



Review

Deep Learning-Based Computer-Aided Diagnosis of Glaucoma from Fundus and OCT Imaging: A Structured Review

Qurat Ul Aen Naqvi ^{1*}, Aneeza Saeed ², Ahmad Muneeb ³, Ume Aiman ¹ and Kashmala Sahar ⁴

¹Government College University Faisalabad, Faisalabad, Pakistan

²National College of Business Administration & Economics (NCBA&E) Bahawalpur, Bahawalpur, Pakistan

³University of Agriculture, Faisalabad, Pakistan

⁴Riphah International University, Faisalabad, Pakistan

* Corresponding Email: anniejarrar@gmail.com (Q. U. A. Naqvi)

Received: 28 June 2026 / Revised: 10 September 2026 / Accepted: 24 September 2026 / Published online: 13 October 2026

This is an Open Access article published under the Creative Commons Attribution 4.0 International (CC BY 4.0) (<https://creativecommons.org/licenses/by/4.0/>). Journal of Engineering, Science and Technological Trends (JESTT) published by Scientific Collaborative Online Publishing Universal Academy (SCOPUA). SCOPUA stands neutral with regard to jurisdictional claims in the published maps and institutional affiliations

ABSTRACT

Glaucoma remains a leading cause of irreversible visual disability worldwide, largely because structural damage to the optic nerve can progress substantially before a patient notices any vision change. Computer-aided diagnosis (CAD) using deep learning has therefore been explored as a strategy to support earlier identification, drawing on fundus photography, optical coherence tomography (OCT), visual field testing, and combinations of these modalities. This structured review discusses recent studies on AI-assisted detection in glaucoma management and considers their practical significance for clinical care. According to the literature reviewed, convolutional neural networks (CNNs) play a clear role in glaucoma screening, explicit segmentation models support assessment of the optic disc and cup, and OCT-based methods represent structural change more directly than surface imaging alone. Multimodal approaches, three-dimensional OCT analysis, transformer architectures, and progression-tracking models represent active and promising directions. However, clinical translation remains constrained by biased or geographically narrow datasets, device-specific variation, limited external validation, and unresolved questions about the reliability of individual predictions. The literature therefore supports AI as a developing decision-support tool that complements, rather than replaces, ophthalmic assessment. Further research should prioritise validation of AI-based tools in multi-centre settings, transparent documentation of errors and uncertainty, and prospective testing of systems intended for real-time clinical use.

Keywords: Glaucoma; computer-aided diagnosis; deep learning; fundus photography; optical coherence tomography; clinical decision support.

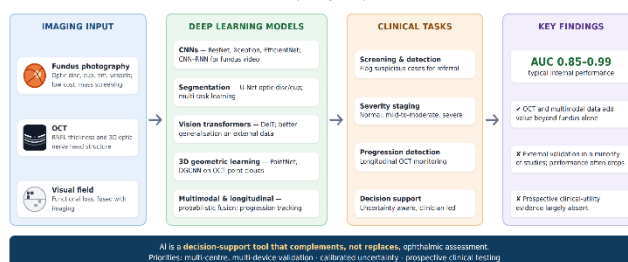
1. Introduction

Glaucoma refers to an optic neuropathy characterised by progressive damage to the optic nerve, causing irreversible loss of visual function. What makes it clinically significant is not only the severity of its consequences but the nature of the disease process: in many cases, structural damage begins well before a patient experiences any symptom that would

prompt them to seek treatment. Tham et al. estimated that the number of glaucoma cases among adults aged 40 to 80 years worldwide could reach 111.8 million by 2040, with the largest increases projected in Asian and African populations [1]. Broader reviews have similarly highlighted the expanding role of artificial intelligence across the glaucoma diagnostic and predictive pipeline, from screening through to prognosis [2].

Deep Learning-Based Computer-Aided Diagnosis of Glaucoma from Fundus and OCT Imaging

