identifiable glaucoma-related task, a quan-titative performance report, and a stable bibliographic record (DOI or indexed identifier). Exclusions comprised: non-peer-reviewed theses or preprints without subsequent peer-reviewed publication; pure hardware or device-engineering papers with-out patient-relevant endpoints; secondary commentary that did not present new evidence or extracted data; and citations not retrievable through DOI or stable bibliographic records. Foundational pre-2015 references were retained where they remain canonical (e.g., epidemiology, classical ML baselines).

A. Information sources and search

We queried PubMed/MEDLINE, IEEE Xplore, and Springer-Link with controlled terms and synonyms for glaucoma, fundus, OCT, machine/deep learning, and screening. Database-specific strings and approximate hit counts are summarized in Appendix A. Snowballing from landmark surveys and challenge reports (e.g., REFUGE) supplemented database retrieval [9], and forward citation tracking through Google Scholar was used to recover recent extensions of seminal works. We also examined the reference lists of recent narrative reviews [9, 10, 18] to ensure recall of frequently cited studies that may have been indexed under non-MeSH keywords.

B. Study selection and data extraction

Two-stage screening (title/abstract, full text) was performed, with disagreements resolved by consensus discussion among the authors. Records were de-duplicated by DOI and by normalized title/author keys before screening. Where multiple publications described the same dataset and pipeline (e.g., conference paper and journal extension), the most complete peer-reviewed version was retained for extraction and the others were treated as companion records.

For each included study we extracted: imaging modality; task (binary glaucoma, referable disease, segmentation, progression); dataset(s) and split strategy; model class and backbone; main metrics with confidence intervals where reported; external validation (yes/no, and if yes, the number of independent cohorts); sample size and class balance; explainability method (if any); and stated limitations. The full field dictionary appears in Appendix B. Where a paper reported both within-dataset and cross-dataset performance, both numbers were retained because the pair is itself the most informative signal of generalization gap.

C. Critical appraisal and quality rubric

Because imaging-AI studies vary widely in design, we scored studies using a transparent heuristic rubric (Table I) rather than a single domain-specific Risk-of-Bias tool. Higher tiers emphasize independent external test cohorts or cross-dataset

Tier Criteria (examples)

A	Independent external or genuinely cross-dataset test cohort; explicit patient-level (not image-level) splits; leakage safeguards stated and audited; primary clinical metrics reported at prespecified operating points; calibration or uncertainty reporting (reliability, ECE/Brier, conformal sets); subgroup performance reported where ethically permissible; limitations explicitly discuss distribution shift, label semantics, and spectrum bias.
B	Single-site hold-out with clear patient/eye-level splits and class balance reported; preprocessing reproducible and documented; comparisons against strong baselines (not only weak ones); some discussion of operating-point selection; partial subgroup or cross-cohort evaluation.
C	Repeated k -fold or random hold-out on a single corpus without external testing; image-level rather than patient-level splits, or split key not stated; limited reporting of prevalence, thresholds, and sample sizes; high risk of optimistic bias for deployment claims; saliency presented as “explainability” without sanity checks.

evaluation, explicit patient-/eye-level splits, leakage avoidance (no patient overlap across splits), and reporting of uncertainty or calibration. The rubric is intentionally simple so that readers can re-apply it at a glance; it is not a substitute for PROBAST-AI or QUADAS-AI in a registered systematic review setting.

D. Threats to validity of this review

We acknowledge several threats to the validity of the review itself. First, English-language indexing biases recall against work in non-English ophthalmology journals, which are relevant for some Latin American, East Asian, and Middle Eastern cohorts. Second, our appraisal rubric is heuristic and may under-penalize subtler forms of leakage (e.g., adversarial validation, augmentation-induced near-duplicates, or threshold tuning on the test set). Third, we deliberately excluded preprint-only work, which biases recent recall against rapidly evolving subfields such as foundation-model-based ophthalmic AI; we treat that as a conservative choice for a peer-reviewed venue and explicitly flag the trade-off in the discussion. Fourth, like any narrative review we cannot fully eliminate selection bias in the choice of representative studies in Tables II and III; the appendix schema is provided so that readers can reproduce or contest the selection.

CLINICAL BACKGROUND AND COMPUTATIONAL TASK FORMULATIONS

A. Disease heterogeneity and diagnostic constructs

Glaucoma is not a single uniform phenotype; clinicians integrate IOP, optic disc appearance, RNFL and ganglion cell-inner plexiform layer (GC-IPL) metrics on OCT, and visual field defects [1, 4]. Some patients develop detectable structural change with “normal” IOP (normal-tension glaucoma), while elevated IOP alone is insufficient for diagnosis [5]. Major sub-types include primary open-angle glaucoma (POAG),

primary angle-closure glaucoma (PACG), normal-tension glaucoma (NTG), congenital and juvenile glaucoma, and various secondary forms (pigmentary, pseudoexfoliative, neovascular, steroid-induced, post-traumatic). These subtypes differ in epidemiology, mechanism, characteristic ONH morphology, and progression rate, and they imply different cost asymmetries between false negatives and false positives in screening: missing an aggressive normal-tension or angle-closure case has different downstream consequences than missing a slowly progressive low-risk POAG case [1, 3].

ML publications often reduce disease status to a binary label on fundus datasets; readers should map

this simplification to the broader clinical construct. In particular, “glaucoma” on a public benchmark may correspond, depending on the dataset, to: (a) clinical diagnosis from chart review by a specialist, (b) referable disease as graded by trained image readers, (c) optic-nerve appearance suggestive of glaucomatous damage even without confirmatory functional testing, or (d) elevated CDR above a fixed threshold. These four definitions are correlated but not equivalent, and merging datasets across them without harmonization injects label noise that can dwarf architecture-level differences [9].

B. Modality-specific signal

Color fundus photography supports assessment of neuroretinal rim, disc size, cupping, localized thinning (notching), parapapillary atrophy (PPA) zones, vessel deflections (bayonet-ing), and retinal nerve fiber layer (RNFL) defects when image quality is sufficient [1, 9]. It is widely available, inexpensive, and integrates naturally into screening at the primary-care level, but is operator- and quality-dependent and is intrinsically a 2-D projection of a 3-D nerve.

Optical coherence tomography provides micron-scale cross-sectional anatomy enabling quantitative RNFL and GC-IPL thickness summaries, B-scan classification, and more recently OCT-A vasculature analysis [19, 20]. OCT is more sensitive to early structural change than fundus photography in many cohorts, but artifact and segmentation errors propagate to downstream ML, and device-specific normative databases complicate cross-vendor pooling [19].