Structured review of deep learning studies published 2020-2023



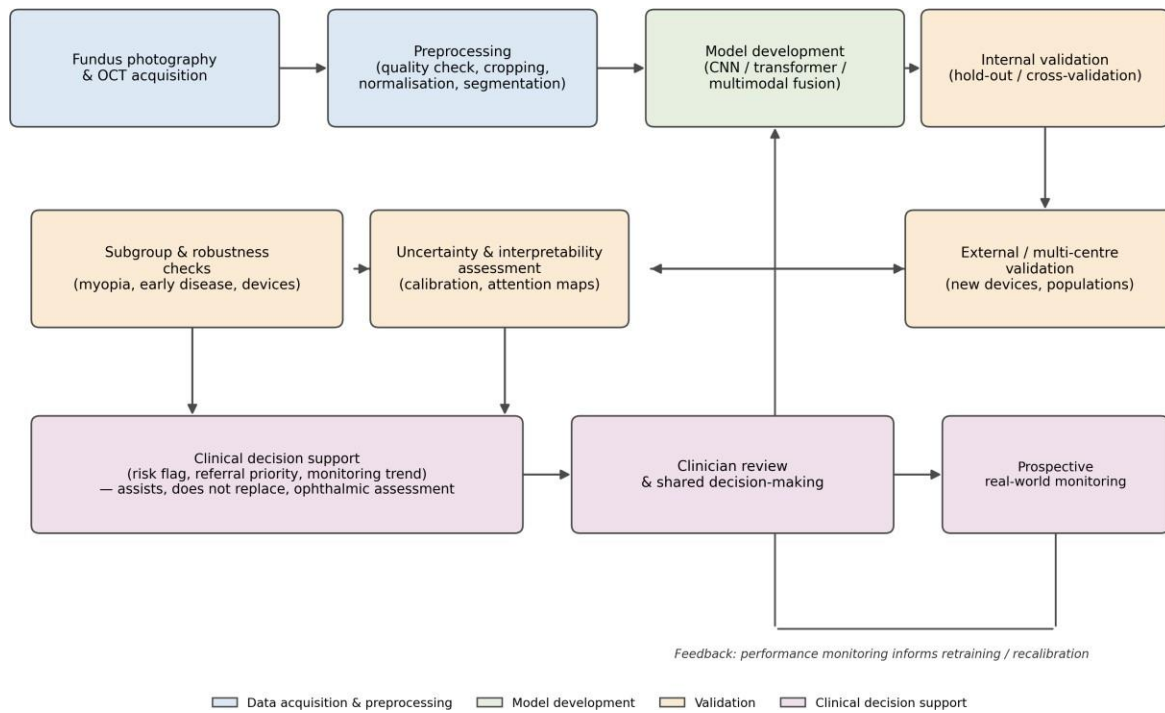


Figure 1: Schematic overview of the AI-assisted glaucoma diagnostic workflow, from fundus/OCT acquisition through model development, internal and external validation, uncertainty and subgroup assessment, and integration into clinician-led decision-making, with a feedback loop for prospective monitoring and retraining. This workflow structures the discussion of imaging modalities (Section 3), representative evidence (Section 4), and clinical translation (Section 5) that follows.

The diagnosis of glaucoma is based on neither a single image nor a single measurement. The optic nerve head, retinal nerve fibre layer (RNFL), intraocular pressure, visual field test, and the overall clinical picture of the patient are all taken into consideration. Colour fundus photography is used to visualise changes in the optic disc and cup, while OCT provides cross-sectional structural information about the retina and optic nerve. Visual field testing, in turn, provides information about the functional status of the eye.

AI has gained popularity in this domain because of its ability to extract patterns from ophthalmic images that can help detect suspicious cases, quantify specific structural features, or track disease progression over time. This paper does not propose or benchmark a new model. Rather, it aims to provide a structured overview of recent developments in deep learning approaches to CAD based on fundus and OCT imaging, with explicit attention to how the datasets, validation practices, and reported limitations of individual studies bear on their readiness for clinical use. Existing reviews of this literature often summarise studies individually without systematically contrasting datasets, validation practices, and clinical readiness across modalities; this review addresses that gap by organising the evidence around imaging modality (Section 3), critically comparing representative studies on those grounds (Section 4), and synthesising what is and is not yet established about clinical translation (Section 5).

2. Scope and Method of the Review

This is a structured narrative review, not a systematic review or meta-analysis; the term “systematic” is deliberately avoided elsewhere in this manuscript to prevent overstating its methodological rigour. Consistent with this, the review does not follow PRISMA reporting, does not include a formal risk-of-bias

assessment, and does not present a PRISMA-style flow diagram of study selection (see Section 6 for a full statement of this limitation). Search strategy: the literature was searched in PubMed, Scopus, IEEE Xplore, and Google Scholar for peer-reviewed journal articles published between January 2020 and June 2025 (final search conducted 30 June 2025), using combinations of the terms ‘glaucoma’, ‘deep learning’, ‘artificial intelligence’, ‘fundus photography’, ‘optical coherence tomography’, ‘convolutional neural network’, ‘transformer’, ‘multimodal’, ‘segmentation’, and ‘progression’.

Inclusion criteria: studies were eligible if they (i) applied a deep learning model to fundus photographs and/or OCT images for glaucoma detection, severity staging, optic disc/cup segmentation, progression tracking, or multimodal fusion; (ii) reported a defined reference standard for glaucoma status; (iii) reported at least one quantitative diagnostic performance metric (e.g., AUC, sensitivity, specificity, accuracy); and (iv) were published in a peer-reviewed journal.

Exclusion criteria: conference abstracts without an associated peer-reviewed article, studies that did not clearly report the imaging modality or reference standard used, and studies limited to non-deep-learning machine learning methods were excluded.