Visual field testing (perimetry) provides functional evidence of glaucomatous damage but exhibits high test-retest variability early in disease, requires a cooperative patient, and is far less scalable than imaging [21, 22]. Visual-field-derived archetypal patterns have been used both as labels and as features in ML pipelines [21].

Tonometry and risk factors. IOP, central corneal thickness, axial length, family history, ethnicity, age, and systemic vascular factors are inputs to clinical risk stratification. These tabular variables are increasingly used as auxiliary features alongside imaging in multimodal models [23].

Free-text and structured EHR. Encounter notes, problem lists, and medication histories carry information about glaucoma diagnoses, severity, treatment intensification, and surgical events that is sometimes more reliable than image-based labels alone, particularly for progression endpoints [24].

C. Computational task taxonomy

The literature can be organized into the following task families:

- **Binary image classification** (glaucoma vs. normal) on static fundus or OCT-derived representations is the dominant framing. It is convenient computationally and aligns with most public benchmarks, but is a coarse abstraction of clinical screening where the operating point and prevalence are critical.
- **Optic disc/cup segmentation** produces masks that yield CDR, ISNT-rule features, rim-to-disc area ratio, and other geometry-based descriptors, which are then fed into classical or shallow classifiers [25–27]. This route is more interpretable and easier to audit.
- **Referable-disease classification**, in which the positive class is operationalized as “should be referred to a specialist” rather than “has glaucoma”, is closer to actual screening deployment but rarer in public datasets.
- **OCT -based modeling**, including B-scan classification, volumetric (3-D) CNNs, and parametric thickness-map classifiers [20, 28, 29].
- **Multimodal fusion** (fundus + OCT + clinical variables) aligned at the patient level [23].
- **Free-text EHR pipelines** for progression and treatment-intensification prediction [24].
- **Federated and privacy-preserving training** across multiple clinics without centralizing raw images [30, 31].
- **Longitudinal progression prediction** from repeated imaging and visual fields, which remains comparatively underrepresented due to data friction (alignment, dropout, censoring).

This taxonomy guides the synthesis in Section IV.

D. Why glaucoma is methodologically harder than some neighbouring tasks

Compared with diabetic retinopathy or cataract grading, glaucoma poses several specific methodological challenges that recur throughout the literature. (i) Definitions: “glaucoma” is not unambiguously visible in any single image; it is a composite construct combining structural and functional signs over time, and image-only ground truth is therefore intrinsically noisier than for vascular pathology in DR. (ii) Normal anatomical variation: physiological cupping, large disc area, oblique insertion, and tilted discs produce CDR values that can mimic glaucomatous appearance without disease, especially in myopic eyes [1]. (iii) Cross-modality complementarity: OCT and visual field add information that is not present in fundus images, so models trained only on fundus are inherently capacity-limited on early disease. (iv) Spectrum

bias: most public datasets are skewed toward moderate or advanced cases, making early-disease performance unreliable to estimate [32]. These factors should temper expectations from headline AUC numbers, regardless of how strong the model architecture is.

RESULTS: SYNTHESIS BY MODALITY AND METHOD FAMILY

This section summarizes recurring design patterns and validity considerations. Because primary studies use incompatible metrics and test splits, we avoid ranking models by headline accuracy; instead we summarize mechanisms and evaluation strength using the rubric in Table I. Where direct numerical comparisons are mentioned, they are intended to illustrate cross-study patterns—not to imply a meta-analytic effect estimate.

A. Fundus imaging: classical machine learning

Classical pipelines extract CDR, neuroretinal-rim area, ISNT-rule violations, parapapillary atrophy descriptors, texture descriptors (e.g., GLCM, wavelet, higher-order spectra), vessel geometry, and shape moments, then train SVMs, random forests, gradient-boosted trees, or shallow neural nets [7, 33]. Strengths include interpretability, data efficiency when accurate segmentation masks are available, and a natural connection to clinical reasoning—a CDR-based decision tree is something an ophthalmologist can directly inspect and audit.

Limitations are also well established. Performance hinges on the quality of upstream segmentation: errors in disc or cup boundary localization propagate directly into CDR and ISNT estimates and dominate downstream classifier error in many cohorts [9]. Cross-dataset transfer is brittle when preprocessing or fundus appearance shifts (camera, illumination, color balance, image cropping). Hand-engineered texture statistics are often more sensitive to acquisition pipeline than to disease, which limits generalization to new clinics. The classical-ML literature nonetheless remains valuable because it provides interpretable baselines that deep models must outperform meaningfully—not just marginally—to justify their complexity and opacity in a clinical setting.

B. Fundus imaging: deep classification and transfer learning

CNN classifiers (often ImageNet-initialized ResNet-, DenseNet-, or EfficientNet-style backbones, and more recently Vision Transformer and ConvNeXt variants) dominate public benchmarks such as RIM-ONE, DRISHTI-GS, ACRIMA, ORIGA, and REFUGE [34–39]. Performance is typically highest under random hold-out from a single dataset distribution. Cross-dataset evaluation frequently reveals larger performance gaps than within-dataset cross-validation, suggesting overfitting to camera, color, and grading conventions [40–42]. Phene et al. [42], for example, demonstrated strong within-cohort performance but observed meaningful drops when the same model was applied to images from cameras and centres not seen during training—a pattern that has since been replicated repeatedly.

Transfer learning and pretraining. ImageNet pretraining remains the default initialization, but the literature increasingly explores domain-specific pretraining on large unlabeled fundus corpora using self-supervision (e.g., contrastive learning, masked image modeling) [43, 44]. Foundation-model-style retinal models trained on millions of fundus images report stronger out-of-distribution behavior on downstream ophthalmic tasks, although peer-reviewed glaucoma-specific evidence is still maturing.

Architectural breadth. Practical surveys observe that the choice of backbone (ResNet-50, DenseNet-121, EfficientNet-B0/B3, ConvNeXt-Tiny, ViT-Base) typically explains less of the variance in cross-dataset behavior than the data, label, and split definitions. Convergence in within-dataset AUC across reasonable backbones is consistent with the view that the bottleneck is signal in the labels and shift in the data, not capacity in the model.

Practical note. Many papers report accuracy on balanced test sets; screening usefulness depends on prevalence, referral threshold, and the cost asymmetry of false negatives versus false positives [9]. AUC alone does not pin down a clinically useful operating point: two models with identical AUC can have markedly different sensitivity at a clinically motivated specificity, and only one of them may be deployable in a low-prevalence screening setting.