Screening: records were screened by title and abstract for relevance to AI-assisted glaucoma diagnosis, followed by full-text assessment against the criteria above. Because the aim was a structured narrative synthesis rather than an exhaustive systematic search, the number of records identified, screened, and excluded at each stage is not reported as a PRISMA flow diagram; this is stated explicitly as a limitation of the review in Section 6, rather than presented as a systematic count.



Table 1:
Clinical role, representative AI approaches, and limitations of common information sources used in glaucoma CAD.

Modality	Information Represented	Potential Role in CAD	Representative AI Approaches	Key Advantages	Key Limitations
Fundus photography	Visible optic disc, cup, neuroretinal rim, and retinal vascular appearance	Screening, segmentation, and risk estimation	CNN classifiers (ResNet, Xception, EfficientNet); CNN-RNN; vision transformers (DeiT); U-Net-based disc/cup segmentation	Low cost, wide availability, large public datasets (REFUGE, AIROGS, ORIGA), fast acquisition suited to mass screening	2D projection of a 3D structure; sensitive to illumination, camera, and media-opacity variation; anatomical variability (e.g., myopic discs) reduces generalizability
OCT	RNFL and optic nerve head structure, including volumetric/3D features	Structural assessment, progression-related modelling, 3D structural phenotyping	3D CNNs; geometric deep learning (PointNet, DGCNN); segmentation-free CNNs on RNFL-thickness maps	Quantitative and structural; closely tied to early glaucomatous change; supports longitudinal/progression analysis	Segmentation errors in atypical anatomy; device-specific normative databases limit cross-device transfer; higher cost and lower availability than fundus photography
Visual field	Functional loss across the patient's field of vision	Functional assessment and fusion with imaging data	CNNs and probabilistic models combining VF indices with imaging features	Directly measures functional impact, the outcome most relevant to patients	Patient-dependent variability (fatigue, learning effect); test-retest variability; longer acquisition time than imaging
Multimodal data	Complementary structural and functional evidence	Decision support, uncertainty-aware assessment, subgroup-specific diagnosis (e.g., high myopia)	Late/intermediate-fusion CNN ensembles; probabilistic multimodal CNNs; multimodal ML ensembles (RF/SVM/DNN on fused features)	Quantifiable performance gains over single-modality models in several studies (Section 4.2); can express uncertainty in borderline cases	More demanding acquisition and workflow integration; requires paired, time-matched data; added modalities increase cost without guaranteed proportional benefit

References describing background epidemiology or the general landscape of AI applications in glaucoma (e.g., refs. [1], [2]) fall outside the 2020–2025 evidence window defined above and were retained for background and context only; they were not included in the evidence synthesis presented in Sections 3–5.

Quantitative extraction: given substantial heterogeneity in datasets, reference labels, imaging modalities, and evaluation protocols across the included studies, a formal meta-analysis of pooled diagnostic accuracy was judged inappropriate. Quantitative evidence was nonetheless extracted consistently where available, including AUC/AUROC, sensitivity, specificity, dataset size, and external-validation status; these are summarised for the representative studies discussed in Section 4 in Table 2.

This distinction between technical feasibility and clinical safety is important for interpreting the literature correctly. A high score on one dataset may indicate that a model is technically capable, but it does not by itself establish that the same performance would be safe or reliable when applied to a new patient population, a different camera or OCT system, or a different hospital environment. This principle recurs throughout Sections 3–5 and underpins the research priorities identified in Table 3.

3. Imaging Information Used in Glaucoma CAD

3.1. Fundus Photography

Colour fundus photography remains the most widely available and lowest-cost imaging modality used for glaucoma screening, and it is the modality on which the largest deep-learning training datasets have been assembled to date (e.g., AIROGS, with over 100,000 labelled images [14]). Its principal diagnostic content is the appearance of the optic disc, neuroretinal rim, and cup-to-disc ratio, together with peripapillary and retinal vascular features. Its main technical limitations for CAD are illumination and quality variation between cameras and operators, sensitivity to media opacities such as cataract, and the fact that it is a two-dimensional projection of an inherently three-dimensional structure, which can obscure subtle depth-related change. Because fundus cameras and acquisition protocols differ across manufacturers and clinical sites, models trained on one camera population frequently show reduced performance when applied to images from a different device or population, a pattern documented repeatedly among the studies reviewed in Section 4.



3.2. Optical Coherence Tomography

OCT provides cross-sectional, quantitative structural information, most importantly RNFL thickness and optic nerve head morphology, that is not directly visible on fundus photography and that correlates more closely with early glaucomatous damage. This makes OCT particularly well suited to

progression tracking and to three-dimensional structural analysis (Section 4.3). Its CAD-relevant limitations differ from those of fundus photography: automated retinal-layer segmentation can fail in atypical anatomy such as high myopia or peripapillary atrophy, each OCT manufacturer applies its own normative reference database, and scan quality is more sensitive to patient fixation and

Table 2:

Selected studies, their dataset scale, modelling approach, key reported performance, and relevance to clinical translation.

Study	Input & Purpose	Dataset (size)	Model / Architecture	Key Performance Metric	External Validation	Caution for Clinical Use
Gheisari et al. (2021) [3]	Fundus videos; glaucoma detection	Proof-of-concept video cohort (frame-level)	CNN–RNN (spatiotemporal)	F-measure 96.2% (vs. 79.2% CNN-only)	Not reported	Video acquisition may not match standard screening practice
Pascal et al. (2022) [4]	Colour fundus; multi-task model	REFUGE, 1,200 images (400/400/400)	Multi-task CNN (segmentation + classification)	AUC 92.6% (vs. 87.4% single-task)	Not reported (within-REFUGE test split)	Requires dependable segmentation annotations; external testing needed
Hemelings et al. (2021) [5]	Fundus; feature attribution	UZL training set; REFUGE external set (1,200 images)	CNN with occlusion-based attribution	Internal AUC 0.94; external AUC 0.87	Yes (REFUGE)	Model attention beyond the disc must be checked for clinical meaning, not assumed
Lim et al. (2022) [6]	Fundus + OCT-derived features; high-myopia setting	1,155 images, 667 individuals	Xception + multimodal ML ensemble (RF/SVM/DNN)	Internal AUROC 95–99%; external AUROC ≈69%	Yes — marked degradation	Illustrates the internal-vs-external performance gap; broader validation essential
Huang et al. (2022) [7]	Fundus + visual field; multimodal model	1,655 eyes development; 196 eyes independent validation	Probabilistic multimodal CNN	AUC: fundus 0.90, VF 0.89, combined 0.94	Yes (independent cohort)	Multiple tests may complicate workflow deployment despite quantifiable AUC gain
Braeu et al. (2023) [8]	3D OCT; geometric deep learning	4,506 scans, 1,725 subjects	PointNet / DGCNN on 3D point clouds	AUC 0.95–0.97	Not reported	Relies on upstream OCT segmentation quality; cross-device validation still needed
Mariottoni et al. (2023) [9]	Longitudinal SD-OCT; progression detection	14,034 scans, 816 eyes, 462 individuals	CNN on paired RNFL-thickness scans	AUC 0.938 (sens. 87.3%, spec. 86.4%)	Not reported (single-centre)	Needs confirmation in prospective, multi-centre monitoring settings
Fan et al. (2023) [10]	Fundus; transformer vs. CNN	OHTS 66,715 images/1,636 participants + 5 external datasets (16,137 images)	DeiT (transformer) vs. ResNet-50	Internal AUROC 0.82–0.91; transformer +0.08–0.20 AUROC externally	Yes (5 datasets)	Generalisation advantage needs confirmation beyond one comparison; higher compute/data demands
Thanapaisal et al. (2025) [13]	Fundus; severity classification	2,940 images, 1,789 patients	EfficientNetB7	Accuracy 87.1%; class-wise AUC 0.93–0.99	Not reported	Staging labels and device transfer need evaluation
Milad et al. (2025) [14]	Fundus; code-free deep learning	AIROGS, 101,442 images	Automated (no-code) CNN classifier	Accuracy 91% internally; external AUC 0.96 / 0.99	Yes (REFUGE, GAMMA)	Accessible development does not by itself establish clinical reliability



media clarity than fundus photography. OCT devices are also less widely available and more costly than fundus cameras, which constrains the size and geographic diversity of OCT datasets relative to fundus datasets and, in turn, the amount of external-validation evidence currently available for OCT-based CAD models.

3.3. Visual Fields and Multimodal Evidence

Glaucoma is characterised by structural as well as functional change; recent research therefore pays increasing attention to multimodal approaches that combine more than one type of information source. Visual field testing is an indicator of functional change, whereas fundus photography and OCT provide structural information. Huang et al. evaluated probabilistic deep learning models using fundus photographs, visual fields, and their combination, and found that accounting for predictive uncertainty, rather than only a point classification, is important for supporting, rather than replacing, a clinician's decision-making process [7]; the quantitative performance of this comparison is discussed in Section 4.2.

4. Machine Learning Classification

4.1. Fundus-Based Classification and Segmentation

Fundus-based modelling is currently the most extensively studied strand of glaucoma CAD in terms of published dataset size and number of studies, reflecting the low cost and wide availability of fundus cameras; this is a statement about research volume rather than about which modality is clinically superior, since OCT-based and multimodal approaches capture structural and functional aspects of the disease that fundus photography alone cannot (Section 5.1). Gheisari et al. compared a combined CNN-RNN model that exploits temporal information from fundus videos against a CNN baseline using static frames, reporting a substantial improvement in F-measure (96.2% vs. 79.2%) [3]. This demonstrates that spatiotemporal features can add diagnostic value beyond single-frame analysis, although video-based acquisition is not yet standard in screening practice. Pascal et al. applied multi-task learning that jointly performs optic disc/cup segmentation and glaucoma classification on the REFUGE challenge dataset (1,200 fundus images), reporting an AUC of 92.6% for the multi-task model compared with 87.4% for an equivalent single-task classifier [4], indicating that shared structural supervision can improve classification even without additional external data.