C. Segmentation-first and disc-aware models

A clinically grounded alternative segments optic disc and cup pixels, derives CDR or rim features, and optionally combines learned embeddings with geometry [25–27, 45]. Disc-aware architectures explicitly attend to neuroretinal rim structure, partially mitigating background interference. The U-Net family (and attention-augmented variants) remains the dominant segmentation backbone, with widely used implementations in open-source libraries [46]. Joint multi-task models that simultaneously predict glaucoma status and optic-disc/cup masks have repeatedly outperformed naive end-to-end classification

on small benchmarks, presumably because the segmentation task acts as an inductive bias toward the clinically relevant region of interest [25].

Failure modes include small training masks, ambiguous boundaries in tilted or crescent-shaped discs, poor generalization to out-of-distribution cameras, and reliance on a single grader's mask definition [9]. Inter-rater disagreement on cup boundaries is non-trivial even between glaucoma specialists; treating any single grader's mask as a hard target overstates the sharpness of the underlying clinical signal.

D. OCT-centric learning

OCT supports glaucoma-related tasks via B-scan classification, volumetric (3-D) CNNs, thickness-map modeling, and feature-level summaries (peripapillary RNFL, GC-IPL) [19, 20, 28, 29]. Hospital-based studies often demonstrate feasibility but emphasize device-specific layer segmentation and scan quality control [19, 23]. Several studies have shown that OCT-derived ML predictions can correlate with future visual-field decline, supporting a longitudinal use-case beyond static classification [20, 29]. Comparison to fundus-only models is not universal: some cohort designs favor OCT-correlated structural labels because the labels themselves are partly derived from OCT, so apparent OCT superiority may be partly tautological. Cross-vendor robustness (Cirrus, Spectralis, Topcon, Nidek, Optopol) remains an open area, particularly for thickness-map normative ranges.

E. Multimodal, textual, and privacy-preserving approaches

Combining fundus and OCT, or imaging with IOP and visual field summaries, aligns more closely with clinical reasoning when patient-level alignment is correct [23, 47]. Multimodal approaches face two recurring difficulties: (i) data alignment—ensuring that fundus, OCT, and clinical variables refer to the same eye at the same time, with consistent anonymization—and

(ii) modality drop-out, since real-world patients often lack one of the modalities, requiring fusion architectures that gracefully handle missing inputs.

Free-text notes enable screening-type NLP tasks (classification of glaucoma severity from chart documentation, prediction of treatment intensification) but introduce documentation bias: variation in how clinicians describe similar findings, censoring of negative findings, and copy-forward artifacts in EHR systems [24]. Federated optimization is attractive for multi-site training without centralizing images, but non-IID data across clinics complicates convergence, evaluation, and privacy auditing [30, 31]. Differential privacy and secure aggregation are increasingly discussed but rarely fully implemented in glaucoma AI pilots.

F. Explainability and saliency

Grad-CAM-style heatmaps, EigenCAM, and attention maps are widely used to visualize fundus regions influencing predictions [18, 48–50]. Clinicians should interpret saliency overlays cautiously: sensitivity maps are not causal explanations and can be unstable under preprocessing changes, model re-initialization, or trivial input perturbations [51]. The honest framing in the literature is increasingly that saliency methods generate hypotheses about what a model attends to, which can then inform targeted failure-mode analyses; saliency alone does not constitute interpretability in any clinical-regulatory sense. Concept-based and prototype-based explanations [52, 53], as well as inherently interpretable decompositions (CDR, ISNT, PPA scores), remain underexplored in the glaucoma literature relative to their potential for clinician acceptance. We expect this to be a productive direction over the next few years, particularly when combined with structured radiology-style prediction reports.

G. Generative models, augmentation, and synthetic data

Generative adversarial networks (GANs) and diffusion models have been applied to fundus image generation, denoising, modality translation, and data augmentation [54–56]. While direct evidence for their value in glaucoma classification is modest, related ophthalmic literature on diabetic retinopathy and AMD suggests that synthetic augmentation can mitigate class imbalance and improve robustness when appearance shift is moderate. Synthetic-data approaches must be validated carefully against the risk of distributional artefacts that the model can latch onto.

Table II lists representative studies illustrating diversity of modality, design, and evaluation rigor.

V. DATASETS, CHALLENGES, AND BENCHMARKING NORMS

A. Public fundus resources

REFUGE popularized simultaneous disc/cup segmentation and glaucoma classification on standardized splits, and remains the most widely cited benchmark for joint screening pipelines [9, 25]. RIM-ONE (in its r1, r2, and v3 releases), DRISHTI-GS, ACRIMA, ORIGA/ORIGA-light, and challenge derivatives (e.g., AIROGS via community usage) provide complementary appearance diversity but differ substantially in sample size, image resolution, label definition, image quality control, and inclusion criteria [10, 59]. As a result,

headline AUCs are not directly comparable across these benchmarks. ACRIMA, for example, is balanced and relatively clean, which inflates accuracy-style metrics; AIROGS-style large corpora are heavily imbalanced and stress sensitivity at clinically motivated specificities.

Dataset-specific label heterogeneity also limits direct pooling without harmonization. Some datasets define “glaucoma” as expert consensus diagnosis from chart review, others as referable optic-nerve appearance graded from the image alone, and still others as elevated CDR above a fixed cutoff. Mixing these definitions in a combined training set without label harmonization injects noise that the model will absorb as feature variance. A defensible strategy, when combining datasets, is to keep them separate during evaluation (so cross-dataset transfer can be observed) and only optionally pool during training with explicit reweighting and per-source loss tracking.

B. OCT and multimodal corpora

Compared with color fundus, fewer large open OCT volumes with consensus glaucoma labels exist; many strong OCT studies rely on institutional repositories with restricted access [19, 20, 23, 29]. The Harvard Glaucoma Detection and Progression dataset [41] is a recent multimodal multi-task benchmark that explicitly couples imaging with demographic attributes for fairness analysis, and represents an encouraging trend toward larger, generalization-focused public resources. Multimodal public sets remain rarer than single-modality corpora, which inherently skews the literature toward fundus CNNs and limits the field’s ability to test imaging-clinical fusion at scale. Cross-vendor OCT pooling is hampered by device-specific normative databases and proprietary scan formats, and remains an under-published area.

C. Adjacent datasets and pretraining corpora

Several adjacent fundus datasets, while not glaucoma-labeled, are routinely used for pretraining or for joint training across ophthalmic conditions. EyePACS-style diabetic retinopathy corpora and the Kaggle DR competition data are commonly used for fundus-image self-supervision; DR detection itself has been a productive proving ground for deployment-grade ophthalmic AI [6, 60]. Adjacent retinal pathology datasets such as those used in foundation-model-style ophthalmic pretraining have produced backbones that transfer well to glaucoma in some peer-reviewed reports, though external validation in glaucoma-specific cohorts is still maturing.

TABLE II REPRESENTATIVE PEER-REVIEWED STUDIES ILLUSTRATING RECURRING MODALITY AND PIPELINE CHOICES (NOT AN EXHAUSTIVE CORPUS; SEE METHODS FOR SELECTION LOGIC). “EV” INDICATES WHETHER CROSS-DATASET OR GENUINELY EXTERNAL VALIDATION WAS A CENTRAL EVALUATION ELEMENT.

Ref.	Modality	Method family	EV	Notes
[33]	Fundus	Classical ML on wavelet / higher-order spectra features	Single-source splits	Early handcrafted-feature baseline; illustrates pre-deep-learning pipelines
[7]	Visual field	Classical ML vs. traditional classifiers	Institutional cohort	Foundational ML-vs-statistical comparison in glaucoma
[25]	Fundus	Disc-aware ensemble CNN	Partial/challenge	Joint segmentation and screening; optic-nerve-centric representations
[26]	Fundus	Attention U-Net for disc-cup segmentation	Cross-dataset	Attention-augmented segmentation backbone
[47]	Fundus	Multi-feature deep model	Varies by setup	Engineered-feature + DL hybrid
[57]	Fundus	CNN + random forest ensemble	Single-source splits	High reported discrimination; underscores need for external audit
[34]	Fundus	Mobile-oriented CNNs	Public sets	Dataset choice and class balance drive reported metrics
[35]	Fundus	Residual-network CNNs	Public single-set	Sensitivity to dataset definition

[42]	Fundus	Deep referable-glaucoma classifier	Multi-cohort	Strong within-cohort, drop on cross-camera evaluation
[23]	Fundus + OCT	Classical ML / fusion on extracted inputs	Institutional	Multimodal tabular fusion feasibility
[19]	OCT	Unsupervised representation / progression modeling	Longitudinal cohort	ML on RNFL-derived structural patterns
[28]	OCT	Feature-based deep learning on OCT	Institutional	Demonstrates OCT-specific deep features
[20]	OCT	Deep learning predicts VF damage from OCT	Longitudinal cohort	Couples structural imaging to functional endpoints
[29]	OCT	Deep learning on RNFL, predicts VF outcomes	Longitudinal cohort	Reinforces structure-function pipeline value
[21]	Visual field	Archetypal pattern analysis	Institutional cohort	Unsupervised VF feature extraction
[41]	Multimodal	Multitask dataset + semi-supervised learning	Cross-domain	Generalization-focused multimodal benchmark
[58]	Multimodal	Fairness in vision-language learning	Cross-domain	Foregrounds demographic fairness in retinal AI
[48]	Fundus	CNN + feature visualization	Single-site typical	Explainability-first screening study
[24]	EHR text	NLP / AI on free-text notes	Varies	Documentation bias and label censoring matter
[30]	Multi-site	Federated learning perspective	N/A	Privacy-preserving multi-site training
[9]	Review	Survey of fundus AI	N/A	Synthesis reference for epidemiology of methods

D. Challenge reports and leakage

Shared challenges reduce arbitrary split variance but can still induce implicit distribution shift if teams extensively tune on public leaderboards [9]. Authors should report patient-level splits and any preprocessing tuned using validation labels in order to avoid understated optimism. Concrete leakage patterns we have observed in informal reading of the literature include:

(i) bilateral eyes from the same patient appearing in both train and test splits when the split key is image-level rather than patient-level; (ii) augmented near-duplicates of training images present in the validation set (e.g., the same image cropped or color-adjusted in both); (iii) test-time threshold tuning on the test set itself, reported as if pre-specified; and (iv) using a dataset's pretrained encoder, derived from the same images, in subsequent "external" evaluation on the same dataset. None of these leakages are unique to glaucoma, but they are all well documented in adjacent medical imaging tasks [61, 62].

E. Reporting standards

Beyond PRISMA for the review level, primary studies in this space increasingly cite TRIPOD/TRIPOD-AI [15], CONSORT-AI/SPIRIT-AI [14], STARD/STARD-AI [63], FUTURE-AI principles for trustworthiness [16], and the Model Cards / Datasheets for Datasets templates [64, 65] for transparency. Adoption is uneven: structured Datasheets remain the exception rather than the norm in glaucoma AI, though uptake is gradually increasing in journals that explicitly require them.

F. Summary table of common public datasets

Table III condenses commonly encountered public datasets and typical task usage; consult primary dataset documentation before training, since licenses, splits, and label definitions are revised over time.

VI. PERFORMANCE PATTERNS, THREATS TO VALIDITY, AND DEPLOYMENT BARRIERS

A. External validity and dataset shift

Models tested only on random splits from a single institution or challenge corpus frequently underperform when applied to new devices, populations, or illumination pipelines [40–42]. The drop is rarely a single phenomenon; it decomposes into:

- **Acquisition shift:** differences in fundus camera optics, sensor noise, illumination, field of view, color balance, and JPEG quantization;
- **Population shift:** differences in age, ethnicity, refractive error distribution, and disease prevalence relative to the training cohort;
- **Label-construct shift:** differences in how “glaucoma” or “referable disease” are defined and adjudicated;
- **Workflow shift:** training data curated by trained image graders versus deployment images captured by primary-care staff under time pressure.

External validation should be reported as a primary result, not a supplementary afterthought [14, 66]. Domain adaptation, test-time training, and fairness-aware training are increasingly recognized as prerequisites rather than optional enhancements

TABLE III SELECTED PUBLICLY AVAILABLE OR COMMUNITY-STANDARD RESOURCES FREQUENTLY APPEARING IN ML GLAUCOMA SCREENING PAPERS. SAMPLE SIZES AND LABEL CONVENTIONS ARE APPROXIMATE AND REFLECT THE MOST COMMONLY CITED RELEASE; CONSULT PRIMARY DATASET DOCUMENTATION FOR CURRENT LICENSING, SPLITS, AND LABEL DEFINITIONS BEFORE TRAINING.

Resource	Modality	Approx. size	Labels / annotations	Common ML uses
REFUGE [25]	Fundus	~1,200 images on fixed splits	Disc/cup masks + glaucoma label	Joint segmentation + classification benchmarks
RIM-ONE family	Fundus	Several hundred images per release	Variable disc/cup masks; binary grading	Cup/disc segmentation; binary grading
DRISHTI-GS	Fundus	101 images	Expert disc-cup segmentations + grading	Geometry-based CDR; segmentation challenges
ACRIMA [59]	Fundus	705 images	Balanced binary glaucoma label	CNN benchmarking; balanced classification
ORIGA / ORIGA-light	Fundus	650 images	Disc/cup masks; binary label	CDR-driven classical ML and DL
AIROGS (community use)	Fundus	>100,000 images	Referable-glaucoma binary labels	High-N training; class imbalance considerations
G1020	Fundus	1,020 images	Disc/cup masks; binary glaucoma label with curated splits	Modern public benchmark
PAPILA	Fundus	~488 images	Glaucoma labels with patient metadata	Patient-level analyses; cross-dataset evaluation
EyePACS / Kaggle DR	Fundus	>80,000 images	DR severity (no glaucoma label)	Pretraining for fundus self-supervision
Harvard-GDP [41]	Fundus + OCT + demog.	Several thousand multimodal records	Glaucoma/progression + demographic attributes	Generalization-focused multitask benchmark

[58, 67]. Published recipes for distribution-robust training include sharpness-aware optimization, group-DRO, and explicit invariance objectives, but their glaucoma-specific evaluations remain limited.