The issue of model attention has been investigated separately. Hemelings et al. showed that a CNN trained on full fundus images achieved an internal AUC of 0.94 for glaucoma detection, and that performance remained significant (AUC 0.87–0.88) even when the optic disc itself was progressively cropped out of the image, indicating that the model was also using peripapillary and other extra-disc regions to make its prediction [5]. This finding is clinically informative but also raises an important caution. Strong performance from regions outside the expected anatomical target can reflect either genuinely distributed disease signal or the model exploiting dataset-specific artefacts and confounding features unrelated to glaucomatous pathology, a pattern generally termed shortcut learning. Attribution maps and occlusion-based visualisations cannot distinguish between these two explanations on their own, since they show where a model is attending, not whether that attention reflects a clinically meaningful mechanism. This limits the evidentiary weight that heat maps and similar visual explanations can currently provide in support of clinical trust, and

argues for corroborating such findings against independent datasets and, where feasible, against known anatomical or pathophysiological correlates. Improved segmentation architectures and explicit optic disc/cup localisation are one response to this concern, since constraining a model to anatomically meaningful regions reduces (though does not eliminate) the risk that predictions rely on incidental artefacts [11],[12].

4.2. The Value of Multimodal and Subgroup-Specific Evidence

In the probabilistic multimodal study by Huang et al., fundus photographs alone achieved an AUC of 0.90 and visual fields alone an AUC of 0.89 for glaucoma detection in a development cohort of 1,655 eyes, while the combined fundus-plus-visual-field model reached an AUC of 0.94, evaluated further in an independent cohort of 196 eyes [7]. This four-to-five-point AUC improvement is a concrete, quantifiable benefit of multimodal fusion in this dataset, although the combined model also required paired acquisition of both modalities, which adds workflow burden relative to a fundus-only pathway; uncertainty could additionally be estimated from the probabilistic model, which is particularly relevant for borderline cases, since models able to express doubt are less risky than models that always return a binary output.

Lim et al. studied a fundus-based multimodal model (image features combined with OCT-derived structural parameters and demographic data) in a population with a high prevalence of myopia, 1,155 fundus images from 667 individuals, reporting cross-validated AUROCs above 95%, but a marked drop to approximately 69% when the same pipeline was applied, without adaptation, to an external public dataset with different acquisition characteristics [6]. This contrast, strong internal performance alongside substantial external degradation, is one of the clearest illustrations in the reviewed literature of the gap between technical feasibility and generalisable clinical performance. It reinforces a broader problem in assessing AI effectiveness: models can perform very well on the population and device they were trained on, yet perform considerably worse on eyes with unusual features or on unfamiliar acquisition equipment, so broader population and device validation remains essential before such models can be assumed to generalise.

4.3. OCT, Three-Dimensional Analysis and Progression

CAD based on OCT extends the scope of CAD beyond the surface appearance captured in colour photographs. Braeu et al. applied geometric deep learning (PointNet and a dynamic graph convolutional network, DGCNN) to 3D point-cloud representations of the optic nerve head derived from OCT, using 4,506 scans from 1,725 participants, and reported AUCs of 0.95 (PointNet) to 0.97 (DGCNN) for glaucoma detection, together with an interpretable, anatomically localised pattern of critical structural points [8]. Because glaucoma is fundamentally a structural disease, this approach is conceptually well matched to the pathology, although its performance depends on the quality of the upstream automated tissue segmentation used to construct the point cloud, and external validation across different OCT devices was not reported.

A separate advance concerns progression detection rather than cross-sectional classification. Mariottoni et al. used a CNN trained on 14,034 longitudinal SD-OCT RNFL-thickness scans from 816 eyes of 462 individuals to detect glaucoma progression, achieving an AUC of 0.938 (sensitivity 87.3%, specificity 86.4%), which outperformed conventional trend-based analyses at matched specificity [9]. Progression detection is a materially different



clinical task from cross-sectional classification: it requires longitudinal, well-aligned imaging from the same instrument over time, and its clinical value lies less in confirming glaucoma status than in flagging patients whose disease trajectory warrants closer monitoring or treatment escalation — a distinction that matters when comparing performance figures across the studies summarised in Table 2.

4.4. Transformers and Practical Deployment

Fan et al. directly compared a Data-efficient Image Transformer (DeiT) with a ResNet-50 CNN for glaucoma detection, training on 66,715 fundus photographs from 1,636 participants in the Ocular Hypertension Treatment Study (OHTS) and testing on five independent external datasets totalling 16,137 photographs [10]. Internal performance was similar between the two architectures (AUROC 0.82–0.91 depending on the reference-standard definition used), but the transformer's AUROC exceeded that of the CNN by 0.08–0.20 on the external datasets, suggesting better generalisation under distribution shift. This gain came with materially different resource requirements: vision transformers typically require larger and more varied training data to reach comparable performance and are more computationally demanding to train than CNNs of similar capacity, and their attention maps, while more spatially localised than typical CNN saliency maps in this study, still require independent validation before they can be treated as evidence of clinically meaningful reasoning rather than a different pattern of shortcut learning. Whether the observed generalisation advantage reflects the transformer architecture itself, the specific training data used, or both, remains an open question given the small number of head-to-head comparisons published to date.

Two further fundus-based studies illustrate complementary directions rather than architectural advances. Thanapaisal et al. trained an EfficientNetB7 model to classify glaucoma severity (normal, mild-to-moderate, severe) using 2,940 fundus photographs from 1,789 patients matched to visual field staging, reporting an overall accuracy of 87.1% and class-wise AUCs of 0.93–0.99 [13]. Milad et al. used a code-free deep learning platform, requiring no programming expertise, to train a fundus classifier on 101,442 images from the AIROGS dataset, achieving 91% accuracy internally and external AUCs of 0.96 and 0.99 on the REFUGE and GAMMA datasets, respectively [14]. The latter result is notable for showing that automated, non-expert model development can achieve external performance comparable to purpose-built architectures; however, ease of model development does not by itself address the validation, calibration, and workflow requirements needed for safe clinical deployment, and both studies used cross-sectional, single-country training cohorts. A real-time model of this kind still requires validation across diverse patients and devices, and its errors must remain recognisable to the clinicians who rely on it.

5. Discussion

5.1. What the Literature Shows Clearly?

Overall, the studies discussed above show that artificial intelligence can effectively recognise information relevant to glaucoma diagnosis in both fundus images and OCT data. Models relying on fundus photographs may facilitate large-scale screening, whereas OCT-based models allow more detailed recognition of structural change. The transition towards multimodality and longitudinal data analysis also appears promising for reflecting the

way glaucoma is actually diagnosed and monitored in clinical practice, where structural and functional evidence are considered together over time rather than as a single snapshot.

It is also important to note that, beyond a binary positive-or-negative output, artificial intelligence may provide additional information of value to a clinician. Segmentation, interpretable attention, and calibrated uncertainty can help convey how plausible a given prediction is, which can be useful for prioritising patients who require further examination rather than for making a final diagnostic determination on its own.

5.2. Why Clinical Adoption Is Still Limited?

Despite these positive results, several recurring issues limit clinical adoption. Many studies use datasets that are restricted either geographically or with respect to device type and patient characteristics (Table 2). Models trained on such datasets do not necessarily perform identically when tested on other populations, for example, other clinics, patients with early-stage disease, normal-tension glaucoma, or high myopia. Diagnostic labels also differ across studies, since not all glaucoma datasets use the same reference standard, which complicates direct comparison between reported performance figures.

It is useful to separate three related but distinct questions when judging a model's readiness for clinical use. Technical validity asks whether a model performs well against a defined reference standard on held-out or external data, the quantity most commonly reported, and summarised for the studies discussed in Section 4 in Table 2. Clinical validity asks whether that performance holds for the patients, devices, and disease stages actually encountered in a target care setting, which requires external and, ideally, prospective validation rather than retrospective internal testing alone. Clinical utility asks whether using the model changes decisions or outcomes for patients, for example, by triaging referrals more accurately, reducing the time to detection of progression, or improving access to screening in under-resourced settings, which requires prospective, ideally comparative, studies embedded in real clinical workflows. Almost all of the evidence reviewed in Section 4 addresses technical validity; clinical validity evidence is present but limited to a minority of studies with reported external testing (Table 2); and clinical utility evidence, in the sense of prospective outcome studies, is largely absent from the current published literature. Calibration, whether a model's predicted probabilities match observed frequencies, referral thresholds, and integration into existing clinical workflows also remain underreported, even though these ultimately determine how a model's output is used by a treating clinician.

In terms of interpretability, a visual presentation such as a heat map or attention map can be a useful aid, but it does not by itself guarantee that the underlying medical decision is correct (Section 4.1). Finally, the term 'real-time' in this literature refers not only to fast data processing but to the efficiency of the whole clinical workflow into which a model is placed, from acquisition through to a clinician acting on its output.