B. Label noise, spectrum bias, and outcome misalignment

Binary “glaucoma” labels on fundus images may not correspond to the same clinical construct across datasets (early vs. moderate disease, referral suspicion vs. definitive diagnosis). Spectrum bias arises when training data disproportionately represent easy cases or when clinically relevant subgroups are

underrepresented in evaluation cohorts [32]. The asymmetry matters for screening: a model trained on cohorts of advanced glaucoma will overstate its sensitivity for early disease, where the marginal value of automated screening is largest. Few papers calibrate predicted risks to local prevalence [68], although temperature scaling and isotonic regression are simple to apply and add little overhead.

Outcome misalignment is also common. A model labelled “glaucoma classifier” may, in practice, be discriminating advanced cupping rather than glaucomatous neuropathy per se; on a tilted-disc, large-disc, or myopic population the same predictor will trigger many false positives. The framing problem here is that performance reports rarely stratify error rates by physiological-cupping or by myopia status, which is exactly the population in which the model is most likely to fail.

C. Fairness and equity

Performance disparities across ethnicity, age, refractive error, or camera settings can exacerbate access inequities if AI is deployed as triage [69, 70]. Reporting stratified metrics (when legally and ethically permissible) is essential for trustworthy deployment planning, and is now explicitly required by several regulatory and reporting frameworks [16]. The Harvard glaucoma multimodal dataset [41] is one of the few large public resources to ship demographic attributes alongside images and to encourage subgroup analyses; broader uptake of this practice across glaucoma datasets is overdue.

D. Calibration, uncertainty, and operating points

A clinically useful screening model must be well-calibrated at its chosen operating point. Modern deep classifiers are systematically over-confident in their predicted probabilities [68], which has direct consequences for screening: a probability that a clinician interprets as “95% glaucoma” may correspond to a true positive rate well below 95% in a low-prevalence setting. Reporting reliability diagrams, expected calibration error (ECE), Brier score, and decision-curve analysis at deployment-relevant prevalence is straightforward and substantially improves the credibility of headline performance numbers. Uncertainty quantification (deep ensembles, Monte-Carlo dropout, conformal prediction [71–73]) provides per-instance confidence intervals that are particularly relevant for triage applications, where a model can defer ambiguous cases to a human grader rather than commit to a hard label.

E. Regulatory, workflow, and human factors

Clinical translation requires integration with PACS or device gateways, audit trails, alert fatigue management, and governance for model updates [11, 66]. Even strong offline performance does not guarantee benefit in prospective workflows without task-appropriate study designs. Practical experience from deployed AI in adjacent ophthalmic tasks (notably autonomous DR screening) has shown that real-world pass-through rates, image-quality rejections, and the cognitive load on referring clinicians are first-order determinants of whether a screening tool is actually used [11, 74]. Glaucoma-specific deployments must reckon with the same variables, which are systematically under-reported in the algorithmic literature.

F. Privacy, security, and governance

Federated learning and differential privacy provide partial answers to multi-site training without raw-image centralization [30, 31], but they do not by themselves solve the harder governance problems of data stewardship: who is responsible for model updates, how drift is detected, how false-negative root-cause analyses are conducted, and how the model registry is audited. These are non-glamorous problems that nonetheless determine whether a deployed glaucoma AI tool degrades safely over the years.

G. Reproducibility

A practical observation across the literature is that even strong studies can be hard to reproduce because preprocessing recipes (CLAHE strength, vessel suppression, disc centering, augmentation policies, normalization) are described only narratively. Releasing not only model weights but also a deterministic preprocessing pipeline, version-pinned dependencies, and an explicit per-fold split manifest would substantially raise the bar of reproducibility in this subfield, in line with general medical-AI guidance [16, 64, 65].

VII. DISCUSSION

The past decade has produced rapid progress in detecting glaucomatous optic neuropathy signatures from color fundus photographs using deep networks, alongside steady methodological advances in segmentation-first pipelines and select OCT and multimodal studies [9, 10, 25]. The field’s central tension is no longer whether neural networks can fit curated public benchmarks, but whether those benchmarks predict safe, equitable behavior in heterogeneous real populations [40, 62].

A. Narrative takeaway

Work that combines (i) clinically meaningful labels, (ii) patient-level splits, (iii) explicit cross-dataset or external testing, and (iv) transparent uncertainty or calibration reporting is still comparatively rare but disproportionately informative for policy-makers and clinics [14]. Conversely, incremental architecture changes evaluated solely by in-dataset accuracy add limited cumulative knowledge—in part because the within-dataset ceiling is now close to the inter-rater agreement on the same images, so additional gains there are not interpretable as gains in clinical performance.

A second pattern that emerges from the synthesis is that combinations of methods (segmentation + classification, fundus

+ OCT, image + tabular, deep + classical ensembles) tend to be more robust under shift than monolithic single-pipeline classifiers, even when their headline within-dataset metrics are similar. This is consistent with the broader medical-imaging literature, where ensembles of heterogeneous backbones with explicit segmentation priors have repeatedly been more resilient to domain shift than a single high-capacity backbone trained end-to-end.

B. Relation to prior surveys

Compared with general-purpose review articles [9, 10], we emphasize evaluation hygiene (leakage, shift, stratified report-ing) as the primary axis for interpreting results. This framing is intended to help journal reviewers and clinical partners discriminate foundational contributions from overfit claims. We also intentionally include the medical-AI methodology literature [14–16, 61–65] alongside ophthalmology-specific references; this cross-pollination is, in our view, a feature rather than a digression, because most of the validity threats observed in glaucoma AI are not glaucoma-specific.

C. What recent surveys converge on

Synthesizing across recent narrative reviews [9, 10, 18], the points of convergence are: (i) within-dataset performance has saturated and no longer discriminates contributions; (ii) external validation is the strongest single predictor of meaningful methodological progress; (iii) explainability is necessary but not sufficient for clinical acceptance, and saliency alone is not interpretability; (iv) fairness and calibration must be evaluated, not just acknowledged; and (v) reproducibility requires more than weight release, including data pipelines, preprocessing recipes, and per-fold splits. The points of divergence are mostly emphasis, not substance.