5.3. Implications for Real-Time CAD Research

The findings of this review point to a reasonably clear direction for future CAD research. It would make sense to build systems that combine high-quality image processing with explicit validation, uncertainty management, and prospective testing, following the workflow outlined in Figure 1. Multimodal systems can be especially useful where the added information provides a measurable increase in confidence (Section 4.2), but the benefit of



additional tests should be weighed against the acquisition and workflow cost they introduce. Lightweight, fundus-only systems may be particularly valuable for screening in resource-limited settings, but the trade-off against the added safety that OCT-based or multimodal information can provide should not be overlooked when the goal is diagnostic confirmation rather than initial triage.

Table 3: Main research gaps identified in the literature and actionable priorities for future work.

Gap Identified in the Literature	Clinical Consequence	Actionable Priority for Future Studies
Limited multi-centre validation	Performance may not transfer across devices or populations	Conduct independent external testing across at least two devices/sites, and prospective clinical validation studies, before deployment claims are made
Underrepresented difficult cases	Early, myopic, or atypical cases may be missed.	Report subgroup-stratified sensitivity/specificity (e.g., by disease stage, axial length, ethnicity) and recruit diverse clinical cohorts prospectively.
Unclear model explanations	Clinicians may not know when a prediction is unreliable	Report calibration metrics (e.g., calibration curves, Brier score) alongside AUC, and evaluate attention/explanation outputs directly with clinicians rather than by visual inspection alone
Real-time claims without workflow evidence	A fast model may still be unsafe or impractical in a clinic	Measure end-to-end latency, failure modes, referral-pathway impact, and clinician/patient usability in prospective workflow studies, not benchmark inference speed alone
Different diagnostic labels across studies	Performance comparisons may be misleading.	Adopt and report a transparent, named reference standard (e.g., combined structural and functional criteria) with sufficient methodological detail for independent reproduction.

6. Limitations of the Review

This literature review is organised in a format consistent with a structured narrative review; it is not a systematic review or a meta-analysis. It does not contain a PRISMA flow chart, pooled results of diagnostic measures, or an explicit risk-of-bias assessment (Section 2). Its aim is to summarise available and pertinent evidence on deep learning models used in glaucoma diagnostics and to highlight questions that still require answers. Its findings should therefore be read as an analysis of selective, published evidence rather than as an exhaustive or quantitatively pooled account of the field.

As with any narrative synthesis, this review is also subject to publication bias, since studies with strong positive results are more likely to be published; to a restriction toward English-language, peer-reviewed sources; and to selective reporting by the original study authors. Each of these factors may make the published record appear more favourable than the underlying evidence base, and readers should interpret the performance figures summarised in Table 2 with this in mind.



7. Conclusion

The evidence reviewed here supports several conclusions with reasonable confidence: deep learning models can achieve high internal diagnostic performance (typically AUC in the range of 0.85–0.99) for glaucoma detection from fundus photographs and OCT (Table 2); structural information from OCT adds diagnostic value beyond fundus photography alone; and multimodal or longitudinal approaches can improve discrimination and enable progression tracking that a single cross-sectional image cannot provide.

What remains substantially uncertain is how these models perform outside their development population and device. External-validation evidence is present in only a minority of the studies reviewed, and where it is reported, performance frequently declines, in some cases markedly (Table 2); whether model outputs meaningfully change clinical decisions and patient outcomes has also not yet been tested prospectively for the majority of these systems (Section 5.2).

Before routine deployment in glaucoma care, we consider the minimum evidence base should include: performance reported separately for external, multi-device cohorts; calibrated and clinically interpreted uncertainty estimates rather than point predictions alone; explicit reporting of performance in known difficult subgroups such as early disease, high myopia, and atypical optic discs; and at least preliminary prospective evidence that using the model changes referral, monitoring, or treatment decisions in a way that benefits patients. Meeting these criteria would move the field from demonstrating technical feasibility toward demonstrating clinical utility, the standard against which any diagnostic technology intended to support ophthalmic care should ultimately be judged. Broader surveys of machine-learning-assisted image analysis for glaucoma detection point in the same direction, reinforcing validation and reproducibility as the field's shared near-term priority [15].

Declaration

Acknowledgment: This study was supported by Riphah Artificial Intelligence Research (RAIR) Lab, Riphah International University, Faisalabad, Pakistan.

AI Disclosure: During the preparation of this work, the authors used grammarly and chatgpt solely for language editing and improving readability. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication. No AI tools were used for literature selection, data extraction, analysis, or interpretation.

Author Contribution Statement: Q. U. A. N. contributed to the conceptualization, methodology, literature search, writing-original draft, and supervision of the study. A. S. contributed to literature screening, data extraction, and writing-review and editing. A. M. contributed to formal analysis, visualization, and writing-review and editing. U. A. contributed to literature search, data curation, and validation. K. S. contributed to investigation, resource acquisition, proofreading, and final review of the manuscript. All authors have read and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable

Competing Interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent to Publish: The authors agree to publish the version of the manuscript in this journal.

Ethical Issues: There are no ethical issues.