D. Open scientific questions

Several scientific (rather than engineering) questions remain genuinely open and would reward focused future work:

- How much of the apparent advantage of OCT over fundus persists once labels are harmonized across cohorts and adjudicated independently of OCT-derived metrics?
- Can large self-supervised retinal foundation models close the cross-dataset gap on glaucoma in a way that supervised ImageNet-initialized backbones have not, and at what data scale?
- Are CDR/ISNT-style geometric features still additive to learned representations in modern multi-task architectures, or are they fully subsumed?
- How well-calibrated are deployed glaucoma AI systems at the actual prevalence of their target screening setting, and how does calibration drift over time?
- What level of fairness across ethnicity, age, refractive error, and camera vendor is achievable in current public datasets, and what minimal demographic metadata is required to evaluate it?

None of these questions can be answered solely by within-dataset architecture comparisons.

E. Threats to validity of this review

Our evidence base depends on English-language, primarily journal-indexed literature; conference abstracts, regulatory filings, and commercial systems without peer-reviewed performance data are underrepresented. Field extraction is necessarily coarse for heterogeneous study designs; formal meta-analysis of AUCs was not attempted due to heterogeneous endpoints and reporting. Our appraisal rubric is heuristic and may miss subtle leakage patterns (e.g., adversarial validation, augmentation-induced near-duplicates, operating-point tuning on the test set). The review's selection of representative studies in Tables II and III reflects the authors' judgment of illustrative diversity, not an exhaustive corpus, and other selections would also be defensible. Finally, author instructions differ by venue—final structural elements (word limits, ethics statements, supplementary material) must follow the target journal's information for authors.

F. Implications for primary studies

The most actionable implication of the synthesis is that clearer, fuller reporting buys more credibility—per page—than additional architectural complexity. A primary study that reports patient-level splits, an external test cohort (even a small one), calibration at a clinically motivated threshold, and stratified error rates by camera vendor and refractive error has, in our view, more residual scientific value than a much larger study with only within-dataset cross-validation, regardless of headline AUC.

VIII. CONCLUSION

Machine learning and deep learning can automate informative patterns in fundus, OCT, and multimodal data relevant to glaucoma detection, yet the gap between leaderboard metrics and dependable clinical assistance remains substantial [9, 40, 62]. Progress will likely depend less on marginal architecture gains and more on larger, diverse, harmonized datasets; routine external validation; calibration-aware reporting; fairness analysis; and prospective workflow evaluations [14, 16, 30]. Researchers should treat cross-dataset robustness and transparent limitations as core scientific contributions, not optional supplements.

Recommendations for primary studies

- 1) **Splits and leakage.** Predefine patient-level (not image-level) train/validation/test splits; prohibit leakage via bilateral eyes or augmented near-duplicates; document the split key and dataset version explicitly.
- 2) **Calibration and operating points.** Report calibration (reliability diagrams, Brier score, ECE) alongside AUC, and choose at least one clinically motivated operating point (e.g., target sensitivity at a specified specificity, or threshold from a target screening prevalence) rather than arg-max accuracy.
- 3) **External validation.** Include at least one external or cross-dataset test when feasible; interpret performance drops as primary evidence of generalization rather than as a footnote.
- 4) **Stratified reporting.** Stratify performance by camera vendor, refractive error (where available), age, sex, and ethnicity (where ethically permissible), and explicitly report subgroups in which the model underperforms.
- 5) **Reproducibility.** Publish preprocessing, training code, and model cards under licenses compatible with reproduction; release per-fold split manifests and version-pinned dependency lists.
- 6) **Honest XAI.** Treat saliency maps as failure-mode hypothesis generators, not causal explanations; subject saliency to sanity checks and report them.
- 7) **Reporting standards.** Follow CONSORT-AI / SPIRIT-AI for prospective evaluations [14], TRIPOD/TRIPOD-AI for prediction-model reporting [15], STARD/STARD-AI for diagnostic accuracy studies [63], and FUTURE-AI principles [16] for trustworthiness.

Outlook

The next phase of glaucoma AI will likely be defined less by architecture and more by data infrastructure: larger and more demographically diverse public corpora, vendor-agnostic OCT pooling, harmonized labels across centres, prospective screening pilots with pre-registered endpoints, and federated training pipelines with auditable privacy guarantees. We expect saturating returns from purely within-dataset architecture innovation, and accelerating returns from infrastructure, governance, and reporting work. These practices align emerging medical-AI standards with the evidentiary needs of ophthalmology screening and longitudinal care.

ACKNOWLEDGMENT

The authors thank reviewers and colleagues for feedback on earlier drafts. Contributions: S.A.K.: conceptualization, methodology, screening, data extraction, narrative synthesis, writing—original draft. S.J.: supervision, methodology refinement, writing—review and editing. The authors declare no competing interests. This literature review did not involve human subjects or new patient data; ethics approval was not required. All sources are peer-reviewed publications retrievable through the cited DOIs and publisher repositories.

REFERENCES

- [1] J. B. Jonas, T. Aung, R. R. A. Bourne, A. M. Bron, R. Ritch, and S. Panda-Jonas, "Glaucoma," *Lancet*, vol. 390, no. 10108, pp. 2183–2193, 2017.
- [2] Y. C. Tham, X. Li, T. Y. Wong, H. A. Quigley, T. Aung, and C. Y. Cheng, "Global prevalence of glaucoma and projections of glaucoma burden through 2040: A systematic review and meta-analysis,"

- Ophthalmology, vol. 121, no. 11, pp. 2081–2090, 2014.
- [3] H. A. Quigley and A. T. Broman, “The number of people with glaucoma worldwide in 2010 and 2020,” *Br. J. Ophthalmol.*, vol. 90, no. 3, pp. 262–267, 2006.
- [4] K. A. Lucy and G. Wollstein, “Structural and functional evaluations for the early detection of glaucoma,” *Expert Rev. Ophthalmol.*, vol. 11, no. 5, pp. 367–376, 2016.
- [5] M. Gandhi and S. Dubey, “Evaluation of the optic nerve head in glaucoma,” *J. Curr. Glaucoma Pract.*, vol. 7, no. 3, pp. 106–114, 2013.
- [6] D. S. W. Ting, L. R. Pasquale, L. Peng, J. P. Campbell, A. Y. Lee, R. Raman, G. S. W. Tan, L. Schmetterer, P. A. Keane, and T. Y. Wong, “Artificial intelligence and deep learning in ophthalmology,” *Br. J. Ophthalmol.*, vol. 103, no. 2, pp. 167–175, 2019.
- [7] K. Chan, T.-W. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, and T. J. Sejnowski, “Comparison of machine learning and traditional classifiers in glaucoma diagnosis,” *IEEE Trans. Biomed. Eng.*, vol. 49, no. 9, pp. 963–974, 2002.
- [8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, “A survey on deep learning in medical image analysis,” *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [9] L. J. Coan, B. M. Williams et al., “Automatic detection of glaucoma via fundus imaging and artificial intelligence: A review,” *Surv. Ophthalmol.*, vol. 68, no. 1, pp. 17–41, 2023.
- [10] D. M. Barros, J. C. Moura, C. R. Freire, A. C. Taleb, R. A. Valentim, and P. S. Morais, “Machine learning applied to retinal image processing for glaucoma detection: Review and perspective,” *Biomed. Eng. Online*, vol. 19, no. 1, p. 20, 2020.
- [11] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, “A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy,” *Proc. CHI Conf. Hum. Factors Comput. Syst.*, pp. 1–12, 2020.
- [12] K. S. Deepak and N. Chitkara, “A review of deep learning approaches in retinal image analysis,” *Procedia Comput. Sci.*, vol. 185, pp. 31–40, 2021.
- [13] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan et al., “The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,” *BMJ*, vol. 372, p. n71, 2021.
- [14] S. Cruz Rivera, X. Liu, M. J. Calvert et al., “Guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension,” *Nat. Med.*, vol. 26, no. 9, pp. 1364–1374, 2020.
- [15] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, “Transparent reporting of a multivariable pre-diction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement,” *BMJ*, vol. 350, p. g7594, 2015.
- [16] K. Lekadir, A. Feragen, A. J. Fofanah, A. F. Frangi, A. Buyx, A. Emelie, A. Lara, A. R. Porras, A.-W. Chan, A. Navarro et al., “FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare,” *BMJ*, vol. 388, p. e081554, 2025.
- [17] R. F. Wolff, K. G. M. Moons, R. D. Riley, P. F. Whiting, M. Westwood, G. S. Collins, J. B. Reitsma, J. Kleijnen, and S. Mallett, “PROBAST: A tool to assess the risk of bias and applicability of prediction model studies,” *Ann. Intern. Med.*, vol. 170, no. 1, pp. 51–58, 2019.
- [18] S. J. Kim, K. Cho, S. Kim, K. Jung, S. Suh, and K. H. Park, “Explainable artificial intelligence for ophthalmology: A review,” *Transl. Vis. Sci. Technol.*, vol. 10, no. 7, p. 14, 2021.
- [19] M. Christopher, A. Belghith, R. N. Weinreb, C. Bowd, M. H. Goldbaum, L. J. Saunders, F. A. Medeiros, and L. M. Zangwill, “Retinal nerve fiber layer features identified by unsupervised machine learning on optical coherence tomography scans predict glaucoma progression,” *Invest. Ophthalmol. Vis. Sci.*, vol. 59, no. 7, pp. 2748–2756, 2018.
- [20] F. A. Medeiros, A. A. Jammal, and A. C. Thompson, “From machine to machine: An OCT-trained deep learning algorithm for objective quantification of glaucomatous damage in fundus photographs,” *Ophthalmology*, vol. 126, no. 4, pp. 513–521, 2019.
- [21] T. Elze, L. R. Pasquale, and L. Q. Shen, “Patterns of functional vision loss in glaucoma determined with archetypal analysis,” *J. R. Soc. Interface*, vol. 12, p. 20141118, 2015.
- [22] R. Asaoka, H. Murata, A. Iwase, and M. Araie, “Detecting preperimetric glaucoma with standard automated perimetry using a deep learning classifier,” *Ophthalmology*, vol. 123, no. 9, pp.

- 1974–1980, 2016.
- [23] G. An, Y. Shiga, K. Omodaka, K. Hashimoto, and T. Nakazawa, “Glaucoma diagnosis with machine learning based on optical coherence tomography and color fundus images,” *Comput. Math. Methods Med.*, vol. 2019, p. 4063136, 2019.
- [24] S. Jalamangala Shivananjaiah, S. Kumari, I. Majid, and S. Y. Wang, “Predicting near-term glaucoma progression: An artificial intelligence approach using clinical free-text notes and data from electronic health records,” *Front. Med.*, vol. 10, p. 1157016, 2023.
- [25] H. Fu, J. Cheng, Y. Zhang, J. Gao, D. W. K. Wong, J. Liu et al., “Disc-aware ensemble network for glaucoma screening from fundus image,” in *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2018, pp. 132–140.
- [26] T. Shyamalee and D. Meedeniya, “Attention U-Net for glaucoma identification using fundus image segmentation,” *Mach. Intell. Res.*, vol. 20, no. 4, pp. 504–518, 2023.
- [27] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2015, pp. 234–241.
- [28] S. Maetschke, B. Antony, H. Ishikawa, G. Wollstein, J. Schuman, and R. Garnavi, “A feature agnostic approach for glaucoma detection in OCT volumes,” *PLoS ONE*, vol. 14, no. 7, p. e0219126, 2019.
- [29] A. C. Thompson, A. A. Jammal, S. I. Berchuck, E. B. Mariottoni, and F. A. Medeiros, “Assessment of a segmentation-free deep learning algorithm for diagnosing glaucoma from optical coherence tomography scans,” *JAMA Ophthalmol.*, vol. 138, no. 4, pp. 333–339, 2020.
- [30] N. Rieke, J. Hancke, W. Li et al., “The future of digital health with federated learning,” *NPJ Digit. Med.*, vol. 3, p. 119, 2020.
- [31] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nat. Mach. Intell.*, vol. 2, no. 6, pp. 305–311, 2020.
- [32] L. Oakden-Rayner, J. Dunnmon, G. Carreiro, C. Ré, and J. Zou, “Hidden stratification causes clinically meaningful failures in machine learning for medical imaging,” in *Proc. ACM Conf. Health Inference Learn.*, 2020, pp. 151–159.
- [33] U. R. Acharya, S. Dua, X. Du, S. V. Sree, C. K. Chua, and E. C. K. Chua, “Automated diagnosis of glaucoma using digitized wavelet higher order spectra feature set,” *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3456–3464, 2015.
- [34] M. Esengönül and A. Cunha, “Glaucoma detection using convolutional neural networks for mobile use,” *Procedia Comput. Sci.*, vol. 219, pp. 1153–1160, 2023.
- [35] S. Ovrelu, I. Cristescu, F. Balaș, A. Sultan, and E. Ovrelu, “Early detection of glaucoma using residual networks,” *IEEE Access*, vol. 8, pp. 179 314–179 327, 2020.
- [36] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.
- [37] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.
- [38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [39] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, “A ConvNet for the 2020s,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 11 976–11 986.
- [40] X. Liu, L. Faes, A. U. Kale et al., “A comparison of deep learning performance against health-care professionals in medical imaging: A systematic review and meta-analysis,” *Lancet Digit. Health*, vol. 1, no. 6, pp. e271–e297, 2019.
- [41] Y. Luo, M. Shi, Y. Tian et al., “Harvard glaucoma detection and progression: A multimodal multitask dataset and generalization-reinforced semi-supervised learning,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023, pp. 20 471–20 482.
- [42] S. Phene, R. C. Dunn, N. Hammel, Y. Liu, J. Krause, N. Kitade, M. Schaeckermann, R. Sayres, D. J. Wu, A. Bora et al., “Deep learning and glaucoma specialists: The relative importance of optic disc features to predict glaucoma referral in fundus photographs,” *Ophthalmology*, vol. 126, no. 12, pp. 1627–1639, 2019.
- [43] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, “Momentum contrast for unsupervised visual