Funding Statement: This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

References

1. 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, doi: 10.1016/j.ophtha.2014.05.013.
2. L. Zhang, L. Tang, M. Xia, and G. Cao, "The application of artificial intelligence in glaucoma diagnosis and prediction," *Frontiers in Cell and Developmental Biology*, vol. 11, 1173094, 2023, doi: 10.3389/fcell.2023.1173094.
3. S. Gheisari et al., "A combined convolutional and recurrent neural network for enhanced glaucoma detection," *Scientific Reports*, vol. 11, 1945, 2021, doi: 10.1038/s41598-021-81554-4.
4. L. Pascal, O. J. Perdomo, X. Bost, B. Huet, S. Otálora, and M. A. Zuluaga, "Multi-task deep learning for glaucoma detection from color fundus images," *Scientific Reports*, vol. 12, 12361, 2022, doi: 10.1038/s41598-022-16262-8.
5. R. Hemelings, B. Elen, J. Barbosa-Breda, M. B. Blaschko, P. De Boever, and I. Stalmans, "Deep learning on fundus images detects glaucoma beyond the optic disc," *Scientific Reports*, vol. 11, 20313, 2021, doi: 10.1038/s41598-021-99605-1.
6. W. S. Lim et al., "Use of multimodal dataset in AI for detecting glaucoma based on fundus photographs assessed with OCT: Focus group study on high prevalence of myopia," *BMC Medical Imaging*, vol. 22, 206, 2022, doi: 10.1186/s12880-022-00933-z.
7. X. Huang et al., "Detecting glaucoma from multi-modal data using probabilistic deep learning," *Frontiers in Medicine*, vol. 9, 923096, 2022, doi: 10.3389/fmed.2022.923096.
8. F. A. Braeu et al., "Geometric deep learning to identify the critical 3D structural features of the optic nerve head for glaucoma diagnosis," *American Journal of Ophthalmology*, vol. 250, pp. 38-48, 2023, doi: 10.1016/j.ajo.2023.01.008.
9. E. B. Mariottoni et al., "Deep learning-assisted detection of glaucoma progression in spectral-domain OCT," *Ophthalmology Glaucoma*, vol. 6, no. 3, pp. 228-238, 2023, doi: 10.1016/j.ogla.2022.11.004.
10. R. Fan et al., "Detecting glaucoma from fundus photographs using deep learning without convolutions: Transformer for improved generalization," *Ophthalmology Science*, vol. 3, no. 1, 100233, 2023, doi: 10.1016/j.xops.2022.100233.
11. R. Kashyap, R. Nair, S. M. P. Gangadharan, M. Botto-Tobar, S. Farooq, and A. Rizwan, "Glaucoma detection and classification using improved U-Net deep learning model," *Healthcare*, vol. 10, no. 12, 2497, 2022, doi: 10.3390/healthcare10122497.
12. M. Nawaz et al., "An efficient deep learning approach to automatic glaucoma detection using optic disc and optic cup localization," *Sensors*, vol. 22, no. 2, 434, 2022, doi: 10.3390/s22020434.
13. S. Thanapaisai et al., "Machine learning technology in the classification of glaucoma severity using fundus photographs," *Scientific Reports*, vol. 15, 26151, 2025, doi: 10.1038/s41598-025-11697-1.
14. D. Milad et al., "Code-free deep learning glaucoma detection on color fundus images," *Ophthalmology Science*, vol. 5, no. 4, 100721, 2025, doi: 10.1016/j.xops.2025.100721.
15. V. Yadav, B. Dey, U. Baruah, S. Das, and O. Prakash, "Machine learning-assisted image analysis techniques for glaucoma detection," *EURASIP Journal on Image and Video Processing*, vol. 2025, 7, 2025, doi: 10.1186/s13640-025-00668-1.

Author(s) Bio

Qurat Ul Aen Naqvi is affiliated with Government College University Faisalabad, Pakistan as a PHD Student. Her research interests include image processing, artificial intelligence, and deep learning, with a focus on developing intelligent methods for image analysis.

Email: anniejarrar@gmail.com

Aneesa Saeed is affiliated with the National College of Business Administration & Economics (NCBA&E), Bahawalpur, Pakistan. Her research interests include artificial intelligence and image processing.

Email: anezasaeedkm@gmail.com

Ahmad Muneeb is affiliated with the University of Agriculture, Faisalabad, Pakistan. His research interests include machine learning and artificial intelligence.

Email: ahmadmuneeb13@gmail.com

Ume Aiman is affiliated with Superior University, Lahore, Pakistan as a PHD researcher. Her research interests include deep learning and medical image analysis.

Email: aimannaqvi514@gmail.com

Kashmala Sahar is affiliated with Riphah International University, Faisalabad, Pakistan. Her research interests include machine learning and artificial intelligence.

Email: aimannaqvi514@gmail.com