- representation learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 9729–9738.
- [44] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. Int. Conf. Mach. Learn. (ICML), 2020, pp. 1597–1607.
- [45] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," in Medical Imaging with Deep Learning (MIDL), 2018.
- [46] P. Iakubovskii, "Segmentation models PyTorch," https://github.com/qubvel/segmentation_models.pytorch, 2019.
- [47] N. Akter, J. Fletcher, S. Perry, M. P. Simunovic, N. Briggs, and M. Roy, "Glaucoma diagnosis using multi-feature analysis and a deep learning technique," Sci. Rep., vol. 12, p. 8064, 2022.
- [48] N. K. Jisy, M. H. Ali, S. Senthil, and M. B. Srinivas, "Early detection of glaucoma: Feature visualization with a deep convolutional network," Comput. Methods Biomech. Biomed. Eng. Imaging Vis., vol. 12, no. 1, p. 2350508, 2024.
- [49] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017, pp. 618–626.
- [50] M. B. Muhammad and M. Yeasin, "Eigen-CAM: Class activation map using principal components," in Proc. Int. Joint Conf. Neural Netw. (IJCNN), 2020, pp. 1–7.
- [51] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2018.
- [52] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in Proc. Int. Conf. Mach. Learn. (ICML), 2018, pp. 2668–2677.
- [53] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su, "This looks like that: Deep learning for interpretable image recognition," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2019, pp. 8930–8941.
- [54] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2014.
- [55] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 8110–8119.
- [56] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.
- [57] A. Aljohani and R. Y. Aburasain, "Hybrid CNN-RF-Based model for early glaucoma detection using retinal fundus images," BMC Med. Inform. Decis. Mak., vol. 24, p. 115, 2024.
- [58] Y. Luo, M. Shi, Y. Tian et al., "FairCLIP: Harnessing fairness in vision-language learning," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024, pp. 12 289–12 301.
- [59] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Köhler, J. M. Mossi, and A. Navea, "CNNs for automatic glaucoma assessment using fundus images: An extensive validation," Biomed. Eng. Online, vol. 18, no. 1, p. 29, 2019.
- [60] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. Cuadros et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," JAMA, vol. 316, no. 22, pp. 2402–2410, 2016.
- [61] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," NPJ Digit. Med., vol. 5, p. 48, 2022.
- [62] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nat. Mach. Intell., vol. 3, no. 3, pp. 199–217, 2021.
- [63] V. Sounderajah, H. Ashrafian, S. Rose, N. R. Shah, M. Ghassemi, R. Golub, C. E. Kahn Jr., A. Esteva, A. Karthikesalingam, B. Mateen et al., "Developing specific reporting guidelines for diagnostic

- accuracy studies assessing AI interventions: The STARD-AI steering group," *Nat. Med.*, vol. 26, pp. 807–808, 2020.
- [64] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 220–229.
- [65] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, 2021.
- [66] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *Nat. Med.*, vol. 25, no. 1, pp. 9–11, 2019.
- [67] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
- [68] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1321–1330.
- [69] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferrante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Sci. Rep.*, vol. 10, p. 11074, 2020.
- [70] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
- [71] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using
- [72] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016,
- [73] A. N. Angelopoulos and S. Bates, "A gentle introduction Models

Source	Core concepts combined with Boolean operators
PubMed	(glaucoma[Title/Abstract]
OR optic nerve	
head[Title/Abstract])	
AND (deep	
learning[Title/Abstract]	
OR machine	
learning[Title/Abstract] OR convolutional[Title/Abstract]) AND (fundus[Title/Abstract] OR retinal	
image*[Title/Abstract]	
OR OCT[Title/Abstract]	
OR optical coherence tomography[Title/Abstract]).	
IEEE Xplore (glaucoma OR optic disc) AND	
(deep learning OR CNN OR machine learning) AND (fundus OR retinal OR OCT).	
SpringerLink (glaucoma) AND ("machine	
learning" OR "deep learning") AND (fundus OR retina OR OCT).	

B. Study extraction schema (fields)

a) Minimum extraction fields per included primary study:

Field	Description
Study ID	First author–year key and DOI
Clinical task	Screening binary / referable disease / segmentation surrogate / progression
Modality	Color fundus, OCT (2-D/2.5-D/3-D), visual
deep ensembles,"	in <i>Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)</i> , 2017, pp. 6402–6413.
Cohort	
field map, EHR text, multimodal	
Public vs. private; sample size; adult vs. pediatric (if stated)	
Labels	Definition of "glaucoma" (e.g., expert consensus, VF-based surrogate)
Splits	Random vs. patient-level; k-fold; external
pp. 1050–1059.	

test set available (Y/N)

Backbone; segmentation stage (Y/N); hy-brid classical stage (Y/N)

to conformal prediction and distribution-free uncertainty quantification," arXiv preprint arXiv:2107.07511, 2021.

[74] M. D. Abràmoff, P. T. Lavin, M. Birch, N. Shah, and J. C.

Metrics AUC, sensitivity/specificity, accuracy, Dice/IoU; operating point rule XAI / fairness Saliency, attention; subgroup metrics re-

Folk, "Pivotal trial of an autonomous AI-based diagnostic Appraisal tier ported (Y/N)

A/B/C mapping to Table I

system for detection of diabetic retinopathy in primary care offices," NPJ Digit. Med., vol. 1, p. 39, 2018.

APPENDIX

SEARCH PROTOCOL AND EXTRACTION SCHEMA

A. Search strings and databases

a) Illustrative keyword search strings.:
Limitations (author) Verbatim limitations as stated